Egocentric Activity Recognition Using HOG, HOF and MBH Features

K. P. Sanal Kumar
Asst. Professor/Programmer
Dept. of ECE
Annamalai University
sanalprabha@yahoo.co.in

R. Bhavani
Professor
Dept. of CSE
Annamalai University
bhavaniaucse@gmail.com

Abstract:- recognizing egocentric actions is a challenging task that has to be addressed in recent years. The recognition of first person activities helps in assisting elderly people, disabled patients and so on. Here, life logging activity videos are taken as input. There are 2 categories, first one is the top level and second one is second level. In this research work, the recognition is done using the features like Histogram of Oriented Gradients (HOG), Histogram of optical Flow (HOF) and Motion Boundary Histogram (MBH). The extracted features are given as input to the classifiers like Support Vector Machine (SVM) and k Nearest Neighbor (kNN). The performance results showed that SVM gave better results than kNN classifier for both categories.

Keywords: *Egocentric; Histogram of Oriented Gradients; Histogram of Optical Flow; Motion Boundary Histogram; Life Logging Activity;*

*****

1. INTRODUCTION

Recognizing human actions from videos is a trend in computer vision. The advancement of wearable devices has led the trend to identify actions from such videos. These videos are egocentric videos which means self-centered. The term egocentric defines that it is concerned with an individual rather than the society. Analysing the activities of a person is to help the elderly people, people who are disabled, patients and so on [1].

In the welfare field, there are several categories in domestic behaviors which are identified as ADL (activity in daily living). A home monitoring system gives a one-day analysis on ADL to monitor the user's lifestyle. ADL is also used for special health care which improves one's strength both physically and mentally. Nowadays, glass-type camera devices have been given for users, which extends its applications into his/her daily life as well. One study on activity analysis in daily life [2] identifies all objects from the egocentric videos. This taxonomy defines the user's behavior. For instance, if "a TV", "a sofa" and "a remote" are the objects in a frame, this circumstance can be defined as "watching TV". This method categories each distinguished object into an "active object" or "passive object" based on whether the consumer manipulates the object or not. "Active object" is considered a key object in activity estimation. However, the above method has an inevitable issue, that is, there is a limitation on applicable scenes and object types [3-13]. In this research work, HOG (Histogram of oriented gradients), HOF (Histogram of optical flow) and MBH (Motion Boundary Histogram) are extracted. Here, HOG gives the static information whereas HOF and MBH provides the motion information.

2. RELATED WORKS

Human behavior recognition from videos is a popular research topic. Recognizing activities have been focused in previous works [14–16]. In [14] the history of tracked keypoints are features for identifying complex kitchen activities. A kitchen environment is chosen in [16]. Recently, ADL analysis is implemented by exploiting RGB-D sensors [15]. Here, the performance is improved based on cameras which are traditional. The issues in analyzing human behavior from wearable camera's data are represented in [17–21]. In [20] an approach is implemented for identifying anomalous events from the chest mounted camera videos. In [21] a video summarization method targeted to FPV is presented. [18] implemented a method for isolating social interactions in egocentric videos obtained in social events. Some latest works have considered different environments like kitchen, office, etc. In [17] some features were pointed based on the output of multiple object detectors. In [18] office environment is taken egocentric activity recognition. Here, motion descriptors are extracted which are combined with user eye movements. In [19] codebook is generated for a kitchen environment which reduces the problem provided by different styles and speed between different subjects. In [22], multi-channel kernels were investigated to combine local and global motion information. This created a new activity recognition method that models the temporal structures of First Person Video data. In [23], First Person Videos are temporally segmented into twelve hierarchical classes. Here, the problem of FPV ADL analysis is implemented for multi-task learning framework.

3. PROPOSED WORK

The input video is fed where the feature extraction is applied. The features are Histogram of Oriented Gradients (HOG), Histogram of optical flow (HOF) and MBH (Motion Boundary Histogram). The extracted features are fed to the classifiers like Support Vector Machine (SVM) and k Nearest Neighbor (kNN) where activity recognition is obtained as output.

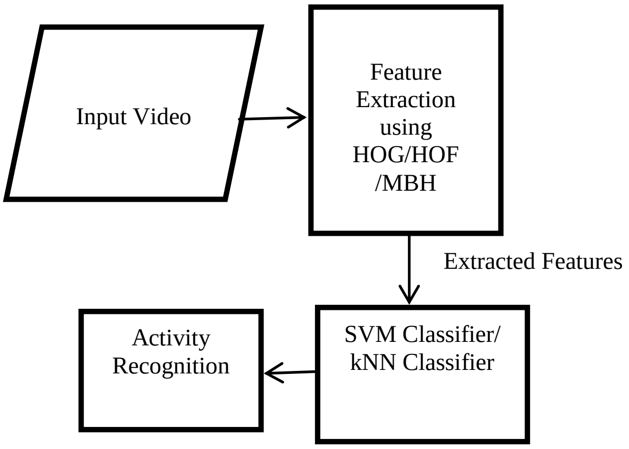
Input Video
Feature Extraction using HOG/HOF /MBH
Extracted Features
SVM Classifier/ kNN Classifier
Activity Recognition

Figure 1. The Proposed model

4. FEATURE EXTRACTION

4.1 Histogram of Oriented Gradients (HOG)

This feature is local shape information described by the distribution of gradients or edge directions. This generally focuses on static appearance information [24].

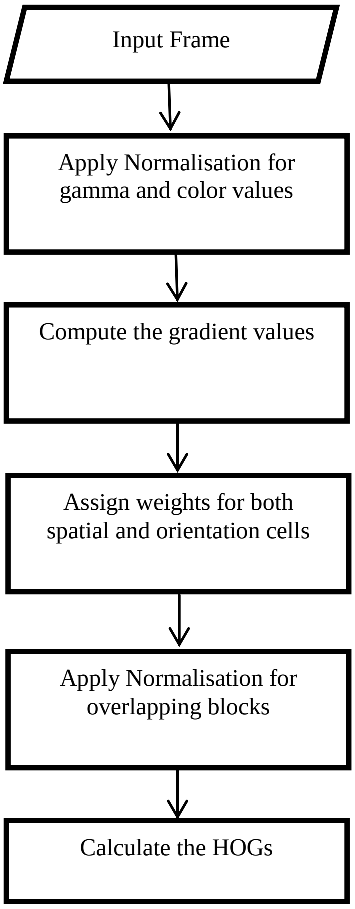
Input Frame
Apply Normalisation for gamma and color values
Compute the gradient values
Assign weights for both spatial and orientation cells
Apply Normalisation for overlapping blocks
Calculate the HOGs

Figure 2. Feature Extraction diagram for HOG

The input frame is chosen where the normalization is applied on both color values and gamma correction is applied. The next step is to apply gradient filter for finding the gradient values. Divide the window into adjacent, non-overlapping cells of size C×C pixels (C = 8). In each cell, compute a histogram of the gradient orientations binned into B bins.

Calculate the weights by using bilinear interpolation. These are subjected to normalization by concatenating overlapping blocks.

The normalized block features are concatenated into a single feature vector which is the HOG.

Initially the videos are divided into frames and then the feature extraction is applied.

Figure 2 depicts the flow diagram for HOG feature extraction.

4.2 Histogram of Optical Flow

In video domain, optical flow is commonly known as the apparent motion of brightness patterns in the images [25]. An optical flow vector is defined for a point (pixel) of a video frame. In optical flow estimation of a video frame, selection of "descriptive" points is important. This selection is done using visual features.

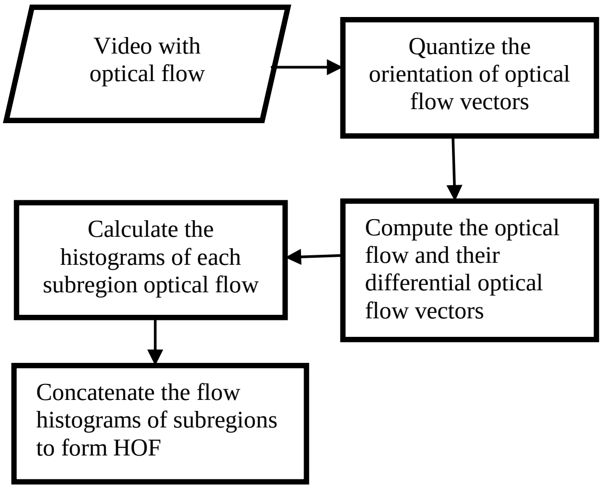
Video with optical flow
Quantize the orientation of optical flow vectors
Compute the optical flow and their differential optical flow vectors
Calculate the histograms of each subregion optical flow
Concatenate the flow histograms of subregions to form HOF

Figure 3. Flow diagram for HOF

4.3 Motion Boundary Histogram

Optical flow represents the absolute motion between two frames, which contains motion from many sources, i.e., foreground object motion and background camera motion. If camera motion is considered as action motion, it may corrupt the action classification. Various types of camera motion can be observed in realistic videos, e.g., zooming, tilting, rotation, etc. In many cases, camera motion is translational and varies smoothly across the image plane. [26] proposed the motion boundary histograms (MBH) descriptor for human detection by computing derivatives separately for the horizontal and vertical components of the optical flow. The descriptor encodes the relative motion between pixels. Since MBH represents the gradient of the optical flow, locally constant camera motion is removed and information about changes in the flow field (i.e., motion boundaries) is kept. MBH is more robust to camera motion than optical flow and thus more discriminative for action recognition. The MBH descriptor separates optical flow ω= (u, v) into its horizontal and vertical components. Spatial derivatives are computed for each of them and orientation information is quantized into histograms. The magnitude is used for weighting. We obtain a 8-bin histogram for each component (i.e., MBHx and MBHy). Compared to video stabilization [27] and motion compensation [28], this is a simpler way to discount for camera motion. The MBH descriptor is shown to outperform significantly the HOF descriptor in our experiments. For both HOF and MBH descriptor computation, we reuse the dense optical flow that is already computed to extract dense trajectories. This makes our feature computation process more efficient.

5. CLASSIFIERS

5.1. Support Vector Machine

In machine learning, support vector machines are models that are associated with learning algorithms for analyzing and classifying data. An SVM algorithm creates a model that assigns examples to one category or another. This model maps the examples which divides into the separate categories [29]. New examples are trained to assign into categories already divided and they are predicted. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using Here, multiclass SVM is used.

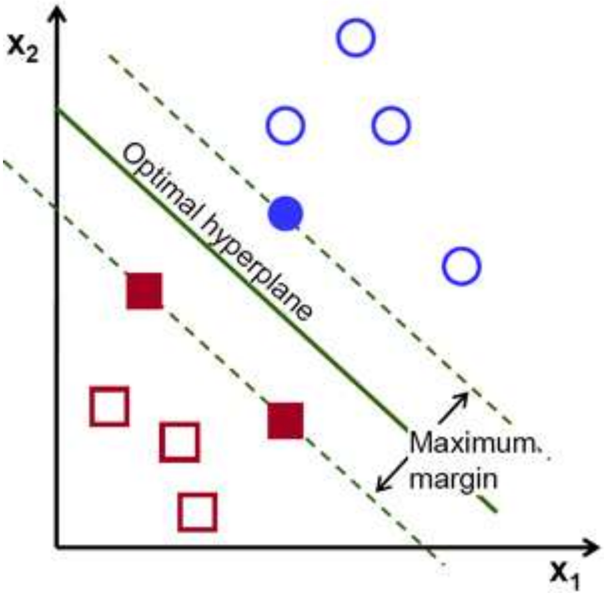

Figure 4: Support Vector Machine

Figure 4 shows the separation of examples using optimal hyperplane which is cited from [30]. The operation of SVM is based on the hyperplane that gives largest minimum distance to the training examples.

5.2 kNN Classifier

The kNN is k-Nearest Neighbor which is supervised algorithm. Here, the result is classified based on majority of k-Nearest Neighbor category. This algorithm classifies a new entity based on attributes and training samples. This algorithm uses neighborhood for classification as the prediction value of the new example. Figure 5 depicts the kNN classifier which separates the objects into different classes. This is cited from [31].

Steps in kNN classifier algorithm:

1. Assign a value for k
2. Compute the distance between the test object and every object in the set of training objects.
3. Select the closest training object with respect to the test object.
4. Select the class with maximum number of matched objects.
5. Repeat until a same class is obtained.

In our research work, Euclidean distance function is used

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

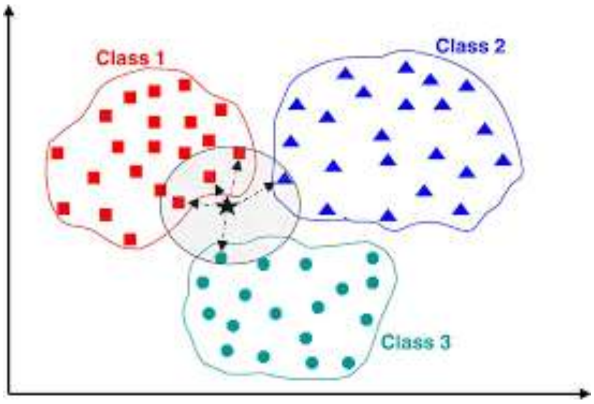

Figure 5: kNN classifier for a vector where k value is 5

6. EXPERIMENTAL RESULTS

In our work, the videos are classified into two levels. They are 5 top level categories and 13 second level categories. The top level categories are motion, social interaction, office work, food and house work. The second level categories are walk straight, walk back and forth, walk up and down, running, talk on the phone, talk to people, watch videos, use internet, write, read, eat, drink and housework. Table 1 shows the accuracy of both classifiers for top level. Table 2 depicts the accuracy of both classifiers for second level.

Table 1: Accuracy for top level categories

Features / Classifiers	Histogram of Oriented Gradients	Histogram of Optical Flow	Motion Boundary Histogram
SVM	80.25	84.38	85.56
kNN	72.15	76.39	79.83

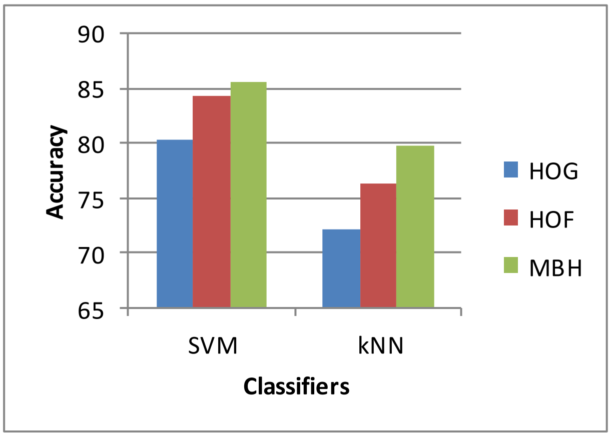

Figure 6: The chart to depict accuracy for top level categories

Table 2: Accuracy for second level categories

Features / Classifiers	Histogram of Oriented Gradients	Histogram of Optical Flow	Motion Boundary Histogram
SVM	69.23	77.14	79.54
kNN	64.34	71.56	73.21

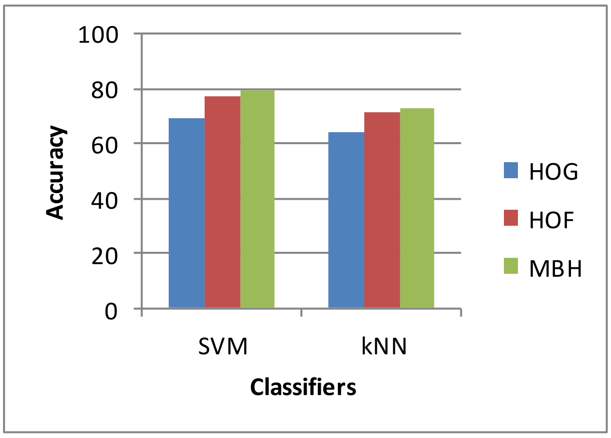

Figure 7: The chart to depict accuracy for second level categories

From the above tables 1 and 2 and figures 6 and 7, SVM performed the best for both categories than kNN. Therefore SVM is better than kNN classifier for all features.

7. REFERENCES

[1] Kenji Matsuo, Kentaro Yamada, Satoshi Ueno, Sei Naito, "An Attention-based Activity Recognition for Egocentric Video", *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2014.

[2] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views", IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012.

[3] M. Cerf, J. Harel, W. Einhauser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. Advances in Neural Information Processing Systems (NIPS), Vol. 20, pp. 241-248, 2007.

[4] L. Itti, N. Dhavale, F. Pighin, et al. Realistic avatar eye and head animation using a neurobiological model of visual attention. In SPIE 48th Annual International Symposium on Optical Science and Technology, Vol. 5200, pp. 64-78, 2003.

[5] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. Advances in Neural Information Processing Systems (NIPS), Vol. 19, pp. 545-552, 2006.

[6] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology, Vol. 4, No. 4, pp. 219-227, 1985.

[7] L. Itti, C. Koch, and E. Neibur. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 20, No. 11, pp. 1254-1259, 1998.

[8] T. Avraham and M. Lindenbaum. Esaliency (extended saliency); Meaningful attention using stochastic image modeling. IEEE Transactions on Pattern Analysis and Machine intelligence (PAMI), Vol. 32, No. 4, pp. 693-708, 2010.

[9] L. F. Coasta. Visual saliency and attention as random walks on complex networks. ArXiv Physics e-prints, 2006.

[10] W. Wang, Y. Wang, Q. Huang, and W. Gao. Measuring visual saliency by site entropy rate. In Computer Vision and Pattern Recognition (CVPR), pp. 2368-2375, IEEE, 2010.

[11] T. Foulsham and G. Underwood. What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. Journal of Vision, Vol. 8, No. 2;6, pp. 1-17, 2008.

[12] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu. Global contrast based salient region detection. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[13] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki. "Attention prediction in egocentric video using motion and visual saliency," in Proc. 5th Pacific-Rim Symposium on Image and Video Technology (PSIVT) 2011, vol.1, pp. 277–288, Nov. 2011.

[14] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in IEEE International Conference on Computer Vision, 2009.

[15] J. Lei, X. Ren, and D. Fox, "Fine-grained kitchen activity recognition using RGB-D," in ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2012.

[16] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[17] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[18] K. Ogaki, K. M. Kitani, Y. Sugano, and Y. Sato, "Coupling eye-motion and ego-motion features for first-person activity recognition," in CVPR Workshop on Egocentric Vision, 2012.

[19] E. Taralova, F. De la Torre, and M. Hebert, "Source constrained clustering," in IEEE International Conference on Computer Vision, 2011.

[20] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in European Conference on Computer Vision, 2012.

[21] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in IEEE International Conference on Computer Vision, 2011.

[22] M. S. Ryoo and L. Matthies, "First-person activity recognition: What are they doing to me?" in IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[23] Y. Poleg, C. Arora, and S. Peleg, "Temporal segmentation of egocentric videos," in IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[24] Kishor B. Bhangale and R. U. Shekokar, "Human Body Detection in static Images Using HOG & Piecewise Linear SVM", International Journal of Innovative Research & Development, vol.3 no.6, June 2014.

[25] Sri Devi Thota, Kanaka Sunanda Vemulapalli , Kartheek Chintalapati, Phanindra Sai Srinivas Gudipudi, "Comparison Between The Optical Flow Computational Techniques",

International Journal of Engineering Trends and Technology, vol 4. No.10, October 2013.

[26] N. Dalal, B. Triggs, and C. Schmid, “Human detection using oriented histograms of flow and appearance,” in European Conference on Computer Vision, 2006.

[27] N. Ikizler-Cinbis and S. Sclaroff, “Object, scene and actions: combining multiple features for human action recognition,” in European Conference on Computer Vision, 2010.

[28] H. Uemura, S. Ishikawa, and K. Mikolajczyk, “Feature tracking and motion compensation for action recognition,” in British Machine Vision Conference, 2008.

[29] M.Suresha, N.A. Shilpa and B.Soumya, “Apples Grading based on SVM Classifier”, in National Conference on Advanced Computing and Communications, April 2012.

[30] OpenCV-Introduction to Support Vector Machines. http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html Accessed: 2016-07-22.

[31] Wanpracha Art Chaovalitwongse, Ya-Ju Fan, Rajesh C. Sachdeo, “On the Time Series K-Nearest Neighbor Classification of Abnormal Brain Activity”, IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, vol.37, no.6 November 2007.