Quantifying and Transferring Contextual Information in Object Detection

Wei-Shi Zheng, *Member, IEEE,* Shaogang Gong, and Tao Xiang

***Abstract*— Context is critical for reducing the uncertainty in object detection. However, context modelling is challenging because there are often many different types of contextual information co-existing with different degrees of relevance to the detection of target object(s) in different images. It is therefore crucial to devise a context model to automatically quantify and select the most effective contextual information for assisting in detecting the target object. Nevertheless, the diversity of contextual information means that learning a robust context model requires a larger training set than learning the target object appearance model, which may not be available in practice. In this work, a novel context modelling framework is proposed without the need for any prior scene segmentation or context annotation. We formulate a polar geometric context descriptor for representing multiple types of contextual information. In order to quantify context, we propose a new maximum margin context (MMC) model to evaluate and measure the usefulness of contextual information directly and explicitly through a discriminant context inference method. Furthermore, to address the problem of context learning with limited data, we exploit the idea of transfer learning based on the observation that although two categories of objects can have very different visual appearance, there can be similarity in their context and/or the way contextual information helps to distinguish target objects from non-target-objects. To that end, two novel context transfer learning models are proposed which utilise training samples from source object classes to improve the learning of the context model for a target object class based on a joint maximum margin learning framework. Experiments are carried out on PASCAL VOC2005 and VOC2007 datasets, a luggage detection dataset extracted from the i-LIDS dataset, and a vehicle detection dataset extracted from outdoor surveillance footages. Our results validate the effectiveness of the proposed models for quantifying and transferring contextual information, and demonstrate that they outperform related alternative context models.**

***Index Terms*— Context modelling, object detection, transfer learning**

I. Introduction

It has long been acknowledged that visual context plays an important role in visual perception of object [7], [3]. Consequently, there has been an increasing interest in recent years in developing computational models to improve object detection in images by exploiting contextual information [20], [13], [45], [25], [31], [28], [9], [48], [38], [24], [22], [13]. These existing studies show that by fusing contextual information with object appearance information, the uncertainty in object detection can be reduced leading to more accurate and robust detection, especially in images with cluttered background, low object resolution and severe occlusions (see Figure 1 for examples). In particular, large amount of false detections can be filtered out when the object contextual information is examined, e.g. luggage is normally carried by or in a vicinity of a person rather than on a wall or train carriage (see Figure 1(a)).

(a) Luggage detection on an underground platform (b) Vehicle detection at an airport forecourt

Fig. 1. Examples of improving object detection using contextual information. The top row shows the detection results by a HOG detector [12] without context modelling, and the second row shows the results obtained by using the same detector with the proposed context model. The green and red bounding boxes indicate false detections and true detections respectively.

However, modelling visual context remains a challenging problem and is largely unsolved mainly due to the following reasons: 1) **Diversity of contextual information –** 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 [25], objects in a visual scene 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 [31], Stuff-Stuff [44], Thing-Thing [38], and Thing-Stuff [25] 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) **Ambiguity of contextual information –** 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. 3) **Lack of data for context learning –** It is well known that visual object detection is a hard problem because of the large intra-class variation of object appearance; learning an object appearance model thus often faces the problem of sparse training data which can lead to

Wei-Shi Zheng is now with School of Information Science and Technology, Sun Yat-sen University, China, and was with School of Electronic Engineering and Computer Science, Queen Mary University of London, UK, wszheng@ieee.org

Shaogang Gong and Tao Xiang are with School of Electronic Engineering and Computer Science, Queen Mary University of London, {sgg,txiang}@eecs.qmul.ac.uk

Digital Object Indentifier 10.1109/TPAMI.2011.164

(a) (b)

Fig. 2. Transferrable knowledge can be extracted and shared between object categories. (a) Cars and motorbikes have similar contextual information. (b) It is not the case for people and bicycles but context in general provides a similar level of assistance in detection. In both cases, transfer learning can help to address the problem of learning context from limited data.

model over-fitting. However, the problem of learning with limited training data is much more acute for context learning because the variation of contextual information and the variation of its degree of relevance to the detection of target object can be larger. For instance, some objects such as people can appear everywhere and certain contextual information (e.g. on top of a sofa) can be more useful than others (e.g. on top of grass).

In this paper, the three aforementioned problems are tackled by a novel context modelling framework with three key components:

- **A polar geometric descriptor for context representation –** We formulate a polar geometric context descriptor for representing multiple types of contextual information. This representation offers greater flexibility in capturing different types of context including Thing-Thing and Thing-Stuff context compared to existing representation methods most of which focus on a single type of contextual information. It avoids the tedious and unreliable process of manual labelling of object context required by most existing methods.
- **A maximum margin context model (MMC) for quantifying context –** More does not necessarily mean better as not all contextual information is equally useful and reliable. To evaluate and measure the relevance and usefulness of different contextual information, we propose a context risk function and formulate a MMC model which is a discriminant context inference model designed to minimize the risk of model misfitting and solve the problem of fusing context information with object appearance information.
- **A context transfer learning model for context learning with limited data –** We exploit the idea of transfer learning based on the observation that although two categories of objects can have different visual appearance, there can be similarity in their context and/or the way contextual information helps to disambiguate target objects from non-target-objects. For instance, as shown in Figure 2(a) cars and motorbikes can look quite different, but due to their similar functionalities (transport tools for human), there can be common contextual information that has a similar effect on detecting cars and motorbikes (e.g. roads underneath a candidate object). The availability of a set of training images of cars can thus be useful for learning a context model for motorbikes and vice versa. It is also noted that even for seemingly unrelated object categories, there can be useful knowledge about the contextual information that can be transferred across categories. For example, people and bicycles, although often appearing together, have very different appearance as well as associated context (see Figure 2(b)). However, it is still possible to exploit the prior knowledge that both can appear in very diverse environments (indoors and outdoors), and thus context in general may provide a similar level of assistance in detecting both categories. In this paper, a novel context transfer learning method is proposed which utilises training samples from object classes of source task to improve the learning of the context model for a target object class based on a joint maximum margin learning framework. Specifically, two transfer maximum margin context models (TMMC) are devised. The first model is applied for knowledge transfer between objects that share similar context (e.g. cars and motorbikes), the second for related objects with different context benefiting from modelling context in general (e.g. people and bicycles).

The effectiveness of our approach is evaluated using the PASCAL Visual Object Classes challenge 2005 dataset [15] and 2007 dataset [16], a luggage detection dataset extracted from the UK Home Office i-LIDS database [27], and a vehicle detection dataset extracted from outdoor surveillance footages. Our results demonstrate that the proposed MMC context model improves the detection performance for all object classes, and our TMMC model is capable of further improving the performance of object detection by incorporating the transferrable contextual information extracted from training data of object categories from source task when the available target data are limited. In addition, it is also shown that our context model clearly outperforms the related state-of-the-art alternative context models, and the improvement is especially significant in the more challenging i-LIDS luggage and surveillance vehicle datasets.

II. Related Work

Most existing context modelling works require manual annotation/labelling of contextual information. Given both the annotated target objects and contextual information, one of the most widely used methods is to model the co-occurrence of context and object. Torralba et al. [46], Rabinovich et al. [38] and Felzenszwalb et al. [18] infer the semantic information about how a target object category co-occurs frequently with other categories (e.g. a tennis ball with a tennis bracket) or where the target objects tend to appear (e.g. a TV in a living room). Besides co-occurrence information, spatial relationship between objects and context has also been explored for context modelling. The spatial relationship is typically modelled using Markov Random Field (MRF) or Conditionally Random Field (CRF) [28], [9], [22], or other graphical models [24]. These models incorporate the spatial support of target object against other objects either from the same category or from different categories and background, such as a boat on a river/sea or a car on a road. Along a similar line, Hoim et al. [26] and Bao et al. [2] proposed to infer the interdependence of object, 3D spatial geometry and the orientation and position of camera as context; and Galleguillos et al. [21] inferred the contextual interactions at pixel, region and object levels and combine them together using a multi-kernel learning algorithm [21], [47].

Although different context representation and models have been adopted in these works, they all suffers from the same drawback that laborious manual efforts are required in order to either label contextual objects or parts of a scene that can provide contextual support for the target object/class, or specify the location and/or assign the spatial relationship between target objects and context. In contrast, our context model does not need any manually labelling of contextual information or its spatial relationship with target objects, thus is able to learn contextual information for improving object detection in a more unsupervised way. Moreover, many existing context learning works first learn an

object appearance model and a context model independently and then fuse them together for detection [48], [36], [37], [13]. On the contrary, our model quantifies contextual information conditioned on the appearance model for a target object category, so that more effective and useful contextual information can be selected explicitly to leverage the detection performance.

Recently Heitz and Koller [25] also investigated the use of context in an unsupervised way in order to reduce the cost of human annotation. The proposed Things and stuff (TAS) model in [25] first segments an image into parts and then infers the relationship between these parts and the target objects detected by a base detector in a Bayesian framework using a graphical model. Compared to TAS, our MMC model differs in that (1) we develop a discriminative rather than a generative context model so that no prior manually defined rules are required to describe the spatial relationship between context and target objects; (2) Our model is not limited to the Thing-Stuff context; (3) No global image segmentation is required which could be unreliable especially for a cluttered scene; (4) Our model can be extended to perform transfer learning for context learning given limited data.

To the best of our knowledge, this work is the first attempt to context transfer learning. However, transfer learning has been exploited extensively for learning a detector by object appearance modelling using training data of both target object category and source object categories. The existing object transfer learning methods mainly fall into three broad categories according to the relationship between the target and source object categories: 1) cross-domain but from the same categories [34], [35], [49], [14] (e.g. detecting fastback sedan cars using hatchback sedan cars as source data), 2) cross-category but relevant using hierarchical category structure [52] (e.g. detecting giraffes using other four-leg animals as source data), and 3) cross category and irrelevant [17], [5], [39] (e.g. detecting people using motorbike as source data). Nevertheless none of the existing transfer learning techniques designed for object appearance transfer learning can be applied directly to the object context transfer learning problem. This is due to the fundamental difference between the two problems: an object appearance model is only concerned with the appearance of a target object category, whilst to learn an object context model one must model both context and object appearance with the emphasis on their relationship, i.e. how different contextual information can assist in the detection of the target object; in other words a context model is not just about context because context is defined with respect to a target object and without the object modelling context itself is meaningless. Correspondingly, object appearance transfer learning aims to extract similarity between the appearance of target object class and source classes, whilst object context transfer learning is concerned with extracting similarities between the ways in which different contextual information can help to detect a target object class and source object classes.

In summary, compared to existing context learning approaches, the proposed framework has two major advantages:

1) Our context model is able to explicitly and directly quantify context by learning a context risk function, which combines the prior detection confidence and contextual information in a selective and discriminant framework.
2) Our context model can be learned with limited training data due to a novel context transfer model which utilises data from related source object classes even when they are visually very different from the target object.

The maximum margin context (MMC) model was first proposed in our preliminary work [51]. In this paper, apart from providing more detailed formulation and in-depth analysis, and evaluating the model using more extensive experiments, the major difference between this paper and [51] is the introduction of the new context transfer learning methods. Our experiments suggest that with this context transfer learning method, the MMC model can be better learned given limited target object data, leading to further improvement of detection performance. In addition, HOG features rather than SIFT features are used in this work for context feature extraction which also improves the performance.

III. Learning a Discriminant Context Model for Quantifying Context

Assume we have a training set images of a target object class with the ground truth locations of the target objects in each image known. First, a detector is learned to model the object appearance which is called a base detector. In this paper the histogram of oriented gradients (HOG) detector [12] is adopted, but any sliding window based object detector can be used. Second, the base detector is applied to all the training images to obtain a set of N candidate object windows (bounding boxes), denoted as $\mathcal{O} = \{\mathbf{O}_i\}_{i=1}^{N}$, which yield the highest scores among all candidate windows using the detector. Without loss of generality we 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 positive ones. Let us denote the context corresponding to $\mathbf{O}_i$ by $\mathbf{h}_i$ and define $\mathcal{H}_p = \{\mathbf{h}_i\}_{i=1}^{\ell}$ and $\mathcal{H}_n = \{\mathbf{h}_i\}_{i=\ell+1}^{N}$. We call $\mathcal{H}_p$ the positive context set and $\mathcal{H}_n$ the negative context set.

For our detection problem, we wish to compute the confidence of an object being the target object based on both its appearance and its visual context. Specifically, given an object $\mathbf{O}_i$ and its context $\mathbf{h}_i$, the confidence of an candidate detection window containing an instance of the target object class is computed as:

$$D(\mathbf{O}_i, \mathbf{h}_i) = D_o(\mathbf{O}_i) \cdot D_c(\mathbf{h}_i), \tag{1}$$

where $D_o(\mathbf{O}_i)$ is the prior detection confidence of an object being the target class obtained based on the output of the object appearance detector (base detector), $D_c(\mathbf{h}_i)$ is the context score which is to be learned, and $D(\mathbf{O}_i, \mathbf{h}_i)$ is the posterior detection confidence which will be used to make the decision on object detection using context. In our work, $D_o(\mathbf{O}_i)$ is computed based on the detection score of the base detector $s_i (\in (0, 1])$ parameterised by α as follows

$$D_o(\mathbf{O}_i) = s_i^{\alpha}. \tag{2}$$

In the above equation, α determines the weight on the prior detection score s_i in computing the posterior detection confidence $D(\mathbf{O}_i, \mathbf{h}_i)$. More specifically, the higher the value of α, the more weight is given to the object appearance model which indicates that context in general is less useful in detecting the target object class. The value of α will be automatically estimated along with context quantification through learning, as described later.

We wish to learn a context model such that the confidence $D(\mathbf{O}_i, \mathbf{h}_i)$ for true positive detections is higher than the false positive detections in the training set so that it can be used for detection in an unseen image. Before describing our context model in details, let us first describe how the context for the i-th candidate window $\mathbf{h}_i$ is computed.

Fig. 3. Examples of the polar geometric structure for context modelling. The target object classes are car (left image) and people (right image).

A. A Polar Geometric Context Descriptor

Given a candidate object window $\mathbf{O}_c$, we use a polar geometric structure [30] expanded from the centroid of the candidate object (see Figure 3) to explore and represent the contextual information associated with the object detection window. With r orientational and $b+1$ radial bins, the context region centred around the candidate object is divided into $r \cdot b + 1$ patches with a circle one at the centre, denoted by $\mathcal{R}_i$, $i = 1, \cdots, (r \cdot b + 1)$. In this paper b is set to 2 and r is set to 16. The size of the polar context region is proportional to that of the candidate object window $\mathbf{O}_c$. 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 $\mathbf{O}_c$. As shown in Fig. 3, our polar context region bins 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, from which more information should be extracted.

The polar context region is represented using the Bag of Words (BoW) method. To build the code book, the HOG features [12] which is robust to partial occlusion and image noise are extracted densely as described in [8]. These features are invariant to scale 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, HOG 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 outside the detection window of the candidate object, a normalised histogram vector [19] 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, denoted as the context associated to the object as described above by $\mathbf{h}_i$. The high order information of the interaction between the context inside and outside of the detection window can then be inferred by the proposed selective context model as described in the next section.

Note that with the proposed polar geometric structure for context modelling, context is always captured from adjacent regions of candidate object, and the potentially useful information in regions farther away may be neglected. There are a number of reasons for choosing the current setting: (1) for object detection in a surveillance environment where the scene background is fixed but objects presented in the scene are small, dynamic and are often in large numbers, the regions adjacent to each object detection window is more relevant. Importantly, in this case, including the regions farther away could have an adverse effect as all candidate object windows will have similar context in those regions which makes the task of distinguishing true detections from false positives harder. For instance, in the underground luggage detection example shown in Figure 1, the local contextual information (objects next to luggage) is more useful than the global one (e.g. other objects on the train platform). (2) increasing the context regions size will also lead to the increase of computational cost during both training and testing. Nevertheless, it could in general be beneficial to explore contextual information from farther away regions when these information is not overly noisy. This can be achieved by simply increasing the context region size.

Our polar context descriptor differs from alternative polar context descriptors [48], [36], which also describe the context expanded from the centre of the object, in that 1) Bag-of-Words method is employed for robustness against noise; 2) Pixels within context region need not be labelled; and 3) Context features are extracted more densely to cope with low image resolution and noise in our method. In contrast, only some predetermined sparse pixel locations were considered for context feature extraction in [48], [36]. The proposed contextual descriptor is also related to a number of existing descriptors for object appearance representation, including the shape context descriptor [6], the correlogram descriptor [42], the spatial pyramid representation [23], [29] and the spatial-temporal descriptor for describing pedestrian activity [10]. In particular, as most polar geometric structure based descriptors, our descriptor is inspired by the shape context work of Belongie et al. [6]. The main difference here is that our descriptor is designed for context representation. It is thus centered at the candidate object location and captures contextual information from mainly the surrounding area of an object. Our context descriptor could incorporate the idea of correlogram [42] to better capture the spatial co-occurrences of features cross different bins, although this would lead to an increase in computational cost. The works by Choi et al. [10], [11] would be a natural way to extend our context descriptor for modelling dynamic context for action/activity recognition. The spatial pyramid matching approaches [23], [29] which were originally formulated for object categorisation could be considered if we want to replace the exponent $\mathcal{X}^2$ distance kernel (to be detailed next) used in our framework with a more sophisticated kernel.

B. Quantifying Context

Without relying on segmentation, our polar context region contains useful contextual information which can help object detection to different extents, 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 and measure the usefulness of different contextual information represented using our polar geometric context descriptor.

A context model is sought to utilise contextual information to leverage the prior detection score obtained using the base detector so that the posterior detection score of the true positive detection is higher than the false positives [1]. Specifically, the objective of context modelling is to minimize the following risk function with the positive and negative context sets $\mathcal{H}_p$ and $\mathcal{H}_n$:

$$\mathcal{L} = \sum_{\mathbf{h}_i \in \mathcal{H}_p} \sum_{\mathbf{h}_j \in \mathcal{H}_n} \delta(D(\mathbf{O}_i, \mathbf{h}_i) \leq D(\mathbf{O}_j, \mathbf{h}_j)), \qquad (3)$$

[1]Note that the detection score of the base detector for a false positive detection window can be higher than that of a true positive window.

where $\mathbf{h}_i$ is the polar context descriptor corresponding to true positive detection windows as described in Section III-A, $\mathbf{h}_j$ is the context descriptor corresponding to the false positives, and 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 value of the risk function is, the more confident the detection would be for unseen data.

In order to compute the posterior detection score $D(\mathbf{O}_i, \mathbf{h}_i)$ defined in Eqn. (1), we need to compute both $D_o(\mathbf{O}_i)$ and $D_c(\mathbf{h}_i)$. $D_o(\mathbf{O}_i)$ is obtained by Eqn. (2), and we compute $D_c(\mathbf{h}_i)$ as

$$D_c(\mathbf{h}_i) = \exp\{f(\mathbf{h}_i)\}, \tag{4}$$

where $f(\mathbf{h}_i)$ is a *leverage function* that outputs the confidence score of the context descriptor $\mathbf{h}_i$, and the higher the value of f is the more positive the contextual information is. We consider to learn the leverage function f as a kernel linear function:

$$f(\mathbf{h}_i) = \mathbf{w}^T \varphi(\mathbf{h}_i) + b, \tag{5}$$

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 here because the descriptor we introduce (i.e. a histogram) is a distribution representation of high dimension; the exponent $\mathcal{X}^2$ distance kernel [19], which is a Mercer kernel, can thus be used to measure the distance between two discrete distributions. Note that the variable b in Eqn. (5) does not have any impact on the risk function up to now, but it will be useful for learning a much better $\mathbf{w}$ in an approximated way. This is because a more flexible solution for $\mathbf{w}$ can be found by utilising b at the training stage, as we shall describe next (see Eqn. (10)). Now Eqn. (3) becomes

$$\mathcal{L} = \sum_{\mathbf{h}_i \in \mathcal{H}_p} \sum_{\mathbf{h}_j \in \mathcal{H}_n} \delta(s_i^\alpha \cdot \exp\{f(\mathbf{h}_i)\} \leq s_j^\alpha \cdot \exp\{f(\mathbf{h}_j)\}), \tag{6}$$

The ideal case to minimize the risk function in Eqn. (6) is to learn a leverage function f fulfilling all the following constraints:

$$s_i^\alpha \cdot \exp\{f(\mathbf{h}_i)\} > s_j^\alpha \cdot \exp\{f(\mathbf{h}_j)\}, \ \forall \mathbf{h}_i \in \mathcal{H}_p, \mathbf{h}_j \in \mathcal{H}_n. \tag{7}$$

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. By taking logarithm on both sides of Eqn. (7), we approach the problem of minimizing the risk function by investigating a solution constrained by a margin $\rho(\geq 0)$ as follows:

$$\begin{aligned} &f(\mathbf{h}_i) + \log s_i^\alpha \geq \rho, \ \forall \mathbf{h}_i \in \mathcal{H}_p, \\ &f(\mathbf{h}_j) + \log s_j^\alpha \leq -\rho, \ \forall \mathbf{h}_j \in \mathcal{H}_n. \end{aligned} \tag{8}$$

Ideally, the constraints in Eqn. (7) 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. Therefore, we aim to find the optimal $\mathbf{w}$, b, and α such that ρ is maximized (or $-\rho$ is minimized) as follows:

$$\begin{aligned} \min \quad & -\rho \\ s.t. \quad & \mathbf{w}^T \varphi(\mathbf{h}_i) + b \geq \rho - \log s_i^\alpha, \ \forall \mathbf{h}_i \in \mathcal{H}_p, \\ & \mathbf{w}^T \varphi(\mathbf{h}_j) + b \leq -\rho - \log s_j^\alpha, \ \forall \mathbf{h}_j \in \mathcal{H}_n, \\ & \rho \geq 0. \end{aligned} \tag{9}$$

Note that without regularization, the margin ρ can be made as large as possible by simply scaling $\mathbf{w}$, b, and α in the above criterion. In order to avoid this problem, for non-negative ν and C, we introduce the following regularized criterion:

$$\begin{aligned} &\{\mathbf{w}_t, b_t, \alpha_t\} = \arg \min_{\mathbf{w}, b, \alpha, \rho} -\nu \cdot \rho + \frac{1}{2}(||\mathbf{w}||^2 + C^2 \cdot \alpha^2) + \frac{1}{N} \sum_{i=1}^{N} \xi_i \\ &s.t. \quad \mathbf{w}^T \varphi(\mathbf{h}_i) + b \geq \rho - \xi_i - \log s_i^\alpha, \ \forall \mathbf{h}_i \in \mathcal{H}_p, \\ &\quad \mathbf{w}^T \varphi(\mathbf{h}_j) + b \leq -\rho + \xi_j - \log s_j^\alpha, \ \forall \mathbf{h}_j \in \mathcal{H}_n, \\ &\quad \rho \geq 0, \ \xi_i \geq 0, \ i = 1, \cdots, N, \end{aligned} \tag{10}$$

where positive slack variables $\{\xi_i\}_{i=1}^N$ are additionally added to the margin ρ for each constraint, because completely satisfying all the constraints without the slack variables in model (10) would be very difficult.

With the criterion above learning our context model becomes a constrained quadratic programming problem. Next, we show that we can reformulate the problem so that the popular SVM [43] technique can be used to find the optimal solution. Let $\alpha' = C \cdot \alpha$ and define $\mathbf{z} = [\mathbf{w}^T \ , \alpha']^T$ and $\psi_C(\mathbf{h}_i, s_i) = [\varphi(\mathbf{h}_i)^T, \frac{\log s_i}{C}]^T$, Eqn. (10) can be rewritten as:

$$\begin{aligned} &\{\mathbf{z}_t, b_t\} = \arg \min_{\mathbf{z}, b, \rho} -\nu \cdot \rho + \frac{||\mathbf{z}||^2}{2} + \frac{1}{N} \sum_{i=1}^{N} \xi_i \\ &s.t. \quad \mathbf{z}^T \psi_C(\mathbf{h}_i, s_i) + b \geq \rho - \xi_i, \ \forall \mathbf{h}_i \in \mathcal{H}_p, \\ &\quad \mathbf{z}^T \psi_C(\mathbf{h}_j, s_j) + b \leq -\rho + \xi_j, \ \forall \mathbf{h}_j \in \mathcal{H}_n, \\ &\quad \rho \geq 0, \ \xi_i \geq 0, \ i = 1, \cdots, N. \end{aligned} \tag{11}$$

Now any SVM solver such as [43] can be used to learn the model parameters $\mathbf{w}$, b, and α. Note that with the formulation above, a new mercer kernel $\dot{\kappa}$ for ψ_C can be defined as $\dot{\kappa}(\{\mathbf{h}_i, s_i\}, \{\mathbf{h}_j, s_j\}) = \kappa(\mathbf{h}_i, \mathbf{h}_j) + C^{-2} \cdot \log s_i \cdot \log s_j$. In this work, for the two free parameters C and ν, we set $C = 1$ and estimate the value of ν using cross-validation.

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

$$D(\mathbf{O}_i, \mathbf{h}_i) = D_o(\mathbf{O}_i) \cdot D_c(\mathbf{h}_i) = s_i^{\alpha_t} \cdot \exp\{\mathbf{w}_t^T \varphi(\mathbf{h}_i) + b_t\}. \tag{12}$$

Context aware object detection can then be performed on a test image by thresholding the posterior detection confidence score for each candidate window.

It should be noted that when a linear kernel is used the proposed MMC can be seen as a feature selector for the simple concatenation of detection score and contextual features. However, since the detection score and the contextual features are lying in different spaces or manifolds, the linear kernel is not a suitable similarity measurement for such kind of combination. Hence, the nonlinear exponent $\mathcal{X}^2$ distance kernel is adopted to measure the similarity between histogram based contextual features.

Comments on α. Intuitively for different target object classes, contextual information has different levels of usefulness in disambiguating the target object class from background and other object classes. For instance, context is much more important for an object class with very diverse appearance but always appearing with a fixed surroundings than one that has uniform appearance but can appear anywhere. As mentioned earlier, the value of α

in our MMC model, which is learned from data, indicates how important context information in general is for detecting the target object class. This is different from the model parameter $\mathbf{w}$ which corresponds to the relevant importance of different high order context information with regard to each other. Specifically, $\alpha = 0$ means that 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. For $\alpha > 0$, the larger it is, the more important the prior detection probability is and the less useful the contextual information in general will be. In particular, a very large α value will mean the contextual information is completely discarded. Note that the value of α is not restricted to be non-negative. When α has a negative value, the smaller the prior detection probability is, the larger the posterior detection score is expected. This is because $s_i \in (0, 1]$ and the leverage function $f(\mathbf{h}_i)$ is always bounded by investigating the dual problem formulated in Eqn. (10). Although theoretically possible, it is unlikely that a negative value of α will be obtained in practice unless a poor base detector is used which completely fails to capture the object appearance information.

IV. Context Transfer Learning

Compared with an object appearance model, a context model requires much more data to learn due to the diversity of contextual information. Context model is thus more likely to suffer from model over-fitting problem caused by the limited availability of training data. In this section, we formulate a novel context transfer learning model which utilises training samples from source object classes to improve the proposed MMC model for a target object class based on a joint maximum margin learning framework. Specifically, two transfer maximum margin context models (TMMC) are devised. The first model is applied for knowledge transfer between objects that share similar context (e.g. cars and motorbikes) and the second for related objects that have different context but similar level of benefit from modelling context in general (e.g. people and bicycles).

Let us first formally define the context transfer learning problem. Assume we have a set of training samples for context aware detection of Q categories, where context samples from each category contain both a positive and negative context sets. Let $\{(\mathbf{h}_i, s_i, y_i, \tau_i)\}_{i=1}^N$ be the training dataset, where $\mathbf{h}_i$ is the associated contextual information of candidate object window $\mathbf{O}_i$, s_i is the corresponding prior detection score (obtained using different base detectors trained for different object categories), $y_i \in \{+1, -1\}$ is the ground truth label of the contextual information (either positive or negative), and τ_i indicates the category label of the candidate object $\mathbf{O}_i$. For Q tasks of object detection, i.e. $\tau_i \in \{1, \cdots, q, \cdots, Q\}$, there are N_q candidate windows for each task (category). Let the first category ($q = 1$) be the target object class and the rest be the source categories ($q > 1$) which are used to facilitate the context quantification for target class. We wish to develop two joint maximum margin learning models for context transfer learning based on the assumption on how contextual information can be shared across the target and source object categories.

A. TMMC-I: Transferring Discriminant Contextual Information

Our first transfer MMC model assumes that the usefulness of different discriminant contextual information is shared between categories; that is different categories can have similar projection $\mathbf{w}$ in Eqn. (5) which weights the usefulness of higher-order contextual information, whilst having different prior importance weight α_q on the detection confidence, different margin ρ_q and constant b_q. Similar to Eqn. (8), the MMC context models for the target object category can thus be learned using the following model with samples from both the target and source categories as training data:

$$\begin{aligned}&\mathbf{w}^T\varphi(\mathbf{h}_i) + b_q + \alpha_q \cdot \log s_i \geq \rho_q - \xi_i,\ \forall\ y_i = 1\ \&\ \tau_i = q,\\ &\mathbf{w}^T\varphi(\mathbf{h}_j) + b_q + \alpha_q \cdot \log s_j \leq -\rho_q + \xi_j,\ \forall\ y_j = -1\ \&\ \tau_j = q.\end{aligned} \tag{13}$$

We then consider the following optimization problem, which we call TMMC-I,

$$\begin{aligned}\{\mathbf{w}_t, b_q^t, \alpha_q^t\} = \arg \min_{\mathbf{w}, b_q, \alpha_q, \rho_q, \xi_i} &\frac{1}{2}(||\mathbf{w}||^2 + \sum_{q=1}^{Q}\alpha_q^2)\\ &+ \frac{1}{N}\sum_{i=1}^{N}\xi_i - \frac{\upsilon}{N}\sum_{q=1}^{Q}N_q \cdot \rho_q\\ s.t.\ \ &y_i(\mathbf{w}^T\varphi(\mathbf{h}_i) + b_q + \alpha_q \log s_i) \geq \rho_q - \xi_i,\ if\ \tau_i = q,\\ &\xi_i,\ \rho_q \geq 0.\end{aligned} \tag{14}$$

To solve Eqn. (14) by convex optimization, we first derive the Lagrange equation of its optimization problem as follows:

$$\begin{aligned}f =& \frac{1}{2}(||\mathbf{w}||^2 + \sum_{q=1}^{Q}\alpha_q^2) + \frac{1}{N}\sum_{i=1}^{N}\xi_i - \frac{\nu}{N}\sum_{q=1}^{Q}N_q \cdot \rho_q\\ &- \sum_{q=1}^{Q}\sum_{\tau_i=q} c_i(y_i(\mathbf{w}^T\varphi(\mathbf{h}_i) + b_q + \alpha_q \cdot \log s_i) - \rho_q + \xi_i)\\ &- \sum_{i=1}^{N}\lambda_i\xi_i - \sum_{q=1}^{Q}\gamma_q\rho_q,\end{aligned} \tag{15}$$

where c_i, λ_i, $\gamma_i \geq 0$. Note that

$$\frac{\partial f}{\partial \mathbf{w}} = 0\ \Rightarrow\ \mathbf{w} = \sum_{i=1}^{N} c_i y_i \varphi(\mathbf{h}_i), \tag{16}$$

$$\frac{\partial f}{\partial \alpha_q} = 0\ \Rightarrow\ \alpha_q = \sum_{\tau_i=q} c_i y_i \log s_i, \tag{17}$$

$$\frac{\partial f}{\partial b_q} = 0\ \Rightarrow\ \sum_{\tau_i=q} c_i y_i = 0, \tag{18}$$

$$\frac{\partial f}{\partial \rho_q} = 0\ \Rightarrow\ \sum_{\tau_i=q} c_i \geq \nu\frac{N_q}{N}, \tag{19}$$

$$\frac{\partial f}{\partial \xi_i} = 0\ \Rightarrow\ c_i \leq \frac{1}{N}, \tag{20}$$

According to the dual and primal problem as well as the Karush-Kuhn-Tucker (KKT) conditions [32], the dual problem of Eqn. (14) can then be formulated as follows:

$$\begin{aligned}\{c_i^t\} = \arg\max_{c_i} &-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j y_i y_j \kappa(\mathbf{h}_i, \mathbf{h}_j)\\ &- \frac{1}{2}\sum_{q=1}^{Q}\sum_{\tau_i,\tau_j=q} c_i c_j y_i y_j \log s_i \log s_j\\ s.t.\ \ &\sum_{\tau_i=q} c_i y_i = 0,\ q = 1, \cdots, Q\\ &\sum_{\tau_i=q} c_i \geq \upsilon \cdot \frac{N_q}{N},\ 0 \leq c_i \leq \frac{1}{N}.\end{aligned} \tag{21}$$

The optimal projection $\mathbf{w}_t$ and weights α_q^t are determined by

$$\mathbf{w}_t = \sum_{i=1}^{N} c_i^t y_i \varphi(\mathbf{h}_i),\ \alpha_q^t = \sum_{\tau_i=q} c_i^t y_i \log s_i. \quad (22)$$

B. TMMC-II: Transferring the Weight of Prior Detection Score

The second context transfer model is designed for the case where the target object category and the related source ones could have little in common in both appearance and context, but contextual information can provide similar level of assistance in detection. As we discussed in the previous section, the usefulness of contextual information in general for a specific object class can also be indicated by the learned model parameter α. This is because that although α is an importance weight on the prior detection confidence, it is not independent of context information because it is learned, not set manually, using both the detector scores and context descriptors from both positive and negative examples. Since the more important (trustworthy) context detector score is, the less important context information is for detection, the learned α value is an indication of both the importance of detector score and the importance of contextual information. The importance of contextual information is thus also quantified by α during the optimisation of the context model in Eqn. (10). In TMMC-II, we aim to learn the maximum margin context with different margin variables ρ_q, different projections $\mathbf{w}_q$ and constant b_q but with the same importance weight α on the prior detection score for different categories as follows:

$$\begin{aligned} &\mathbf{w}_q^T \varphi(\mathbf{h}_i) + b_q + \alpha \cdot \log s_i \geq \rho_q - \xi_i,\ \forall\ y_i = 1\ \&\ \tau_i = q, \\ &\mathbf{w}_q^T \varphi(\mathbf{h}_j) + b_q + \alpha \cdot \log s_j \leq -\rho_q + \xi_j,\ \forall\ y_j = -1\ \&\ \tau_j = q. \end{aligned} \quad (23)$$

We then consider the following optimization function

$$\begin{aligned} \{\mathbf{w}_q^t, b_q^t, \alpha_t\} = \arg \min_{\mathbf{w}_q, b_q, \alpha, \rho_q, \xi_i} &\frac{1}{2}\Big(\sum_{q=1}^{Q} ||\mathbf{w}_q||^2 + \alpha^2\Big) \\ &+ \frac{1}{N}\sum_{i=1}^{N} \xi_i - \frac{\upsilon}{N}\sum_{q=1}^{Q} N_q \cdot \rho_q \\ s.t.\ \ y_i(\mathbf{w}_q^T \varphi(\mathbf{h}_i) + b_q + \alpha \log s_i) &\geq \rho_q - \xi_i,\ if\ \tau_i = q, \\ \xi_i,\ \rho_q &\geq 0. \end{aligned} \quad (24)$$

Following a similar derivation as for TMMC-I, the dual problem of Eqn. (24) can be formulated as follows:

$$\begin{aligned} \{c_i^t\} = \arg\max_{c_i} &-\frac{1}{2}\sum_{q=1}^{Q}\sum_{\tau_i,\tau_j=q} c_i c_j y_i y_j \kappa(\mathbf{h}_i, \mathbf{h}_j) \\ &- \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j y_i y_j \log s_i \log s_j \\ s.t.\ \ &\sum_{\tau_i=q} c_i y_i = 0,\ q = 1, \cdots, Q \\ &\sum_{\tau_i=q} c_i \geq \upsilon \cdot \frac{N_q}{N},\ 0 \leq c_i \leq \frac{1}{N}. \end{aligned} \quad (25)$$

The optimal projections $\mathbf{w}_q^t$ and weight α_t are learned by

$$\mathbf{w}_q^t = \sum_{\tau_i=q} c_i^t y_i \varphi(\mathbf{h}_i),\ \alpha_t = \sum_{i=1}^{N} c_i^t y_i \log s_i. \quad (26)$$

In the above formulations for TMMC-I and TMMC-II, we could obtain the MMC model parameters for all Q categories jointly by solving a single optimization problem, which is why our TMMC is a joint maximum margin learning framework. Our TMMC model is also closely related to the multi-task learning [1]. However, note that there is one free parameter υ in our model which needs to be estimated via cross-validation. Since the model we are after is the one for the target object category, υ is estimated using the training samples from the target category only. Therefore, our TMMC model is different from the conventional symmetric multi-task learning which treats all tasks equally. Nevertheless, there is sometimes no clear boundary between transfer learning and general multi-task learning. According to [50], our context transfer models can be seen as a kind of asymmetric multi-task learning, which has a target task among the learned tasks. Compared to existing multi-task learning methods, TMMC is specifically designed for transferring the useful way/manner how contextual information help detect related source categories (tasks) for object detection to target category (task), and thus is more appropriate for solving the data sparsity problem for our context transfer learning for target object detection. It is also worth pointing out that there are different flavours of transfer learning. Since TMMC aims to transfer useful context information from related object categories for improving the detection performance on target class, we follow the terminology in [33] and consider our TMMC as an inductive transfer learning method.

V. Experiments

A. Datasets and settings

We evaluate the proposed context model and transfer learning framework against alternative models using four datasets: the PASCAL Visual Object Classes (VOC) 2005 challenge dataset [15] and 2007 dataset [16] for detecting a total of 10 different categories of objects, a subset of the UK Home Office i-LIDS [27] called i-LIDS Luggage for detecting two types of luggage (suitcases and bags) (see Fig. 10), and a dataset captured at an airport forecourt called Forecourt Vehicle for detecting vehicles (including private cars, buses, vans, taxis) (see Fig. 11). Among these datasets, the Forecourt Vehicle dataset is a new dataset captured by us. The Forecourt Vehicle dataset is mainly featured with low resolution images taken from cameras mounted near and far away from the forecourt of an airport at different times of a day. Compared with the VOC2005 and VOC2007 datasets, the i-LIDS Luggage and Forecourt Vehicle datasets are much harder due to much more crowded scenes causing more severe occlusion, and lower image resolution with smaller object size. In addition, the Forecourt Vehicle dataset suffers from challenging outdoor lighting and image blurring caused by dirty camera lens.

- **PASCAL VOC**. The PASCAL VOC2005 dataset includes four object categories: car, motorbike, people and bicycle [15]. All four categories were used in our experiments. Among the PASCAL VOC2007 object categories, six categories that are different from the four in VOC2005 dataset were chosen. They include aeroplane, bus, cat, cow, horse, and train. The setting of the experiments for our MMC model was the same as that in [25]. That is a HOG detector [12] was first learned as a base detector and applied to the training set to get a set of candidate detection windows with associated prior detection scores, based on which the MMC context model was then learned.
- **i-LIDS Luggage**. For i-LIDS, we selected 1045 image frames of an image size of 640×480 from the i-LIDS

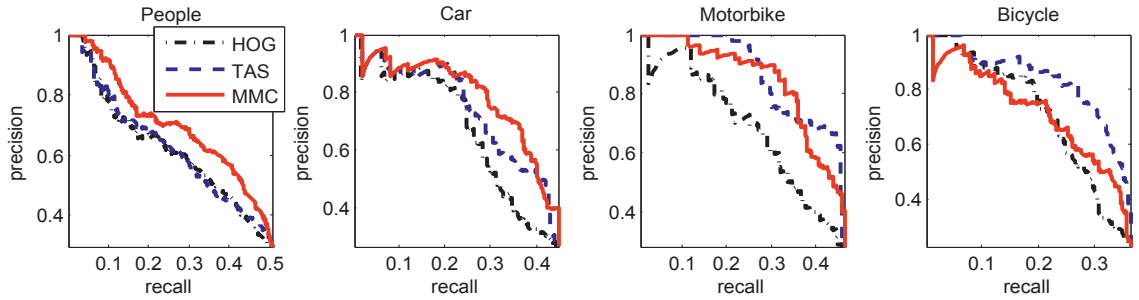

Fig. 4. Precision-Recall curves for the detection of four object categories in PASCAL VOC2005.

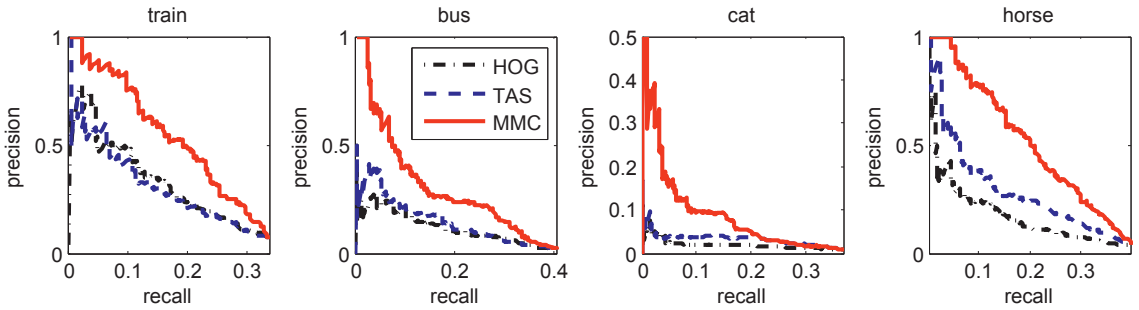

Fig. 5. Precision-Recall curves for the detection of four object categories in PASCAL VOC2007.

underground scenario, with 656 for training and the rest for testing. For context model training on i-LIDS for luggage detection, we first trained a pair of HOG luggage detectors using 540 positive samples for each of the two types of luggage (suitcases and bags), and 7278 and 5047 negative samples for the two detectors respectively. Separate i-LIDS testing image frames consisting of 1170 true luggage instances were selected with ground truth manually annotated for performance evaluation.

- **Forecourt Vehicle**. For the Forecourt Vehicle dataset [2], we selected 275 image frames of 720×576, from which 104 images were used for training. For context model training on Forecourt Vehicle dataset for vehicle detection, we first trained a HOG vehicle detector using 1038 vehicle images and 2300 background (non-vehicle) images. The context model was then evaluated on a separate testing set consisting of 1583 true vehicle instances from the 171 testing images.

The parameter ν in the MMC and TMMC models was estimated by five-fold cross-validation in a candidate set $\{\nu = \eta^2 | \eta \in [0.01 : 0.01 : 1]\}$. The threshold of the overlap rate between the correct object detection bounding box and the ground truth one was set to 0.5 according to the PASCAL VOC protocol [15]. We evaluate the detection performance by average precision rate and precision-recall curves [15].

B. Evaluation of Context Models

MMC vs. no context (HOG) and using only context

Our MMC model utilises both the object appearance information (via the base detector score) and contextual information for both fusion and contextual information selection. To evaluate its effectiveness, we first compare its performance with the base detector (HOG) without context modelling and a detector learned using only contextual information represented by our proposed contextual descriptor (termed as Context Only). Specifically, for

[2]The dataset has been made publically available and can be downloaded at http://www.eecs.qmul.ac.uk/~jason/forecourt/.

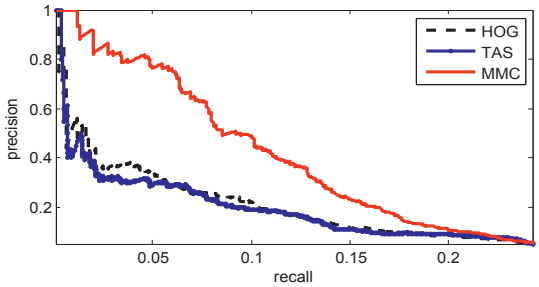

Fig. 6. Precision-Recall curves for luggage detection on i-LIDS dataset.

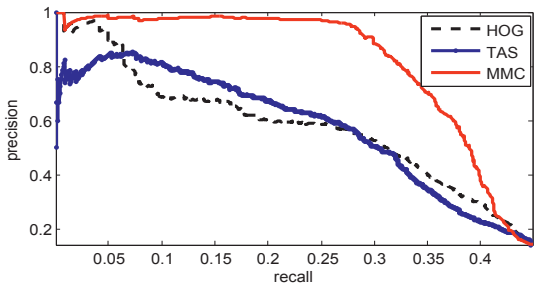

Fig. 7. Precision-Recall curves for vehicle detection on the Forecourt Vehicle dataset.

the latter, we train a SVM classifier using the positive context set $\mathcal{H}_p$ and the negative context set $\mathcal{H}_n$ without utilizing the prior detection score of the base detector.

The results for the PASCAL VOC2005 dataset are presented in Table I and Figure 4, whilst the results for the PASCAL VOC2007 dataset can be seen in Table II and Figure 5. For PASCAL VOC2005 dataset, it is evident from Table I and Figure 4 that by modelling context, MMC significantly improves the detection performance of the base detector especially on car, motorbike and people. For PASCAL VOC2007, Table II and Figure 5 show that even bigger improvements are obtained for all six classes except aeroplane using our MMC model over the base detector without context modelling. It is noted that in the case of aeroplane, the candidate detection windows produced by the base detector

tend to include a large proportion of background. This is because since the aeroplane shape is not rectangular, the annotated training samples of aeroplane in rectangular boxes contain lots of context information (mostly sky). As a result, context information has already been utilised by the base detector. MMC as well as other context models are thus unlikely to offer significant help.

Table I and Table II also show the result of the Context Only detector. It is observed that although for a number of object classes (e.g. motorbike in VOC2005 and Horse in VOC2007), better performance over the base HOG detector can be obtained using context only, its performance is much weaker compared to MMC. Overall, the results show that without combining with the prior detection score for context evaluation and selection, the contextual information itself is not reliable enough for detection. This is because contextual information inevitably contains irrelevant information for detecting the target object category, and without utilizing the prior detection score, the most discriminant context could not be identified.

TABLE I
AVERAGE PRECISION RATES ON PASCAL VOC2005.

Object Class	HOG [12]	TAS [25]	Context Only	HOG+Context	MMC
Car	0.325	0.363	0.3135	0.3437	**0.3741**
Motorbike	0.341	0.390	0.3594	0.3981	**0.4020**
People	0.346	0.346	0.3528	0.3710	**0.3862**
Bicycle	0.281	**0.325**	0.2503	0.2621	0.2878

TABLE II
AVERAGE PRECISION RATES ON PASCAL VOC2007.

Object Class	HOG [12]	TAS [25]	Context Only	HOG+Context	MMC
Aeroplane	0.0915	**0.0930**	0.0926	0.0922	0.0926
Bus	0.0817	0.0834	0.1475	**0.1711**	0.1674
Cat	0.0147	0.0242	0.0312	0.0696	**0.1056**
Cow	0.0234	0.0193	0.0562	**0.0937**	0.0929
Train	0.1471	0.1619	0.1847	0.2123	**0.2209**
Horse	0.1312	0.1606	0.2227	**0.2479**	0.2472

The comparative results on two visual surveillance datasets, i.e. the i-LIDS Luggage and Forecourt Vehicle datasets are shown in Table III and Table IV respectively in the form of average precision rate, and Figures 6 and 7 in terms of precision-recall curve. The results show that for these more challenging datasets, the improvement of our MMC model over detection using base detector only and context only is more significant compared with most object categories in the two PASCAL VOC datasets. This suggests that contextual information is more useful for disambiguating the target objects from background and other objects for these two datasets. This is mainly due to the fact that there is less distinctive appearance information extractable for the target object categories in i-LIDS and Forecourt because of the low image resolution and lack of colour and texture information in the case of luggage. As a consequence, the contextual information is more useful, which also explains why the context only detector outperforms the base detector for both i-LIDS Luggage and Forecourt Vehicle detection.

MMC vs. TAS

We compared MMC with a state-of-the-art context model TAS [25] which is closely related to our model in that both do not require annotation of contextual information. The results of MMC against the best reported results of TAS on the PASCAL VOC2005 and VOC2007 datasets are shown in Table I and Table II respectively. The results of TAS on PASCAL VOC2005 has been reported in [25]. In our experiments, we re-ran the TAS model provided by the authors[3]. Note that TAS is an EM based method thus sensitive to initialisation. Our results using TAS (see the blue-dashed plots in Figures 4 and 5) are either very similar or slightly better (e.g. motorbike) than those originally reported in [25]. To test TAS on PASCAL VOC2007, we segmented each image frame using the superpixel technique [40] and represented each region using 44 features (color, shape, energy responses) similar to the ones used in [25], [4], and then we run the TAS model provided by the authors several times and the best results are shown, where the model parameters were set according to the values given by the authors in their code available on the web. The results show that MMC outperforms TAS on the detection of 8 out of 10 categories in the two datasets.

TABLE III
AVERAGE PRECISION RATE ON LUGGAGE DETECTION ON I-LIDS.

HOG [12]	TAS [25]	Context Only	HOG+Context	MMC
0.1195	0.1167	0.1348	0.1435	**0.1460**

TABLE IV
AVERAGE PRECISION RATE ON THE FORECOURT VEHICLE DATASET.

HOG [12]	TAS [25]	Context Only	HOG+Context	MMC
0.2818	0.2927	0.3591	0.3806	**0.3838**

In particular, MMC improved the detection of people with a fairly large margin (a 4.22% increase in the average precision rate). As acknowledged by the authors in [25], the TAS model struggles with people detection in PASCAL VOC2005. This can be caused by two factors. First, as people appear more randomly compared to other rigid objects such as cars on a road, 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 models such as TAS could not utilise effectively the ambiguous contextual information for object detection improvement. Second, the TAS model focuses on Thing-Stuff context, i.e. the context between people and the background regions. The useful contextual information between people and other objects could be thus ignored (e.g. luggage and people). In contrast, our model is able to utilise any contextual information that is relevant and captured by the polar context descriptor regardless the type of the context. The performance of MMC is also much superior to TAS on the detection of bus, cat, cow, train, and horse in the VOC2007 dataset and almost equal to TAS on aeroplane.

Note that MMC achieves lower average precision rate than TAS on bicycle. The bicycle class is unique with no clear boundary between the object and background (one can see the background through a bicycle). In this case, alternative models such as TAS with scene segmentation may be less affected, although segmentation itself is challenging in a cluttered scene.

We also implemented TAS for luggage detection using the i-LIDS dataset and vehicle detection using the Forecourt dataset. We performed TAS on i-LIDS and Forecourt Vehicle datasets as similarly done on PASCAL VOC2007. The results are shown in Tables III, IV, Figures 6 and 7. As can seen clearly, MMC outperforms TAS on both datasets with a significant margin. In particular, the detection performance of luggage using TAS is worse than that of a HOG detector which was also used as the

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

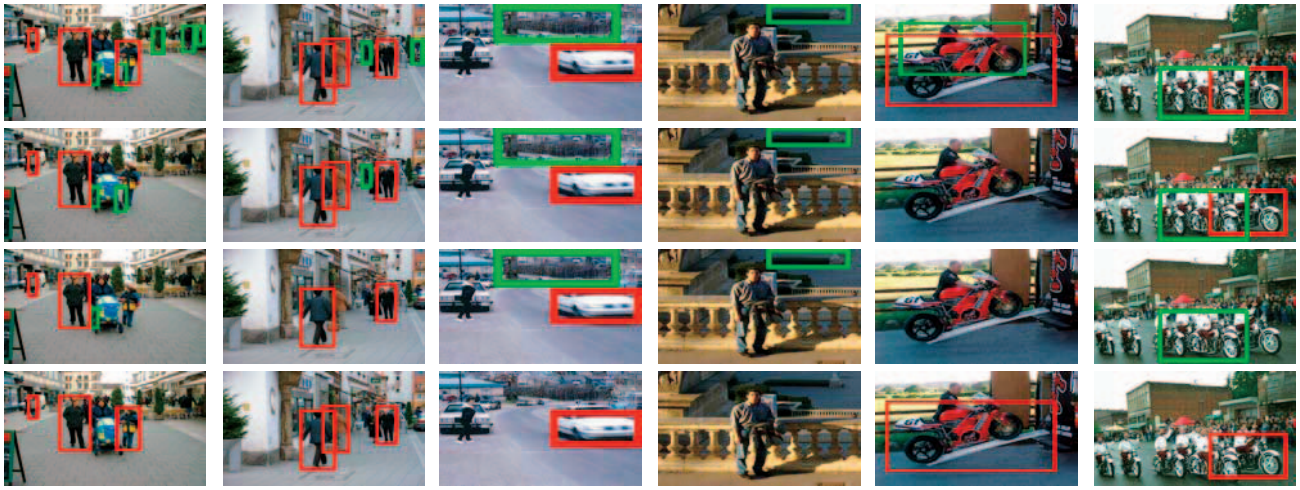

Fig. 8. Examples of object detections using HOG, TAS and MMC models on PASCAL VOC2005. The left-hand side two columns are for people detection, the middle two are for car detection, and the right-hand side two are for motorbike detection. The following setting of illustration applies to Figures 8, 9, 10 and 11: 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.

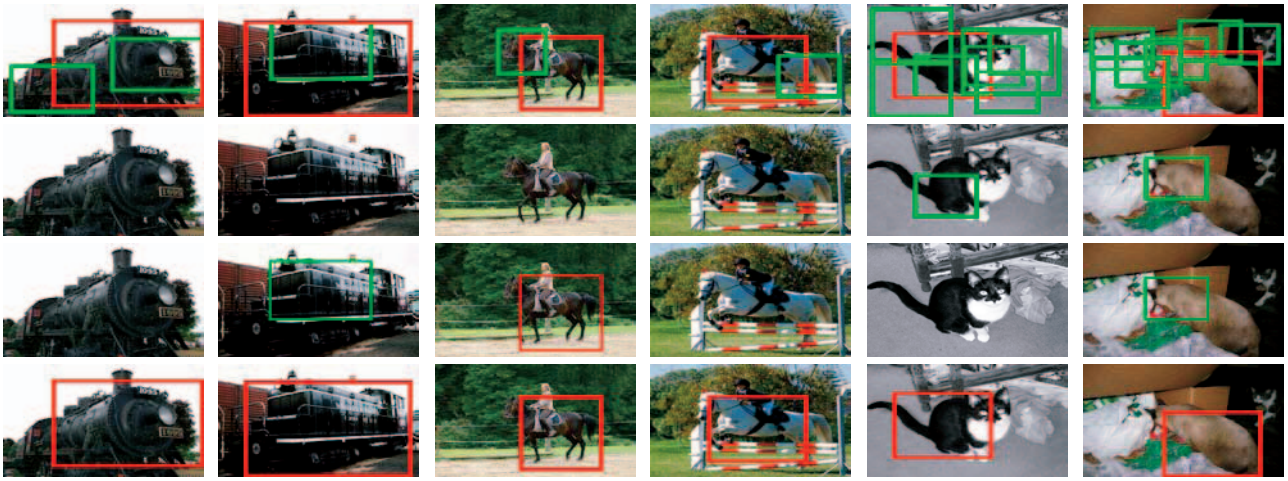

Fig. 9. Examples of object detections using HOG, TAS and MMC models on PASCAL VOC2007. The left-hand side two columns are for train detection, the middle two are for horse detection, and the right-hand side two are for cat detection.

Fig. 10. Examples of object detections using HOG, TAS and MMC models for luggage detection on i-LIDS.

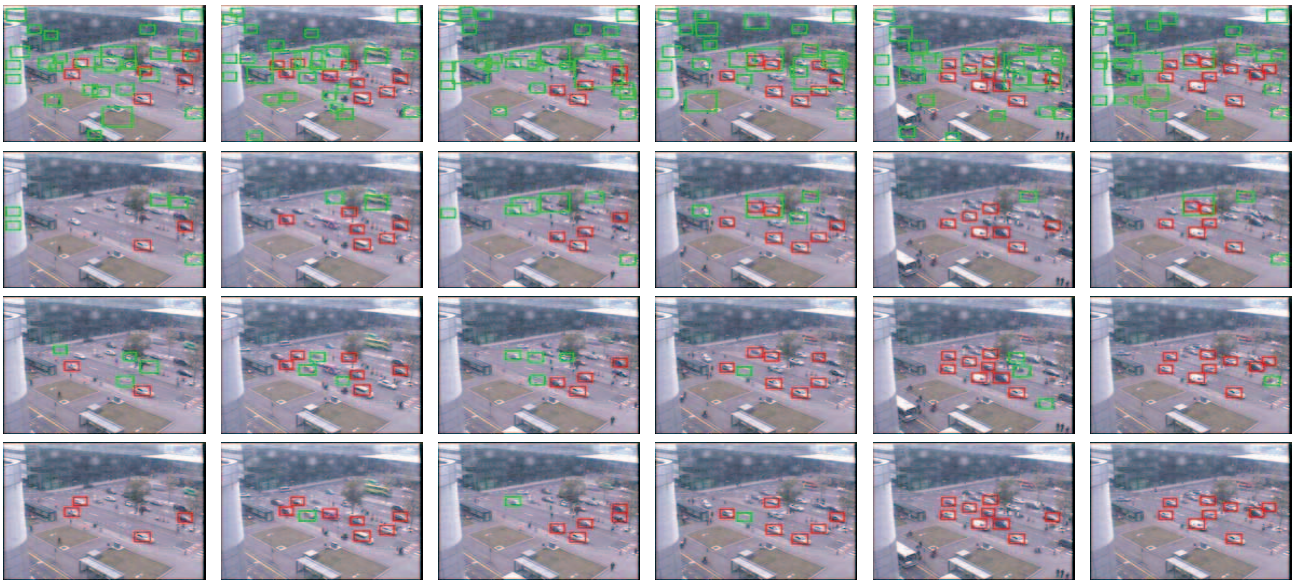

Fig. 11. Examples of object detections using HOG, TAS and MMC models for vehicle detection on the Forecourt dataset.

base detector in TAS. This again demonstrates that without any segmentation of a whole image, more effective context information can be learned when contextual information is evaluated and selected explicitly and directly. In addition, it demonstrates the benefit of utilising multiple types of context.

MMC vs. HOG+Context

One of the existing context modelling strategy is to learn a context only detector and an object appearance based detector independently and then fuse these two information by multiplying the two detector scores to compute the final detector score [45], [36], [37], [13]. It essentially performs naive score level fusion of context and object appearance. It assumes that the context and object are independent and treats 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. Directly comparing with [45], [36], [37], [13] is unfair because different context only detector and object appearance detector were used. We thus use the same HOG detector as the appearance based object detector and fuse its score with that of our context only detector using our polar context descriptors for context representation (termed as HOG+Context model). The difference in performance thus is solely due to the different context modelling strategies adopted. The results in Tables I, II, III and IV show that overall fusing the inferred contextual information conditioned on the prior detection score, whose importance weight is automatically estimated, is more effective than direct and blind fusion using HOG+Context. In particular, the results of MMC is either markedly better than or similar to those of HOG+Context. A closer examination reveals that the advantage of performing context selection is more apparent when the contextual information is more diverse, e.g. the context for the cat, people, and bicycle categories. In this case, the explicit and direct evaluation of contextual information becomes more critical. Blindly fusing contextual information with appearance information with equal weight assigned to each is thus less likely to assist in detection and may even have an adverse effect, as in the case of bicycle in the PASCAL VOC2005 dataset (see Tables I). On the contrary, if contextual information is relatively more dominant than object appearance, either because the object always appears within a certain context or the object appearance is more diverse, a blind fusion of context and appearance could give comparable results. This is expected because when context is more dominant, context selection becomes less critical. Examples of these object categories include, e.g. i-LIDS luggage, bus, cow and horse. For these object classes, the Context Only detector also performs stronger than the basic appearance only detector as can be seen in Tables I-IV.

Examples of reducing false positive detections using context.

We now show some visual examples to illustrate the benefit of our MMC model on reducing false positive detections. Figures 8, 9, 10 and 11 give some typical examples of false positive removal in PASCAL VOC2005, VOC2007, i-LIDS, and Forecourt respectively. For all methods, we illustrate the detection results when the recall-rate is at 0.3 for PASCAL VOC2005 and Forecourt and 0.1 for PASCAL VOC2007 and 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 modelling, 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 note that TAS tends to either remove both false positive and true positives or preserves more false positives, in particular in the case of luggage detection in i-LIDS. Again, this is because the crucial contextual information between luggage and other objects (people in this case) could not be effectively captured by TAS. Figure 12 shows some examples of failed detections by all three models. This is mainly due to drastic illumination condition that is not captured in the training data, and severe occlusion.

Our results (Tables I-IV) show that, in some categories, our method only achieves limited improvement over alternative methods. For instance, the performance of MMC and HOG+context can be close and HOG+context even fares slightly better in a few cases. As we discussed earlier, this is because when the contextual information is relatively dominant, context selection becomes less critical and a blind fusion could be equally effective. However, it is worth pointing out that one of the main

Fig. 12. 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.

TABLE V
EVALUATION OF THE EFFECTIVENESS OF TMMC-I. THE RESULTS OF OUR MMC MODEL WITHOUT TRANSFER LEARNING ARE IN BRACKETS FOR COMPARISON.

Target Category	Source Category	Model	Average Precision Rate
car	motorbike	TMMC-I	0.4006 (0.3741)
motorbike	car	TMMC-I	0.4253 (0.4020)
vehicle	car	TMMC-I	0.3952 (0.3838)
people	car	TMMC-I	0.3963 (0.3862)
people	bicycle	TMMC-I	0.3887 (0.3862)
people	motorbike	TMMC-I	0.3916 (0.3862)
bicycle	people	TMMC-I	0.2703 (0.2878)
car	people	TMMC-I	0.3607 (0.3741)
motorbike	people	TMMC-I	0.3954 (0.4020)

strengths of MMC is that it yields consistent improvement over detection without context, over all tested object categories and regardless of the usefulness of context, due to its ability to select context. In contrast, for some categories in PASCAL VOC2005 and VOC2007, particularly those with very diverse context, the TAS and HOG+Context models failed to improve the detection performance even with context modelled (e.g. people for TAS and bicycle for HOG+Context). Overall, our experiments suggest that the performance of those alternative models are much less stable.

C. Evaluation of Context Transfer Learning Models

We compare the two proposed context transfer learning models (TMMC-I and TMMC-II) with our MMC model to evaluate the effectiveness of transferring contextual information from source object categories to a target object category when the target data are limited. Specifically, among the 5 object categories in the three datasets used in our experiments that consist of limited target data (people, car, motorbike and bicycle in PASCAL VOC2005, and vehicle in Forecourt), we select one as the target category and another as an source category and perform detection using both TMMC-I and TMMC-II. The performance is then compared with that obtained by our MMC model using the target category data only. The results of the two transfer learning models are shown in Table V and Table VI respectively.

TABLE VI
EVALUATION OF THE EFFECTIVENESS OF TMMC-II. THE RESULTS OF OUR MMC MODEL WITHOUT TRANSFER LEARNING ARE IN BRACKETS FOR COMPARISON.

Target Category	Source Category	Model	Average Precision Rate
bicycle	people	TMMC-II	0.3063 (0.2878)
people	bicycle	TMMC-II	0.3964 (0.3862)
vehicle	car	TMMC-II	0.4019 (0.3838)
car	motorbike	TMMC-II	0.3724 (0.3741)
motorbike	car	TMMC-II	0.3835 (0.4020)
car	people	TMMC-II	0.3720 (0.3741)
motorbike	people	TMMC-II	0.3932 (0.4020)

Recall that the two models are designed for transferring contextual information when it is shared between the target and source categories in two different ways. In particular, TMMC-I should be used when the objects have similar context, i.e. likely to appear in similar environment or next to similar objects due to, e.g. similarity in functionality. TMMC-II, on the other hand, should be deployed when the objects have different context but their detections have a similar level of benefit from context, e.g. both are likely to appear in specific (albeit different) context or very diverse context. Table V and Table VI show that for different target and source pairs, different performance was achieved. Specifically, we have the following findings:

- The result in Table V suggests that when the target and source objects share similar context, TMMC-I does the job it was designed for, that is, improving the detection performance by utilising contextual information from source object categories. For instance, for car and motorbike, the AP rate is increased by 7% when car is the target category and 6% with motorbike as the target category.
- It is interesting to note that when people is the target category, its detection can benefit from transferring context from various other object categories including car, bicycle and motorbike using TMMC-I. But the same cannot be said when it is the other way around, i.e. people as source category. For example, the people detection performance is increased by about 3% when car is the source category, whilst the detection of car is decreased by about 4% (called negative transfer) when people is used as the source data. It can be because people often appear next to car, bicycle, or motorbike so they do share context. However, people also appear in much more diverse context, e.g. on a sofa. Therefore, the context of car, bicycle or motorbike can be considered as a subset of that of people. Consequently it is not a problem to transfer the context of car, bicycle or motorbike to people, but the effect can be adverse if the opposite is done, e.g. a car rarely appears on top of a sofa.
- As expected, TMMC-II improves the detection performance when the usefulness of context for the target and source categories are similar. For instance, people and bicycle not only share similar context but also similar weight of context. The detection of people is thus improved using both TMMC-I and TMMC-II. However, when the assumption made for TMMC-II does not hold, negative transfer is observed. For instance, car and motorbike share similar context but the context for motorbike could be more diverse than car probably due to its smaller size, resulting in negative transfer between them using TMMC-II. However, since TMMC-I is able to select the most common high-order contextual information shared between cars and motorbikes, TMMC-I is more effective in this case. Similarly transferring context weighting from people to car or motorbike would not help as shown in Table VI.

VI. DISCUSSIONS AND CONCLUSION

In this paper, we argue that contextual information should be quantified and selected explicitly before combining it with object appearance information for detection. To that end, we introduced

a context risk function and formulated a maximum margin context (MMC) model to quantify the contextual information of a candidate object, which is modelled by an object centred polar geometric context descriptor. In order to overcome the problem of lack of training context samples for context learning, we further proposed two transfer maximum margin context (TMMC) models under a joint maximum margin learning framework for context transfer learning. Compared to the state-of-the-art context models, the proposed MMC model utilises a novel context risk function 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 in that our MMC model directly addresses the maximization of the confidence of true positive detections defined by the 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. More importantly our TMMC models are able to transfer the useful related contextual information from other source categories in order to further reduce the ambiguity of context for target object detection. The effectiveness of the proposed models have been validated using both public benchmark datasets and datasets extracted from surveillance videos of busy public spaces.

It is worth pointing out that although in this work contextual information is represented by the proposed polar geometric context descriptor in order to capture multiple types of context, the MMC model is not restricted to any context representation. Due to the context selection ability, one may consider integrating different context representations combined with PGCD in the proposed MMC framework. For instance, our context descriptor may not be suitable enough to capture Scene-Thing context due to its object centred nature. However, a Scene-Thing context representation such as the one in [31] can be easily combined with our descriptor and selected in the same MMC model. Similarly, the HOG feature used in our descriptor sometimes may not be good enough for capturing 'Stuff' context (e.g. sky, road). One could thus combine HOG features with colour features to better represent both Thing-Thing and Thing-Stuff context. In addition, a potential improvement of the proposed method is to exploit contextural information from farther away regions. However, care also needs be taken when the farther away regions are non-stationary and distractive, or overly cluttered and noisy, e.g. in a crowded public scene, resulting in diminished benefit whilst increasing the computational cost.

One of the key contributions of this work is that for the first time a context transfer learning model is developed to address the over-fitting problem caused by lack of training data for context learning. Our experiments show both the potential of the proposed models and a limitation of the current models, that is, one has to use prior knowledge to select manually a suitable model to apply given the available training object categories. Overcoming this limitation by automatically selecting source categories is necessary for applying the proposed method to address a large-scale object detection problem when the number of object categories can be over thousands. This, however, is a very challenging problem. In particular, when a unsuitable model is applied, negative transfer learning which leads to unsatisfactory detection performance can happen. This is not a unique problem. Existing popular transfer learning methods for object appearance learning [34], [35], [49], [14], [52] also rely on prior knowledge to manually select suitable source object categories in order to avoid negative transfer. Although there are unsupervised methods that are applicable using any object categories as source data [17], [5], [39], as pointed out in [33], the problem of unsupervised transfer learning with negative transfer prevention given any auxiliary data is far from being solved. Developing such a method for context transfer could be even more challenging. This is because, as we explained in the related work, there are fundamental differences between object appearance and context transfer learning and none of these methods can thus be directly used for our problem. One obvious option is to detect and minimise negative transfer learning via cross validation. However, the very reason for using transfer learning is because of the lack of training data which will pose challenges for using cross validation to avoid negative transfer. Among the few existing unsupervised transfer learning work, the idea in self-taught learning [39] can be considered which infers the sparse coding for target contextual information over source context anchors. However, how this kind of sparse coding can be derived optimally for assisting object detection without negative transfer still needs more investigation. Another possible solution is to directly measure the similarity between different categories in order to identify whether certain aspects of the context of the two categories can be shared. Nevertheless, the challenge is about which or what technique should be selected or developed to compute the similarity and how the similarity score can be integrated into the TMMC models. Again it is more straightforward to find out whether two object categories are related in their appearance. For instance, one could perform attribute correlation by mining tags of Fickr images [41]. However, to infer the relationship automatically between the contexts of two object categories is much harder. We believe that context transfer remains an open problem and we wish that this work will help to attract more interests on this problem from the computer vision community.

ACKNOWLEDGEMENT

This research was mainly funded by the EU FP7 project SAMURAI with grant no. 217899. Dr. Wei-Shi Zheng was also additionally supported by the 985 project in Sun Yat-sen University with grant no. 35000-3181305 and NSFC (U0835005).

REFERENCES

[1] R. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.
[2] S. Y.-Z. Bao, M. Sun, and S. Savarese. Toward coherent object detection and scene layout understanding. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 65 –72, 2010.
[3] M. Bar and S. Ullman. Spatial context in recognition. *Perception*, 25:343–352, 1993.
[4] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3, 2003.
[5] E. Bart and S. Ullman. Cross-generalization- learning novel classes from a single example by feature replacement. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
[6] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 509–522, 2002.
[7] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14:143–177, 1982.
[8] A. Bosch, A. Zisserman, and X. M. noz. Scene classification via plsa. In *European Conference on Computer Vision*, 2006.

[9] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *European Conference on Computer Vision*, 2004.
[10] W. Choi, K. Shahid, and S. Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1282–1289. IEEE, 2009.
[11] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2011.
[12] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
[13] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[14] L. Duan, I. W. Tsang, D. Xu, and S. J. Maybank. Domain transfer svm for video concept detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[15] M. Everingham. The 2005 pascal visual object classes challenge. In *MLCW*, 2005.
[16] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.
[17] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.
[18] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.
[19] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nystrom method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004.
[20] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712–722, 2010.
[21] C. Galleguillos, B. McFee, S. Belongie, and G. Lanckriet. Multi-class object localization by combining local contextual interactions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
[22] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
[23] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *International Conference on Computer Vision*, pages 1458–1465, 2005.
[24] A. Gupta and L. S. Davis. Beyond nouns: exploiting prepositions and comparative adjectives for learning visual classifier. In *European Conference on Computer Vision*, 2008.
[25] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *European Conference on Computer Vision*, 2008.
[26] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15, 2008.
[27] HOSDB. Imagery library for intelligent detection systems (i-lids). In *IEEE Conf. on Crime and Security*, 2006.
[28] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *International Conference on Computer Vision*, 2005.
[29] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
[30] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.
[31] K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the tree: a graphical model relating features, objects and the scenes. In *Advances in Neural Information Processing Systems*, 2003.
[32] J. Nocedal and S. Wright. Numerical optimization, 2006. 2nd ed., Springer.
[33] S. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.
[34] S. J. Pan, J. T. Kwok, and Q. Yang. Transfer learning via dimensionality reduction. In *Proceedings of the 23rd AAAI conference on Artificial Intelligence*, pages 677–682, 2008.
[35] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. In *International Joint Conferences on Artificial Intelligence*, 2009.
[36] R. Perko and A. Leonardis. Context driven focus of attention for object detection. In *WAPCV*, 2007.
[37] R. Perko, C. Wojek, B. Schiele, and A. Leonardis. Probabilistic combination of visual context based attention and object detection. In *WAPCV*, 2008.
[38] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *International Conference on Computer Vision*, 2007.
[39] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766, 2007.
[40] X. Ren and J. Malik. Learning a classification model for segmentation. In *International Conference on Computer Vision*, 2003.
[41] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where and why? semantic relatedness for knowledge transfer. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.
[42] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.
[43] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.
[44] A. Singhal, J. Luo, and W. Zhu. Probabilistic spatial context models for scene content understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.
[45] A. Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2), 2003.
[46] A. Torralba, K. Murphy, W. Freeman, and M. Rubin. Context-based vision system for place and object recognition. In *International Conference on Computer Vision*, 2003.
[47] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *IEEE International Conference on Computer Vision*, 2009.
[48] L. Wolf and S. Bileschi. A critical view of context. *International Journal of Computer Vision*, 69(2), 2006.
[49] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *ACM Multimedia*, pages 188–197, 2008.
[50] Y. Zhang and D. Yeung. A convex formulation for learning task relationships in multi-task learning. In *Proceedings of the Twenty-fourth Conference on Uncertainty in AI (UAI)*, 2010.
[51] W.-S. Zheng, S. Gong, and T. Xiang. Quantifying contextual information for object detection. In *International Conference on Computer Vision*, 2009.
[52] A. Zweig and D. Weinshall. Exploiting object hierarchy: Combining models from different category levels. In *IEEE 11th International Conference on Computer Vision*, 2007.

Wei-Shi Zheng has joined Sun Yat-sen University under the one-hundred-people program. Prior to that, he received his Ph.D. degree in Applied Mathematics at Sun Yat-Sen University, China, 2008, and has been a Postdoctoral Researcher on the European SAMURAI Research Project at Queen Mary University of London, UK. His current research interests are in object association and categorization in visual surveillance.

Shaogang Gong is Professor of Visual Computation at Queen Mary University of London, a Fellow of the Institution of Electrical Engineers and a Fellow of the British Computer Society. He received his D.Phil in computer vision from Keble College, Oxford University in 1989. He has published over 200 papers in computer vision and machine learning, a book on Visual Analysis of Behaviour: From Pixels to Semantics, and a book on Dynamic Vision: From Images to Face Recognition. His work focuses on motion and video analysis; object detection, tracking and recognition; face and expression recognition; gesture and action recognition; visual behaviour profiling and recognition.

Tao Xiang received the Ph.D degree in electrical and computer engineering from the National University of Singapore in 2002. He is currently a lecturer in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include computer vision, statistical learning, video processing, and machine learning, with focus on interpreting and understanding human behaviour.