Computationally Efficient Navigation System for Unmanned Ground Vehicles

Peyman Moghadam[1], Saba Salehi[2], Wijerupage Sardha Wijesoma[1]
[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
peym.mo@gmail.com

Abstract—**This paper proposes to enhance the existing methods of Self-Supervised Learning (SSL) with application to autonomous navigation systems through efficient computational approaches that are the principal requirements in a practical system. First, confidence-based auto labeling for self-supervised learning is introduced which identifies and eliminates the input samples with low confidence level that are susceptible to be mislabeled. Then, a biologically inspired saliency detection approach for feature biasing is presented which is able to detect the salient features through top-down task specific guidance. The proposed methods are general and can be applied to a variety of applications. Finally, experimental results on real datasets from the DARPA-LAGR program are given to illustrate the effectiveness of the proposed approaches.**

Keywords-self-supervised learning; feature selective attention; Support Vector Machine (SVM).

I. INTRODUCTION

Autonomous vehicle navigation for Unmanned Ground Vehicles (UGVs) in outdoor unstructured environments have been studied for several decades for a variety of applications such as planetary exploration [1], military [2], and hazardous areas due to natural or unnatural events. Fundamentally, the objective is to have an autonomous vehicle that can move reliably at a desired speed from a starting location to a goal point while it avoids obstacles. This scenario in outdoor unstructured environments leads to many difficulties such as different ground cover variation, uncontrolled lighting, weather conditions, and shadows [3].

The autonomous navigation systems are often facilitated with color stereo vision systems to sense their environments in order to predict the traversability of the terrain (i.e. to discriminate traversable path and non-traversable path regions). Color stereo vision cameras can provide not only 3D perception of terrain geometry but also color and texture information. However, the problem with these stereo vision systems is the limited cameras' field of view and maximum range [4]. As a result, they can only provide accurate geometry information from their near field. This limitation of near-sightedness of typical off-the-shelf stereo vision systems can be overcome by associating the geometry of terrain segments close to the vehicle (near field) with their visual appearance attributes (e.g. color and/or texture) and use this association to segment terrain and obstacles in the far field. However, finding a global correlation between terrain geometry characteristics and appearances that can be broadly applied is formidable due to the complex variability of appearance with terrain geometry, lighting, shadows, and weather conditions, especially in outdoor unstructured environments. This problem was addressed by several research groups in the DARPA-LAGR program by exploiting self-supervised, near-to-far learning algorithms [4].

The self-supervised, near-to-far learning refers to an approach of generating near field traversability labels (road and non-road) from each incoming pair of stereo images using only the near field stereo information. Next, these labels are associated with the visual appearance features extracted from the image to train a classifier, and the model obtained is used to classify the far field terrain well beyond the stereo range (in this case greater than 7 meter). Subsequently, this model is discarded and therefore, one full training and classification cycle are completed on every incoming pair of stereo images [4]. Therefore, self-supervised, near-to-far learning algorithms enable the autonomous vehicle to adapt to any unknown environments concurrently without using any hand-labeled training data. However, the major limitation of the current algorithms is that they generally assume that the information in the training data is noise free. As the near field traversability labels are provided by stereo supervision autonomously, the class labels can be contaminated with mislabeling errors. The classifier performance degrades significantly when this error rate increases. Therefore, it is essential to address such mislabeling errors in order to have an efficient learning method.

In addition, the system performance criteria including classification accuracy, learning convergence, and computational efficiency are highly dependent to feature selection step. The current self-supervised, near-to-far learning algorithms assume that all the visual features extracted from the input images (such as intensity, color, or texture) have the same level of prominence. Generally, a set of features is extracted from the input image and they are added to a feature vector with the same or different predefined weights for each pixel location of the input image. However, in most outdoor and unstructured environments, one or some of the attributes of the feature vectors may be more informative for discrimination of the terrain from non-terrain regions. Thus, identification of those

978-1-61284-481-7/11/$26.00 ©2011 IEEE

important features that have intuitive physical interpretation is another critical requirement [5].

To overcome the shortcomings of the current self-supervised, near-to-far learning algorithms, we propose two approaches. First, we present a confidence-based auto labeling module that explores the training data and identifies the examples that are susceptive to be mislabeled. Second, a biologically inspired saliency detection approach through task specific guidance is proposed to detect the most dominant features in the input images.

The rest of the paper is organized as follows: Section 2 introduces a basic auto labeling step and it is followed by our proposed confidence-based auto labeling method, which identifies and eliminates the noisy mislabeled data. Next, in section 3, for the input samples, the optimal features are extracted and then weighted through our new proposed technique, which is based on features' saliency. Section 4 presents SVM learning method that trains the autonomous navigation systems efficiently and section 5 provides the experimental results. Finally, section 6 concludes the paper.

II. AUTO LABELING

A. Basic Auto Labeling

In basic auto labeling method, first a pair of stereo color images is used as an input to the auto labeling module that performs stereo matching technique to create a disparity image (disparity value is inversely proportional to the range at each pixel location). Next, a ground plane estimation method is applied directly to this disparity image to detect the most dominant ground plane in the near field of the vehicle. In outdoor unstructured environments, where there is no dominant flat ground, a robust estimator is required that not only tolerates a large outlier percentage but also several discontinuities. For this purpose, we apply fast robust Random Sample Consensus (RANSAC) algorithm directly to this disparity image to estimate the dominant ground plane. The details of this method can be found in [6]. Once the ground plane has been estimated, any point whose measured disparity is not within some threshold of the expected ground plane disparity range will be considered as an obstacle. This threshold is selected through several validation runs trying to get the minimum error on a man-labeled dataset. Consequently, for each pixel of the input image with valid stereo data in the near field of the vehicle we have,

$$y_i = \begin{cases} 1 & d_m(i) \geq d_e(i) + \varepsilon \\ -1 & d_m(i) < d_e(i) + \varepsilon \end{cases} \quad (1)$$

for $i = 1, 2, ..., N$, where, N is the number of valid pixels in the image, respectively. Moreover, $d_m(i)$ shows the measured disparity, and $d_e(i)$ shows the estimated disparity by RANSAC ground plane fitting at each pixel location. Label 1 corresponds to an obstacle or non–traversable pixel and -1 to a ground or traversable, and ε represents the threshold.

In this method, the label propagation depends on disparity differences. Thus, when the disparity difference is not significant enough, such as where the lower parts of obstacles (i.e. footlines) meet the ground points, the label propagation has some uncertainty regarding the correct labels of the those points. This uncertainty becomes more severe when the vehicle is maneuvering in the outdoor, unstructured environments where the assumption of flat ground planes cannot be applied. This type of error in ground and obstacle detection is not crucial in the field of robotics as the labeled ground plane and obstacle points in the image are further projected into a 2-dimensional Cartesian grid map centered on the vehicle. Therefore, the upper part of an obstacle, which is most probably labeled correctly, will cover the grid map cell representing the obstacle location even if its lower part is mislabeled as ground plane. However, in our case, since these mistaken labels are going to be employed in learning procedure, difficulties will be imposed.

Machine learning methods' performances are degraded by mislabeled data which cause highly nonlinear decision surface, over-fitting of the training set, and poor generalization ability of the classifier [7]. For instance, the generalization ability of SVMs is achieved by finding a large margin between two classes and since the optimal hyper-plane obtained by the SVM depends on only a small part of the data points, it may become sensitive to noises in the training set [8]. Accordingly, we will next present the novel confidence-based auto labeling method to deal with the discussed issue.

B. Confidence-Based Auto Labeling

In this section, we propose a novel approach which explores the training data in order to identify the susceptive examples. For this purpose, the distance between the measured and estimated disparity, called the "score value", is calculated for each input instance.

Suppose a set of training data is given. Each data point belongs to a pixel in an input image. Depending on which class the data points belong to, they are labeled by $y_i \in \{1, -1\}$, which were determined through the procedure of stereo supervision auto labeling mentioned before.

To identify the susceptive examples, first we separate the training dataset into two sets based on their estimated labels y_i. Next, we define two score functions for each of these sets with their values determined by the Euclidean distance between the measured and estimated disparity for each labeled data point,

$$\begin{cases} f_1 = \left\| d_m(i) - \left(d_e(i) + \varepsilon \right) \right\| & \text{for i such that } y_i = 1 \\ f_2 = \left\| d_m(i) - \left(d_e(i) + \varepsilon \right) \right\| & \text{for i such that } y_i = -1 \end{cases} \quad (2)$$

The longer distance a data point has from the estimated ground plane, it gains higher score value. Subsequently the data points, whose score function values are smaller than a

threshold, are recognized to have high probability of being mislabeled. The major problem is that finding a predefined global threshold that can be applied to different situations is almost impossible since we do not have any prior knowledge about the scene and distribution of the f_1 and f_2. One solution is to define an adaptive threshold value based on the statistical information of each score function. The simplest and the most efficient statistical value is the *mean* of the score function values:

$$T_k = \text{mean}\{f_k\} \qquad \text{for } k = 1, 2. \tag{3}$$

Once the mislabeled data points are identified, we need to handle the mislabeled examples. One way of improving the quality of the training data is to remove the unclear instances. Those points that have low score values calculated from equation (2) are discarded prior to applying to the classification method. Each score function value is compared with its class related adaptive threshold found by (3) and if it is smaller than the threshold, its respective point is removed from the training data.

Fig. 1 shows an example of mislabeled data during auto labeling process. The lower parts of the hay bales in the image are detected as ground plane, while they are eliminated in the proposed auto labeling method correctly. The ground plane (shown in green) and obstacles (shown in red) detected in the near field are projected back to the original RGB image.

III. Biological-Based Feature Selective Attention

Once the ground plane and obstacles are estimated in the disparity image, and mislabeled data are identified and discarded through confidence-based auto labeling procedure, their corresponding image pixels in the reference image (right image) are labeled as either traversable (negative class -1) or non-traversable pixel (positive class +1). These labels are then associated with their visual appearance features to generate feature vectors for classifier training. In order to reduce computational time and increase the robustness of extracted feature, an input image is decomposed into non-overlapping patches (windows) of size 10x10 and a set of features are extracted within each labeled patch.

The color information output from the camera is in RGB color space, which is highly correlated and varies drastically with illumination changes. Thus, for color features, the RGB color space is transformed to the CIE LAB, which is nearly a uniform color space and close to human visual perception. An advantage of CIE LAB color space is that the intensity (channel L) is separated from chromatic channels (A, and B). The CIE LAB statistics consisting of mean and standard deviation for each patch are chosen as color descriptors. However, color and intensity information are not enough for discriminating regions in outdoor environments. Texture is also a fundamental property of surfaces. Image texture features represent a structure in terms of smoothness, coarseness, directionality and regularity of an object.

Figure 1. Input image (Top left), hand labeled ground truth (green: ground plane, red: obstacle, and blue: unknown) (top right), basic auto labeling output image (bottom left), proposed auto labeling algorithm output image (bottom right).

Using the mean and standard deviation of a bank of Gabor filters for four different orientations and four different scales, the texture attributes are constructed for each patch. Hence, a set of 38 features, which consist of intensity (2 features), color (4 features) and texture (32 features) are extracted for all patches of the input image..

The current self-supervised, near-to-far learning algorithms assume that all the visual features extracted from the input images (such as intensity, color, or texture) have the same level of prominence [4]. Generally, a set of attributes is extracted from the input image and they are added to a feature vector with the same or different predefined weights. However, in most outdoor and unstructured environments, one or some of the attributes of the feature vectors may be more informative for discrimination of the terrain from non-terrain regions. Recently, Blas et.al [9] assumed that the chrominance is more distinct than luminance in outdoor environments, therefore, they balanced their feature vector to rely more on color than texture or intensity, whereas, in an outdoor unstructured environment, one feature is not constantly dominant compared to others for all types of terrains and weather conditions.

To compensate the mentioned limitation of the current self-supervised, near-to-far learning algorithms, we present a novel approach which exploits the biological-based procedure of selecting the most salient features based on top-down goal driven, resulting in more efficient and effective classification.

An important mechanism in human visual perception is to detect regions of interest with visual attention. While a human observer looks at a scene, attention is driven to both bottom-up salient regions [10] and top-down relevant tasks [11]. Bottom-up depends on various factors such as line orientation or color of an object, and the dissimilarity between the object and the nearby distracters [12]. When these factors are not strong enough, it is possible to bias the

attention by top-down processes among the multiple regions. Top-down attention enhances the target's visual features to make it more salient compared to its background [13]. For instance, human drivers pay more attention to road (i.e. target) when driving rather than non-road even if the non-road regions have more distracting features. However, as the neurobiological foundations are not completely understood, the top-down influences, despite their importance in human visual system, are rarely considered in computational attention systems [12].

A. Proposed Top-Down Feature Biasing

Top-down task-specific guidance influences the human visual system based on nature of the tasks or goals. Inspired from this biological process in human, we now propose a top-down biasing method (briefly described in block diagram of fig. 2), which generates a set of adaptive weights representing the saliency of each feature. A salient feature is defined as a feature that is highly distinctive between traversable and non-traversable terrain. The more salient a feature is, the greater weight it gains. As a consequence of the larger adaptive weights, the classifier devotes more attention to the salient features.

The top-down task-specific guidance part of our approach is organized as three phases. First, all the possible features from a scene or a region are extracted. Then, the samples of each class are separated and the feature mean values (μ_m) are calculated within each class, where m varies from 1 to M (the total number of features). Finally, the ratio of each feature's mean value within first class is compared to the corresponding mean value of that feature from the second class. This provides information about whether a feature is salient in one class with regard to other existing classes or not. Thus, the bigger the ratio for a feature, the more consideration is needed to be allocated on that. This proposed technique presents a superior method for salient feature detection compared to those that assign predefined constant weights to balance the role of each attribute based on prior knowledge about the scene.

The next part of our approach is to feedback top-down attention weights to the feature vector. For this purpose, a weight proportional to the ratio obtained in previous step is calculated for each feature as,

$$\bar{w}_m = \begin{cases} \dfrac{\mu_{m|y=1}}{\mu_{m|y=-1}} & \text{if } \mu_{m|y=1} \geq \mu_{m|y=-1} \\ \dfrac{\mu_{m|y=-1}}{\mu_{m|y=1}} & \text{if } \mu_{m|y=1} < \mu_{m|y=-1} \end{cases} \quad (4)$$

where, μ_m is the mean value of each feature and $\bar{w}_m$ is the corresponding feature weight where m varies from 1 to M (the total number of features). In order to make a uniform interpretation of the weights, we place them in a same interval through mean value normalization such that they sum to a constant,

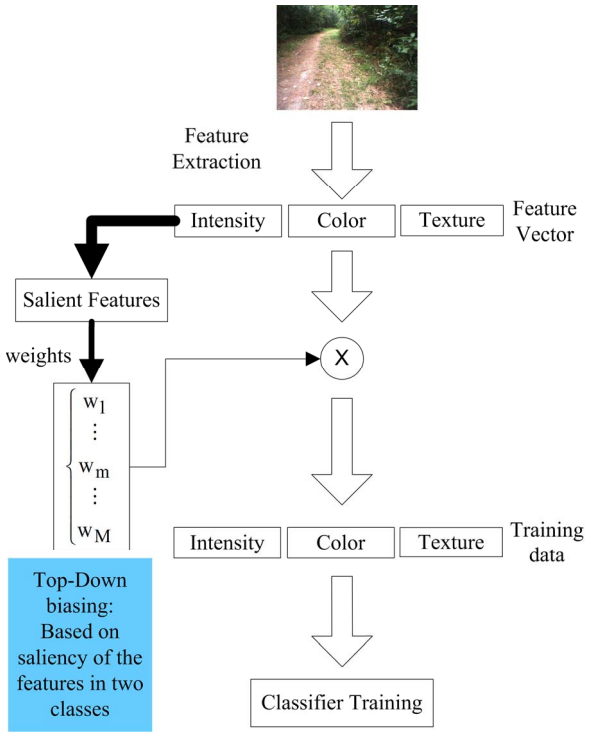

Figure 2. Block diagram of our proposed biological-based feature selective attention model.

$$w_m = \frac{\bar{w}_m}{\frac{1}{M}\sum_{l=1}^{M} \bar{w}_l} \quad (5)$$

Finally, the weights obtained from (5) are multiplied by their corresponding feature values to magnify the role of the most informative attributes. Note that these weights are calculated for each incoming input image.

IV. CLASSIFIER TRAINING

After the visual features with associated labels are extracted and biased through top-down task-specific guidance in the near field of the vehicle, they are used as inputs to train a classifier. Then, the obtained model is used to classify the entire image. In this work, we have used Support Vector Machine (SVM), which has become the reference for many classification problems because of their flexibility, computational efficiency and capacity to handle high dimensional data [14]. In real time applications, in which computational time is a critical issue and thus it is essential for the learning algorithm to be fast in recognition, SVMs have shown to be efficient. SVM constructs an optimal hyper plane such that the space points are classified into separate categories by a clear gap which is as wide as possible. The hyper plane parameters are obtained through an optimization method which tries to maximize the margins between different classes. This maximization is equivalent to minimization of weight matrix norm described as,

$$\text{Minimize: } \frac{1}{2}\|W\|^2 + C\sum_p \zeta_p$$
$$\text{Subject to: } \begin{cases} y_p\left(W^T X_p + b\right) \geq 1 \\ \zeta_p \geq 0 \end{cases} \quad (6)$$

where, p is the number of input samples, W is the weight matrix of the optimal hyper-plane, d is the desired label, and ζ are the slack variables, which basically allow some misclassification in the training aiming at better overall misclassification in test and train dataset. Value of $C > 0$ reflects the cost of violating constraints. A large C generally leads to smaller margin but also fewer misclassifications of training data and vice versa. The data points lying on the optimal margin are called support vectors, which contain the most relevant and sufficient information of the SVM algorithm [15].

V. Experimental Results

To evaluate the performance of the proposed techniques, in this section, we present the experimental results obtained from applying the approaches to the dataset in [16] which contains the image frames logged during live runs of the robot in the DARPA-LAGR program, which is the current gold standard for vision-based autonomous navigation systems using machine learning techniques. The datasets used here were taken from two different scenarios, each with two dissimilar lighting conditions. Each scenario has 100 image sequences and each frame consists of a raw RGB image, raw disparity information and a hand labeled image. Each pixel in a hand-labeled image indicates one of three classes: obstacle, ground plane, or unknown.

First, the confidence-based auto labeling approach is assessed. For quantitative evaluation, one widely used evaluation metric is the overall classification accuracy; however, this is not a suitable metric when the datasets are highly imbalanced. For example, for an autonomous navigation system, the number of instances for traversable terrains (i.e. negative class) outnumbers the number of instances in the positive class (non-traversable terrain) during most of its course [17]. For evaluation of our results, we use G-means, which is a commonly used performance metric for imbalanced data classification. G-means metric combines both the sensitivity and specificity by taking their geometric mean as follows,

$$\begin{cases} \text{Sensitivity} = \dfrac{TP}{TP + FN} \\ \text{Specificity} = \dfrac{TN}{TN + FP} \end{cases} \quad (7)$$

$$\text{G-means} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \quad (8)$$

where TP is the number of True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The results for all four datasets are given in Table I. As it can be seen, the proposed approach improves the performance of the learning algorithm on all data sets by increasing G-means , as well as, significantly reducing the number of SVs which results in less computational time and is memory efficient by removing the mislabeled samples from training data. It should be mentioned that since the classification time is primarily determined by the number of support vectors (SVs) [18], sparsity in the number of support vectors provides the opportunity to considerably reduce the memory and time requirements of SVM solvers [19]. Moreover, fig. 3 is provided as an example to measure the effectiveness of our method by means of visual perception.

Next, we evaluate the performance of our navigation system with the proposed biological based feature biasing method. Table II and fig. 4 provide the outcomes gained for the integration of confidence-based auto labeling with the proposed feature biasing approach, which indicates an improvement compared to the results of basic SSL technique. It is obvious from fig. 4 that the side road and far field of the image, which were classified inaccurately in bottom left image, are correctly classified by our proposed method. For this particular example, the intensity and color features are more distinct than texture information. Hence, they are signified by larger weights.

TABLE I. Results for Confidence-Based Auto Labeling Approach

Auto Labeling		**Data Sets**			
		DS1	***DS2***	***DS3***	***DS4***
SSL with Basic Auto Labeling	***G-means***	93%	76.6%	64%	78.5%
	No. of SVs	58	40	17	38
SSL with confidence-based Auto Labeling	***G-means***	94.6%	84%	77%	83.9%
	No. of SVs	27	21	13	21

Figure 3. Input image (Top left), hand labeled ground truth (green: ground plane, red: obstacle, and blue: unknown) (top right), basic self-supervised SVM learning (bottom left), proposed confidence-based auto labeling approach (bottom right), DS4.

TABLE II. RESULTS FOR BIOLOGICAL-BASED FEATURE BIASING APPROACH

Feature Biasing		Data Sets			
		DS1	*DS2*	*DS3*	*DS4*
SSL with No Feature Biasing	***G-means***	93%	76.6%	64%	78.5%
	No. of SVs	58	40	17	38
SSL with Feature Biasing	***G-means***	95%	85.6%	75%	84%
	No. of SVs	18	14	10	18

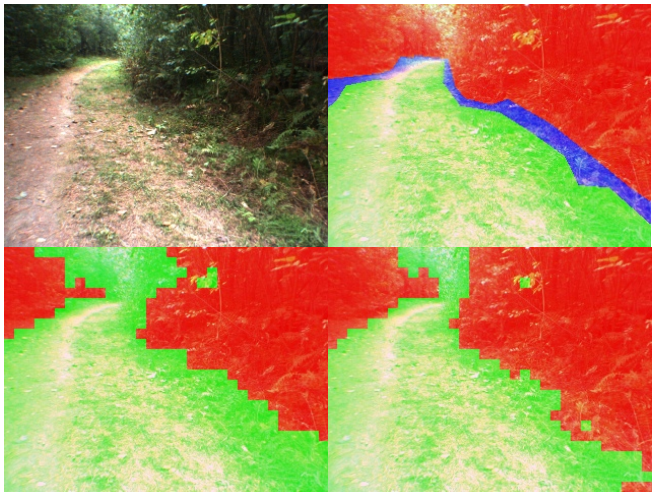

Figure 4. Input image (Top left), hand labeled ground truth (top right), output image before feature biasing (bottom left), proposed saliency-based feature selective attention algorithm output image (bottom right), DS1.

VI. CONCLUSION

In this study, novel approaches have been proposed to improve the performance of self-supervised near-to-far learning algorithms. First, confidence-based auto labeling for self-supervised online learning was introduced which detected and eliminated the input samples with low confidence level that were susceptible to be mislabeled. This technique resulted in reduction of support vectors in addition to an increase in prediction G-means rate. Then, a biologically inspired saliency detection approach through task specific guidance for feature weighting was presented. The method was able to detect the salient features and devote higher attention to them through top-down weighting. Worth to mention that top-down weights were considered to mimic the role of the brain's feedback discussed in biological contexts and is designated for discriminating the regions of interest. These weightings led to a higher G-means rate and lower number of SVs. Finally, the proposed methods were validated by the real-time application of autonomous navigation systems confirming the reported results.

REFERENCES

[1] P. Schenker, T. Huntsberger, P. Pirjanian, S. Dubowsky, K. Iagnemma, and V. Sujan, "Rovers for intelligent, agile traverse of challenging terrain," in 11th International Conference on Advanced Robotics, Portugal, 2003.

[2] C. Shoemaker and J. Bornstein, "The Demo III UGV program: A testbed for autonomous navigation research," in Proceedings of the IEEE International Symposium on Intelligent Control, pp. 644–651, 1998.

[3] P. Moghadam and W. Wijesoma, "Online, self-supervised vision-based terrain classification in unstructured environments," in Proceedings of the 2009 IEEE international conference on Systems, Man and Cybernetics, TX, USA, pp. 3100-3105, 2009.

[4] L. Jackel, E. Krotkov, M. Perschbacher, J. Pippine, and C. Sullivan, "The DARPA LAGR program: Goals, challenges, methodology, and phase I results," Journal of Field Robotics, vol. 23, pp. 945-973, 2006.

[5] A. Oliva, A. Torralba, M. Castelhano, and J. Henderson, "Top-down control of visual attention in object detection," in IEEE Proceedings of the International Conference on Image Processing, Barcelona, Spain, pp. 253-256, 2003.

[6] P. Moghadam, W. S. Wijesoma, and J. F. Dong. "Improving path planning and mapping based on stereo vision and lidar," ICARCV, pp. 384-389, 2008.

[7] W. Hu, Y. Liao, and V. Vemuri, "Robust support vector machines for anomaly detection in computer security," in Proceedings of the International Conference on Machine Learning and Applications (ICMLA), LA, USA, 2003.

[8] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in Proceedings of the fifth annual workshop on Computational learning theory, pp. 144-152, 1992.

[9] M. Blas, M. Agrawal, A. Sundaresan, and K. Konolige, "Fast color/texture segmentation for outdoor robots," in International Conference on Intelligent Robots and Systems, Nice, France, pp. 4078-4085, 2008.

[10] L. Itti, C. Koch, and E. Niebur, ''A model of saliency-based visual attention for rapid scene analysis,'' IEEE Trans. Pattern Anal. Mach. Intell. 20, 1254–1259,1998.

[11] Hopfinger, J.B., Buonocore, M.H., and Mangun, G.R. "The neural mechanisms of top-down attentional control". Nat. Neurosci. 3, 284–291, 2000.

[12] S. Kastner and L. Ungerleider, "Mechanisms of visual attention in the human cortex," Annual Review of Neuroscience, vol. 23, pp. 315-341, 2000.

[13] V. Navalpakkam and L. Itti, "Search goal tunes visual features optimally," Neuron, vol. 53, pp. 605-617, 2007.

[14] M. Nguyen and F. de la Torre, "Optimal feature selection for support vector machines," Pattern Recognition, vol. 43, pp. 584-591, 2010.

[15] N. Syed, H. Liu, and K. Sung, "Handling concept drifts in incremental learning with support vector machines," in Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, USA, , pp. 317-321, 1999.

[16] M. Procopio, "Hand-labeled DARPA LAGR data sets," *Available at* http://ml.cs.colorado.edu/~procopio/labeledlagrdata/, 2007.

[17] P. Moghadam, W. S. Wijesoma, M.D.P. Moratuwage, "Towards A Fully-Autonomous Vision-based Vehicle Navigation System in Outdoor Environments," ICARCV, pp. 597-602, 2010

[18] V. Vladimir and V. Vapnik, "The nature of statistical learning theory," Springer, 1995.

[19] L. Bottou and C. Lin, "Support vector machine solvers," Large scale kernel machines, pp. 1–27, 2007.