Abstract — We propose a method for improving object recognition in street scene images by identifying and filtering out background aspects. We analyse the semantic relationships between foreground and background objects and use the information obtained to remove areas of the image that are misclassified as foreground objects. We show that such background filtering improves the performance of four traditional object recognition methods by over 40%. Our method is independent of the recognition algorithms used for individual objects, and can be extended to generic object recognition in other environments by adapting other object models.

Index Terms — Object recognition, background detection, semantic modelling, scene understanding.

I. INTRODUCTION

SIGNIFICANT research has been dedicated to generic object recognition in static scene images. Robust object recognition, relating low level image features to content semantics, has become a key goal in computer vision. Classic content-based information retrieval (CBIR) systems such as QBIC [1] and VisualSeek [2] offer particular mechanisms for interactive search over images using different kinds of queries relating to the features and compositions of features of the images.

More recent CBIR systems [3] extend the scale and focus to online retrieval, fusing images with surrounding textual, geometric or even verbal information. Research is gradually moving from context-based object recognition towards a more generic knowledge-based scene recognition [4, 5], deriving conceptual information through recognition of scene objects.

Context-based recognition involves the analysis of the visual appearance of target objects and retrieval of identical visual patterns. This alone is insufficient to achieve recognition in generic natural scenes. Knowledge-based recognition identifies not only an object’s visual appearances, but also the contextual relationships with the environment, which can be used to reinforce recognition.

Hierarchical representation is one of the common ways to encapsulate domain knowledge in object recognition. Face recognition [6] is a typical example that benefits from such a representation, comparing the visual appearance of each facial component and the spatial relationships between them to differentiate one face from another. Domain knowledge is applied to guide the recognition, specifying the visual properties and spatial location for detecting a specific visual pattern. Each recognised pattern has a meaning associated with it to form a component, which can be used to derive the higher level semantic – the face.

Hybridised recognition, combining top down and bottom up approaches [7], has recently attracted attention as a systematic way to combined local feature identification with global object recognition. This hybrid recognition method often involves organising images into structural hierarchies of segments and searching for desired objects, represented by the intermediate nodes, within the hierarchy [8]. Schindler’s [9] discriminative approach is generally more computationally efficient then the hierarchical approach, as it is initiated from potential segments.
and recursively merges them with the surrounding regions or splits them into sub-regions to achieve an optimised matching with the template.

Qin and Vrusias [10] proposed a component-based shape frequency approach to recognising vehicles in real life scenes, using a generic model to monitor vehicle characteristics at global and component level. Lin et al [11] proposed a component decomposition method, in which different objects are represented and monitored in basic geometric shapes. Both methods require human intervention in component selection to form models, so are not directly applicable to generic object recognition. Leibe et al [12] proposed a generative approach, identifying key features and their co-location and co-activation. This method minimises manual intervention by symmetrically matching visual features extracted from image patches in a prebuilt feature library. However, the key visual features may not necessarily represent the key semantic features.

In this paper, we are proposing a method to improve the identification of foreground objects in street scene images by automatically filtering background objects. As shown in Fig 1, we introduce independent detection methods for foreground and background objects; and filter out the foreground misclassification based on an integrated background map. The method is evaluated here in relation to vehicles, but has been tested with pedestrian objects, with similar results achieved.

II. FOREGROUND OBJECT DETECTION

For foreground object detection, four recognition methods are evaluated: shape contour matching; segmentation splitting & merging; Edge-Surface frequency modelling and Saliency frequency modelling.

A. Shape Contour Matching

Shape contour matching is a simple sampling approach, which works efficiently when the shape of the detectable object is visually separable from its surroundings. As shown in Fig 2, matching is a process of minimizing distance differences between the sampled contour points on the template and the corresponding points on the processing image patches.

B. Segmentation Splitting & Merging

The Splitting & Merging approach involves decomposing the image into segments based on visual continuity, then iteratively merging or splitting these segments to form a desired object, guided by an object model [7, 8]. By combining adjacent segments with similar visual features, following Xu, the processing [8] image can be converted into a hierarchical representation based on merge order. The image hierarchy is searched top-down for the target object.

C. Edge-Surface Frequency Modelling

Edge and surface (i.e. non-edge) approaches are robust against photometric variations. The Edge/Surface frequency approach [10] analyses edge distributions to establish a generic model, which can be used to identify similar patterns in the image dataset.

The generic object model is an edge/surface frequency map, generated by statistically assembling the shapes of individual sample objects. As shown in Fig. 3, regions with concentrated points indicate the likelihood that an edge/surface point is present in these regions.
The frequency map is a unique way to monitor the common edges and surfaces shared among individual instances for a specific object. Template matching is performed to locate image patches with similar edge and surface distributions. The edge/surface ratio between the vehicle patch and the template is expected to be stable.

D. Saliency Frequency Modelling

Instead of building the frequency map using shape information, the frequency map can be built by measuring saliency distribution. The saliency frequency map is formed by assembling normalised saliency maps generated from individual samples. The top salient areas within the frequency map is used as a template in the recognition process.

III. BACKGROUND AREA DETECTION

It is difficult to achieve robust generic object recognition in natural environments due to the visual complexity of such scenes. However, it is a simpler task to recognize specific background objects in a scene. By filtering out such background objects, recognition can be focussed on the remaining areas, potentially increasing recognition precision and improving computational efficiency.

### Table 1

**BACKGROUND OBJECTS IN STREET SCENE IMAGES**

<table>
<thead>
<tr>
<th>Frequency of Occurrence (total of 500 images)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Road</td>
</tr>
<tr>
<td>493</td>
</tr>
</tbody>
</table>

Table 1, showing the top five frequent appearing background objects, is an annotation summary of randomly collected images from the ‘Street’ sub-category in MIT LabelMe dataset [13]. We implemented three identification methods to cover the five common background objects: foliage, road, and structural blocks; then the identification from the three methods are combined to form a background map to guide the background filtering.

A. Foliage Detection

Hamme claims that foliage tends to have a large amount of random but strong edges and distinctive colour and texture distribution [14]. Following Hamme, we combine edge intensity, orientation, and colour distribution for foliage detection.

B. Road Detection

According to Hu [15], photometric road regions have high illumination and stable texture distributions. Analysing the colour distribution in R*G*B space reveals that road regions contain average distribution across all colour bands with a slight blue predominance.

C. Structural Block Detection

Street scene is often filled with artificial objects containing basic geometric shapes, which can be detected using basic shape extraction algorithms such as Prewitt, and Hugh.

IV. BACKGROUND FILTERING

Road regions normally co-appear together and occupy large areas [15]. Thus, any isolated regions, which are smaller than the size threshold, are discarded from being recognised as road region. In addition, vehicle hypotheses should co-locate closely with the road regions.

By analysing the structural representation of the detecting objects, the irrelevant structures can be filtered out. In the case of vehicle recognition, vertical edge distribution has an eligible presence. Thus, filtering out vertical structural patterns should have minimum impact on vehicle recognition.

Combining the detected foliage, road and structural regions forms a background map for street scene. Shown in Fig 5, the regions in black are regions that are not background, thus they can be used as the target regions for vehicle recognition.

V. EXPERIMENTS

The training dataset are randomly selected from the ‘street’ subset of the MIT LabelME dataset [13]. It comprises 64 image samples containing 100 vehicles (i.e. each image may contain 0 up to 4 vehicles). Testing images (73 images with 150 vehicles) are randomly selected from MIT StreetScene [16] database. The four traditional image processing methods
generally provide poor performance. In particular, Shape contour and edge/surface frequency matching alone are insufficient to distinguish vehicle objects from regions with large amount of random edge distribution.

After introducing background filtering, recognition results (W/O BKGD) improved by over 40% (dF) for all methods, shown in Table 2. Furthermore, background filtering is able to filter out obvious false positives (average 70% reduction in all methods) which, as explained previously, can be difficult to distinguish using sole foreground detection methods.

<table>
<thead>
<tr>
<th></th>
<th>F(_{w}) W BKGD</th>
<th>F(_{w}) W/O BKGD</th>
<th>dF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape Contour</td>
<td>0.012468</td>
<td>0.469036</td>
<td>0.456566</td>
</tr>
<tr>
<td>Segment Merging</td>
<td>0.225594</td>
<td>0.681198</td>
<td>0.455604</td>
</tr>
<tr>
<td>Edge Frequency</td>
<td>0.209774</td>
<td>0.649833</td>
<td>0.440058</td>
</tr>
<tr>
<td>Saliency Frequency</td>
<td>0.257714</td>
<td>0.684605</td>
<td>0.426891</td>
</tr>
</tbody>
</table>

VI. CONCLUSION & FURTHER WORK

Comparing component-based recognition [17] that focuses solely on vehicle, it is more computationally efficient to pre-filter out easy identified non-vehicle areas. More importantly, it is able to reduce false positives based on extracting mutually exclusive visual features.

In this work, we have presented a method for background filtering that improves vehicle recognition in street scenes. Compared to recognition algorithms that focus solely on objects, significant improvement can be achieved by analysing the semantic relationships between foreground objects and background objects. This method could be applied to generic object recognition by generating appropriate object models. Further improvement on filtering can be achieved if a systematic approach can be formed to capture the semantic relationships between background objects and between all objects in images could be derived. Further extension of this work can lead to multimedia ontology for particular scenes.

REFERENCES


