# Contextual smoothing of image segmentation

Jonathan Letham & Neil M. Robertson
Heriot-Watt University
Edinburgh, UK
{jal13,n.m.robertson}@hw.ac.uk

Barry Connor
Thales Optronics
Glasgow, UK
barry.connor@uk.thalesgroup.com

## Abstract

*This paper presents a new method for improving region segmentation in sequences of images when temporal and spatial prior context is available. The proposed technique uses elementary classifiers on infra-red, polarimetic and video data to obtain a coarse segmentation per-pixel. Contextual information is exploited in a Bayesian formulation to smooth the segmentation between frames. This is a general framework and significantly enhances segmentation from the classifiers alone. The method is demonstrated by classifying images of a rural scene into 3 positive classes: sky, vegetation and road, and one class of all other unlabelled data. Priors for the probabilistic smoothing in this scene are learned from ground-truth images. It is shown that an overall improvement of around 10% is achieved. Individual classes are improved by up to 30%.*

## 1. Introduction

In a dynamically changing environment it is often difficult to detect objects due to occlusion, shade and clutter. It is our ultimate aim to be able to detect objects that can be at times hard to find and in variable environments. This will be achieved with multiple sensors on a moving platform. Context is being used increasingly in computer vision techniques to help perform scene recognition [10], region categorisation [1, 5, 7, 12] and object detection [6, 11, 13, 15, 16]. We believe that context can be used to aid the accuracy of our objectives and, generally, make the deployment of image processing operations more effective.

Context will be used in a variety of different ways to achieve our goal. By extracting the general regions of an image, we can use this spatial context to assist object detection by applying image processing to regions where one would expect to find the object. (Or, conversely, to regions where the object would not be expected to be found.) In this paper, methods for extracting informative image regions are described. These methods themselves use two different types of contextual information to achieve the results - learnt prior context of the regions for a given scene and temporal context from regions extracted in previous frames.

The principal contribution of this work is the use of prior and temporal contextual knowledge for the improved classification of regions. This is a general technique and we demonstrate its efficacy via the combination of infra-red, polarimetric and electro-optical sensor data. The accuracy of classification of image regions is improved compared to when the classifiers are used without a contextual framework by using a probabilistic formulation which fuses temporal and spatial information.

### 1.1. Related work

Earlier work investigating how humans recognise objects and scenes has been presented [9]. This found that objects were more easily recognised in a scene when in proper spatial relation i.e. when in the correct context. These principles of human vision can be applied to computer vision as shown by Torralba *et al.* [13] using scene recognition to improve object detection. Context of the scene has been used elsewhere to improve object recognition [15]. Regions are learnt that are spatially associated with an object. The learnt regions that surround an object are then used to help identify it. Although this only found marginal improvements compared to when the detector was used with no contextual knowledge, context came in useful in scenes where the object was difficult to find.

A more successful approach by Heitz and Koller uses regions in an image to serve as context for the detection of objects [6]. This is a similar concept to what we want to eventually build. There has been much work into using context to improve automated annotation of image regions. Li, Socher and Fei-Fei use a top down approach to improve annotation of segmented regions of an image [7]. In the work of Barnard *et al*. image regions are learnt in order to associate text with segmented regions of an image [1]. Rabinovich *et al*. use context at a semantic level to improve region labelling of an image [12]. They take the technique further by adding another type of context to their process [5]; semantic context is used along with spatial context to

further improve the labelling of a segmented image.

We build on the concepts of using learnt priors for a given environment and using more than one type of context in order to improve region classification. In particular, we are extending on work done of Matzka *et al*. where a Bayesian probability framework to improve vehicle detection on different road types is introduced [8] . Matzka uses prior learnt knowledge of what type of vehicles are likely to be on a certain road type and temporal contextual knowledge of previous detections. We are inspired by this framework and extending it to region classification by using prior context of regions and previous classified regions.

The work we have done thus far has laid the ground for a final, complete contextual framework with many more classes. In doing so, we make a new contribution in at least two ways: (a) using two different forms of context, prior and temporal, to improve the accuracy of region classification; (b) fusing multiple sensor data via contextual smoothing after classification.

In Section 2, the classifiers used to extract image regions are described. The contextual framework that we use is explained in Section 3. And finally, the results of the experiments are shown in Section 4 where it is demonstrated that context improves classification.

## 2. Classifiers

The data used in this paper was collected by an array of visible and thermal cameras mounted on a moving vehicle in a rural environment. Example images taken from our different data sources are shown in Figure 1. Classification methods are illustrated too.

Since our data is collected from an array of individual sensors - infra-red (in the thermal waveband $8\mu m$ to $12\mu m$), polarimetric and electro-optical ($450nm$,$550nm$,$650nm$ and $880nm$), it is necessary to register the gathered images in order to fuse the information at data level. The long-wave infra-red camera has a resolution of $636 * 513$ pixels and the eloctro-optical camera have a resolution of $1024 * 768$ pixels. Registration of the visual cameras is achieved automatically by finding corresponding control points using the Speeded Up Robust Features (SURF) [2]. Thermal to visual camera registration is performed by selecting control points manually. As well as needing to be spatially registered, the visual and thermal data had to be temporally registered as the thermal camera runs at a faster frame rate (3.1 times faster). All experiments were carried out at the visual frame rate and the exact corresponding thermal frames are used.

### 2.1. Region areas and classification methods

We extract four regions from the images: sky, road, foliage and the "other" class. The sky is defined as the region above the horizon including clouds; road is any visible tarmac road; foliage is defined as bush, tree or cut grass (BTG from now) in the image; and the other class is any pixels that do not fall into these regions. Described below is the data and methods used to classify the regions. It is important to stress that we have not expended enormous effort to make these classifiers very robust or accurate. Rather the purpose of this paper is to show that classification accuracy can be improved by using context, regardless of classifier. The classifiers are binary detectors, classifying a pixel as either a region or not a region via a predefined threshold. (It will be seen that the Bayesian formulation presented in Section 3 can take full probabilistic distributions and this is a topic of current work.)

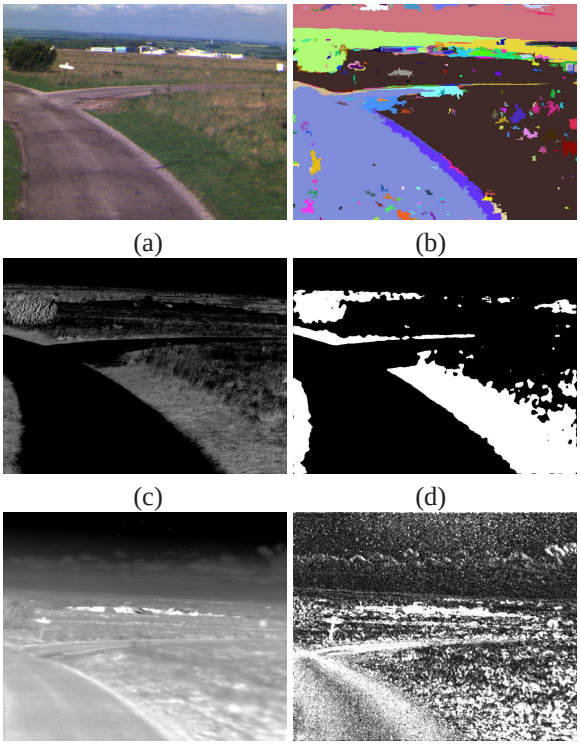


Figure 1. (a) and (b) show a colour image and a graph-based segmented image respectively. The sky is classified as the largest segmented region in the top of the image. (c) is the vegetation index of the same scene. A threshold and median filter are then applied to classify bush tree and grass as is shown in (d). The thermal image of the scene is shown in (e) and the corresponding polarised data in (f). The image in (f) is thresholded using an adaptive threshold and then a median filter is applied to classify the road.

**Sky.** Graph-based image segmentation is used to segment a RGB image [4]. The sky is then classified by taking the largest top region of the segmented image. This is a heuristic assumption valid only for this dataset.

**Road.** The thermal camera is designed to be sensitive to polarised radiation. Connor *et al*. explain the operation and benefit of a long wave infra red polarimetric imager [3].

The phenomenon of polarisation causes man-made objects, such as metal, glass, tarmac, to have a different polarisation signature to that of natural vegetation. Therefore, polarisation has the potential to discriminate man-made objects from background clutter. Polarimetric information, combined with conventional thermal imaging, provides a powerful means of detecting objects in applications such as situational awareness. Many factors affect an object's polarisation signature such as texture and orientation. Stokes images [3] are used to quantify the polarisation signature. The $Q$ Stokes image, defined as the amount of linear polarisation in the horizontal direction, is useful in segmenting out roads. $Q$ is computed using the following equation,

$$Q = i_0 - i_{90} \tag{1}$$

where $i_0$ and $i_{90}$ are the intensity images at $0^o$ and $90^o$ polarisation, respectively. An adaptive threshold then a median filter are applied to the $Q$ data in order to finally classify the road.

**BTG.** Live green plants have evolved to absorb solar radiation in the photosynthetically active radiation (PAR) spectral region which includes the red waveband. They have also evolved to scatter light in the near infra-red (NIR) region. Therefore, NIR and red wavebands can be combined in the Vegetation Index in order to highlight vegetation [14]. We use the Vegetation Index to help classify bush, tree and cut grass. It is defined as:

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{2}$$

where $NIR$ is the 880nm waveband of light (for these experiments), $RED$ is the red band (650nm) and $NDVI$ is the normalised vegetation index. NDVI highlights vegetation and is used to discriminate bush, tree and cut grass from the rest of the scene in our experiments.

**Other.** As there is no classifier built for the other class, it is identified simply by all the pixels that have not been detected by the three active classifiers for a particular frame.

### 2.2. Classifier results

Images are ground-truthed by hand so that the accuracy of the classifiers can be calculated. Fifteen frames were ground-truthed stretching over the entire driving sequence. The classifiers were tested on a set of images and the true positive rates, false positive rates and accuracies of the detectors were computed over each image. Accuracy is calculated as the total number of true detections divided by the total number of pixels. The medians of the results are shown in Table 1. Some classification results are shown in Figure 2. These show the colour image, the ground truth and the detection results for two frames.

As can be seen from the results, the classifiers have varying accuracies. The sky classifier performs very well with a high true positive rate and low false positive rate. The road classifier has a very good true positive rate but poor false positive rate which makes the accuracy of this detector poor. This is due to movement affecting the polarisation calculation in the thermal camera. The BTG detector performs reasonably well. Finally, the other classifier has poor true and false positive rates mainly due to the false detections of the road classifier. We aim to improve these accuracies by using context (which is, after all, the purpose of this work) and this is shown in the following section.

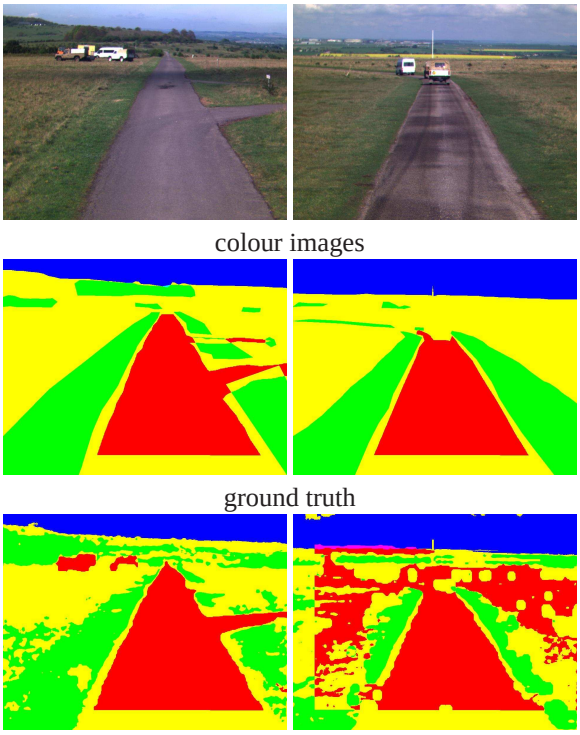


Figure 2. Classification results of the detectors showing the varying results of the classifiers when used in isolation. In the ground truth and classification images, blue is sky, red is road, green is BTG and yellow is other. The left column is an example where the classifiers work well. The right column is an example where the system has performed poorly mainly due to the large false positive rate of the road classifier.

| Classifier | TPR | FPR | Acc |
|---|---|---|---|
| Sky | 0.97 | 0.0 | 0.99 |
| Road | 0.98 | 0.35 | 0.71 |
| BTG | 0.81 | 0.07 | 0.90 |
| Other | 0.54 | 0.53 | 0.76 |

Table 1. Median true positive rate, false positive rate and accuracy of the individual classifiers.

## 3. Contextual Smoothing

Now having the binary classifiers for the regions of our images, we aim to improve their accuracy using context. We compute the probability that a region exists given a detection of a region from prior and temporal knowledge via the following:

$$P(R_l|D_l) = \frac{P(D_l|R_l)P(R_l)}{P(D_l)} \quad (3)$$

where, for a given region type and classifier $l$, $P(R_l|D_l)$ is the probability of a region $R$ to exist given a detection $D$, $P(D_l|R_l)$ is the true positive rate of the classifier, $P(R_l)$ the prior probability and $P(D_l)$ a normalising constant. If the region is not detected at a pixel, the probability for the region to exist is given conversely by:

$$P(R_l|\neg D_l) = \frac{P(\neg D_l|R_l)P(R_l)}{P(\neg D_l)} \quad (4)$$

The prior probability $P(R_l)$ is defined as a weighted summation of the prior probability dependent on the current scene $P(R_l|S)$ and the posterior probability of the previous frame $P_{k-1}(R_l|D_l)$

$$P(R_l) = (1-w)P_{k-1}(R_l|D_l) + wP(R_l|S) \quad (5)$$

where $w \in R|0 \leq w \leq 1$. This factor controls the degree of adaptiveness of the contextual knowledge. For $w = 0$, the prior knowledge derived from the scene type $S$ is only used for the first calculation and the previous posterior probability is used for all future calculations. For $w = 1$, only the learnt prior probabilities for a region to exist are considered for the prior probability $P(R_l)$ and any previous calculations are disregarded. The weighting factor controls the amount of temporal or prior contextual knowledge that is used.

For these tests, only one scene is considered so only one set of spatial prior probabilities $P(R_l|S)$ are learnt. The consideration of different scenes (e.g. urban, rural, outdoor, indoor etc.) is easily incorporated provided the classifiers adapt to the changing region types (e.g. buildings appear constantly in an urban location). Prior probabilities can be learned for each scene and a change of scene can be detected by a Global Positioning System (GPS).

## 4. Results and Discussion

To calculate the region priors for our scene $P(R_l|S)$, the ground truth is used. The prior is the average of the ground truth for a given region. For the true positive rate $P(D_l|R_l)$, the median true positive rates calculated in Section 2.2 are used (see Table 1). In the final classification of a pixel, the maximum posterior probability out of the four regions is taken to be the class.

Tests were carried out on the data gathered from a moving vehicle on a rural road. The number of visual frames in the sequence was 2035. The priors were built on 12 ground truthed images for each class. The contextual framework was tested for $w = 0, 0.25, 0.5, 0.75, 1$ of Equation 5.

Receiver operator characteristic (ROC) plots for our results are shown in Figure 5. Median results are shown for different values of $w$ and the classifier ROC with no context used is also shown for comparison. The percentage increase in accuracy for the detectors is shown in the individual plots in Figure 4. We now discuss these results, briefly.

**Sky.** The sky classifier is a good classifier before contextual smoothing is applied (accuracy of 0.99). It can be seen from Figure 5 that there is little effect on the TPR of this classifier with a less than 1% difference between the best and worst TPR. The detector accuracy levels remain high apart from when $w = 0$. At this weight no learnt priors are used and the system relies only on temporal context. The lack of priors for this weight cause errors made by sky classifier to propagate and grow frame-by-frame. This is responsible for increasing the FPR and decreasing accuracy. For all the other weights, the accuracy remains as high as when the classifier was used outwith the framework.

**Road.** The context significantly improves the road classifier. There is a small decrease in TPR for any $w$, but this is more than compensated for by the huge decrease in FPR.This means that there is at least an increase in accuracy of 30% for all $w$. Interestingly, the lowest FPR rate occurs when $w = 0$ - the weight at which the sky classifier performs worst. The large FPR of the road classifier is due to miscalulation of $Q$ caused by large movements of the camera. These false detections of polarised materials occur sporadically while the true detections of the polarised road happen continuously. The temporal framework only builds on the true detections of the road as these occur constantly which creates the lower FPR.

**BTG.** The BTG classifier performs worst in the contextual smoothing. However, FPR improves for all weights apart from $w = 0$. As for sky, the poor FPR rate at this weight is caused by the propagation of initial errors in the BTG detector. Theses are intensified by the temporal context being used without any constraining spatial priors. Learnt priors for BTG are not as strong as the other classes leading to a decrease in TPR. Where the sky, road and other classes are likely to be in certain parts of the image, bush and tree tree locations are far more varied. These weaker learnt priors lead to lower posterior probabilities of BTG in Equation 3. Thus at classification some BTG regions are neglected.

**Other.** The "other" classifier is improved by contextual smoothing. There is an increase of about 80% in TPR for all

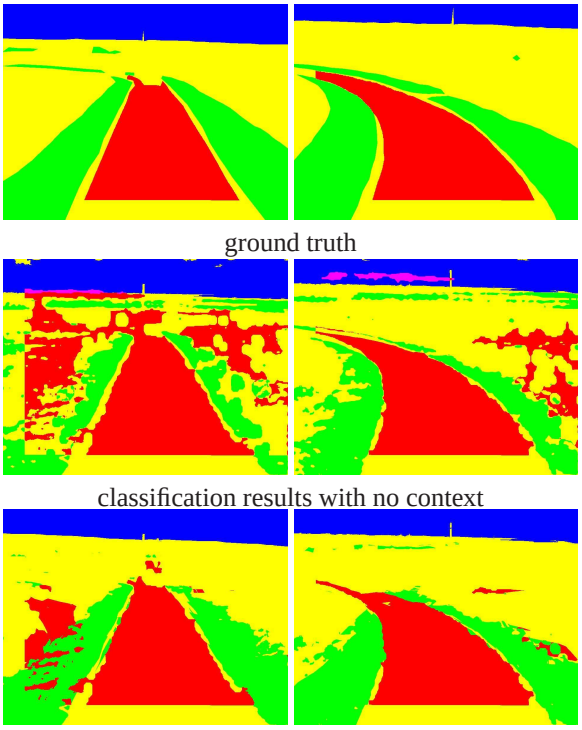


Figure 3. Context improves the segmentation vs. ground truth. As in Figure 2 we choose an example where the classifier performs poorly (*left col.*) and well (*right*). In both cases context plays an important role in improving the result.

$w$ except $w = 0$. This is a result of the huge improvement in the road classifier. There is a slight decrease in accuracy at $w = 0$ caused by the poorer performances of the sky and BTG classifiers. The other class has strong priors in areas where the BTG priors are weaker leading to false detection of other class in BTG regions. However, this small increase in FPR is outweighed by the large increase in TPR meaning the accuracy is enormously improved for $w \neq 0$.

As can be seen from the graph in Figure 4, the optimum value of $w$ is 0.25. The performance of the whole system is improved by about 10% when at least a proportion of prior contextual information is included. Further investigation is required into whether varying $w$ for the different classifiers shows a more optimum performance. These results indicate that contextual smoothing improves the classification of this data set. There is the potential for this method to be applied to diverse data sets with different classes. In that case, priors have to be learnt for the classes of the data and the optimum $w$ can be found from testing as we have done here.

## 5. Conclusion and Future Work

We have shown in this paper that using a combination of temporal and prior context does improve the accuracy of

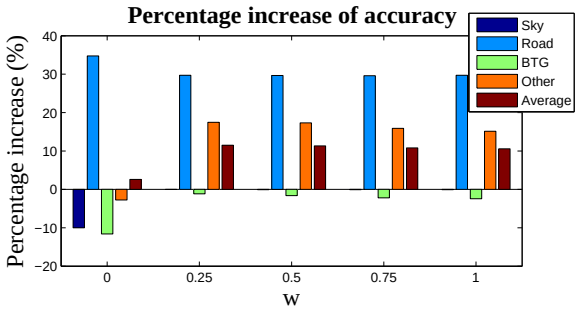


Figure 4. The percentage increase in accuracy for each class for different values of $w$. The average increase in accuracy is also shown.

image segmentation in video from various sensor modalities. This approach is a novel way of combining contextual information to improve region labelling. Our method is especially effective for those regions which have strong spatial priors. The method serves as a very useful tool in decreasing the FPR of our classifiers.

In the future we aim to build priors for differing scenes $P(R|S)$ so that the contextual framework can be tested between different environments. In such a scheme GPS will provide the information about a change of scene and the learnt priors can be applied according to the scene type. We will develop classifiers that are not binary but return a probability of class per pixel. The framework does not need to be adapted to adopt this extension. Our grand aim is to use the regions extracted in the images to enhance object detection. The regions will provide the context of the scene and this information will be used to make our object detection more accurate and targeted (thus making more efficient use of computing resources in real-time) by applying specific algorithms to the regions where objects are more likely to appear.

## References

[1] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *Machine Learning Research*, 3:1107–1135, 2003. 1

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. *ECCV*, pages 404–417, 2006. 2

[3] Barry Connor, Iain Carrie, Robert Craig, and John Parsons. Discriminative imaging using a lwir polarimeter. *SPIE*, 2008. 2, 3

[4] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59:167–181, 2004. 2

[5] Carolina Galleguillos, Andrew Rabinovich, and Serge Belongie. Object categorization using co-ocurrence, location

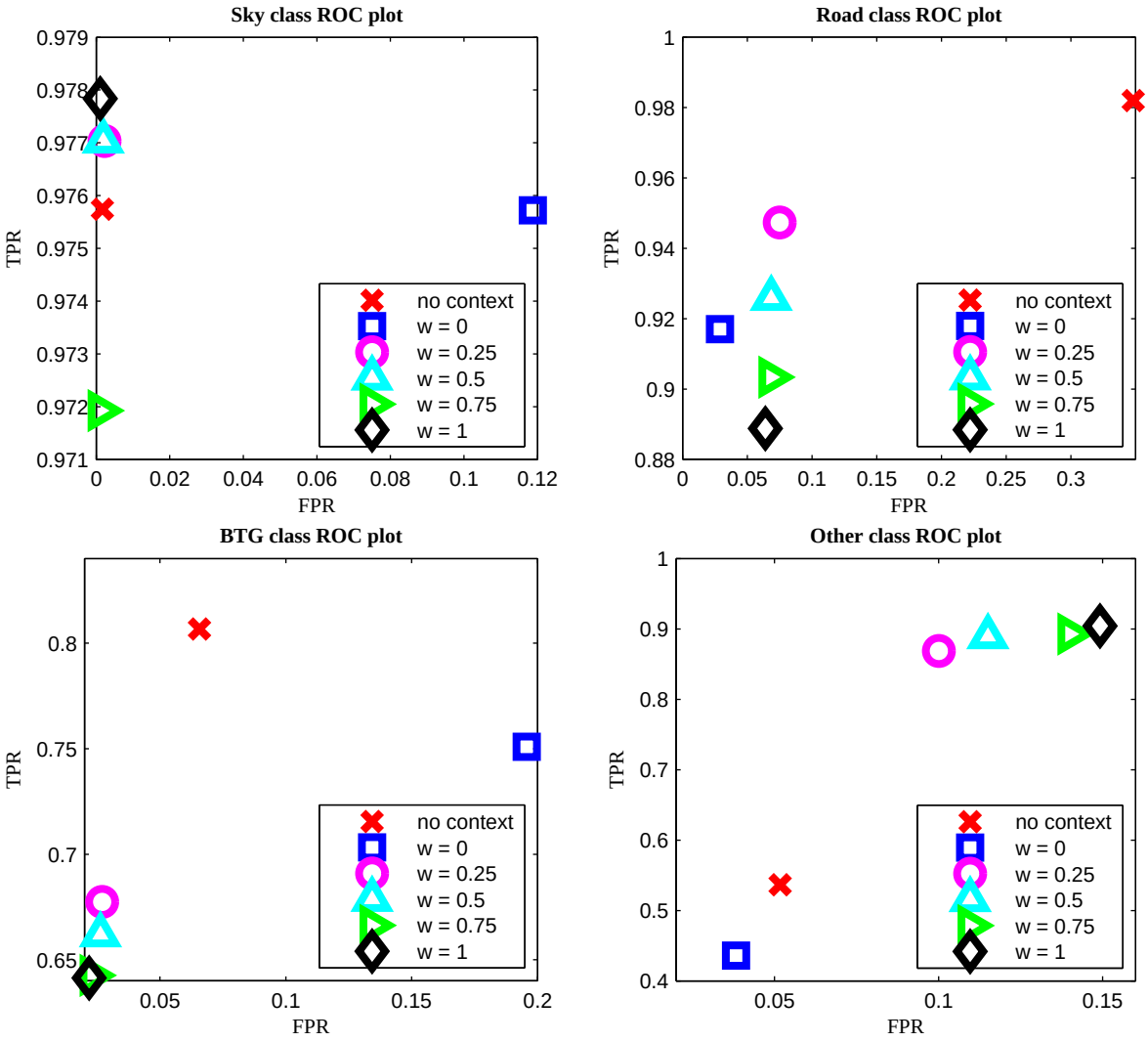


Figure 5. ROC plots for classes (*clockwise from top left*), (a) sky, (b) road, (c) other, (d) BTG (vegetation)


and appearance. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 1

[6] Geremy Heitz and Daphne Koller. Learning spatial context: Using stuff to find things. *ECCV*, 10:30–43, 2008. 1

[7] Li-Jia Li, Richard Socher, and Li Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 1

[8] Stephan Matzka. *Effcient Automotive Resource Allocation*. PhD thesis, Heriot-Watt University, 2009. 2

[9] Simon Ullman Moshe Bar. Spatial context in recognition. *Perception*, 23:343–352, 1996. 1

[10] Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–26, 2006. 1

[11] Roland Perko, Christian Wojek, Bernt Schiele, and Aleš Leonardis. Probabilistic combination of visual context based attention and object detection. *International Workshop on Attention in Cognitive Systems*, 2008. 1

[12] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. Objects in context. *ICCV*, 2007. 1

[13] Antonio Torralba, Kevin P. Murphy, William T. Freeman, and Mark A. Rubin. Context-based vision system for place and object recognition. *ICCV*, 2003. 1

[14] C. J. Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8:127–150, 1979. 3

[15] Lior Wolf and Stanley Bileschi. A critical view of context. *IJCV*, 69:251–261, 2006. 1

[16] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Quantifying contextual information for object detection. *IEEE International Conference on Computer Vision*, 2009. 1