Efficient Human Motion Detection Feature Set by Using HOG-LPQ Technique

Arwa Alzughaibi
Taibah University
University of Technology Sydney
Faculty of Engineering and Information Technology

Zenon Chaczko
University of Technology Sydney
Faculty of Engineering and Information Technology

***Abstract*—Human Motion detection is a challenging task due to a number of factors including variable appearance, posture and a wide range of illumination conditions and background. So, the first need of such a model is a reliable feature set that can discriminate between a human and a non-human form with a fair amount of confidence even under difficult conditions. By having richer representations, the classification task becomes easier and improved results can be achieved. The Aim of this paper is to investigate the reliable and accurate human motion detection models that are able to detect the human motions accurately under varying illumination levels and backgrounds. Different set of features are tried and tested including Histogram of Oriented Gradients (HOG), Deformable Parts Model (DPM), Local Decorrelated Channel Feature (LDCF) and Aggregate Channel Feature (ACF). However, we propose an efficient and reliable human motion detection approach by combining Histogram of oriented gradients (HOG) and local phase quantization (LPQ) as the feature set, and implementing search pruning algorithm based on optical flow to reduce the number of false positive. Experimental results show the effectiveness of combining local phase quantization descriptor and the histogram of gradient to perform perfectly well for a large range of illumination conditions and backgrounds than the state-of-the-art human detectors. Area under the ROC Curve (AUC) of the proposed method achieved 0.781 for UCF dataset and 0.826 for CDW dataset which indicate that it performs comparably better than HOG, DPM, LDCF and ACF methods.**

***Index Terms*—Human Motion Detection, Histograms of Oriented Gradient, Local Phase Quantization.**

I. Introduction

In order for the machine to interact well with the humans in its workspace, it is very important that it interacts safely and naturally with them. Detecting and classifying a human with a fair amount of confidence is crucial for application areas such as robotics, surveillance, entertainment, assistive technology and car safety. Among several challenges that an intelligent machine is required to deal with in order to differentiate a human shape cleanly are: varying levels of illumination and background clutter.

Over the last ten years lots of research has been done to improve the detection performance and accuracy. Benenson et al. studied more than 40 detectors and compared them based on their feature-set, classifier, detection details and dataset, [8] most of the results of these methods are very similar to each other. Whereas the main focus of all of these methods is to detect human in individual monocular colored frames, some of them also make use of additional information such as context, stereo images and optical flow. And, not surprisingly, the success achieved by [3] on Daimler dataset suggests that exploiting such information does improve detection figures significantly.

Throughout the history of human detection, it can be seen that whenever immense improvement in detection quality has been achieved, it has been accomplished through diversification of the feature set. Higher dimensional representations incorporating features such as edge, color, texture, shape have been shown to make the classification task easier and thus improve the detection performance manyfolds. While developing features manually through extensive trial and error has been a method-of-choice for the last decade, recent advances in computing power and deep learning architectures call on us to exploit deep learning approaches for extracting better features.

Although a significant research has been conducted in the direction of human motion detection and classification, performance is still far from being perfect even under most suitable conditions due to variations in illumination levels and background clutter.

The aim of this paper is to further explore the domain of human motion detection and look for feature that can describe human motion better than the existing feature sets and thus allow us to make notable gains in the detection accuracy.

Rest of the paper is organized as follow: in section II the related work is presented, while the overview of the proposed human motion detection method, datasets ,methodology and result are presented in section III ,and finally section IV discusses conclusion.

II. Related Work

several methods have been proposed last decade for detecting human in videos and images [1], [4], [8].

The most successful approaches used for human detection relies on using low-level features for image representation, in particular the grids of Histogram of Oriented Gradient (HOG) descriptor [5] has shown to provide excellent results. Recently combining multiple cues has been investigated [2], [10], [12]. Wang et al. [12] combines shape and texture cues in one representation for human detection where HOG is computed for shape and LBP is employed to incorporate

the texture information. However, as a studied shows [9] that implementing LPQ method as texture classification is significantly more invariant to uniform illumination changes than other methods such as LBP and this improve the overall performance of the human detection algorithm.

III. Overview of the proposed Human Motion Detection Method

The proposed feature set is based on computing local histograms of image gradient orientations (HOG) in a dense grid of uniformly spaced cells, evaluating a histogram of local phase quantization (LPQ) and then concatenating them both into a complete feature set. Among a number of factors that motivated this approach are:

- Local appearance and shape of the object can be described very well by the distribution of local intensity gradients or directions of the edges [5].
- Low frequency phase components are known to be highly insensitive to blur and uniform illumination changes [9].

Extraction of HOG features involves dividing the image window into small cells, extracting a local 1-D histogram of gradient directions over the pixels of cells, and flattening all local histograms into a single feature vector. This feature set captures gradient structure of a local shape and is thus invariant to change in appearance and sideways motion of limbs and body segments. In order to compute LPQ features, phase information is evaluated locally at every pixel location of the image. The phases of the four low-frequency coefficients are then decorrelated and quantized uniformly into an eight dimensional space. The decorrelation is done to make sure that the samples to be quantized are statistically independent and thus information is maximally preserved. The quantized coefficients are integers with values between 0 and 255. A histogram of these integers from all pixels of the image is then extracted as a 256-dimensional feature set. This feature set is invariant to blur and uniform illumination changes. Once these feature sets are extracted, they are concatenated to form a feature set that is not only insensitive to a wide range of illumination conditions but also invariant to change in appearance and sideways motion.

A. Data Sets and Methodology

1) Data Sets: For training of the proposed detector, INRIA person dataset [5] has been used. This dataset contains 1805 cropped images of size 64x128 taken from a number of personal photos. In these images, the humans are usually standing but appear in different orientation and against a wide range of backgrounds including crowds. The dataset has been downloaded from http://pascal.inrialpes.fr/data/human/. The downloaded version of the dataset contains two different formats of images original images with relevant annotation files and positive images in normalized 64x128 pixels format with original negative images. Normalized images are further split into two folders i.e. those containing positive and negative examples for training of the model and those containing positive and negative examples for testing of the detector.

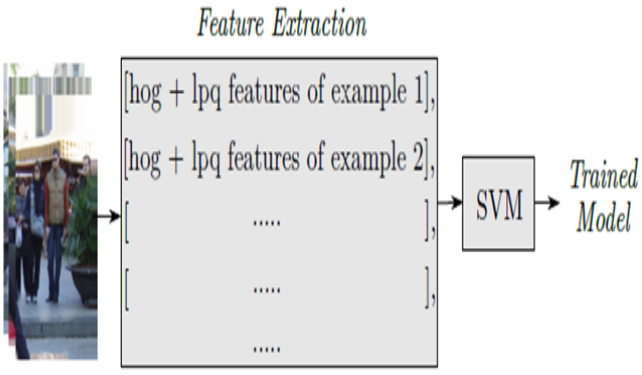


Fig. 1: This figure presents a few steps involved in training of the proposed human detector

Positive images in the training folder are of size 96x160 pixels and those in the test folder are of size 70x134 pixels. To generate negative examples, a negative window of a specific size has been sampled randomly from 1218 negative training photos.

For testing of the detector, UCF sports action and CDW datasets are used. these datasets represent a collection of actions recorded from a wide range of scenes and view points. In this paper, our main focus is on the development of a reliable and accurate human detection model for human motion detection so, the detector has been tested on the most relevant videos of these datasets. Each of these videos contain one or more person/s walking in a certain direction with their faces toward camera.

The details related to the preprocessing of these video frames, feature extraction and classification are presented in the following sections.

2) Methodology: For positive examples, all images of INRIA dataset containing one or more human are resized to 70x134 pixels. For negative examples, a random sample of size 70x134 pixels is cropped out of each of the negative images. In the feature extraction stage, HOG and LPQ features are extracted from each of the positive and negative examples and concatenated into a feature vector. Once these feature vectors are extracted, they are labeled appropriately (1) if the feature vector describes a human form and (0) if it describes a non-human form. These feature vectors and labels are then used to train an SVM classifier(see Figure.1).

The goal of SVM is to produce a model which predicts if an unseen video frame contains a human form or not based on the HOG and LPQ features of that frame. Among a few hyper-parameters of an SVM classifier are kernel, penalty parameters e.g. C and kernel parameters e.g. γ . For this project, radial basis function (RBF) kernel is preferred over linear, polynomial and sigmoid kernels because it nonlinearly maps samples into a higher dimensional space and thus can cope with any case where the relation between feature vectors and labels is nonlinear. Another reason for choosing RBF kernel is the number of hyper-parameters involved, polynomial

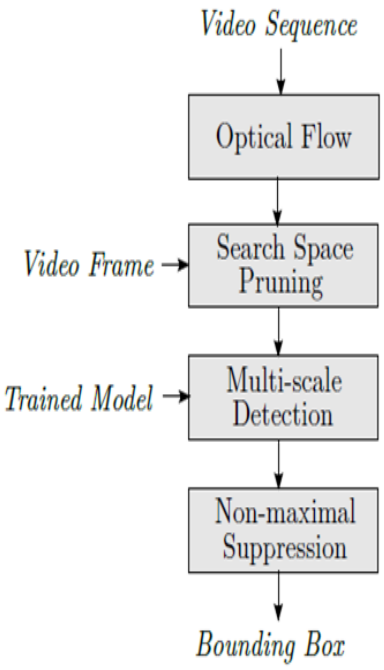


Fig. 2: This figure shows a flow-chart of a few steps involved in detecting one/more human region/s in a video frame

kernel requires more hyper-parameters than RBF kernel and is thus more complex with respect to model selection. Last but not the least, RBF kernel is selected because it has fewer numerical difficulties. For parameters C and γ , default values of 1.0 and 0.001 respectively are used.

In testing of the trained human detector, front walking video sequences of UCF sports action and shadow sequences of CDW 2014 dataset are used. The detection system takes a video frame and returns bounding boxes as well as confidence scores for each of the detections. The flowchart of a few steps involved in the detection of one or more humans in a video frame is presented in the Figure2 and the details related to each of the basic components are presented in the following sections of this paper.

Optical Flow. Optical flow refers to the distribution of velocities of patterns in an image. It can arise from relative motion of objects and the viewer thus can be helpful in understanding of the rate of change in the spatial arrangement of the objects. In this paper, the simplest of optical flow based motion detectors is used. The detector is based on finding the amount of difference in every 5 frames of the video sequence. The greater the size of the difference vector, the greater is the motion. In the very first stage, this detector is used for eliminating video sequences involving camera motion.

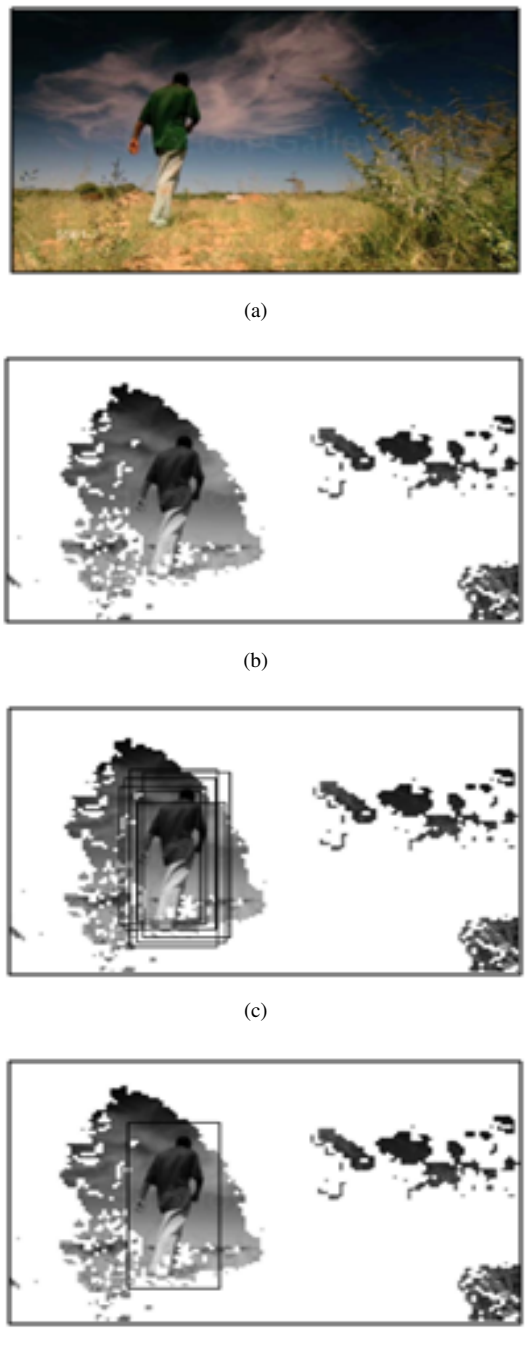
Search Space Pruning. Search space pruning block takes a test video frame as an input and makes use of the motion analysis performed in the last step to identify the regions of maximal movement. Since all of the videos in both UCF sports action and CDW 2014 datasets involve one or more humans walking toward or away from the camera, such a motion analysis can be exploited so as to prune the search space to the regions of maximal movement (see Figure.3b). This preprocessing technique not only allows us to reduce the number of false positives and thus improve the accuracy of the detector but also reduce the computational time. In order to further reduce the computational time, all video frames were resized by a factor of 0.4.

Multi-Scale Detection. Multi-scale detection block takes a test image and a trained human detection model as inputs and returns bounding boxes of the windows where one or more humans are detected with a sufficiently large probability (see Figure.3c). Heres how multi-scale detection is implemented in this paper so as to achieve the best possible human detection performance based on HOG and LPQ features:

1) While moving this window from left to right by a step size of 8 pixels, also known as the window stride, it extracts HOG + LPQ features and predicts the label of the feature vector based on the trained model.
2) If the label is detected as 1, the x, y-coordinates, the width and the height of the corresponding bounding box is recorded; otherwise, the bounding box is discarded and the sliding window continues to move further.
3) If the sliding window encounters the right end of the image, it moves down vertically by a step size of 8 pixels and then steps 2-3 are repeated for the second row of the image.
4) Once this sliding window is moved over the whole test image, the image is scaled down by a scale factor of 1.1, smoothed via a Gaussian filter and steps 1-4 are repeated for the scaled down image.
5) The same process is repeated until the size of the scaled down image becomes lesser than or equal to the size of the sliding window lesser than or equal to 70x134 pixels.

Non-maximal Suppression. Multi-scale detection usually results in a very common problem i.e. detection of multiple bounding boxes around the object of interest. From Figure 3.c it can be seen that in each one of these figures, multiple bounding boxes have correctly detected the human more than one bounding boxes refer to the same human. In order to cope with this issue, we need to apply a suppression algorithm that removes all the redundant bounding boxes except the largest one. Among well-known methods used for this purpose are mean-shift algorithm [5] and non-maximal suppression. In this paper, non-maximal suppression is used. Figure.3d presents a result of implementation of non-maximal suppression on images. Non-maximal suppression removes redundant bounding boxes based on a parameter known as the overlap threshold. If the area of overlap of a detected bounding box exceeds this threshold, then this box is removed from the final set of bounding boxes. In this human detection framework, the value of the overlap threshold is chosen to be equal to 0.85. So, if a value greater than this value is used, then the number of redundant bounding boxes increases and if a value lesser than 0.85 is used, the miss-rate of the detector

ISSN: 2229-6913 (Print), ISSN: 2320-0332 (Online)



(a)

(b)

(c)

(d)

Fig. 3: shows the implementation of the proposed detection methodology as applied on some video sequences in UCF sports action dataset.

increases.

B. Evaluation and Expermintal Results

In this research, AUC-ROC curves are used to evaluate the performance of the proposed detector comparing to state-of-the-art human detectors which are HOG [5], DPM [6], LDCF [11] and ACF [7]. So The lower value of AUC

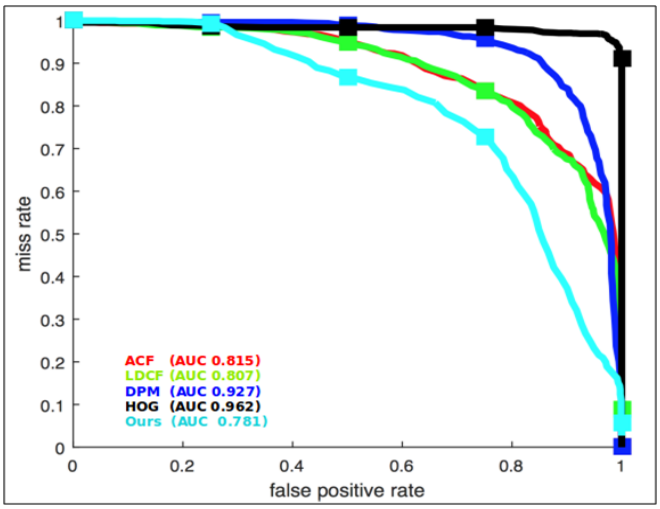


Fig. 4: An overview of the performance of various human detectors. All of these detectors are trained on INRIA person dataset and tested on a few shadow sequences of UCF sports action dataset

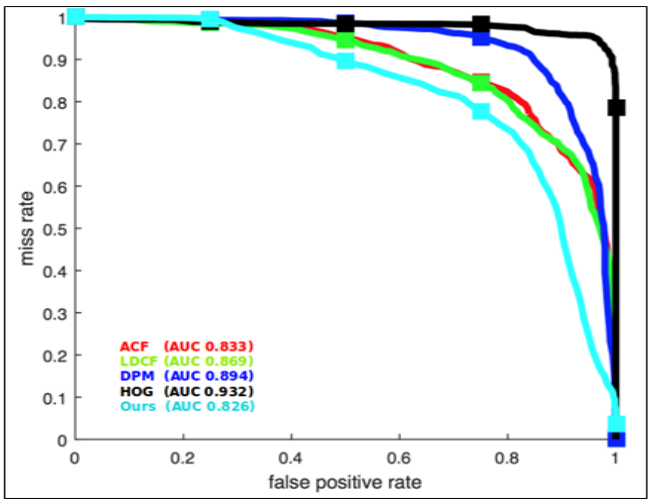


Fig. 5: An overview of the performance of various human detectors. All of these detectors are trained on INRIA person dataset and tested on a few shadow sequences of CDW 2014 dataset

indicates the better performance.

False positive rates and false negative rates are collected and are used to plot Receiver Operator curves (ROC). Area under these curves is found to be the best for HOG+LPQ feature set. From Figure.4 AUC of the proposed detector based on HOG+LPQ is found to be equal to 0.781 on UCF dataset which is comparatively better than AUC of state-of-the-art detectors such as HOG (0.962), DPM (0.927), LDCF (0.807) and ACF (0.815). In Figure.5 similar performance is observed for CDW 2014 dataset, our detectors AUC i.e. 0.826 is found to be comparatively better than those of state-of-the-art detectors such as HOG (0.932), DPM (0.894), LDCF (0.869) and ACF (0.833). Overall, these results

TABLE I: An overview of the benchmark results on UCF and CDW datasets

Method	AUC result	
	UCF dataset	CDW dataset
ACF	0.815	0.833
LDCF	0.807	0.869
DPM	0.927	0.894
HOG	0.962	0.932
Ours	0.781	0.826

show good performance of proposed model on real world datasets (see Table I).

IV. Conclusion

The performance of any human motion detection system is highly dependent on feature sets, features that can accurately discriminate between a human and a non-human and are insensitive to illumination condition and background clutters. In this paper Different set of features are tried including Histogram of Oriented Gradients (HOG), Deformable Parts Model (DPM), Local Decorrelated Channel Feature (LDCF) and Aggregate Channel Feature (ACF). However, by combining HOG and LPQ as a feature set and by implementing search pruning algorithm based on optical flow to reduce the number of false positive, our proposed detector perform comparably better than the state-of-the-art methods on real world datasets.

References

[1] David Geronimo, Antonio M Lopez, Angel D Sappa, and Thorsten Graf. Survey of pedestrian detection for advanced driver assistance systems. IEEE Transactions on Pattern Analysis and Machine Intelligence, (7):12391258, 2009.

[2] GR Rakate, SR Borhade, PS Jadhav, and Milind S Shah. Advanced pedestrian detection system using combination of haar-like features, adaboost algorithm and edgelet-shapelet. In Computational Intelligence and Computing Research (ICCIC), 2012 IEEE International Conference on, pages 15. IEEE, 2012.

[3] Markus Enzweiler and Darieu M Gavrila. Monocular pedestrian detection: Survey and experiments. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 31(12):21792195, 2009.

[4] Markus Enzweiler and Dariu M Gavrila. A multilevel mixture-of-experts framework for pedestrian classification. Image Processing, IEEE Transactions on, 20(10): 29672979, 2011.

[5] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 886893. IEEE, 2005.

[6] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32(9):16271645, 2010.

[7] Piotr Dollar, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 36(8):15321545, 2014.

[8] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? In Computer Vision-ECCV 2014 Workshops, pages 613627. Springer, 2014.

[9] Ville Ojansivu and Janne Heikkil. Blur insensitive texture classification using local phase quantization. In Image and signal processing, pages 236243. Springer, 2008.

[10] William Robson Schwartz, Aniruddha Kembhavi, David Harwood, and Larry S Davis. Human detection using partial least squares analysis. In Computer vision, 2009 IEEE 12th international conference on, pages 2431. IEEE, 2009.

[11] Woonhyun Nam, Piotr Dollr, and Joon Hee Han. Local decorrelation for improved pedestrian detection. In Advances in Neural Information Processing Systems, pages 424432, 2014.

[12] Xiaoyu Wang, Tony X Han, and Shuicheng Yan. An hog-lbp human detector with partial occlusion handling. In Computer Vision, 2009 IEEE 12th International Conference on, pages 3239. IEEE, 2009.