Spatial Context for Moving Vehicle Detection in Wide Area Motion Imagery with Multiple Kernel Learning

Pengpeng Liang[a] **Dan Shen**[b] **Erik Blasch**[c] **Khanh Pham**[c] **Zhonghai Wang**[b]
Genshe Chen[b] **Haibin Ling**[a]
[a]Computer & Information Science Department, Temple University, Philadelphia, PA, USA
[b]Intelligent Fusion Technology, Inc, Germantown, MD, USA
[c]Air Force Research Lab, USA
{*pliang,hbling*}*@temple.edu,* {*dshen,zwang,gchen@intfusiontech.com*}*, erik.blasch@rl.af.mil, Khanh.Pham@kirtland.af.mil*

ABSTRACT

Moving vehicle detection in wide area motion imagery is a challenging task due to the large motion of the camera and the small number of pixels on the target. At the same time, this task is very important for surveillance applications, and the result can be used for urban traffic management, accident and emergency responder routing. Also, the effectiveness of the context in object detection task can be further explored to increase target tracking accuracy. In this paper, we propose to use Spatial Context(SC) to improve the performance of the vehicle detection task. We first model the background of 8 consecutive frames with median filter, and get candidates by using background subtraction. The SC is built based on the candidates that have been classified as positive by Histograms of Oriented Gradient(HOG) with Multiple Kernel Learning(MKL). The region around each positive candidate is divided into m subregions with a fixed length l, then, the SC, a histogram, is built based on the number of positive candidates in each region. We use the publicly available CLIF 2006 dataset to evaluate the effect of SC. The experiments demonstrate that SC is useful to remove false positives, around which there are few positive candidates, and the combination of SC and HOG with multiple kernel learning outperforms the use of SC or HOG only.

Keywords: Vehicle detection, Spatial context, Multiple kernel learning

1. INTRODUCTION

Moving vehicle detection in wide area motion imagery is much more challenging than in traditional imagery, because of large amount of camera motion, small number of pixels on the targets, and the low frame rate of the video.[1] At the same time, more and more research focus on the understanding of such kind of imagery. A frame work to detect and track large number of cars in wide area image is introduced in,[1] and the performance of several state-of-the-art visual trackers are evaluated on Columbus Large Image Format (CLIF) dataset.[2] Zhao et al.[3] used the boundary of the car, the boundary of the front windshield and the shadow as features and integrated these features in the structure of the Bayesian network to detect vehicles. Prokaj et al.[4] used a similar method as[1] to perform vehicle detection, where the detected vehicles are refined by using the tracking result. Xiao et al.[5] used the road information from an additional co-registered geospatial information system (GIS) database as a constraint to refine the detected vehicles from the tracking results. Shi et al.[6] first used the vehicle detection result to construct trajectories, then they use these trajectories to estimate the road information which is used to remove false positives. Liang et al.[7] proposed to combine the HOG and Haar feature to improve the vehicle detection performance.

In order to use the framework introduced in,[1] we need to do background subtraction first. Due to the camera motion, we do registration for a set of consecutive frames. The more the consecutive frames we use, the smaller the common area is. So, in this task, the Gaussian mixture model[8,9] is not suitable, since it requires a relatively large number of frames to get a satisfying background estimate. In our approach, we choose to use the median background model. However, due to the large camera motion, 3D parallax, the registration and the background model is not robust enough to generate accurate foreground estimates, and there are a lot of false positives, so we need to classify these candidates as true positives and false positives.

For classification, since the target in wide area motion imagery is very small as shown in Fig. 1, there is a limit to use shape, appearance, color, or texture models as used in.[10] Beyond using the appearance information, the effectiveness

of context information has been demonstrated in many recent works,[11–13] and is useful to improve the performance of the object detection task. Based on the idea that if a candidate is indeed a vehicle, there are usually other candidates which are also vehicles around it as shown in Fig. 1, we propose the Spatial Context (SC). Given the candidates obtained from background subtraction, we build SC based on the candidates classified as positive by Histograms of Oriented Gradient (HOG).[14] HOG is based on Scale-invariant feature transform (SIFT)[15] and uses oriented gradient histogram to delineate objects, which has been successfully used with support vector machines (SVM) in human detection.

Figure 1. Part of the original image, green and red boxes indicate the positive and negative candidates classified by HOG with GMKL respectively.

In order to benefit from both the appearance information and the context information, we use multiple kernel learning (MKL)[16–19] to combine these two kinds of features. The MKL can learn the trade-off between the appearance and context automatically. In this paper, we use Spatial Projected Gradient -Generalized MKL (SPG-GMKL)[16] which is very efficient and can handle millions of kernels.

The rest of the paper is organized as follows: In Section 2, we give a brief introduction to HOG descriptor. In Section 3, we describe the proposed Spatial Context (SC).In Section 4, the SPG-GMKL is introduced. In Section 5, we describe the dataset and give the experiment results. Finally, we conclude this paper in Section 6.

2. HISTOGRAMS OF ORIENTED GRADIENTS (HOG)

Histograms of Oriented Gradients (HOG)[14] use the distribution of the oriented gradients to describe a local part of an image. The gradient of a pixel in an image is calculated by masks, and for color images, HOG calculates gradients for each color channel, and chooses the one with the largest normal as the pixel's gradient vector. Several masks were tested in,[14] including a 1-D point derivative (un-centered [-1,1], centered [-1,0,1] and cubic-corrected [1,-8,0,8,-1]), 2-D derivatives ($\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$) and a 3×3 Sobel mask. Test results show that the simplest mask [-1,0,1] works best.

For each pixel, $[-1, 0, 1]$ and $[-1, 0, 1]^T$ are used to calculate a gradient horizontally and vertically respectively. Based on the horizontal and vertical gradients, the orientation and weight can be calculated by voting. The votes are accumulated into orientation bins over local spatial regions called cells composed by pixels. The weight of a vote is a function of the gradient magnitude and the function can be the magnitude itself, it's square, it's square root, etc..

In order to alleviate the effect of the variation of illumination and foreground-background contrast, effective local normalization is required. Normalization is done by grouping cells into larger spatial blocks and normalizing each block separately. In practice, the blocks are overlapped so that each cell response contributes to several components of the final descriptor. In,[14] four different normalization scheme are evaluated. They are (a) *L1-norm*,$\boldsymbol{v} \rightarrow \boldsymbol{v}/(\|\boldsymbol{v}\|_1 + \epsilon)$; (b) *L1-sqrt*,$\boldsymbol{v} \rightarrow \sqrt{\boldsymbol{v}/(\|\boldsymbol{v}\|_1 + \epsilon)}$; (c) *L2-norm*,$\boldsymbol{v} \rightarrow \boldsymbol{v}/\sqrt{\|\boldsymbol{v}\|_2^2 + \epsilon^2}$; (d) *L2-Hys*, L2-norm followed by limiting the maximum value of $\boldsymbol{v}$ to 0.2. It turns out that the performance of L2-Hys, L2-norm, and L1-sqrt are almost the same, whereas the L1-norm cannot compete with the other three schemes.

3. SPATIAL CONTEXT(SC)

Given a candidate p in frame I, which has been classified as positive by HOG[14] with GMKL,[16] we first divide the region around p into m subregions with length l. Each subregion is a fan (e.g., center with rotating blades) and all the subregions have the same center angle θ. m is determined by θ, and $m = 360/\theta$. Next, we calculate the number of candidates which are also classified as positive by HOG and GMKL. Then, we build a histogram with m bins, where each bin corresponds to a specific subregion. The value of each bin is the number of positive candidates in the corresponding subregion. The value of the kth bin of candidate p is

$$h_p(k) = \#\{i \in C \wedge i \neq p \wedge i \in R(k)\} \tag{1}$$

where C is the set of the candidates classified as positive by HOG with GMKL, and $R(k)$ is the subregion corresponding to the kth bin. The approach to build SC is described in Algorithm 1.

Algorithm 1 Building SC for a Candidate

Require:
p: a candidate
C: the set of candidates of I
Ensure:
h: the SC, a histogram, for p

for each bin k of h **do**
  $count \leftarrow 0$
  for each $i \in C$ **do**
    if $i \in C \wedge i \neq p \wedge i \in R(k)$ **then**
      $count \leftarrow count + 1$
    end if
  end for
  $h(k) \leftarrow count$
end for
return h

4. MULTIPLE KERNEL LEARNING WITH SPECTRAL PROJECTED GRADIENT (SPG)

The objective of multiple kernel learning is to learn the optimal combination of the given base kernels. In,[17] the optimal kernel is approximated as $\mathbf{K}_{opt} = \sum_k d_k \mathbf{K}_k$ where $\mathbf{d}$ corresponds to the trade-off among the base kernels. However, the sum of base kernels is too constraining, and is the just the concatenation of individual kernel's feature space. In order to relieve this constraint, in,[18] the generalized multiple kernel learning (GMKL) is proposed, so that the combination of base kernels can be more flexible. The primal formulation for optimizing the GMKL is the following:

$$\min_{\boldsymbol{d}} \quad T(\boldsymbol{d}) \quad \text{subject to} \quad \boldsymbol{d} \geq 0$$

$$\text{where} \quad T(\boldsymbol{d}) = \min_{\boldsymbol{w},\boldsymbol{d}} \frac{1}{2}\boldsymbol{w}^t\boldsymbol{w} + \sum_i l(y_i, f(\boldsymbol{x}_i)) + r(\boldsymbol{d})$$

where w are the weights, l isis the loss function, and both the regularizer r and the kernel can be any general differentiable functions of $\boldsymbol{d}$ with a continuous derivative. The optimization of $\boldsymbol{d}$ is carried out in the outer loop, while the optimization of $\boldsymbol{d}$ is performed in the inner loop.

Though the GMKL in[18] allows more flexible combination of kernels, the projected gradient descent optimizer is inefficient, since the computation of the step size and a reasonably accurate gradient direction are all expensive. In order to alleviate this problem, using Spectral Projected Gradient (SPG) descent to solve the GMKL is proposed in.[16] The SPG has the following advantages:[18] (a) second order information is explored to determine the step size; (b) employs a non-monotone step size selection criterion requiring fewer function evaluations; (c) it is robust to gradient noise, and (d) it can take steps when far away from the optimum. The pseudo code of SPG-GMKL is given in Algorithm 2.

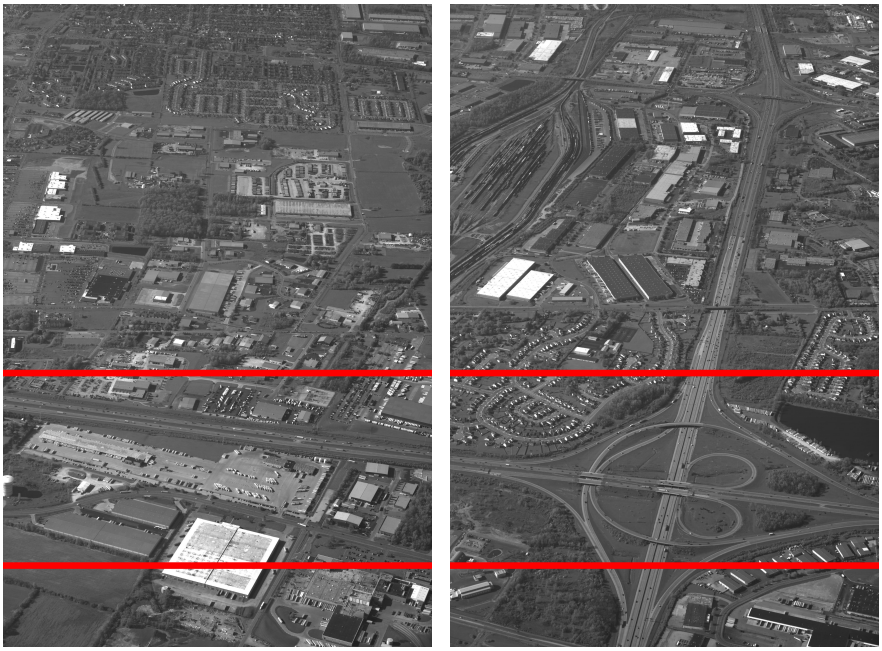
Algorithm 2 SPG-GMKL

$n \leftarrow 0$
Initialize $\boldsymbol{d}^0$ randomly
repeat
  $\boldsymbol{\alpha}^* \leftarrow \text{SolveSVM}_\epsilon(\boldsymbol{K}(\boldsymbol{d}^n))$
  $\lambda \leftarrow$ SpectralStepLength
  $\boldsymbol{p}^n \leftarrow \boldsymbol{d}^n - \boldsymbol{P}(\boldsymbol{d}^n - \lambda \nabla \boldsymbol{W}(\boldsymbol{d}^n))$
  $s^n \leftarrow$ Non-Monotone
  $\epsilon \leftarrow$ TuneSVMPrecision
  $\boldsymbol{d}^{n+1} \leftarrow \boldsymbol{d}^n - s^n \boldsymbol{p}^n$
until converged

5. EXPERIMENT

5.1 Dataset

We use Columbus Large Image Format (CLIF) 2006[2] dataset to evaluate the proposed Spatial Context (SC). The scene of this dataset is a flyover of the Ohio State University (OSU) from a large format monochromatic electro-optical platform which is comprised of a matrix of six cameras and the size of each image is 2672(width) by 4008(height) pixels. Since our focus is vehicle detection, we use a 2672×1200 subregion which contains an expressway as shown in Figure 2 and Figure 3. The subregion not only includes horizontal and vertical roads, but also an overpass. We labeled the vehicles in 102 frames of camera 3, and there are 9364 vehicles in total which are used as test data. We also labeled 1730 candidates obtained from background subtraction from 16 frames of camera 1 which are used as training data.



(a) (b)

Figure 2. (a) and (b) are the original images, where the region between the two red lines is the subregion.

5.2 Quantitative Evaluation

For quantitative evaluation, we use the distance between a positive candidate and the groundtruth to decide whether it is a true positive. Given the groundtruth $G = \{g_1, g_2, ..., g_S\}$, if the candidate c is true positive, there exists $g \in G$ and the distance between the center of g and the center of c is not greater than ten pixels. Since we classify the candidates obtained

(a)

(b)

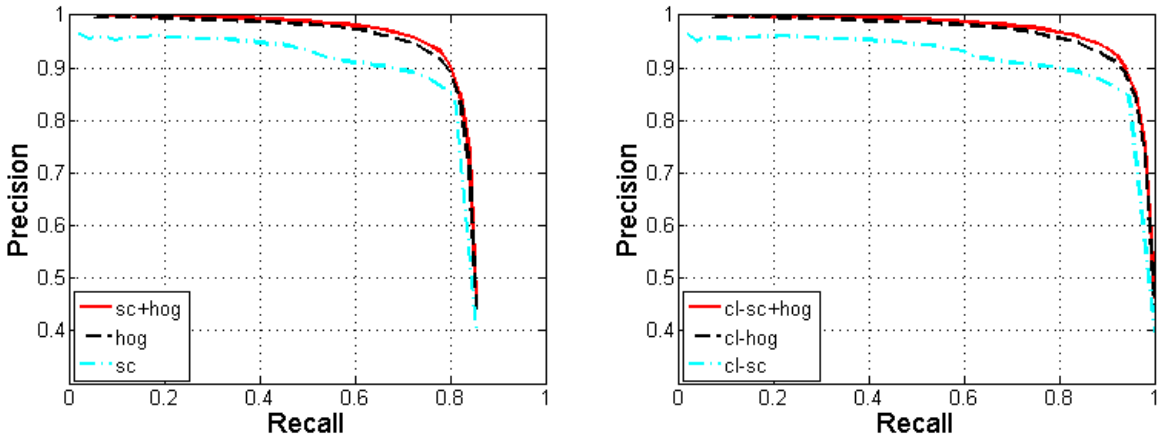
Figure 3. (a) and (b) are subregions for 2(a) and 2(b) respectively.

from background subtraction, we use two different ways to calculate the recall. One way is to use the number of the actual groundtruth, S, i.e., the actual number of vehicles in each image; another way is to use the number of candidates which are indeed vehicles as groundtruth, S'. We give precision-recall curves for both. Without classification, the performance of background subtraction is not satisfying, the precision is only 0.398 at the recall 0.854. For parameters of SC, we test several configurations of m and l, the difference is tiny, and we use 6 and 200 for m and l respectively in the following experiment.

To demonstrate the usefulness of SC, the performance of SC, HOG and the combination of SC and HOG are evaluated. The precision-recall curves are given in Fig. 4. From the results, we can see that HOG outperforms SC, but the combination of SC and HOG performs best, especially at a high recall rate.

5.3 Qualitative Evaluation

Some qualitative results are given in Fig. 5 and Fig. 6. From the results, we can see that SC is useful to remove false positives when there are few candidates that are classified as positive by HOG with GMKL around them. Usually, this kind of candidates are away from the road. So, SC is useful to remove the false positives that are away from the road. At the same time, the number of misclassified vehicles on the road is very small.



(a) (b)

Figure 4. The precision-recall curve for SC, HOG and the combination of SC and HOG.(a) is plotted based on S, (b) is plotted based on S'.

6. CONCLUSION

In this paper, we propose Spatial Context (SC) to improve the vehicles detection task in wide area motion imagery. In order to make use of both the appearance and context information, we use SPG-GMKL which is very efficient to combine SC and HOG. Experiments on the CLIF 2006 dataset verified that SC is useful to remove false positives that are away from the road.

REFERENCES

[1] Reilly, V., Idrees, H., and Shah, M., "Detection and tracking of large number of targets in wide area surveillance." in *European Conference on Computer Vision*, pp. 186–199 (2010).

[2] O. Mendoza-Schrock, J. A. Patrick, and E. P. Blasch, "Video image registration evaluation for a layered sensing environment." in *IEEE Nat. Aerospace Electronics Conf. (NAECON)* (2009).

[3] T. Zhao and R. Nevatia, "Car detection in low resolution aerial image," in *ICCV* (2001).

[4] J. Prokaj, M. Duchaineau, and G. Medioni, "Inferring tracklets for multi-object tracking," in *Workshop of Aerial Video Processing joint with IEEE CVPR* (2011).

[5] J. Xiao, H. Cheng, H. S. Sawhney, and F. Han, "Vehicle detection and tracking in wide field-of-view aerial video," in *CVPR* (2010).

[6] X. Shi, H. Ling, E. Blasch, and W. Hu, "Context-driven moving vehicle detection in wide area motion imagery," in *Int'l Conf. on Pattern Recognition (ICPR)* (2012).

[7] P. Liang, G. Teodoro, H. Ling, E. Blasch, G. Chen, and L. Bai, "Multiple kernel learning for vehicle detection in wide area motion imagery ." in *International Conference on Information Fusion* (2012).

[8] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking." in *Computer Vision and Pattern Recognition*, pp. 2246–2252 (1999).

[9] P. KaewTraKulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection." in *European Workshop on Advanced Video Based Surveillance Systems* (2001).

[10] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Computer Vision and Pattern Recognition (2)*, pp. 1447–1454 (2006).

[11] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *ECCV (1)* (2008).

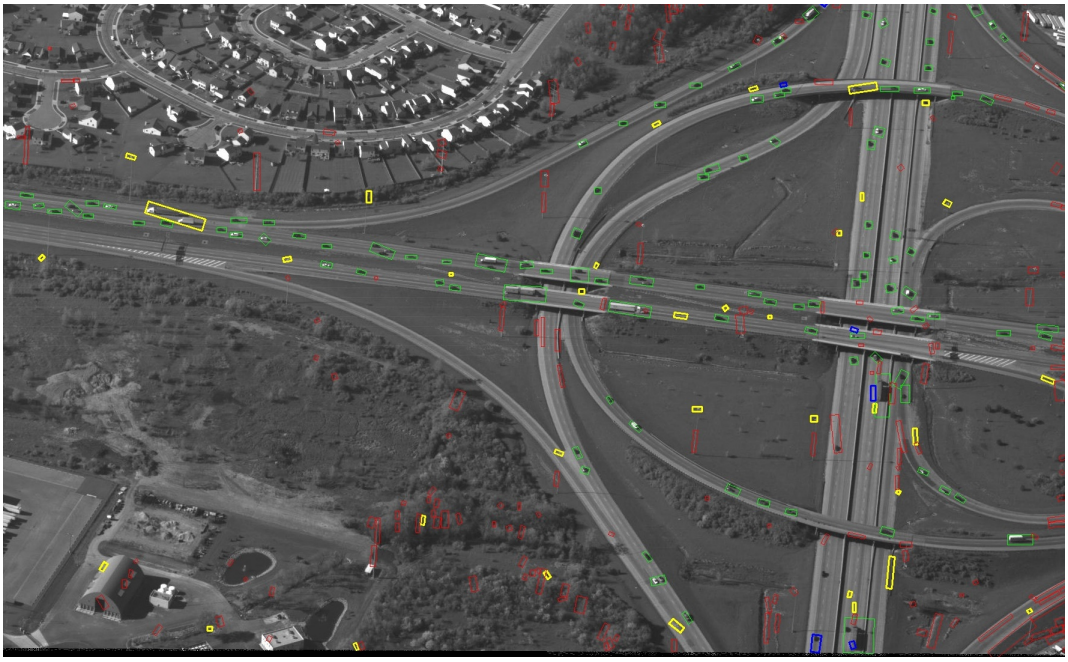
[12] A. Jain, A. Gupta, and L. S. Davis, "Learning what and how of contextual models for scene labeling," in *ECCV (4)* (2010).

[13] S. K. Divvala, D. Hoiem, J. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *CVPR* (2009).

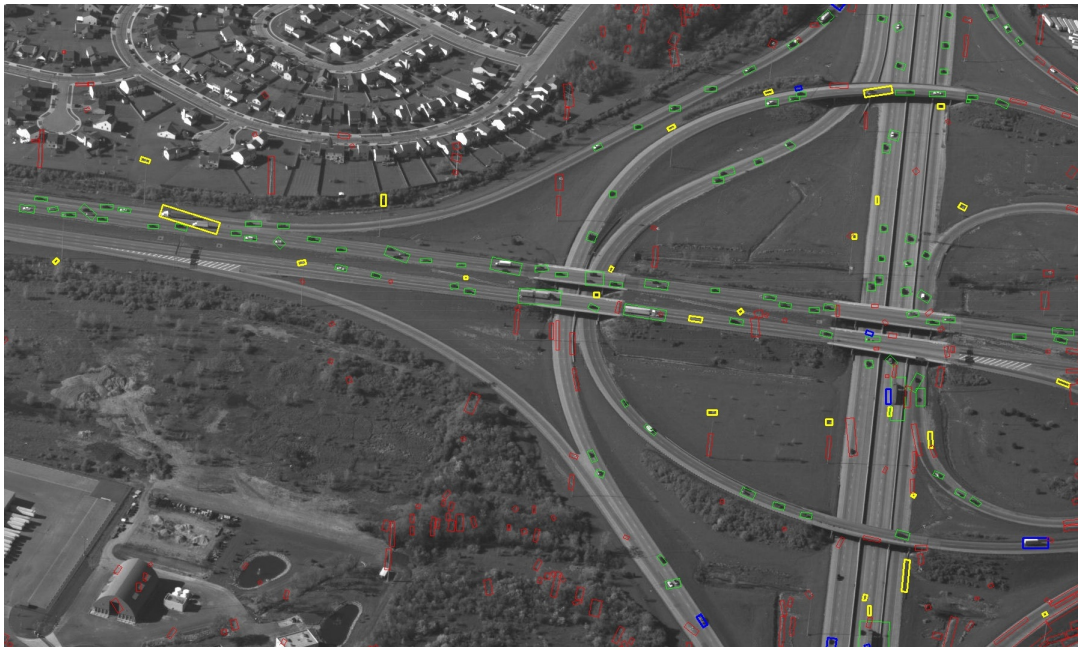
(a)

(b)

Figure 5. The classification result. The green, yellow, red, and blue bounding boxes indicate true positive, false positive, true negative, false negative respectively. (a), (b) are the results of SC, TC+HOG respectively.



(a)



(b)

Figure 6. The classification result. The green, yellow, red, and blue bounding boxes indicate true positive, false positive, true negative, false negative respectively. (a), (b) are the results of SC, TC+HOG respectively.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection." in *Computer Vision and Pattern Recognition*, pp. 886–893 (2005).

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110 (2004).

[16] A. Jain, S. V. N. Vishwanathan, and M. Varma, "Spg-gmkl: generalized multiple kernel learning with a million kernels," in *KDD* (2012).

[17] M. Varma and B. R. Babu, "More generality in efficient multiple kernel learning," in *ICML* (2009).

[18] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *ICCV* (2007).

[19] M. Gönen and E. Alpaydin, "Multiple kernel learning algorithms," *Journal of Machine Learning Research*, vol. 12, pp. 2211–2268 (2011).