Human Detection From Images

Saichand Bandarupalli, Dept of Computer Science, colorado school of mines, Golden, Colorado,
bandarupalli@mines.edu

***Abstract*—The advancements in computer vision and the capabilities of machines to process the data is creating number of possibilities for researchers. Machine vision is now being used in many applications. The applications range from a laparoscopic device maneuvering inside the body trying to grasp a foreign tissue (Medical field) to a Rover on mars planning its maneuvers by trying to understand its environment. Some of the applications are mission critical for example, surveillance, disaster response and etc. Many applications like the examples stated above require a machine vision algorithm capable of distinguishing the objects in the image or the video sequence supplied. This project is one such effort to detect human subjects from images and follow them using a aerial vehicle mounted with a first person view camera (FPV). The existing techniques in object detection will be studied by using the available human data sets and soon will be coming up with a application. Firstly, the detection of human subjects will be done by processing images from a stationary camera. Then will be extended to applications in which the camera is moving.**

I. Introduction

The main goal of Human detection is to use machine vision as a tool to detect humans from the pictures and videos. The detection of humans from pictures using machine vision is gaining importance because an increasing number of applications where the customary technique used for surveillance, gathering information and etc. have become too harsh or routine for the human eye to operate on. So, to impart vision to these systems several methodologies have been devised based on state of the art technologies like machine learning, neural networks. The human detection techniques use a number of characteristic features of humans like human gait, color, and etc. to detect humans by imparting vision to the machine. Detection of humans from video and or images can be done in two stages, object detection and classification.

In this project the detection of humans is carried out by extracting features using histogram of oriented gradients from the images and the feature vector is supplied to the linear support vector machine for classifying the data. All the windows containing humans is suppressed into single window using support vector machine. Before giving the task of detection the support vector machine is trained with positive samples and negative samples. The details of which can be seen in the methods section. Further to reinforce the detection of humans from images especially from far distances the infrared data of the area of interest is used to detect human subjects from long distances. The INRIA human dataset is used for testing and training the support vector machine. /par

The HOG descriptor has a few key advantages over other descriptors. Since it operates on local cells, it is invariant to geometric and photometric transformations, except for object orientation. Such changes would only appear in larger spatial regions. Moreover, as Dalal and Triggs discovered, coarse spatial sampling, fine orientation sampling, and strong local photometric normalization permits the individual body movement of pedestrians to be ignored so long as they maintain a roughly upright position. The HOG descriptor is thus particularly suited for human detection in images.

II. Related work

Existing Object Detection and classification methods The detection of humans from a given set of frames can be done using a number of techniques. Lets see in brief some of the important and most in use techniques for detection. The background subtraction technique detects humans in the image by comparing the current frame to the reference frame (which is updated periodically). A number of techniques have also been developed, for detecting the object without updating the background. Some of them are, Mixture of Gaussian model, nonparametric background model, temporal differencing, Warping background, Hierarchical background model. Optical flow is a vector based approach in which the points on the image is matched between the frames to detect the motion of the object. It basically tracks the characteristic flow vectors of the moving objects over time to detect the moving region in the image. The motion detection in Spatio-temporal filter is done by mapping the entire 3D Spatio temporal data spanned by the object in motion.

The object in motion in the sequence of frames need to be characterized as human. The detection techniques discussed gives the information on the moving objects in the video to differentiate the human subjects from the moving objects in the field of study there are three methods in use, Shape based Method, Motion based method, Texture based method: In shape based approach the moving regions of the image are represented as points, blocks and blobs and then by the use of pattern recognition techniques the objects are classified. But however the pattern recognition techniques only cannot make the classification of subjects in the field of study. Because it is difficult to distinguish the humans from moving objects because there can be a large possibilities of patterns possible from the points on the moving objects in the scene. So in general the shape based approach is combined with other methods like texture based and motion based to the object classification. In motion based classification of humans from moving objects the periodic property of the captured images is used to distinguish humans from other moving objects.

A. The human Dataset Benchmarks

KTH human motion dataset: The most popularly used dataset for human action classification. The dataset contains

six activities boxing, hand waving, hand clapping, running, jogging and walking performed by 25 different subjects in four different scenarios: outdoors, outdoors with scaled version, outdoors with different clothes and indoors. All sequences were taken over homogeneous backgrounds with static camera of 25 frames per second. The sequences were then down sampled to 160*120 pixels and have a length of 4 seconds in average.

Weizmann human action dataset: The dataset contains a total of ten actions performed by nine people and uses static camera unlike that of the KTH dataset, where some videos had zooming and also simple background. This dataset contains ten activities, so it is used as attest approach in which the number of activities are increased.

PETS dataset: PETS datasets is a collection of data sets for different purpose of vision based research. The PETS run an evaluation framework on specific datasets with specific objective. The PETS2000 and PETS2001 datasets are designed for tracking outdoor people and vehicles. PETS2002 has indoor people tracking and hand posture classification data. PETSICVS2003 includes datasets for facial expressions, gaze and gesture/action. VS-PETS 2003 has outdoor people tracking.PETS ECCV2004 has number of video clips recorded for CAVIAR project. PETS2006 has surveillance data of public spaces and detection of left luggage events. PETS 2007 considers both volume crime and threat scenario. The datasets for PETS2009, PETS2010, and PETS 2012 consider crowd image analysis and include crowd count and density estimation tracking of individuals within a crowd and detection of separate flows and specific crowd events. INRIA XMAS multi-view dataset: It contains the actions captured from five viewpoints. A total of eleven persons perform 14 actions. The actions are performed in arbitrary direction from the camera direction with regard to camera setup.

Other datasets: Institution of automation, Chinese Academy of Sciences provides the CASIA gait database for Gait recognition. The Hollywood human action dataset contains eight actions which are extracted from movies and are performed by a variety of actors. The UCS sports action dataset contains 150 sequences of sport motions.

III. METHODS

A. Histogram of Gradients

Histogram of Oriented Gradients Before implementing a feature descriptor the first and foremost step to is to preprocess the images normalization in color and gamma values. But the implementation developed by Dalal and triggs does consist the normalization of the descriptor which is aimed at achieving the same. So, this step can be omitted in HOG descriptor computation. So from this we can say the image pre-processing.

The image which is supplied to the feature descriptor is subsampled into 64 by 128 detection window or a detection window with aspect ratio of 1:2. Then the detection window is further subsampled to form cells which constitute 8 by 8 pixels of the image. Each pixel within the cell casts a weighted vote for an orientation-based histogram channel based on the values found in the gradient computation. Let us now look into the details of gradient vector computation for each pixel in the cell. As shown in the figure below the difference of pixel values of pixels adjacent to the current pixel in x direction is computed and in a similar way the gradient in y direction is computed. The gradients in x and y direction is used to compute the magnitude and direction of the gradient at that pixel. This gradient vector now represents the change in pixel values. The difference in the pixel values can range from -255 to 255. Which is large range to store values in a byte. So we map the pixel values to the 0 to 255. When we carry out the pixels with large negative values will be represented as black, large positive values as white and which doesnt possess gradient will be left gray.

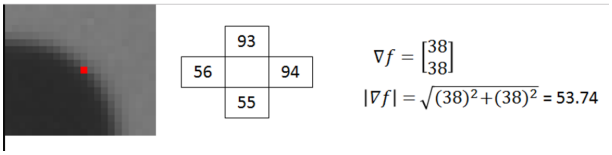

Figure1: the figure shows how the gradient vector is caliculted for ecah pixel value in a image and its representation.

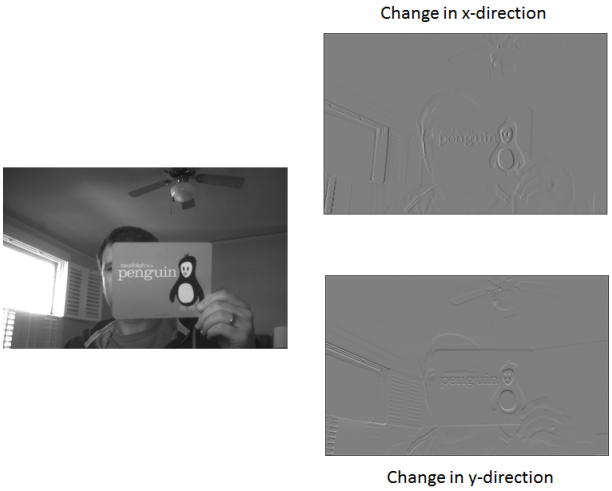

Figure 2: The figure shows the representation of pixels mapped to 0 to 255 values in white, gray and black

From the mapping above we can easily see how helpful are the gradient vectors for edge detection and feature extraction. The gradient vectors with large gradient values represent edges. So based on this we can understand how important gradient vactors are in image processing applications. Many applications which base their object detection on edge detection compute the gradient vectors and map them to arrive at edge features. Let us now see another such useful property of gradient vectors to changes in illumination. When a same image with different pixel intensity is computed for gradient vectors at different pixel values the gradient vector magnitude and direction remains the same. This is given the reason

that we compute the diffrence of pixel values which have same increase in illumination. This helps us to create a value which is invariant to illumintion changes. That helps for object detction applications to detect objects in changing lighting conditions without extra effort.

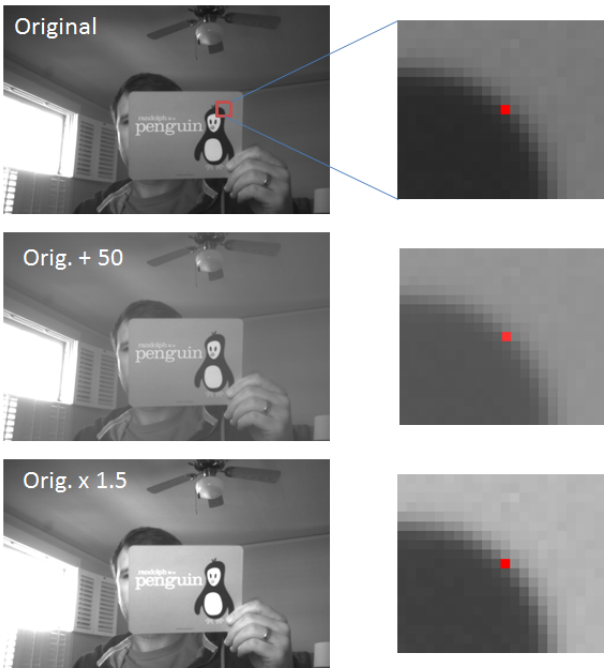

Figure 3: The figure shows the changes made to the original image

Now the gradient vectors obtained can be normalized to make them invariant to pixel density changes. The gradient vectors are by default normalized to illumination changes. But we have to normalize the gradient vectors to be scale invariant. In the above example the same image is subjected to illumination change and pixel multiplication. When we calculated the magnitude of gradient vector we can observe the case in which the pixels have a multiplication factor show a different magnitude values from the other cases. So to improvise the gradient vectors we normalize the vectors by dividing the vectors with their magnitudes. Which will make the gradient vectors invariant to pixel multiplications.

Now the normalized gradient vectors for pixels in the cell are used to form the histogram for each cell. The cells themselves 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. When unsigned gradients used in conjunction with 9 histogram channels performed best in their human detection experiments. As for the vote weight, pixel contribution can either be the gradient magnitude itself, or some function of the magnitude. In tests, the gradient magnitude itself generally produces the best results. Other options for the vote weight could include the square root or square of the gradient magnitude, or some clipped version of the magnitude.

The HOG descriptor is then the concatenated vector of the

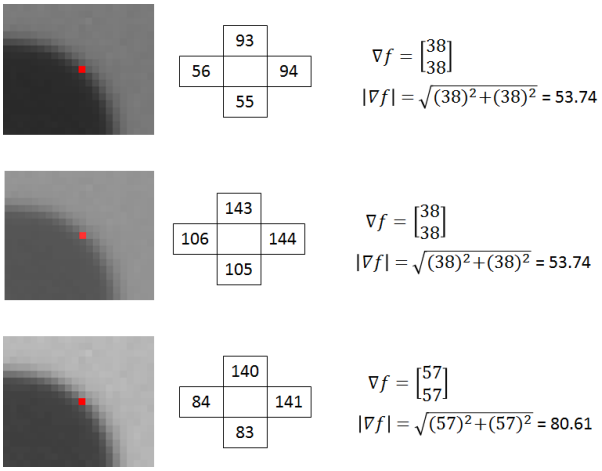

Figure 4: The figure shows the magnitudes of gradient vector for change in intensity, multiplication value

components of the normalized cell histograms from all of the block regions. Blocks are collection of cells, each block is representation of 2 by 2 cells. These blocks typically overlap, meaning that each cell contributes more than once to the final descriptor. The histogram of gradients used a 50 percent overlap between blocks. Two main block geometries exist: rectangular R-HOG blocks and circular C-HOG blocks. R-HOG blocks are generally square grids, represented by three parameters: the number of cells per block, the number of pixels per cell, and the number of channels per cell histogram. Based on experimentation the HOG method human detection experiment, the optimal parameters were found to be four 8x8 pixels cells per block (16x16 pixels per block) with 9 histogram channels.

Moreover, they found that some minor improvement in performance could be gained by applying a Gaussian spatial window within each block before tabulating histogram votes in order to weight pixels around the edge of the blocks less. The R-HOG blocks appear quite similar to the scale-invariant feature transform (SIFT) descriptors; however, despite their similar formation, R-HOG blocks are computed in dense grids at some single scale without orientation alignment, whereas SIFT descriptors are usually computed at sparse, scale-invariant key image points and are rotated to align orientation. In addition, the R-HOG blocks are used in conjunction to encode spatial form information, while SIFT descriptors are used singly. The feature vector is formed using the values from the bins from all the blocks in the detection window. By combining all the histogram values we create a feature descriptor which consists of 3780 values per detection window. Now the feature descriptor is fed to a trained SVM to detect for humans in the detection window. The same thing is repeated for each detection window.

B. Support vector Macines

Support vector machine is a machine learning algorithm that can be used for both classification and regression challenges.

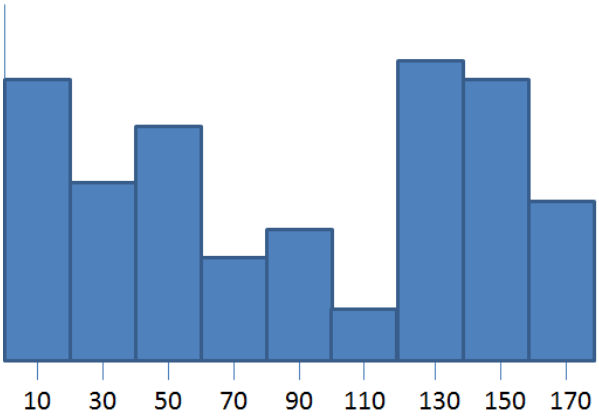

Figure 5: figure shows how the histogram values are plotted into bins.

The feature descriptor vector is plotted as a point on the support vector graph. The INRIA data set consists of training images and the test images. The training images further consists of positive images and negative sample images. The positive and negative samples are fed to the support vector machine to classify them into clusters. The support vector machine based on the trained data creates a hyper plane that separates the two classes. In most of the cases a number of hyper planes can be possible. But to avoid this we choose a hyper plane which has a largest marginal distance to most near data points in the graph. Thereby we arrive at the optimal hyper plane which separates the classes of data. This is how we create the linear SVM.

Figure 6: The figure shows the negative training images for INRIA dataset

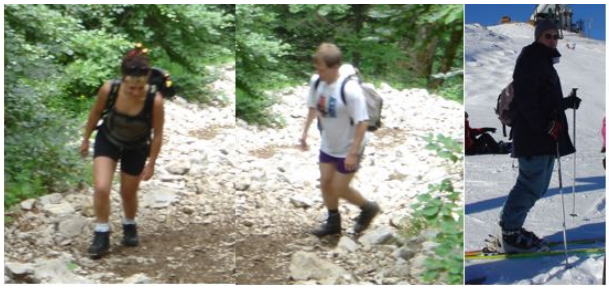

Figure 7: the figure shows the positive training images from INRIA dataset

To elaborate on the training methodology employed for this particular application, let us discuss in detail how we run a number of iterations to make the linear SVM more Robust. The images from the INRIA dataset with positive images (which have humans) and negative images are used for training the SVM initially and then the trained SVM is again used to classify the negative samples. Then the positives observed in the negative samples are again used to train the classifier (which are actually false positive). The same steps are repeated for a couple of times which will aid in making the algorithm more robust, efficient. This process of training the SVM on false positives is called as hard mining.

C. Non Maximum Suppression

The SVM which id trained to detect humans from images returns the positive samples of detection windows from the image. The samples which turned positive for human presence are combined to form a single window using the non-maximum suppression technique.

D. Infrared Sensor

In the far-infrared domain the image of an object depends on the amount of heat the object emits namely on its temperature. Generally, the temperature of a pedestrian is higher than the environment and a persons heat radiation is sufficiently high compared to the background. Such human shapes appear brighter than the background in infrared images easing the detection process. In fact, the developed approach exploits this feature to detect pedestrians. which will enable the with the capabilities to detect humans from far distance and night vision.

IV. Summary

Human detection using Histogram of oriented gradients (HOG) for object detection, non-maximum suppression for result convergence, and linear support vector machine (SVM) for object classification: In this method of implementation we will sample P positive samples from your training data of the object you want to detect and extract the HOG descriptors. And in the same way sample N samples from training data of the objects you will not want to detect. Then train the linear SVM method on the positive and negative samples.

Then we find out the false positives out of the image by scanning each image and all the scale variants of the image using window scanning method. At each window of the image you will be extracting HOG descriptors and will be fed to classifier. If the classifier classifies incorrectly the window record the feature vector of the false positive patch of image along with the probability of classification. Then train the SVM with the false positive feature vectors by their confidence.

Now the trained SVM will be able to classify the human subjects in the image. But the problem is the trained SVM results in multiple windows of the object. So now we use the non maximum suppression to combine all the windows generated with possibility of human objects.

V. Results

The test images from the INRIA data set are tested to check the robustness of the object detection algorithm developed. The detection algorithm is able to do the detection of subjects from

images in maximum of the cases. The extra tool developed that runs parallel with the vision system, uses the data from the far infrared camera. Based on the temperature threshold the human subjects are classified from the group of the objects. The figures below show the working of the detection algorithm both as a vision sensor and infrared sensor.

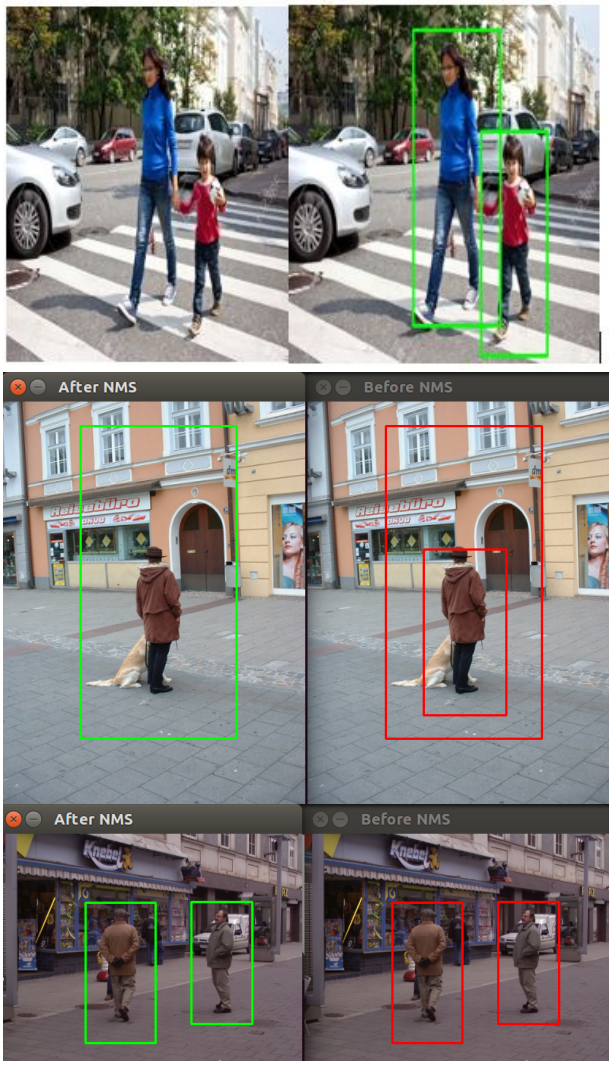

Figure 7: The figure shows the objects detected from images

Figure 8: The figure shows the detected human subjects from infrared images

REFERENCES

[1] Histogram of oriented Gradients Navneet dalal and Bill triggs

[2] C Stauffer, W Grimson, Adaptive background mixture models for real-time tracking, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR 1999) (IEEE, Piscataway, 1999), pp. 246252

[3] YL Tian, RS Feris, H Liu, A Hampapur, M-T Sun, Robust detection of abandoned and removed objects in complex surveillance videos. Syst. Man Cybern. Part C Appl. Rev. IEEE Trans. 41(5), 565576 (2011)

[4] DS Lee, Effective Gaussian mixture learning for video background subtraction. IEEE Trans. Pattern Anal. Mach. Intell. 27(5), 827835 (2005)

[5] A Shimada, D Arita, Dynamic control of adaptive mixture-of-Gaussians background model, in IEEE International Conference on Video and Signal Based Surveillance (AVSS'06) (IEEE, Piscataway, 2006), p. 5

[6] J Wang, G Bebis, R Miller, Robust video-based surveillance by integrating target detection with tracking, in IEEE Computer Vision and Pattern Recognition Workshop (CVPRW '06) (IEEE, Piscataway, 2006), p. 137

[7] C Ridder, O Munkelt, and H. Kirchner, Adaptive Background Estimation and Foreground Detection Using Kalman-Filtering, Proc. Int'l Conf. Recent Advances in Mechatronics, ICRAM 95, pp. 193199 (1995)

[8] B Stenger, V Ramesh, N Paragios, F Coetzee, JM Buhmann, Topology free hidden Markov models: application to background modeling, in IEEE International Conference on Computer Vision (ICCV 2001) (IEEE, Piscataway, 2001), pp. 294301

[9] S Jabri, Z Duric, H Wechsler, A Rosenfeld, Detection and location of people in video images using adaptive fusion of color and edge information, in 15th International Conference on Pattern Recognition (ICPR2000) (IEEE, Piscataway, 2000), pp. 627630

[10] W Zhang, X Zhong, FY Xu, Detection of moving cast shadows using image orthogonal transform, in 18th International Conference on Pattern Recognition (ICPR'06) (IEEE, Piscataway, 2006), pp. 626629

[11] M Heikkil, M Pietikinen, A texture-based method for modeling the background and detecting moving objects. IEEE Trans. Pattern Anal. Mach. Intell. 28, 657662 (2006)

[12] W Kim, C Kim, Background subtraction for dynamic texture scenes using fuzzy color histograms. Signal Process. Lett. IEEE 19(3), 127130 (2012)

[13] A Elgammal, D Harwood, L Davis, Non-parametric model for background subtraction, in 6th European Conference on Computer Vision - Part II (ECCV '00) (Springer, London, 2000), pp. 751767

[14] A Elgammal, R Duraiswami, L Davis, Efficient kernel density estimation using the fast Gauss transform with applications to color modeling and tracking. IEEE Trans. Pattern Anal. Mach. Intell. 25, 1499 (2003)

[15] B Han, D Comaniciu, L Davis, Sequential kernel density approximation through mode propagation: Applications to background modeling, in Asian Conference on Computer Vision Jeju Island, Korea (2004)

[16] Person following and gesture recognition with a quadcopter Tayyab Naseer, Jurgen Sturm, and Daniel Cremers.

[17] A Real-Time Method to Detect and Track Moving Objects (DATMO) from Unmanned Aerial Vehicles (UAVs) Using a Single Camera Gonzalo R. Rodrguez-Canosa, Stephen Thomas 2, Jaime del Cerro, Antonio Barrientos and Bruce MacDonald 2.

[18] Skin detection of various animation characters Kazi Tanvir Ahmed Siddiqui and abu Wasif.