https://doi.org/10.2352/ISSN.2470-1173.2018.10.IMAWM-336

Depth and Superpixel Extraction for Augmenting Human Detection

Hussin K. Ragb, Theus H. Aspiras, and Vijayan K. Asari; Vision lab, University of Dayton; Dayton, Ohio/USA

Abstract

Various human detection algorithms are limited in capability due to the lack of using other supplemental algorithms for enhancing detection. We propose using two different algorithms to extract vital information to augment human detection algorithms for increased accuracy. The first algorithm is the computation of depth information. Information needed to obtain depth is based on the specific location of the camera based from frame to frame. Most calibrated stereo cameras can develop accurate depth information, but the motion that takes place from frame to frame can be utilized for developing rough depth perception of the objects in the scene. Block-matching and optical flow algorithms can be used to provide these disparities that happen in the image, which will provide depth information for the human detection algorithm. The second algorithm is superpixel segmentation. This algorithm determines a rough over-segmentation of the imagery, which well defines the boundaries as larger pixels that are within the imagery. This information can be used to distinguish background and foreground information to create a specific segmentation around the human detection, rather than a bounding box detection that may include various background information. The fusion of these algorithms with human detection has been shown to increase detection accuracy and providing better localization of the human in the imagery.

Introduction

Human detection is one of the widely-used applications in the pattern recognition and computer vision systems. Over the last decade, detection of human beings in a visual surveillance system is a significant task due to its extended applications including human computer interaction, person identification, event detection, counting people in crowded regions, gender classification, automatic navigation, safety systems, etc. Many single feature extraction algorithms are proposed for depicting and describing the human appearance. One of the earliest algorithms used for human detection system is proposed by Papageorgiou et al [1]. This technique used sliding window detector and Harr-like features [2] for describing the person. Shape features such as Edgelet [3], Shapelet [4], and Histogram of Oriented Histogram (HOG) [5] are other feature extraction algorithms proposed for human descriptor. In addition, texture features such as Local Binary Pattern (LBP) [6], as well as color features like color-self-similarity [7], and color histograms are also applied for human detection tasks. Multi-feature descriptors are also proposed for improving the human detection performance [8]. Wojek and Schiele [9] fused HOG, Haar-like features, shapelets and shape context in one descriptor to improve the detection performance. Wang et al [10] combined HOG and LBP features and used the linear Support Vector Machine (SVM) to train the human detector. Zhang and Ram [11] improved the detection of the IR images by the combination of the Edglets and HOG features. Dollar et al [12] developed the integral channel features (ICF) that combined HOG, gradient magnitude and LUV color, etc. The outputs of all these algorithms are represented as bounding boxes surrounding the humans in the scene. These bounding boxes give only the general region where the person is located. The information inside the bounding box may include too much background which decreases the quality of detection. As the foreground in the bounding box is separated from the background as the detection performance increased. In this paper we propose a human detection algorithm enhanced by supplemental algorithm based on depth and super-pixel techniques for increased accuracy. In this approach, a vital information of the true positive detection bounding box is extracted and then separated from the background as shown in Figure 1.

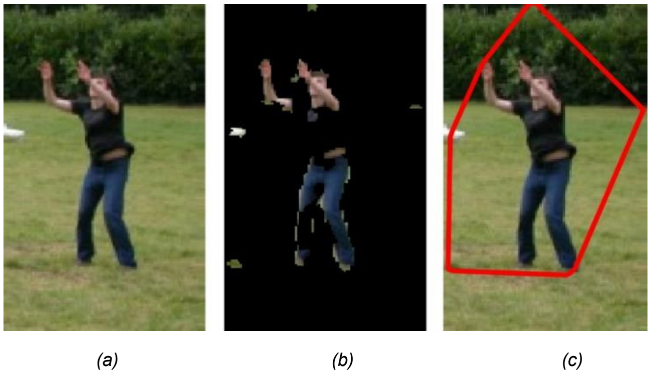


Figure 1. (a) Original detected image. (b) Foreground detection based on depth and super-pixel segmentation. (c) augmented human detection.

The feature extraction algorithm of the human detection system presented in this paper is consisted of the following information: the image gradients, the local phase features based on phase congruency, the phase congruency magnitude, and color features of the input image. The framework of the augmented human detection system based on the proposed descriptor and the superpixel extraction approach is shown in Figure 2. The phase congruency magnitudes and orientations as well as the gradients of the input image are computed for each pixel in the input image with respect to its neighborhood. The resultant images are divided into local overlapped regions called blocks of the size 16×16 pixels. The block region is formed from 2×2 sub-local regions called cells, where each cell is 8×8 pixels. The histogram of oriented phase (HOP) [17] and the histogram of oriented gradient (HOG) [5] are determined for each cell region. These histograms are fused together to form the HOP and HOG features for each block region of the input image. The same is done for the rest of blocks to form the HOP and HOG features for the entire image. A maximum pooling of the candidate features is randomly generated for a one channel of the phase congruency magnitude and the same is done for the three LUV color channels. These features are fused with HOP and HOG features to form the proposed descriptor. This descriptor is fed to an Adaboost classifier (depth two decision tree)

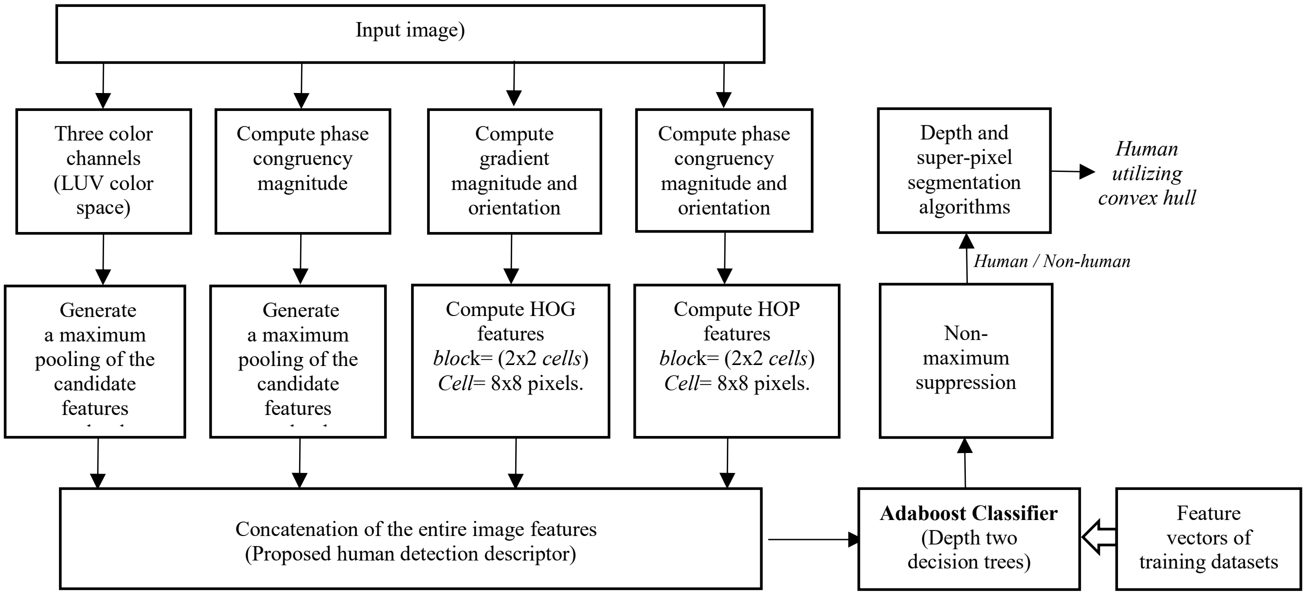

Input image)
Three color channels (LUV color space)
Compute phase congruency magnitude
Compute gradient magnitude and orientation
Compute phase congruency magnitude and orientation
Depth and super-pixel segmentation algorithms
Human utilizing convex hull
Human / Non-human
Generate a maximum pooling of the candidate features
Generate a maximum pooling of the candidate features
Compute HOG features block= (2x2 cells) Cell= 8x8 pixels.
Compute HOP features block= (2x2 cells) Cell= 8x8 pixels.
Non-maximum suppression
Concatenation of the entire image features (Proposed human detection descriptor)
Adaboost Classifier (Depth two decision trees)
Feature vectors of training datasets

Figure 2. Framework of the augmented human detection system based the proposed descriptor and superpixel extraction algorithm.

to select the strongest features and classify between the classes. The proposed human detection system is implemented based on the scanning window approach that used to detect the presence of humans in an image. This method returns a set of detection window and the detection overlap that may occur due to sliding window is resolved using non-maximum suppression [8], [21].

Depth information and image segmentation can help to augment the human detection algorithm. Depth algorithm can provide information about scale and relative size of the detections. Superpixel Segmentation can provide the enclosed region of the person without much background information. Fusion of the human detection algorithm with the depth and superpixel segmentation will provide better localization of humans in various scenes.

Image Gradient computation

Image Gradient is defined as the directional change in the color or an image intensity. The horizontal gradients $G_x(x, y)$ can be obtained simply by convolving the input image with the mask templet $(-1\ 0\ 1)$ and vertically $G_y(x, y)$ by convolving the image with $(-1\ 0\ 1)^T$[19].

The gradient magnitude $G(x, y)$ and the orientation $\phi(x, y)$ for the image $I(x, y)$ is computed as following [19], 20];

$$G_x(x, y) = I(x + 1\,, y) - I(x - 1\,, y) \tag{1}$$

$$G_y(x, y) = I(x\,, y + 1) - I(x\,, y - 1) \tag{2}$$

$$G(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \tag{3}$$

$$\phi(x, y) = \tan^{-1}\left(\frac{G_y(x,y)}{G_x(x,y)}\right) \tag{4}$$

Phase Congruency computation

Phase congruency is an algorithm that was developed to localize the edges and corners of the image. Oppenheim and Lim [13] [14] have shown that the most significant information within an image is provided by the phase rather than amplitude [15]. Phase congruency provides a measure that is independent of the overall magnitude of the signal making it invariant to illumination and contrast variations [16]. The phase congruency function in terms of the Fourier series expansion of a signal at some location x is given as [20], [18], [8]:

$$PC(x) = max_{\bar{\phi}(x)\epsilon[0,2\pi]} \frac{\sum_n A_n \cos(\,\phi_n(x) - \bar{\phi}(x))}{\sum_n A_n} \tag{5}$$

where, A_n is the amplitude of the nth Fourier component [18]. $\phi_n(x)$ represents the local phase of the Fourier component. $\bar{\phi}(x)$ is the mean local phase angle of all the Fourier terms being considered at the point.

An alternative and interpretation of phase congruency, Venkatesh and Owens [23] show that local energy is equal to phase congruency scaled by the sum of the Fourier amplitudes and given as [22], [18], [16], [24], [8]:

$$E(x) = PC(x) \sum_n A_n \tag{6}$$

For a one-dimensional input signal $I(x)$, the local energy function $E(x)$ can be defined as [16], [25], [19], [8]:

$$E(x) = \sqrt{F(x)^2 + F_H(x)^2} \tag{7}$$

where, $F(x)$ is the input signal filtered from a DC component and $F_H(x)$ is Hilbert Transform (90° phase shift of $F(x)$).

The relationship between the local energy, phase congruency and the sum of the Fourier amplitudes can be seen geometrically in Figure 3. To compute the phase congruency, we should first extract the local frequencies and phase information by convolving the input image with a pair of quadrature filters. Log-Gabor filter is an efficient bandpass filter used in this paper to extract the local phase information spread over a wide spectrum [17]. The transfer function of the log-Gabor filter is given by [15], [25], [19]:

$$G(\omega,\theta) \ = exp\left(\frac{-\left(log(\omega/\omega_o)\right)^2}{2\left(log(k/\omega_o)\right)^2}\right) exp\left(\frac{-(\theta-\theta_o)^2}{2{\sigma_\theta}^2}\right) \qquad (8)$$

where ω_o is the center frequency of the filter. k/ω_o is kept constant for various ω_o [26], [27]. θ_o is the center orientation of the filter, and σ_θ is the standard deviation of the Gaussian function in angular direction [16], [26], [25], [19].

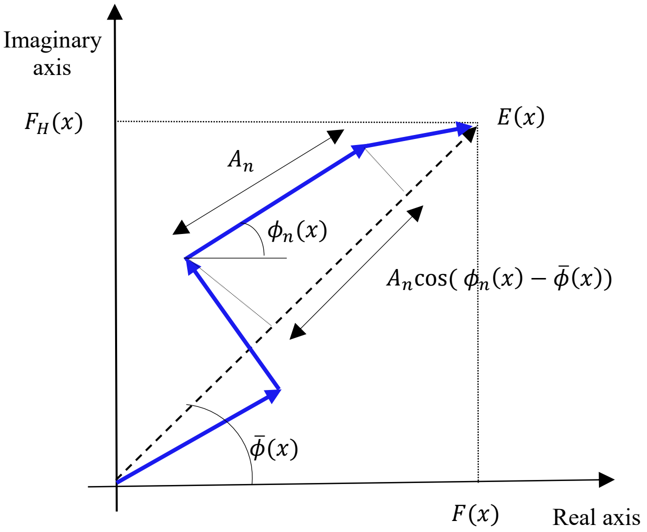


Figure3. The relationship between phase congruency, local energy and the sum of the Fourier amplitudes.

Consider M_{no}^{o} and M_{ne}^{e} are the odd symmetric and even symmetric components that represent the quadrature pair of the Log-Gabor filter at scale n and orientation o. The response vector at scale n and orientation o is obtained by the convolution of each quadrature pair with the input signal $I(x,y)$ and is given by [25], [19], [24], [8]:

$$[e_{no}(x,y)\,,o_{no}(x,y)] = [I(x,y)*M_{no}^{e}\ ,I(x,y)*M_{no}^{o}] \qquad 9)$$

the response A_{no} and the phase angle ψ_{no} at scale n and orientation o are given by:

$$A_{no} = \sqrt{(e_{no}^2\ (x,y)+o_{no}^2\ (x,y))} \qquad (10)$$

$$\psi_{no}(x,y) = tan^{-1}\left(\frac{o_n(x,y)}{e_n(x,y)}\right) \qquad (11)$$

$F(x,y)$ and $F_H(x,y)$ for a 2D signal are given by [17], [24]:

$$F(x,y) = \ \textstyle\sum_o\sum_n e_{no}(x,y) \qquad (12)$$

$$F_H(x,y) = \ \textstyle\sum_o\sum_n o_{no}(x,y) \qquad (13)$$

From Eq. (7), the energy of the two-dimensional signal is computed as;

$$E(x,y) = \sqrt{F(x,y)^2+F_H(x,y)^2} \qquad (14)$$

Therefore, phase congruency $PC(x,y)$ of the 2D signal is given as [25], [19], [17], [24], [8]:

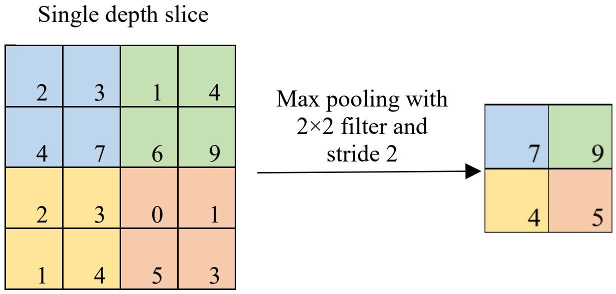
$$PC(x,y) = \frac{\sum_o\sqrt{(\sum_n e_{no}(x,y))^2+(\sum_n o_{no}(x,y))^2}}{\varepsilon+\sum_o\sum_n A_{no}(x,y)} \qquad (15)$$

The orientation $\varphi(x,y)$ is given by:

$$\varphi(x,y) = tan^{-1}\left(\frac{\sum_o\sum_n o_{no}(x,y)}{\sum_o\sum_n e_{no}(x,y)}\right) \qquad (16)$$

Maximum pooling

Max pooling is a sample-based discretization process that used to down-sample an input image representation. Max pooling is done by applying a *maxfilter* to non-overlapping sub-regions of the initial input image. With max pooling, size of the resultant image gets reduced and retaining the image information. For a 4×4 matrix representation, the output of the max pooling is 2×2 as shown in Figure 4.



2	3	1	4
4	7	6	9
2	3	0	1
1	4	5	3

7	9
4	5

Figure 4. max pooling of 4×4 matrix representation.

Depth

Depth estimation from a single image is an important component of many vision systems, including robot navigation, motion capture and video surveillance. It provides a three-dimensional representation for giving the system a reference for distance and scale. The algorithm we have chosen to use is the structure forest framework to infer depth information (Fang et al. 2016) [28]. The algorithm exploits the structural properties that are exhibited in local patches of the depth map. It then provides a structured learning framework based on random decision forests to determine the depth map.

To create the input depth features for the algorithm, the features are Saxena, Chung, and Ng (SCN) [29] features, who developed features in various scales to determine depth. The algorithm obtains 9 Laws [30] masks, 2 color channels (YCbCr), and 6 oriented edges (3x3 patch) for developing the feature set. One specific channel that can be used for depth is the dark channel, which utilizes the assumption that most local patches contain some pixels with very low intensities. These low intensities are a great cue for determining depth. The entire algorithm uses color features and neighborhood features to differentiate depth in the image and computed as:

$$J^{dark} = \min_{c\in\{r,g,b\}}\left(\min_{y\in\Omega(x)}(J^c(y))\right) \qquad (17)$$

where, J is the color channel, and Ω is the neighborhood image at pixel location x.

The structured random forest used for depth information recursively splits data down to a decision tree to reach a leaf node. Allowing multiple independent decision trees creates better generalizability of the data. Also, random subsampling of data allows a diversity of these trees, providing better classification ability. Figure 5, and Figure 6 show the depth information obtained from the image in many different environments. The heat map shows the depth information based on all the relevant features.

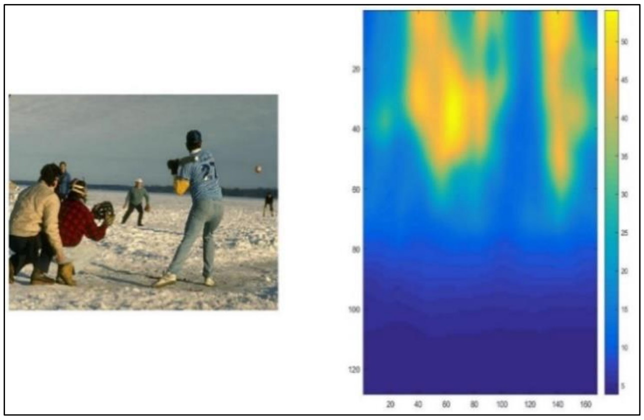


Figure 5. Depth information and the heat map obtained from the image.

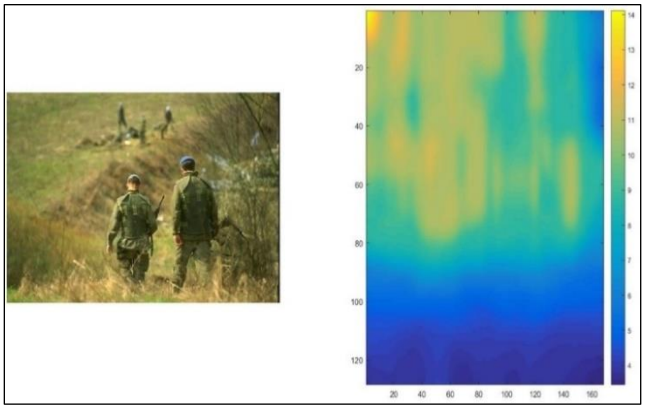


Figure 6. Depth information and the heat map obtained from the image.

Superpixel Segmentation

Many human detection algorithms create a bounding box detection of the human in the scene. These algorithms provide an accurate detection of the person but does include much of the background. By utilizing a segmentation algorithm, we should be able to provide a better detection of the human, which gives a contour around only the human in the detection. The information that we know about a detection of a human is that there are clusters of image pixels that correspond to the human, which are usually centered in the detection box. Therefore, we can utilize a super-pixel segmentation algorithm to aid in the segmentation of the human in the detection box. For super-pixel segmentation, similar image pixels can usually be grouped in a surrounding neighborhood, limited by edge information. Though this will over-segment the image regions, it will keep true the necessary contour information for the image, especially for human detection. The algorithm used for the human detection segmentation is Simple Linear Iterative Clustering (SLIC), which uses K-means clustering approach with a distance measure to create pixel groups that adheres to natural boundaries in the image. Like mean shift, each pixel is associated to a feature vector which is characterized by K-means. SLIC then moves the centroid created by K-means away from edges by using the Lloyd algorithm.

Once the superpixels are created, we can group all of the superpixels as human and background, thus creating a segmented output, which is a better representation than a bounding box representation. To provide human detection segmentation we must gather the boundary information from the image, which uses the assumption of foreground superpixels near the center of the detection bounding box and background superpixels near the outside of the detection bounding box. We then develop a feature set for each superpixel and determine the match distance between superpixels, which classifies the pixels for foreground/background. Once the distances are thresholded for classification, we can remove the background superpixel and create a convex hull around the foreground pixels, thus creating a better segmentation output. Figure 7 shows the superpixel human detection pipeline. The superpixels are found in the imagery and matched to the corresponding background and foreground patches. Small unconnected superpixels are removed due to no relevance to the foreground object. A convex hull is placed over the group of superpixels, creating a segmented object. Figure 8, Figure 9, and Figure 10 show other examples of the superpixel human detection results. It illustrates the ability of the superpixel algorithm to detect foreground information in various resolutions.

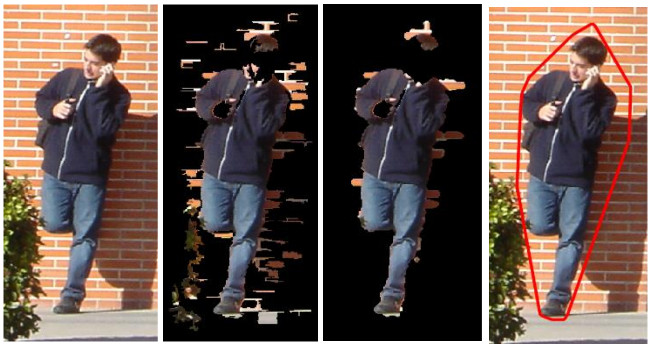


Figure 7. The super pixel human detection pipeline. The superpixels are found in the imagery and matched to corresponding background and foreground patches. Small unconnected superpixels are removed due to no relevance to the foreground object. A convex hull is placed over the group of superpixels, creating a segmented object.

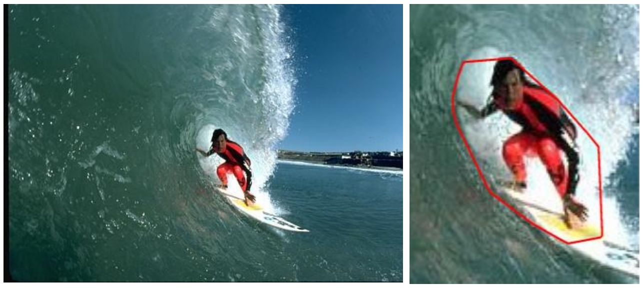


Figure 8. Human detection results augmented by the proposed depth and superpixel algorithms.

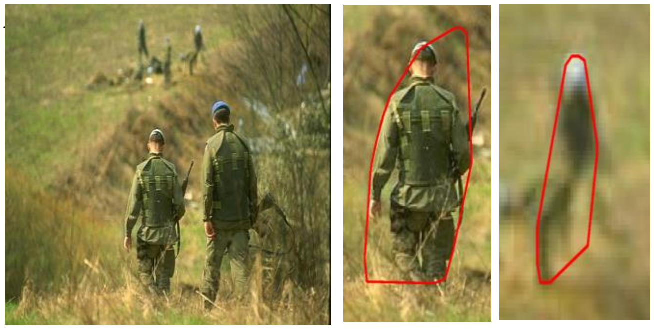


Figure 9. Superpixel human detection results. It shows ability of the superpixel algorithm to detect foreground information in various resolutions.

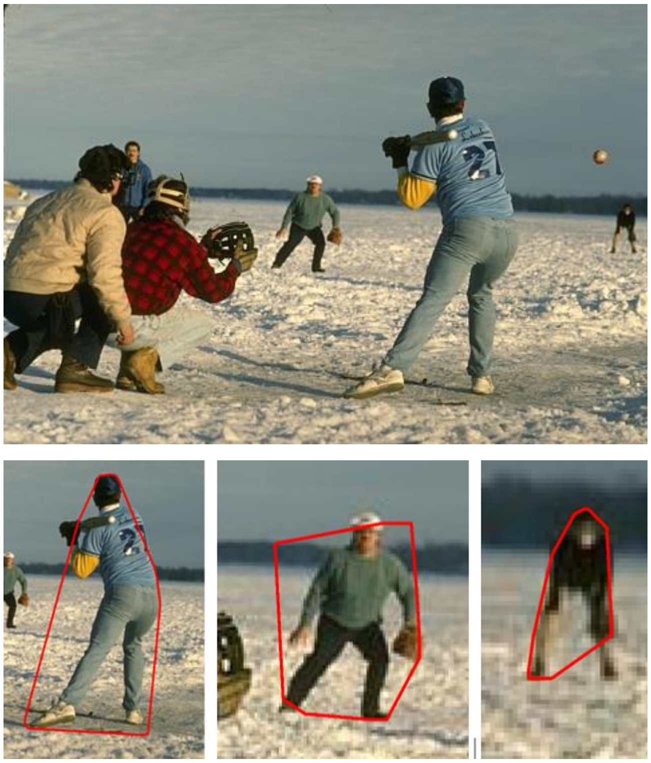


Figure 10 Superpixel human detection results. It shows ability of the superpixel algorithm to detect foreground information in various resolutions.

Conclusion

In this paper we presented an augmented human detection system. The feature extraction algorithm used in this system are based on the gradient, local phase, and color information. Adaboost classifier (depth of two decision tree) is used to select the strongest features to be used for training and the classification between the classes. The detection accuracy of this system is augmented by using the depth and super-pixel approaches. The experiments conducted based on the proposed approach showed an optimistic result and provided better localization of the human region in the imagery.

References

[1] C. Papageorgiou and T. Poggio, "A trainable system for object detection," International Journal of Computer Vision, vol. 38, no. 1, pp. 15-33, 2000.

[2] P. Viola, M. Jones, "Rapid object detection using a boosted cascade of simple features," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 1 (2001) 511-518.

[3] B. Wu, R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," Proceedings of IEEE International Conference on Computer Vision 1 (2005) 90-97.

[4] P. Sabzmeydani, G. Mori, "Detecting pedestrians by learning shapelet features," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2007) 1-8.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," In IEEE Conf. Computer Vision and Pattern Recognition (CVPR), vol. 1, pp 886–893, 2005.

[6] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray- scale and rotation invariant texture classification with local binary Patterns,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971–987, Jul. 2002.

[7] H. Liu , T. Xu , X. Wang, Y. QianA "Novel Multi-Feature Descriptor for Human Detection Using Cascaded Classifiers in Static Images", 31 December 2014# Springer Science+Business Media New York 2014.

[8] H. Ragb and V. Asari, "Multi-feature fusion and PCA-based approach for efficient human detection," IEEE Computer Society Workshop on Applied Imagery and Pattern Recognition - AIPR 2016: Imaging and Artificial Intelligence: Washington DC, Oct. 2016. (IEEE AIPR).

[9] C. Wojek and B. Schiele, "A performance evaluation of single and multi-feature people detection," Proceedings of DAGM Symposium on Pattern Recognition (2008) 82-91.

[10] X. Wang, T. Han, and S. Yan, "An HOG-LBP human detection with partial occlusion handling," proceedings of IEEE International Conference on Computer Vision (2009) 2-29.

[11] Li Zhang, Bo Wu and Ram Nevatia, "Pedestrian Detection in Infrared Images Based on Local Shape Features," In CVPR, June 2007.

[12] P. Dollar, Z. Tu, P. Perona, S. Belongie, "Integral channel features," Proceedingsof British Machine Vision Conference (2009) 1-11.

[13] A.V. Oppenheim, J.S. Lim, "The importance of phase in signals," Proc. IEEE 69 (5), 529–541. 1981.

[14] A. Oppenheim, V., Lim, J.S., Kopec, G.E., Pohlig, S.C., 1979. Phase in speech and pictures. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, April, pp. 632–637.

[15] X. Yuan, P. Shi,"Iris Feature Extraction Using 2D Phase Congruency," Proceedings of the Third International Conference on Information Technology and Applications (ICITA’05), 0-7695-2316-1/05.

[16] S. Gundimada and V. Asari, "A Novel Neighborhood Defined Feature Selection on Phase Congruency Images for Recognition of Faces with Extreme Variations", International Journal of Information Technology Volume 3 Number 1, pp. 2074-2080, 2007.

[17] H. Ragb and V. Asari, "Histogram of oriented phase (HOP): a new descriptor based on phase congruency," SPIE Conference on Commercial + Scientific Sensing and Imaging: Mobile Multimedia/Image Processing, Security, and Applications 2016, Baltimore, MD, USA, 2016.

[18] P. Kovesi, "Image Features from Phase Congruency," Journal of Computer Vision Research, summer 1999, Volume 1, Number 3.

[19] H. Ragb, V. Asari, "Histograms of oriented phase and gradient (HOPG) descriptor for improved pedestrian detection," IS&T International

Conference on Electronic Imaging: Video Surveillance and Transportation Imaging Applications, San Francisco, Feb. 14, 2016.

[20] P. Kovesi, "A dimensionless measure of edge significance." In the Australian Pattern Recognition Society, Conference on Digital Image Computing: Techniques and Applications, pages 281–288, Melbourne, 4-6 December 1991.

[21] H. Ragb, V. Asari, " A feature fusion strategy for human detection in omnidirectional camera imagery," IS&T International Conference on Electronic Imaging and Multimedia Analytics in a Web and Mobile World, San Francisco, 2018.

[22] R. A. Owens, "Feature-free images," Pattern Recognition Letters, 15:35–44, 1994.

[23] S. Venkatesh and R. Owens, "On the classification of image features," Pattern Recognition Letters, 11:339–349, 1990.

[24] H. Ragb and V. Asari, "Color and local phase-based descriptor for human detection," *National Aerospace & Electronics Conference & Ohio Innovation Summit (NAECON-OIS*), Dayton, 26 - 29 July 2016.

[25] H. Ragb, V. Asari, "Fused structure and texture (FST) features for improved pedestrian detection," International Journal of Computer and Information Engineering, (World Academy of Science, Engineering and Technology), vol. 3, no. 1, 2016.

[26] G. Basavaraj, G. Reddy, "An Improved Face Recognition Using Neighborhood Defined Modular Phase Congruency Based Kernel PCA," (IJERA) ISSN: 2248-9622 Vol. 2, Issue 2, Mar-Apr 2012, pp.528-535.

[27] T. Nguyen, S. Kim, I. Seop, "Fast Pedestrian Detection Using Histogram of Oriented Gradients and Principal Components Analysis," International Journal of Contents, Vol.9, No.3, Sep 2013.

[28] S. Fang, R. Jin, and Y. Cao, "Fast depth estimation from single image using structured forest," IEEE International Conference on Image Processing (ICIP), 2016.

[29] A. Saxena · S. H. Chung, and A. Y. Ng, "3-D Depth Reconstruction from a Single Still Image," International Journal of Computer Vision (IJCV), 76:53-69, 2008.

[30] S. Dash and U. R. Jena, "Multi-resolution Laws' Masks based texture classification," Journal of Applied Research and Technology, Volume 15, Issue 6, December 2017, Pages 571-582.

Author Biography

Hussin K. Ragb received his BS in Electrical and Electronic Engineering from the University of Tripoli (1991). Since then he has worked in the optical research center in, Tripoli-Libya. His work has focused on the development of the laser-based systems. He received his MS in Electrical Engineering from the University of Belgrade (2000). Since then he worked as a lecturer at the University of Tripoli. Currently, he is a Ph.D. student in Vision Lab in the University of Dayton-USA.

Dr. Theus Aspiras is a Research Engineer in the Electrical and Computer Engineering Department at the University of Dayton. He is currently working under Dr. Vijayan Asari in the University of Dayton Vision Lab Center of Excellence for Computer Vision and Wide Area Surveillance Research and has worked with Vision Lab for 6 1/2 years. Dr. Aspiras received his bachelor's degree in electrical and computer engineering from Old Dominion University in 2009, his Master's degree in electrical engineering from University of Dayton in 2012, and his Doctoral degree from University of Dayton in 2015. He has published 2 book chapters, 2 journals, and 12 conference papers. His current research areas include object detection and tracking, brain signal analysis, machine learning, and neural networks. Dr. Aspiras was awarded with best paper awards in International Conference on Information Processing 2011 and 2012 in India, International Conference on Information, Communications and Signal Processing 2011 in Singapore, and Applied Imagery Pattern Recognition in 2015.

Vijayan K. Asari is a Professor and Endowed Chair in electrical and computer engineering at the University of Dayton, USA. He is the Director of the Vision Lab at UD. Dr. Asari received his Ph.D. in Electrical Engineering from the Indian Institute of Technology, Madras. He holds three patents and has published more than 500 papers, including 97 journal papers. Dr. Asari is a Member of IS&T and a Senior Member of IEEE and SPIE.