Semantic Object Detection for Human Activity Monitoring System

Nor Surayahani Suriani, Fadilla 'Atyka Nor Rashid and Mohd Hafizrul Badrul
Department of Computer Engineering, Faculty of Electrical and Electronics Engineering
Universiti Tun Hussein Onn Malaysia, Batu Pahat 86400, Johor, Malaysia
nsuraya@uthm.edu.my

***Abstract*—Semantic object detection is significant for activity monitoring system. Any abnormalities occurred in a monitored area can be detected by applying semantic object detection that determines any displaced objects in the monitored area. Many approaches are being made nowadays towards better semantic object detection methods, but the approaches are either resource consuming such as using sensors that are costly or restricted to certain scenarios and background only. We assume that the scale structures and velocity can be estimated to define a different state of activity. This project proposes Histogram of Oriented Gradient (HOG) technique to extract feature points of semantic objects in the monitored area while Histogram of Oriented Optical Flow (HOOF) technique is used to annotate the current state of the semantic object that having human-and-object interaction. Both passive and active objects are extracted using HOG, and HOOF descriptor indicate the time series status of the spatial and orientation of the semantic object. Support Vector Machine technique uses the predictors to train and test the input video and classify the processed dataset to its respective activity class. We evaluate our approach to recognise human actions in several scenarios and achieve 89% accuracy with 11.3% error rate.**

***Index Terms*—Semantic Object Detection; Activity Recognition; Image Understanding;**

I. INTRODUCTION

The semantic object is a rendition of a group of attributes that describes an identifiable object in the specific environment. When we look at an image, we build a representation that supported by the understanding of "what are they doing?" by looking at what is involved such as "tools" or so-called as "semantic object" to recognise the activity. Image understanding is important in order to support the advancement of human-centric data, e.g. photo sharing in social networking websites or from surveillance cameras. The needs to develop smart algorithms that are semantically aware of our actions is significant for numerous applications, e.g. indexing and retrieving based on semantic descriptions, sports analysis, health and home applications.

In addition, semantic objects detection can help in spotting any abnormalities that occurred by determining the objects' displacement. For example, a standard office room usually contains a table, a chair, and a computer. If there is a person in the office, the person would be either standing up or sitting down, which represents that the person is working. However, if the person is lying down on the floor, it is an abnormal because of an office usually does not contain a bed. Thus, this case is called object displacement.

The monitoring system has been improving since the invention of the camera to serve humanity in order to reduce our daily workload and improving the daily output. There are some applications based on semantic objects detection that is being applied in everyday life. Such systems are capable in monitoring daily life activities of the elders using various types of sensors [1], Tele-rehabilitation for remote motor rehabilitation, and self-management healthcare that can be found in hospitals [2] also Human-Computer Interaction (HCI) that addresses the problems of semantic gap between human and computer's understandings towards mutual behaviours.

Therefore, in this project, a system using a camera as visual based application to detect semantic objects used in daily living activities is proposed. The system will be focusing on semantic objects recognition and classify the daily activities related to the objects being interacted with. This system is helpful in assisting a human in taking care of a person who needs regular attention such as elders or disabled person. A caregiver will be able to monitor the daily life activities of the patients through the system. It will be a great change for the future of daily human living if we have a monitoring system which can check on the family members once in a while, without having to give constant supervision on them. The system itself can record and detect any abnormalities in the area of interest, should anything happen, in real-time.

II. RELATED WORKS

The massive increase of multimedia information recently has the worldwide demanding a better way to process certain data needed by users. This leads to the research of semantic identification of objects in the multimedia content where the process of identifying objects and recognising the objects as well as the events in the content, extracts only the data needed by users [3]. Research direction in the scope of semantic object recognition has been progressing towards a direction such as detection of overlapping objects, occlusions and reappearing of objects in the multimedia contents [4]. Although the semantic object recognition technology is noncurrent in the community, a lot of researches and works are still being perfected and still improving from time to time to meet user needs.

Previous work has presented semantic activity recognition using Hidden Markov Models (HMMs) [1, 5]. Different shot of cameras was used such as head-mounted camera and stationary camera respectively. HMMs applicable in modelling the temporal evolution of human gait patterns for action recognition. However, the assumption of Markov model restricted to relatively simple and stationary temporal patterns.

Nowadays, local features (e.g. Scale-invariant feature transform (SIFT) [6], Speeded up robust features (SURF) and Spatiotemporal interest points (STIPS) [7]) and deep representations are favoured. This is due to holistic features are too rigid to capture variations of actions (e.g. viewpoint, appearance, occlusions). Meanwhile, SIFT features avoid local minima, and each SIFT flow considers all possible matches point at each pixel. This will lead to high complexity with low optimisation.

Histogram of Optical Flow (HOF) which inspired by Histogram of Gradient (HOG) is a descriptor that spanned to the spatiotemporal domain to obtain the local descriptor at an interesting point [8]. Local representation with the help of bag-of-words features learnt by extracting spatiotemporal interest points and clustering of the features to a model semantic object in a daily living dataset [9, 10]. These interest points used with machine learning methods such as Support Vector Machine (SVM), Bayesian Network and graphical models to recognise complex activities which involve several actors in the scene. Computational complexity in bag-of-words (BOW) model [11, 12] is another common issue especially to approximate the connections of the model while preserves computational efficiency at the same time.

Prior knowledge of general human body shape and movement patterns are necessary to categorise different types of motion. In many situations, the objects contributed to the actions become important as certain actions are defined by the related objects. Therefore, this paper aims to extract contextual information of person location and their interacting objects. Both information is expected to describe the current situation.

III. System Overview

This section describes the project design; methods applied as well as the software and hardware requirements that will be implemented. Figure 1 shows several stages required in order to fulfil this project's requirements. These phases are reinforced to ensure that the project is developed based on the objectives and the scope of projects.

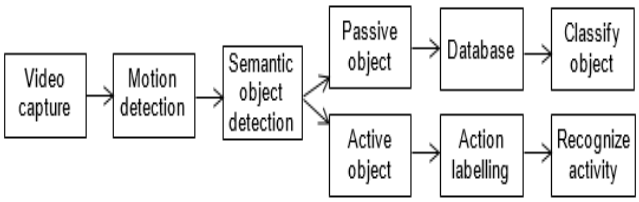

Figure 1: The overview of the system

Input video streams are taken from a standard dataset and recorded webcam. The first phase of object detection process is feature extraction, which will detect and recognise semantic objects in the area of interest as passive objects by applying Histogram of Oriented Gradient (HOG) method. The extracted HOG features of passive objects stored in the database for classification purposes. Therefore, during testing of the overall system, Histogram Oriented of Optical Flow (HOOF) method used to identify an active object. Motion detection captured active motion patterns on the scene and labelling the activities on the current frames. Meanwhile, object classification involves the use of Support Vector Machine (SVM) method which in the early stage learns about the semantic object detected using feature detection. SVM will then classify the images captured into different types of activity.

A. Dataset Collection

The datasets used were downloaded from University of Rochester Activities of Daily Living Dataset. Each of the datasets consists of a video stream of a person doing activities in daily life which was recorded using a stationary webcam camera.

The resolution of all the datasets used is 1280x720 while the fps is constant at 30 fps. The total number of frames in the videos are 10582 frames for 25 videos. Due to the variable size of blocks, the total feature points present inside each of the object frames are converted into an average of 3456 feature points. Table 1 shows the duration of each action. Some action has large variability in duration thus making action detection consumption period longer and difficult.

Table 1
Action Name, Duration and Detected Semantic Objects in The Video Frames.

Action Name	Duration (sec)	Semantic Object Detected
Using cell	15.67	cellphone
Making Tea	80.33	kettle, cup
Cutting banana	88.73	knife, plate
Moving dishes	100.50	plate
Brushing teeth	98.23	toothbrush, tap
Combing hair	31.58	comb
Washing hands	75.00	soap, tap
Drinking water	87.25	glass, bottle
Cleaning floor	120.14	broom
Watching TV	48.72	TV, sofa
Washing dishes	79.58	span, liquid soap
Using computer	103.56	computer, chair, lamp table, book

B. Passive Object Models

Our approached attempt to detect objects and human interaction with the object to define their spatial relationship. The fact that objects may significantly change in appearance during the interaction, hence determine semantically by the interactions object relation with their activity in the frames. Initially, passive object frames are extracted from the datasets and saved in the database for later use in the activity classification phase. Passive object means that the object is not being used and is in its default state.

We applied simple algorithm of Histogram of Oriented Gradient (HOG) by Dalal and Triggs. Each frame is divided into cells of 8 x 8 size of pixels, and each group of 2 x 2 is integrated into a block. Each cell consists of 9-bin of HOG and each block composed of a concatenated vector of all cells. To ease the computational time, we form HOG descriptor using integral images. The Gaussian filter with $\sigma = 0.8$ in orientation direction is applied to reduce quantisation effects. The extracted feature vector was normalised to L2 unit length. Each 64 x 128 window is represented by 7 x 15 blocks, giving a total of 3780 feature per window. The HOG descriptor is similar with SIFT as long as we neglect the scale and rotation invariance of the SIFT detector. Reducing the spatial extent of the descriptor will also reduce the blurring effects at motion discontinuities.

Figure 2 shows the sample frames of the passive state of the objects in the video. With the implementation of HOG, the green markers are being used to identify the states of the objects. All of the objects listed in Table 1 can be classified as a passive object when the initial feature points have not change before the objects interacted with the person in the

video. We also used HOG features to understand and learn the attributes and appearance of the related objects in the still image.

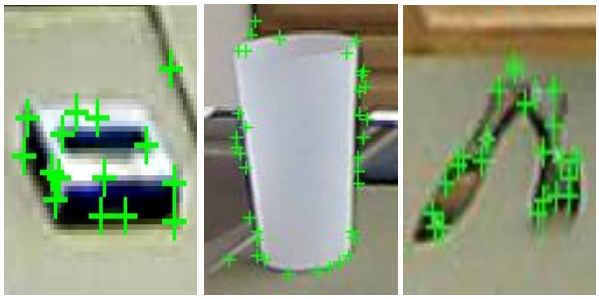

Figure 2: Passive object in the frame (From left: phone, cup, and silverware)

C. Active Object Models

Sudden changes in motion direction, velocity, and magnitude of the optical flow will indicate that previously passive object now being in an active state. Thus, active object refers to the new state of the object where it is has been touched or picked up by the person in the video frames.

We present our proposed HOOF descriptor which uses optical flow information (orientation and magnitude) to indicate the active object in the scene. To perform this, HOOF is used to define the magnitude range of motion flow. For an image window, the HOOF features are represented as a histogram $h_{b,t} = [h_{t,1}, h_{t,2}, \dots, h_{t,b}]$ is produced by extraction of HOOF at each time t, for each block b in the frame.

$$h_i = \frac{h_i'}{\sum_{k=1}^{n} h_k'} \tag{1}$$

$$h_i' = \sum_{\forall\theta\in(i^{th}bin)} \rho_j \tag{2}$$

Each flow vector is binned according to its primary angle the horizontal axis and weighted according to its magnitude. Therefore, every optical flow vector, v and direction, θ contributes its own magnitude, m to the *i*-th bin of the histogram at each frame of the video. To perform classification of actions, we exploit temporal evolution of these histograms to distinguish between passive and active objects. The magnitude of the optical flow indicates the velocity of the moving pixel. Then, we define a scoring function to visualise high dense of motion flow and visualise the scoring with the red bounding areas. Our active object models scoring function is given as follows:

$$score(q) = w.[score(q)\ x\ y\ t\ x^2\ y^2\ t^2]^T \tag{3}$$

The score function used to refer our object-centric model which defined as an active object. The red box visualisation indicates high scored image region with motion patterns which also have similar attributes with a passive object.

Information that can be gathered from the markers inside the active object regions is showing that the features detected are different from the initial state. This condition is called as the active state where the feature points of the objects inside the frames have changed. With the implementation of HOOF and scoring function properties, the red markers are being used to display the visual appearance of motion on the active state of the frames.

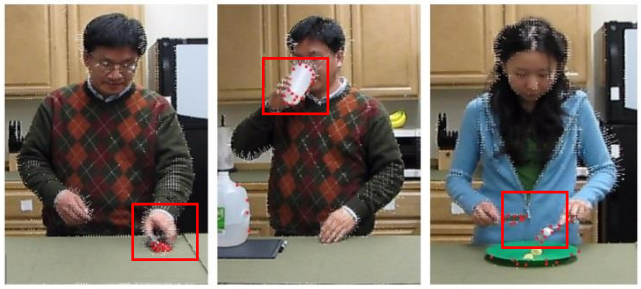

Figure 3: Sample of an active object in the frame for a different activity. From left: using phone, drinking, cutting banana.

Figure 3(a) shows the sample frame of the active state from the dataset of a person picking up the mobile phone to answer a call. Figure 3(b) shows the sample frame of the active state from the dataset of a person using the cup to drink. While Figure 3(c) shows the sample frame of the active state from the dataset of a person using the silverware to eat banana slices on a plate. The objects are being interacted with a human for various types of activity denoted as "active objects". Active objects tend to have larger dense of motion flow compared to passive objects. High dense of motion patterns detected as an active object for similar feature attributes as in the passive object.

To represent features in temporal form, the scale factors used to define finest temporal resolution at which a model feature matches with semantic object features. Linear SVM classifiers used to models our activity recognition based on the detected semantic object in the video frames. A simple linear kernel of multi-SVM works well in classifying the features.

$$argmax\ _{j=1\dots M} g^j(x) \tag{4}$$

where $g^j(x)$ interpret the distance of hyperplane of point *x*. All of the active frames from each activity were tested and trained using the SVM.

IV. Results and Analysis

Initially, we used training data for 20 object categories and carried out leave-one-out cross-validation for testing data. Leave-one-out cross-validation used to ensure that the same footage does not appear in both training and testing data. All passive objects were trained using SVM and stored in the database for further evaluation. We form temporal evaluation by adding spatiotemporal features to the local appearance to learn the active model. Both active and passive features were tested to evaluate action classification and the error rate.

In the classification phase, the SVM classifier made a total of 2000 frames with 3456 feature points as the predictor in the frame. The classifier correctly predicted all of the frames to their respective classes. This can be observed from the confusion matrix where the predicted class, as well as the true class, contains all tested frames in the ADL dataset. The overall classification rate was computed by averaging the diagonal of the confusion matrix and weighing all class of actions equally.

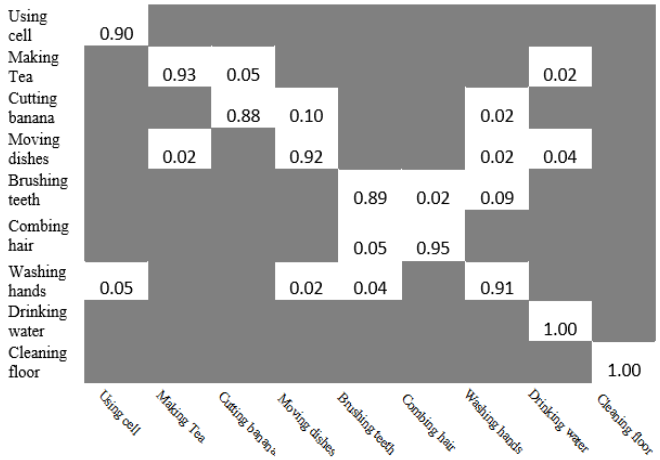

Figure 4: Confusion matrix for HOOF based active object detectors.

Figure 4 shows a confusion matrix for active objects detectors using SVM classifier. Some actions are related to each other which have similar instances for actions that involve interactions with the same objects. For example, "washing hands" and "brushing teeth" have similar objects instances (tap) functional interaction. Similarly, an action like "moving dishes" and "cutting banana" likely has similar objects (plate) during an interaction. Meanwhile, some actions are mismatched due to the size of the objects are small and often occluded by the hands and the human body. Therefore, the accuracy of the obtained results is limited due to difficulties in object annotation.

Table 2
Classification Accuracy Using SVM Classifier for Action Recognition Using Different Detectors. We Compare Results Using Active Object Models (Selected Semantic Objects and All Semantic Objects) in The Scene Based on Different Types Of Detectors.

Approach	SVM	AUC	EER %
BOW	82.3%	0.69	22.5
Temporal pyramids	87.5%	0.81	18.3
STIPS	88.7	0.71	32.3
HOOF	89.4%	0.85	11.2

Based on Table 2, our proposed method HOOF demonstrate that temporal state of the active object produces significant motion patterns of human and object interaction which semantically describe the actual activity in the video. HOOF features achieve 89.4% compared to BOW, temporal pyramids, and STIPS. Overall, the results are encouraging, and HOOF features accurately detect significant motion patterns that interact related actions in the scene.

Additionally, the SVM classifier able to classify all actions accurately based on the active object detected by using HOOF features. Although the simulated data composed of single and multiple semantic objects, the learning mechanism for all active objects definitely increase higher accuracy performance compared using only selected semantic object. Overall, we get chance performance (the area under the ROC curve) to successfully classify correct actions for an active semantic object according to the relationship of object-human interaction. The AUC of our proposed method is 0.85 with 11.2 % EER.

V. CONCLUSIONS

The purpose of this project is to propose a system that detects changes in the semantic objects being monitored. This system proved that the changes in the semantic object being monitored could be detected using HOG method as the feature descriptor and AOM method for object association with human interaction. The developed system provides a smart environment system for human monitoring where it can classify the activities of objects normally found inside the house when the objects are being interacted using SVM method. The system's performance was evaluated, and the outcomes are very outstanding. In future, this work will extend towards focusing on the integration of Internet of Things (IoT) where it can be used to monitor daily activities of the person in an area of interest. In addition, the standard dataset contains human action in controlled conditions. For higher complexity, more dataset which mostly recorded by nonprofessional contains camera motion, viewpoint variations, and resolve inconsistencies. An advanced solution is needed to compensate for the variations above.

ACKNOWLEDGEMENT

This research is funded by MOHE and Universiti Tun Hussein Onn Malaysia under grant FRGS vot 1584.

REFERENCES

[1] Debes C., Merentitis A., Sukhanov S., Niessen M., Frangiadakis N., and Bauer A. Monitoring Activities of Daily Living in Smart Home. *IEEE Signal Processing Magazine*. 2016. 81 - 94.

[2] Pirsiavash H. and Ramanan D. Detecting Activities of Daily Living in First-person Camera Views. *IEEE International Conference*. Computer Vision and Pattern Recognition (CVPR): IEEE. 2012. 2847 - 2854.

[3] Dasiopoulou S., Mezaris V., Kompatsiaris I., Papastathis V. K. and Strintzis, M. G. Knowledge-assisted Semantic Video Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*. 2005. 5(10): 1210 - 1224.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in CVPR, 2014.

[5] Ke S., Thuc H., Lee Y., Hwang J., Yoo J. and Choi K. A Review On Video-Based Human Activity Recognition. *Computers*. 2013. 88 - 131.

[6] D. Lowe. Object recognition from local scale-invariant features. In ICCV, 1999.

[7] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In BMVC, 2009

[8] Adriana Kovashka and Kristen Grauman, Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition , proc. IEEE COnference on Computer Vision and pattern Recognition, 2010. 1-8.

[9] González-Díaz I., Buso V., Benois-Pineau J., Bourmaud G., and Mégret R. Modelling Instrumental Activities of Daily Living in Egocentric Vision as Sequences of Active Objects and Context for Alzheimer Disease Research. *1st ACM MM Workshop*. Multimedia Indexing And Information Retrieval: ACMMM'13. 2013. 11 - 14.

[10] Rybok L., Friedberger S., Hanebeck U. D. and Stiefelhagen R. The KIT Robo-Kitchen Data set for the Evaluation of View-based Activity Recognition Systems. *IEEE-RAS International Conference*. Humanoid Robots (Humanoids): IEEE. 2011. 128 - 133.

[11] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79:299–318, 2008.

[12] M. Sadeghi and A. Farhadi. Recognition using visual phrases. In CVPR, 2011.