Quantifying Contextual Information for Object Detection

Wei-Shi Zheng, Shaogang Gong and Tao Xiang
School of Electronic Engineering and Computer Science
Queen Mary University of London, London E1 4NS, UK
{jason,sgg,txiang}@dcs.qmul.ac.uk

Abstract

Context is critical for minimising ambiguity in object detection. In this work, a novel context modelling framework is proposed without the need of any prior scene segmentation or context annotation. This is achieved by exploring a new polar geometric histogram descriptor for context representation. In order to quantify context, we formulate a new context risk function and a maximum margin context (MMC) model to solve the minimization problem of the risk function. Crucially, the usefulness and goodness of contextual information is evaluated directly and explicitly through a discriminant context inference method and a context confidence function, so that only reliable contextual information that is relevant to object detection is utilised. Experiments on PASCAL VOC2005 and i-LIDS datasets demonstrate that the proposed context modelling approach improves object detection significantly and outperforms a state-of-the-art alternative context model.

1. Introduction

The role of visual context in object detection has been studied recently [4, 2, 23]. It has also been shown that by exploiting contextual information, object detection performance can be improved [12, 15, 13, 6, 25, 18, 11, 10]. However, the general problem of modelling visual context remains a challenging problem and is largely unsolved mainly due to two reasons: 1) There are many different types of context often co-existing with different degrees of relevance to the detection of target object(s) in different images. Adopting the terminology in [12], objects can be put into two categories: monolithic objects or "things" (e.g. cars and people) and regions with homogeneous or repetitive patterns, or "stuffs" (e.g. roads and sky). Consequently, there are Scene-Thing [15], Stuff-Stuff [22], Thing-Thing [18], and Thing-Stuff [12] context depending on what the target objects are and where the context comes from. Most existing work focuses only on one type of context and ignores the others. It remains unclear how different types of contextual information can be explored in a unified framework. 2) Contextual information can be ambiguous and unreliable, thus may not always have a positive effect on object detection. This is especially true in a crowded public scene such as an underground train platform with constant movement and occlusion among multiple objects. How to evaluate the usefulness and goodness of different types of context in a robust and coherent manner is crucial and has not been *explicitly* addressed.

In this paper, the two aforementioned problems are tackled by a novel context modelling framework which has four key (innovative) components: 1) a polar geometric histogram context descriptor for constructing a spatial relational graph between each candidate object and its context; 2) a context risk function that evaluates the effect of context; 3) a maximum margin context (MMC) model to minimize the risk of model misfitting and solve the problem of fusing context information with object appearance information; 4) a confidence function to directly and explicitly evaluate the goodness of the inferred context information.

The proposed approach has four major advantages over the existing models: 1) Rather than focusing on a single type of contextual information, our polar context descriptor offers greater flexibility in capturing different types of context including Thing-Thing, Thing-Stuff, and Thing-Scene contexts. 2) More does not necessarily mean better as not all contextual information is equally useful and reliable. Our discriminant context inference model addressing the context risk function with a context confidence measure evaluates explicitly and directly the available contextual information through learning. This differs significantly from the existing work where such an evaluation is either nonexistent or only done in an implicit and ineffective way. 3) Most existing approaches [18, 10, 24, 13, 6] rely on the laborious and often arbitrary manual annotation/labelling of both target objects and context with the exception of the Thing-and-Stuff (TAS) model proposed in [12]. In contrast, our approach only needs the labelling of the target object class(es) and thus is able to exploit data for contextual object detection in a more unsupervised way. Comparing with

[12], our approach does not require global image segmentation for contextual information extraction, which could be unreliable especially for a cluttered scene. Moreover, our approach is not limited to only Thing-Stuff context. 4) Most existing approaches relate contextual information with object appearance information using a graphical model, either directed [12, 11] or undirected [13, 6]. Our approach differs significantly in that the MMC model is based on discriminant analysis and thus computationally more efficient.

The effectiveness of our approach is evaluated using the PASCAL Visual Object Classes challenge datasets [8] and the UK Home Office i-LIDS data [1]. The latter is featured with a busy underground station where the task is to detect different types of luggages. Our results demonstrate that the proposed MMC context model can improve the detection performance for all object classes. In addition, it is also shown that our context model clearly outperforms a state-of-the-art alternative model from [12], and the improvement is especially significant in the more challenging i-LIDS dataset.

2. A Polar Geometric Context Descriptor

In our approach, contextual information is extracted mainly from the surrounding area of each candidate object. The candidate objects can be obtained using any existing sliding window object detection method, and in this paper the histogram of oriented gradients (HOG) detector [7] is adopted. A low threshold is used to ensure that the candidate object detection windows contain most of the true positives.

Figure 1. Examples of the polar geometric structure. The target object classes are car and people respectively in the left and right images.

Given a candidate object window $\mathbf{W}_c$, a polar geometric structure is expanded from the centroid of the candidate object (see Fig. 1) to represent the context information surrounding the object detection window. With a orientational and $b+1$ radial bins, the context region centered around the candidate object is divided into $a \cdot b+1$ patches with a circle one at the centre, denoted by $\mathcal{R}_i$, $i = 1, \cdots, (a \cdot b + 1)$. In this paper b is set to 2 and a is set to a value between 1 and 16 depending the object categories. The size of the polar context region is proportional to that of the candidate object window $\mathbf{W}_c$. Specifically, the lengths of the bins along the radial direction are set to 0.5σ, σ and 2σ respectively from inside towards outside of the region, where σ is the minimum of the height and width of the candidate detection window $\mathbf{W}_c$. As shown in Fig. 1, our polar context region bins are expanded from the centroid of the object, and thus have two key characteristics: 1) It can potentially represent many existing spatial relationships between objects and their context used in the literature, including inside, outside, left, right, up, down, co-existence. 2) The regions closer to the object are given bins with finer scale. This makes perfect sense because intuitively, the closer the context is, the more relevant it is, and from which more information should be extracted.

The polar context region is represented using the Bag of Words (BoW) method which is robust to partial occlusion and image noise. To build the code book, SIFT features [14] are extracted densely [5]. These features are invariant to scale, orientation, and robust to changes in illumination and noise. They are thus well suited for representing our polar context region. More specifically, given a training dataset, SIFT features are extracted from each image and clustered into code words using K-means with K set to 100 in this paper. Subsequently for each bin in the polar context region, a normalised histogram vector [9] is constructed, entries of which correspond to the probabilities of the occurrences of visual words in that bin. These histogram vectors are then concatenated together with the context inside the detection window which is represented using a single histogram vector to give the final context descriptor for the object candidate window $\mathbf{W}_c$, denoted as $\mathbf{h}_c$. The interaction between the context from inside and outside of the detection window would be inferred by the proposed context model.

Our polar context descriptor differs from alternative polar context descriptors [25, 16] in that the Bag-of-Words method is employed. Additionally, context features are extracted densely to cope with low image resolution and noise in our method, while only some predetermined sparse pixel locations were considered for context feature extraction in [25, 16].

3. A Discriminant Context Model

3.1. Quantify Context

Without relying on segmentation, our polar context region contains useful contextual information as well as information that is irrelevant to the detection task. Therefore for constructing a meaningful context model, these two types of information must be separated. To that end, we introduce a risk function to evaluate context with the help of a prior detection score given by the sliding window HOG detector.

Consider a training set of N candidate detections $\mathcal{O} = \{\mathbf{O}_i\}_{i=1}^{N}$ and each of the detection window has an associ-

ated probability of the target object class being contained in the window $s_i = P(\mathbf{O}_i|\mathbf{W}_i)$, where $\mathbf{W}_i$ is the corresponding detection window. Suppose that the ground truth information about the target object class is available at the training stage, and without loss of generality let the first ℓ candidate detections $\mathcal{O}_p = \{\mathbf{O}_i\}_{i=1}^{\ell}$ be true positive detections and the last $N-\ell$ detections $\mathcal{O}_n = \{\mathbf{O}_i\}_{i=\ell+1}^{N}$ be false positives. Let us denote the polar context descriptor corresponding to $\mathbf{O}_i$ by $\mathbf{h}_i$ and define $\mathcal{H}_p = \{\mathbf{h}_i\}_{i=1}^{\ell}$ and $\mathcal{H}_n = \{\mathbf{h}_i\}_{i=\ell+1}^{N}$.

A context model is sought to increase the confidences of those true positive detections. Specifically, we aim to learn a leverage function g to leverage any prior detection confidence s such that the posterior confidence of the true positive detection is larger than the false positive, i.e. $s_i^{\alpha} \cdot g(\mathbf{h}_i) > s_j^{\alpha} \cdot g(\mathbf{h}_j)$, where $\mathbf{h}_i \in \mathcal{H}_p$, $\mathbf{h}_j \in \mathcal{H}_n$, and α measures the importance of the prior detection probability. To that end, a leverage function g which minimizes the following context risk function is learned as follows:

$$\mathcal{L} = \sum_{\mathbf{h}_i \in \mathcal{H}_p} \sum_{\mathbf{h}_j \in \mathcal{H}_n} \delta(s_i^{\alpha} \cdot g(\mathbf{h}_i) \leq s_j^{\alpha} \cdot g(\mathbf{h}_j)), \quad (1)$$

where δ is a boolean function with $\delta(true) = 1$ and 0 otherwise. This risk function measures the rank information between true positives and false positives. The smaller the risk function is, the more confident the detection would be expected. In our current model, we define the leverage function as:

$$g(\mathbf{h}_i) = \exp\{f(\mathbf{h}_i)\}, \quad (2)$$

where $f(\mathbf{h}_i)$ is the score of context descriptor $\mathbf{h}_i$, the larger the f is the more positive the context information tends to be. Here, f is formulated as a kernel linear function:

$$f(\mathbf{h}_i) = \mathbf{w}_t^T \varphi(\mathbf{h}_i) + b_t, \quad (3)$$

where φ is a nonlinear mapping implicitly defined by a Mercer kernel κ such that $\varphi(\mathbf{h}_i)^T \varphi(\mathbf{h}_j) = \kappa(\mathbf{h}_i, \mathbf{h}_j)$. Kernel trick is used because the descriptor we introduce (i.e. a histogram) is a distribution representation . In this way, the popular exponent $\mathcal{X}^2$ distance kernel [9] for the estimation of the distance between two discrete distributions can be employed, as it is a Mercer kernel. Note that the variable b_t does not have any impact on the risk function up to now, but it will be useful for learning a much better $\mathbf{w}_t$ in an approximate way. This is because a more flexible solution for $\mathbf{w}_t$ can be found by utilising b_t at the training stage, as we shall describe next (see Eq. (5)).

An ideal case for minimizing the risk function is to find $\mathbf{w}_t^T$ and b_t such that the following inequalities are satisfied:

$$f(\mathbf{h}_i) + \log s_i^{\alpha} > f(\mathbf{h}_j) + \log s_j^{\alpha}, \forall \mathbf{h}_i \in \mathcal{H}_p, \mathbf{h}_j \in \mathcal{H}_n. \quad (4)$$

Directly solving this problem is hard if not impossible and would also be a large scale optimization problem. For example, if $\#\mathcal{H}_p = 100$ and $\#\mathcal{H}_n = 100$, there will be 10000 inequalities for consideration. Therefore, an approximate solution is required. We approach the problem of minimizing the risk function by investigating a solution constrained by a margin $\rho(\geq 0)$ as follows:

$$\begin{aligned} f(\mathbf{h}_i) + \log s_i^{\alpha} &\geq \rho, \ \forall \mathbf{h}_i \in \mathcal{H}_p, \\ f(\mathbf{h}_j) + \log s_j^{\alpha} &\leq -\rho, \ \forall \mathbf{h}_j \in \mathcal{H}_n. \end{aligned} \quad (5)$$

Ideally, the constraints in Eq. (4) would be satisfied if the above constraints are valid. For approximation, we would like to learn the function such that the margin ρ is as large as possible. Also, by revisiting Eq. (3), the inference of useful and discriminant context information is actually performed by the projection $\mathbf{w}$ in the high dimensional kernel feature space. We therefore also like to maximize the margin between positive and negative context, so the ℓ_2-norm on $\mathbf{w}$ would be minimized. To this end, we turn to the following optimization problem:

$$\begin{aligned} &\{\mathbf{w}_t, b_t\} = \arg\min_{\mathbf{w},b,\rho} \frac{1}{2}||\mathbf{w}||^2 - \nu\rho \\ &s.t. \quad \mathbf{w}^T\varphi(\mathbf{h}_i) + b \geq \rho - \log s_i^{\alpha}, \ \forall \mathbf{h}_i \in \mathcal{H}_p, \\ &\qquad \mathbf{w}^T\varphi(\mathbf{h}_j) + b \leq -\rho - \log s_j^{\alpha}, \ \forall \mathbf{h}_j \in \mathcal{H}_n, \\ &\qquad \rho \geq 0. \end{aligned} \quad (6)$$

The model is similar to but differs from the ν-SVM [21] in the extra constants $\log s_i^{\alpha}$ and $\log s_j^{\alpha}$ as α is not considered as a variable. It would also result in a slightly different dual problem for solution.

In order to automatically learn the importance factor α, we define $\dot{\mathbf{w}}_t = [\mathbf{w}_t^T, \alpha]^T$ and $\psi(\mathbf{h}_i, s_i) = [\varphi(\mathbf{h}_i)^T, \log s_i]^T$. Subsequently, inequalities Eq. (5) is reformulated as

$$\begin{aligned} \dot{\mathbf{w}}_t^T\psi(\mathbf{h}_i, s_i) + b_t &\geq \rho, \ \forall \mathbf{h}_i \in \mathcal{H}_p, \\ \dot{\mathbf{w}}_t^T\psi(\mathbf{h}_j, s_j) + b_t &\leq -\rho, \ \forall \mathbf{h}_j \in \mathcal{H}_n. \end{aligned} \quad (7)$$

Similar to Eq. (6), the optimization problem becomes:

$$\begin{aligned} &\{\dot{\mathbf{w}}_t, b_t\} = \arg\min_{\mathbf{w},b,\rho} \frac{1}{2}||\mathbf{w}||^2 - \nu\rho \\ &s.t. \quad \mathbf{w}^T\psi(\mathbf{h}_i, s_i) + b \geq \rho, \ \forall \mathbf{h}_i \in \mathcal{H}_p, \\ &\qquad \mathbf{w}^T\psi(\mathbf{h}_j, s_j) + b \leq -\rho, \ \forall \mathbf{h}_j \in \mathcal{H}_n, \\ &\qquad \rho \geq 0. \end{aligned} \quad (8)$$

where the positive ν will be learned by cross-validation. To obtain the solution, a new mercer kernel $\dot{\kappa}$ for ψ can be defined as $\dot{\kappa}(\{\mathbf{h}_i, s_i\}, \{\mathbf{h}_j, s_j\}) = \kappa(\mathbf{h}_i, \mathbf{h}_j) + \log s_i \cdot \log s_j$. In addition, satisfying all the constraints in model (8) could be hard and in practice some positive slack variables ξ_i should be introduced as auxiliary variables.

We refer the above model as the *maximum margin context model* (MMC). It utilises the prior detection results obtained by the sliding window detector and enables the model to selectively learn useful discriminant context information so that the confidence of those marginal true positive detections are maximised. After selecting and quantifying contextual information from the context descriptor using the MMC model, a posterior detection score sc_i for the candidate detection $\mathbf{O}_i$ with the context model is defined as:

$$sc_i = s_i^{\alpha} \times \exp\{\mathbf{w}_t^T \varphi(\mathbf{h}_i) + b_t\}. \tag{9}$$

Remark. In the above models, the importance factor α is not restricted to be non-negative. Automatically learned by the model, α determines not only the importance of the prior detection probability but also whether the approximation Eq. (5) is sufficiently valid. Specifically, when α equals zero, it means the prior probability would not have any effect on the maximum margin model and should also be ignored in the risk function and the posterior detection score. When α is less than zero, the smaller the prior detection probability is, the larger the posterior score is expected. This is because $s_i \in (0, 1]$ and the leverage function $g(\mathbf{h}_i)$ is always bounded by investigating the dual problem of model Eq. (8). For $\alpha > 0$, the larger it is, the more important the prior detection probability is and the less useful the contextual information will be. However, a very large α value will mean the contextual information is completely discarded. Hence if the system outputs a very small or large α, both cases could imply that either the approximation for minimization of context risk function by our MMC model cannot be made, or the context descriptors do not contain sufficiently useful contextual information.

3.2. A Confidence Measure on Context

So far whatever contextual information is inferred by our MMC model is considered to be reliable. This may not always be the case. In some cases, without explicitly measuring the goodness of context, ambiguous contextual information could even reduce the performance of object detection performance. In this section, we introduce a confidence function $c(\mathbf{h}_i)$ which quantifies the goodness of the polar geometric histogram context descriptor $\mathbf{h}_i$. To that end, the posterior detection score in Eq. (9) is rewritten as:

$$sc_i = s_i^{\alpha} \times \exp\{c(\mathbf{h}_i) \cdot \mathbf{w}_t^T \varphi(\mathbf{h}_i) + b_t\}, \tag{10}$$

A low value of $c(\mathbf{h}_i)$ indicates unreliable contextual information.

To learn the confidence function, we first explore the ambiguous context information that is hard to be identified as positive or negative context, i.e. context associated with positive and negative detection windows respectively. In our case, the ambiguous context information $\mathcal{H}_c^c$ is simply extracted by using a threshold ϵ on the $\mathcal{X}^2$ distances between context descriptors as follows:

$$\begin{aligned}\mathcal{H}_c^c = \{\mathbf{h}_i \in \mathcal{H}_p,\ \min\{\mathcal{X}^2(\mathbf{h}_i, \mathbf{h}_j), \forall \mathbf{h}_j \in \mathcal{H}_n\} < \epsilon\} \\ \bigcup \{\mathbf{h}_j \in \mathcal{H}_n,\ \min\{\mathcal{X}^2(\mathbf{h}_i, \mathbf{h}_j), \forall \mathbf{h}_i \in \mathcal{H}_p\} < \epsilon\}.\end{aligned} \tag{11}$$

Detecting whether $\mathbf{h}_i$ belongs to $\mathcal{H}_c^c$ can be considered as outlier detection. Since the proposed context descriptor is of high dimension, we then learn $c(\mathbf{h}_i)$ using one-class SVM [20], which is a non-parametric technique and thus suitable for high dimensional distribution estimation, as follows:

$$\begin{aligned}\{\mathbf{w}_c, b_c\} = \arg\min_{\mathbf{w}, b, \xi_i} \frac{1}{2}||\mathbf{w}||^2 + b + \frac{\sum_i \xi_i}{\nu \cdot \#\mathcal{H}_c^c} \\ s.t.\ \ \mathbf{w}^T \varphi(\mathbf{h}_i) + b \geq -\xi_i,\ \forall \mathbf{h}_i \in \mathcal{H}_c^c,\ \ \xi_i \geq 0.\end{aligned} \tag{12}$$

where ν is a cross-validated positive value. The confidence function $c(\mathbf{h}_i)$ is then modeled as follows:

$$c(\mathbf{h}_i) = -1 \times (\mathbf{w}_c^T \varphi(\mathbf{h}_i) + b_c) \times \delta(\mathbf{w}_c^T \varphi(\mathbf{h}_i) + b_c < 0). \tag{13}$$

Threshold ϵ. The ϵ in Eq. (11) can be specified by

$$\epsilon = \mathcal{X}_{pn}^{2,min} + d \cdot (\mathcal{X}_{pn}^{2,max} - \mathcal{X}_{pn}^{2,min}), \tag{14}$$

where $d \in (0, 1]$, $\mathcal{X}_{pn}^{2,min} = \min\{\mathcal{X}^2(\mathbf{h}_i, \mathbf{h}_j),\ \mathbf{h}_i \in \mathcal{H}_p, \mathbf{h}_j \in \mathcal{H}_n\}$ and $\mathcal{X}_{pn}^{2,max}$ is formulated in a similar way. In addition, for convenience, we will treat the MMC model discussed in the last section as the case when $d = 0$.

Training. With the confidence function, the learning process in the last section is modified so that only confident context descriptors are used to learn our MMC model. More specifically, we specify confident positive context information $\mathcal{H}_p^c$ and negative $\mathcal{H}_n^c$ as:

$$\mathcal{H}_p^c = \mathcal{H}_p - \mathcal{H}_{fp},\ \ \mathcal{H}_n^c = \mathcal{H}_n - \mathcal{H}_{fn}. \tag{15}$$

where $\mathcal{H}_{fp} = \{\mathbf{h}_i \in \mathcal{H}_c^c \bigcap \mathcal{H}_p\} \bigcup \{c(\mathbf{h}_i) = 0, \mathbf{h}_i \in \mathcal{H}_p\}$ and $\mathcal{H}_{fn} = \{\mathbf{h}_i \in \mathcal{H}_c^c \bigcap \mathcal{H}_n\} \bigcup \{c(\mathbf{h}_i) = 0, \mathbf{h}_i \in \mathcal{H}_n\}$. Then the maximum marginal context model is trained just by updating $\varphi(\mathbf{h}_i)$, $\mathcal{H}_p$ and $\mathcal{H}_n$ in Eq. (8) as:

$$\varphi(\mathbf{h}_i) \leftarrow c(\mathbf{h}_i) \cdot \varphi(\mathbf{h}_i),\ \ \mathcal{H}_p \leftarrow \mathcal{H}_p^c,\ \ \mathcal{H}_n \leftarrow \mathcal{H}_n^c. \tag{16}$$

4. Experiments

Datasets and settings – We evaluate the proposed MMC model against a state-of-the-art context model TAS [12] and a traditional HOG model [7] using two datasets: PASCAL Visual Object Classes challenges 2005 [8] for detecting four object categories (car, motorbike, people, bicycle), and the UK Home Office i-LIDS [1] for detecting two types of luggages (suitcases and bags) (see Fig. 5).

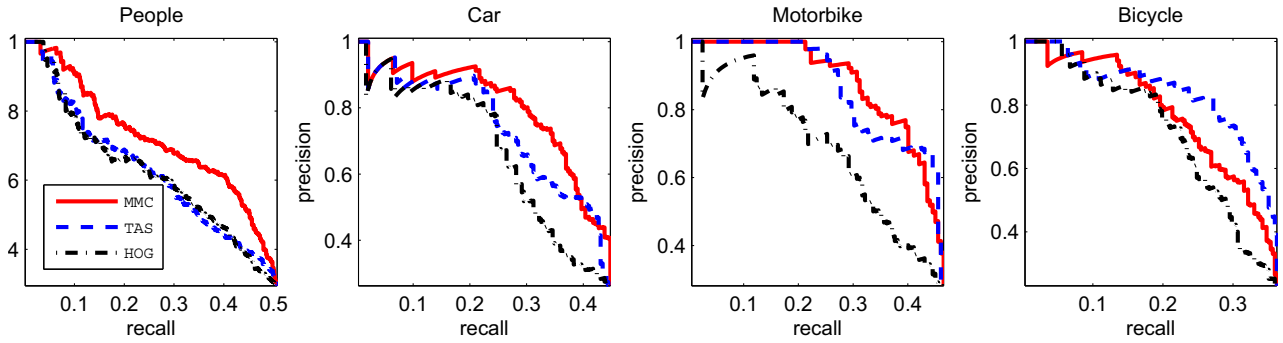

Figure 2. Precision-Recall Curves: Detection of four object categories in PASCAL VOC2005.

For PASCAL, we adopted the same data and settings from [12]. For i-LIDS, we selected 658 image frames of 640×480 from the i-LIDS underground scenario, where 269 for training and the rest for testing. For context model training on i-LIDS, we first train a pair of HOG luggage detectors (refer to as the HOG detector) using randomly selected 540 positive samples for each, 7278 and 5047 negative samples for the two detectors respectively. We then manually cropped and annotated 328 positive luggage sample windows from i-LIDS image frames. Due to frequent luggage partial occlusion in the i-LIDS scenario, we further used the trained HOG detector to select automatically another 628 near-miss negative luggage sample windows. Here we utilsed both positive and near-miss negative sample windows from above as candidate luggages for learning a MMC context model with each sample associated with a prior confidence value obtained by converting the HOG detector output into a probability value using a logistic regression function [12]. By assigning the prior confidence to s_i in Eq. (1), a MMC model is learned. Separate i-LIDS testing image frames consisting of 1170 true luggage instances were selected with groundtruth manually annotated for performance evaluation. The threshold of the overlap between the predicted object bounding box and the groundtruth one the was set to 0.5 [8].

Compare MMC to no context (HOG) and local template context (HOG+SVM) modelling – We first compare the performance of our MMC model to that of a standard HOG detector [7]. We set the number of orientation bin to 16, 16, 1 and 1 for people, cars, motorbike and bicycles respectively for computing the polar geometric histogram context descriptor. For PASCAL, the red and dashed-black plots in Fig.2 show the precision-recall curves for MMC and HOG respectively. Columns 2 and 4 in Table 1 give their average precision rates defined by the PASCAL protocol [8]. Both show that MMC significantly improves the detection accuracy especially on car, motorbikes and people. We also show a clear advantage of modeling explicitly object-centred (local) and location dependent (dynamic) inter-relationships between each object candidate and its context using MMC over a direct template-based context modelling approach. Column 5 (HOG+SVM) of Table 1 shows the detection precision rate by using a normalised joint product of the HOG detection score and the score from a template-based context classifier trained on SVM [21] using our polar context descriptors as features. This method essentially performs naive score level fusion of context and object appearance. It is very similar to the methods in [16][17] in that they all assume the context and object are independent and treat them equally during learning, rather than inferring the most useful and reliable contextual information conditioned on the prior object detection score as our method does. The results show that fusing the inferred contextual information using our MMC is more effective than direct and blind fusion using HOG+SVM.

Object Class	HOG [7]	TAS [12]	MMC	HOG+SVM
Car	0.325	0.363	**0.3773**	0.3467
Motorbike	0.341	0.390	**0.4238**	0.4019
People	0.346	0.346	**0.3924**	0.3706
Bicycle	0.281	**0.325**	0.3025	0.2692

Table 1. Average Precision Rates on PASCAL VOC2005.

Luggage detection results on i-LIDS are shown in Table 2 and Fig. 3. Here the number of orientation bin is set to 16 for luggages. It is evident that MMC outperforms HOG when context confidence is measured, where $(d = 0)$ indicates MMC without confidence function (Eq. (13)) and $(d = 0.1)$ with confidence function (see more details on the effect of confidence function latter). These detection rates are relatively poorer than those from PSACAL. In general reliable luggage detection in public space is more challenging as color and texture features are ineffective especially in low-resolution, poorly lit underground scenes (Fig. 5). Objects in PASCAL are more distinct with more discriminative appearance. Furthermore, context information surrounding luggages in i-LIDS is less well-defined than that of objects in PASCAL VOC2005 (see the effect of confidence measure on false positive detection later).

HOG [7]	TAS [12]	HOG+SVM	MMC (d=0)	MMC (d=0.1)
0.1195	0.1180	0.1185	0.1184	**0.1271**

Table 2. Average Precision Rate on luggage detection in i-LIDS.

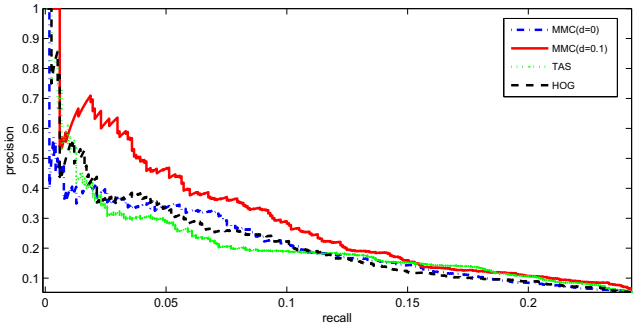

Figure 3. Precision-Recall Curves: Luggage detection in i-LIDS.

MMC vs. TAS – On the PASCAL data, we compared the results from MMC with the best reported results of TAS [12] in Table 1. For obtaining the precision-recall curve for TAS, we re-run the TAS model provided by the authors [1] several times and the best results are shown as blue-dashed plots in Fig 2. Note that TAS is an EM based method and these curves are either very similar or slightly better (e.g. motorbike) than those originally shown in [12]. Overall, MMC outperforms TAS on the detection of car, motorbike and people whilst TAS is better for bicycle.

In particular, MMC improved the detection of people with a fairly large margin. As acknowledged by the authors in [12], the TAS model struggles with people detection in the PASCAL data. This can be caused by two factors. First, as people appear more randomly compared to other rigid objects such as cars on a street, the contextual information for people is more ambiguous and uncertain than the other three object classes. Without measuring the risk of using contextual information for detection explicitly, the existing context model such as TAS will not be able to exploit effectively the ambiguous contextual information for object detection improvement. Second, the TAS model is focused on Thing-Stuff context, i.e. the context between people and the background regions. The useful contextual information between people and other objects is thus ignored. In contrast, our model is able to utilise any contextual information that is relevant regardless the type of the context.

Note that MMC achieves slightly lower average precision rate than TAS on bicycle class, because the bicycle class is unique with no clear boundary between the object and background. In such a case, alternative models such as TAS with scene segmentation may be less affected, although segmentation itself is challenging under occlusion in a cluttered scene.

We also implemented TAS for the i-LIDS dataset using the same HOG detector as base detector for initialising object candidate locations in image frames. For TAS, we segmented each image frame using the superpixel technique [19] and represented each region using 44 features (color, shape, energy responses) similar to the ones used in [12, 3] and implemented TAS with the suggested parameter values given by the authors in their toolkit available on the web. The results are shown in Table 2 and Fig. 4 and Fig.5. As shown, MMC outperforms both TAS and standard HOG with clear margin. It is also evident that the result of TAS resembles that of PASCAL people detection. This can be explained by the same two reasons described above. This also demonstrates that without any segmentation of a whole image, more effective context information can also be learned.

Confidence function evaluation – We further evaluated the effect of the confidence function (Eq. (13)) on regulating context ambiguity in MMC applied to the i-LIDS data (d set to 0.1). The last two columns in Table 2 compare the results and show a sizeable improvement on average precision rate by the introduction of our confidence function. Table 2 also shows that without this confidence measure, the performance of MMC on i-LIDS is very similar to that of TAS and HOG+SVM. This suggests that the 0.87% improvement due to modelling confidence on MMC is significant on reducing false positives, as illustrated by the examples in Fig. 5. This is further supported by the precision-recall curves shown in Fig. 3 where MMC with $d = 0.1$ is significantly better than the other three when the recall rate is less than 0.1.

Reducing false positive detections – Finally, we show some visual examples to illustrate the benefit of our MMC model on reducing false positive detections. Fig. 4 and Fig 5 give some typical examples of false positive detection in both PASCAL and i-LIDS. For all methods, we illustrate the detection results when the recall-rate is at 0.3 for PASCAL and 0.1 for i-LIDS. It is evident from these examples that our MMC model is more capable of removing false positives whilst keeping true positives compared to both TAS and HOG. More specifically, without context modeling, HOG often cannot differentiate true positives and false positives. Although both TAS and MMC can filter out false positive detections, MMC is more effective. Particularly, it is evident that TAS tends to generate false positives or fail to detect when luggages are near people. Again, this is because the crucial contextual information between luggage and other objects (people in this case) cannot be captured by TAS. Fig. 6 shows some examples of failed detections by all three models. This is mainly due to drastic illumination variation and severe occlusion.

5. Conclusion

In this paper we introduced a novel object centred polar geometric histogram context descriptor to represent lo-

[1] http://ai.stanford.edu/∼gaheitz/Research/TAS/

Figure 4. Compare examples of object detections using HOG, TAS and MMC models on PASCAL. The left-hand side two columns are for people detection, the middle two are for car detection, and the right-hand side two are for motorbike detection. The first row corresponds to results from HOG without threshold, the second, third and fourth rows correspond to HOG, TAS and MMC with threshold respectively. The red bounding box indicates true positive detections and the green one is for false positives.

Figure 5. Compare examples of object detections using HOG, TAS and MMC models on i-LIDS. The first row corresponds to results from HOG without threshold. The second, third and fourth rows correspond to HOG, TAS and MMC with threshold respectively. The red bounding box indicates true positive detections and the green one is for false positives.

cal context surrounding of a candidate object. In order to quantify this context, we formulated a new context risk function and a maximum margin context (MMC) model to solve the minimization problem of the risk function. Our

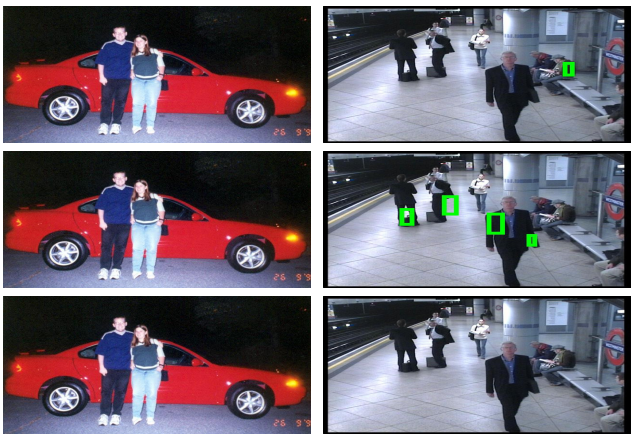

Figure 6. Examples of failed detections. The first, second and third rows correspond to results of HOG, TAS and MMC with threshold respectively. The green bounding box shows false positive detections.

MMC is effectively solved by convex quadratic programming. Compared to the state-of-the-art context models, the proposed MMC model utilises a novel confidence on measuring the goodness of context in order to selectively employ context for more robust object detection. The proposed MMC model also differs from existing models that utilise graph based context information mining. To that end, our MMC model directly addresses the maximization of the confidence of true positive detections defined by a context risk function, whilst a graph model addresses indirectly by classification without any knowledge or measurement on the rank information between true and false positive detections. Moreover, our MMC model does not require any prior image segmentation and labelling of image patches. We demonstrated the superior performance of our MMC model through extensive comparative evaluation against alternative models using both PASCAL VOC2005 and UK Home Office i-LIDS datasets. We also showed that a quantitative measure of the goodness of context is critical in reducing false positive detections in challenging scenes.

Acknowledgement

This research was partially funded by the EU FP7 project SAMURAI with grant no. 217899.

References

[1] UK Home Office. i-LIDS Multiple Camera Tracking Scenario Definition. 2008. 2, 4

[2] M. Bar and S. Ullman. Spatial context in recognition. *Perception*, 25:343–352, 1993. 1

[3] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. Jordan. Matching words and pictures. *JMLR*, 3, 2003. 6

[4] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14:143–177, 1982. 1

[5] A. Bosch, A. Zisserman, and X. M. noz. Scene classification via plsa. In *ECCV*, 2006. 2

[6] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *ECCV*, 2004. 1, 2

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2, 4, 5, 6

[8] M. Everingham. The 2005 pascal visual object classes challenge. 2005. 2, 4, 5

[9] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nystrom method. *PAMI*, 26(2):214–225, 2004. 2, 3

[10] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, 2008. 1

[11] A. Gupta and L. S. Davis. Beyond nouns: exploiting prepositions and comparative adjectives for learning visual classifier. In *ECCV*, 2008. 1, 2

[12] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008. 1, 2, 4, 5, 6

[13] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *ICCV*, 2005. 1, 2

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2(60):91–110, 2004. 2

[15] K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the tree: a graphical model relating features, objects and the scenes. In *NIPS*, 2003. 1

[16] R. Perko and A. Leonardis. Context driven focus of attention for object detection. In *WAPCV*, 2007. 2, 5

[17] R. Perko, C. Wojek, B. Schiele, and A. Leonardis. Probabilistic combination of visual context based attention and object detection. In *WAPCV*, 2008. 5

[18] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007. 1

[19] X. Ren and J. Malik. Learning a classification model for segmentation. In *ICCV*, 2003. 6

[20] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7):1443–1471, 2001. 4

[21] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Comput.*, 12(5):1207–1245, 2000. 3, 5

[22] A. Singhal, J. Luo, and W. Zhu. Probabilistic spatial context models for scene content understanding. In *CVPR*, 2003. 1

[23] A. Torralba. Contextual priming for object detection. *IJCV*, 53(2), 2003. 1

[24] A. Torralba, K. Murphy, W. Freeman, and M. Rubin. Context-based vision system for place and object recognition. In *ICCV*, 2003. 1

[25] L. Wolf and S. Bileschi. A critical view of context. *IJCV*, 69(2), 2006. 1, 2