What Makes an On-road Object Important?
Eshed Ohn-Bar and Mohan M. Trivedi
Computer Vision and Robotics Research Laboratory
University of California San Diego
{eohnbar, mtrivedi}@ucsd.edu

Abstract—Human drivers continuously attend to important scene elements in order to safely and smoothly navigate in intricate environments and under uncertainty. This paper develops a human-centric framework for object recognition by analyzing a notion of object importance, as measured in a spatio-temporal context of driving a vehicle. Given a video, a main research question in this paper is - which of the surrounding agents are most important? The answer inherently requires complex reasoning over the current driving task, object properties, scene context, intent, and possible future actions. Therefore, we find that various spatio-temporal cues are relevant for the importance classification task. Furthermore, we demonstrate the usefulness of the importance annotations in evaluating vision algorithms (specifically, for the task of object detection) in an application where trust in automation is imperative and errors are costly. Finally, we show that importance-guided training of object detection models results in improved detection performance of surrounding objects of higher importance. Hence, such models may be better suited for use in representing safety-critical situations, predicting surrounding agents’ intentions, and in human-robot interactivity. The dataset and code will be made publicly available.

I. INTRODUCTION

Consider the image in Fig. 1(a). Can you try to detect all of the vehicles in the image? Some of the vehicles in the proximity of the ego-vehicle are easy to detect while other vehicles with increased distance, occlusion, and truncation may be more challenging. On the road, human drivers attend to only a subset of the on-road occupants which are most relevant to the current navigation task, given the situational spatio-temporal context. Motivated by this phenomenon, the goal of this paper is to learn object importance ranking models which can automatically predict the importance score of objects in a driving scene.

Next, please re-consider the image in Fig. 1(a). Although knowledge of all obstacles is important for obstacle avoidance while navigating, some objects may require more of a driver’s attention than others (Fig. 1(b)). For instance, remote objects may be less relevant than near ones to the immediate navigation task, but other object properties may also play a role. The answer to the question of what makes an object important inherently requires reasoning over the current driving task, object properties, scene context, intent, previous experience, and expected future actions [2]. Humans rely on situational awareness in order to continuously analyze the scene and its salient elements [2], [3], and intelligent and self-driving vehicles are expected to perform similar reasoning skills for smooth and safe navigation (as well as earn the trust of their users). This study is also concerned with learning human-centric models for scene perception, which can be valuable to a wide range of applications in computer vision including saliency [4]–[7], robotics [8], and ego-centric vision [9], [10].

A. Contributions

Our paper makes the following contributions in the important field of computer vision and pattern recognition for intelligent vehicles.

Dataset and cue analysis: We collect object-level importance annotations on KITTI videos [1] from a varied subject pool. The dataset is used to learn importance ranking models which provide insights into what makes an object important. The proposed prediction task allows for evaluating a complex spatio-temporal reasoning skill in real-world settings, and a wide range of cues is shown to impact spatio-temporal prediction of importance. We emphasize that this study is not concerned with ethical issues in autonomous driving, but with obtaining a better understanding of the limitations and requirements for on-road object recognition, safe navigation, and human-centric AI.
Fig. 2: (a) The interface used to obtain object-level importance ranking annotations. The cyclist is highlighted as it is the currently queried object to annotate, colored boxes have already been annotated with an importance level by the annotator, and blue boxes are to be annotated. (b-d) Example frames in the dataset with overlaid object-level importance (averaged over subjects). Object type, state, position, orientation, scene context, and other cues are all shown to play a role in determining an importance score by annotators.

Importance-guided training and evaluation of models: Our analysis demonstrates that the large majority of road occupants on the KITTI benchmark [1] are consistently categorized as low important objects. As highly important objects are rare, the importance annotations give rise to novel performance metrics and error/bias analysis in an application where errors made by an algorithm are safety-critical. Specifically for the task of object detection, we demonstrate that importance-guided training can significantly improve detection of highly important objects. As such joint detection-attention models are encouraged to emphasize saliency and contextual cues in detection, they can provide more useful output for driver assistance and safe navigation tasks. Interestingly, we show that such an important insight would have been difficult to identify with traditional detection performance evaluation metrics.

II. IMPORTANCE ANNOTATIONS COLLECTION

The KITTI dataset [1], [11] is widely used for evaluating a variety of vision tasks for autonomous driving, including object detection and tracking. We supplement the object annotations with an object importance label. In addition to image-level tracklet annotations of road occupants (pedestrians, cyclists, and vehicles), KITTI provides a rich set of other modalities and semantic object attributes, including a 3D box and orientation (annotated in the LIDAR data), IMU/ego-vehicle dynamics, GPS, and an occlusion state. The sensorization and annotation makes KITTI an ideal dataset for studying cues related to importance in driving settings.

Importance annotations: Experiments were done in a driving simulator with KITTI videos shown on a large screen. A set of 8 videos was selected for annotation. Subjects initially watched each video twice, and then annotated every 10th frame by providing an object-level integer between 1-3 (1 being high and 3 being low importance). A visualization of the annotation interface is shown in Fig. 2. The choice of three importance levels was chosen in order to reduce ambiguity as much as possible without overly restricting the experiments. For instance, allowing for only two levels (yes/no) of importance is somewhat restrictive as there may be ambiguous cases where a decision can’t be confidently made. Furthermore, it is reasonable to expect that some objects will fall under a middle between high and low importance. On the other hand, a continuous ranking score may have been used, but could have lead to large confusion among subjects and more guessing, which we aimed to reduce. Subjects were asked to imagine driving under similar situations, and mark objects by the level of attention and relevance they would’ve given the object under real driving.

Fig. 3 visualizes the output provided by the subjects. Although the instructions were kept fixed among the experiments, the resulting annotations contained variations due to the task’s inherent subjectivity. Upon a closer inspection, we found a strong correlation between annotation output and subject driving experience. Interestingly, greater driving experience implied a higher percentage of moderate and high importance annotations (Fig. 3(b)). Subjects reported inspecting object orientation (useful for knowledge of future activities [12]), relative position, existence of traffic barriers, and relationship among objects for determining the importance label. We note that despite the subjective nature of the task, a subset of the annotated objects does contain high consistency. For instance, scenarios of dense scenes with tens of road occupants that are heavily occluded or are across a barrier consistently generated low importance annotations. This applies to a large number of objects in KITTI. Three importance classes are obtained and used throughout the paper by taking the median vote among subjects for each sample. The dataset used in the experiments
A. Object attributes model, $M_{\text{attributes}}$

For an instance $s$ and class importance $c \in \{1, 2, 3\}$, we train the following prediction model

$$
M_{\text{attributes}}(s) = w_{c, 2D-\text{obj}}^T \phi_{2D-\text{obj}}(s) + w_{c, 3D-\text{obj}}^T \phi_{3D-\text{obj}}(s) + w_{c, \text{ego}}^T \phi_{\text{ego}}(s) + w_{c, \text{temporal}}^T \phi_{\text{temporal}}(s) \tag{1}
$$

where each term is defined below.

**2D object features**: After projection of the annotated 3D object box to the image plane, the $\phi_{2D-\text{obj}} \in \mathbb{R}^4$ features are the concatenation of the height in pixels, aspect ratio, occlusion state (either none, partial, and heavy occlusion) and truncation percentage.

**3D object features**: $\phi_{3D-\text{obj}} \in \mathbb{R}^6$ is composed of the 3D left-right and forward-backward range coordinates $(x, z)$ given by the LIDAR, Euclidean distance from the ego-vehicle, orientation in bird’s eye view, and object velocity components, $[V, \theta V]$.  

**Ego-vehicle features**: Since ego-vehicle speed may impact which objects are considered important and this information is presented to the annotators, the attribute model includes ego-vehicle velocity and orientation features $\phi_{\text{ego}} = [\text{ego} | V, \theta \text{V}]$.

**Temporal attributes**: We hypothesize that past information has an influence on driver expectation as to what is happening and will happen next. Hence, temporal evolution of attributes may contain relevant information for importance ranking. This assumption is captured in $\phi_{\text{temporal}}$, where the object and ego-vehicle attributes described above are concatenated over a past time window. Additionally, max-pooling over the time window and the Discrete Cosine Transform (DCT) coefficients [13] are also included in computing $\phi_{\text{temporal}}$.

B. Visual cues model, $M_{\text{visual}}$

Given the 2D bounding box annotation of objects in video, we propose a visual prediction model for mapping an image region to an importance class,

$$
M_{\text{visual}}(s) = w_{\text{c, obj}}^T \phi_{\text{obj}}(s) + w_{\text{c, spatial}}^T \phi_{\text{spatial}}(s) + w_{\text{c, temporal}}^T \phi_{\text{temporal}}(s) \tag{2}
$$

**Object visual features**: For $\phi_{\text{obj}} \in \mathbb{R}^{4096}$ features, we employ the activations of the last fully connected layer of the OxfordNet (VGG-16) [14] convolutional network. The network was pre-trained on the ImageNet dataset [15] and fine-tuned on KITTI using Caffe [16].

**Spatial context features**: In order to better capture spatial context, such as relationship with other objects in the scene or occlusion state, each object instance is padded by an amount of $\times 1.75$ for generating $\phi_{\text{spatial}} \in \mathbb{R}^{4096}$.  

**Temporal context features**: Given an object tracklet of box positions and a temporal window, temporal $\phi_{\text{temporal}}$ features are extracted. The previous object and spatial context features, $\phi_{\text{obj}}$ and $\phi_{\text{spatial}}$, are computed over a time window, concatenated, and max-pooled.
IV. EXPERIMENTAL EVALUATION OF IMPORTANCE MODELS

For each of the three importance classes, a precision-recall curve is computed and the area under the curve (AP) provides a summary metric for classification quality of the model. Finally, AP is averaged over the three important classes (mAP) for an overall classification quality metric (higher mAP value implies better classification performance). We note that by treating each importance class as a separate class, the proposed evaluation procedure handles the large imbalance in sample size of the importance classes. The dataset is split approximately in half for a 2-fold cross validation. The split is such that no samples from the same video are used for both training and testing.

Analysis of $M_{attributes}$: Table I shows the results of the classification experiments with varying combinations of features and an object class-agnostic importance prediction model. As $M_{attributes}$ employs clean annotation (occlusion state, truncation level using ground truth 3D boxes, etc.) and sensor data, it provides a strong classification baseline. In addition to providing a comparison with $M_{visual}$, it is also useful for analysis of cue relevance. Fig. 4(a) shows the classification power of each of the features and a combination of all, which leads to a performance of 53.70% mAP. We note that the training and testing sets are different for each of the object classes in Fig. 4(a), either limited to a specific object class or over all objects shown in the ‘All’ category. When training a class-agnostic model, all object classes are treated as one and selected cues for importance prediction must be learned such that they generalize over the different object classes. For the vehicle object class, occlusion state is shown to be a strong cue when determining an object’s importance class. This is to be expected, as occlusion by another object may generally reduce the object’s importance class. Nonetheless, a combination of the occlusion state with the other attributes provides a better representation of the object’s importance class (increasing mAP from 47.15% to 53.70%). Vehicle orientation is also shown to be particularly useful when considering the vehicle object class, as it relates to traffic direction and future action. 2D object properties, such as height in pixels, are important as they are related to distance perception. For the pedestrian object class, occlusion state, longitudinal placement, and distance from the ego-vehicle are all shown to be particularly useful in predicting importance. The patterns for the cyclist object class are less clear as it contains a smaller number of object samples, yet similar observations hold.

Table I also demonstrates the impact of temporal cue modeling for importance class prediction. When considering a class-agnostic model, the temporal window features are shown to improve performance over a window of up to 2.5 seconds. Incorporation of $\phi_{temporal}$ leads to a significant classification improvement, from 53.70% to 60.35%. Therefore, a good temporal context model is essential for the task of importance prediction. On the vehicle object class, incorporation of temporal attribute features leads to the largest improvement, from 48.49% to 56.18%. For the pedestrian object class, the performance significantly improves as well, from 36.20% to 40.09%. The improvement for the cyclist object class is minor, but obtaining further instances is required for drawing reliable conclusions.

Analysis of $M_{visual}$: The analysis of $M_{attributes}$ revealed which cues are useful for importance prediction. The visual prediction model must also be able to accurately capture object attributes, including object occlusion state, orientation, and distance from the ego-vehicle. The $\phi_{obj}$ baseline, of importance classification using a local object window, provides a performance of 51.06%. While the network is able to capture...

<table>
<thead>
<tr>
<th>Model</th>
<th>mAP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_{visual}(\phi_{obj})$</td>
<td>51.06</td>
</tr>
<tr>
<td>$M_{visual}(\phi_{obj} + \phi_{spatial})$</td>
<td>55.53</td>
</tr>
<tr>
<td>$M_{visual}(\phi_{obj} + \phi_{temporal})$</td>
<td>53.30</td>
</tr>
<tr>
<td>$M_{visual}(\phi_{obj} + \phi_{spatial} + \phi_{temporal})$</td>
<td>56.34</td>
</tr>
<tr>
<td>$M_{attributes}$ (w/o $\phi_{temporal}$)</td>
<td>53.70</td>
</tr>
<tr>
<td>$M_{attributes}$ (with $\phi_{temporal}$)</td>
<td>60.35</td>
</tr>
</tbody>
</table>
TABLE II: Evaluation of object detection models (average-precision) using the proposed set of importance metrics and the Faster-RCNN framework (FRCN) [17]. ‘IG’ stands for importance-guided cost-sensitive training.

<table>
<thead>
<tr>
<th>Method</th>
<th>Test Settings</th>
<th>Importance Test Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy</td>
<td>Moderate</td>
</tr>
<tr>
<td>FRCN-ZF</td>
<td>89.26</td>
<td>79.70</td>
</tr>
<tr>
<td>FRCN-ZF-IG</td>
<td>91.09</td>
<td>80.86</td>
</tr>
<tr>
<td>∆AP</td>
<td>+1.83</td>
<td>+1.16</td>
</tr>
<tr>
<td>FRCN-VGG</td>
<td>95.63</td>
<td>88.98</td>
</tr>
<tr>
<td>FRCN-VGG-IG</td>
<td>94.54</td>
<td>88.71</td>
</tr>
<tr>
<td>∆AP</td>
<td>-1.09</td>
<td>-0.27</td>
</tr>
</tbody>
</table>

Motivating importance metrics for object detection: Traditional evaluation on vision datasets (PASCAL [24], Caltech [25], KITTI [1], etc.) separate objects by size, occlusion, and truncation. Specifically for KITTI, three evaluation procedures of ‘easy’ (above 40 pixels in height, truncation under 15%), ‘moderate’ (above 25 pixels in height, with partial occlusion, truncation under 30%), and ‘hard’ (with heavy occlusion and truncation under 50%) are employed. In all the aforementioned datasets, challenging instances of size, occlusion, and truncation are often entirely excluded from training/evaluation, yet these choices are arbitrary in the context of driving where such instances may be potentially relevant to safety-critical events (e.g. a highly truncated vehicle which is overtaking the ego-vehicle). This motivates evaluation on importance classes for providing complementary analysis to the traditional metrics. The importance metrics can also reveal dataset bias, as the rarity of high importance instances may bias models both in training and evaluation. Specifically, in training time, the model may emphasize visual attributes found in the most common objects (e.g. vehicles of low relevance), and evaluation with traditional metrics may not reveal such bias.

Importance-guided training: As a final experiment, we employ the Faster-RCNN [17] framework for training a vehicle detector on KITTI. First, fine-tuning is performed on an auxiliary dataset of KITTI objects taken from videos not used in the importance prediction experiments. Next, we continue fine-tuning on the importance dataset (performed twice for each fold in the cross validation), but modify the loss function from [17] to weigh objects of higher importance more heavily. We refer to this training process as ‘importance-guided’ training. The process is performed both for the ZF [26] and VGG network architectures. In all test settings, an overlap of 0.7 is required between a predicted and a ground truth box for a true positive. Furthermore, images are up-sampled by a factor of 2.6 in training and testing, which we found necessary for detection of smaller objects.

Analysis with importance metrics: The results of the experiments for object detection are shown in Table II. Several conclusions can be drawn from the analysis. While importance metrics may be correlated with traditional, importance-agnostic metrics, the two types of metrics contain complementary information. For instance, comparing easy and high importance test settings, or hard and low importance test
settings, the overall AP numbers differ. The low importance category is shown to be particularly challenging. Furthermore, the importance test settings contain vehicles with higher truncation ratio than the hard test settings.

The deeper VGG model is shown to significantly improve over the ZF model in all test settings, but the main takeaway is demonstrated by inspecting the difference in AP (ΔAP) between importance-guided and importance-agnostic training and evaluation. As shown in Table II, importance test settings are essential for measuring improvement in detection performance over objects of higher importance, which is otherwise not clear. For instance, the ZF-IG model achieves an AP of 87.19% on high and moderately important objects, an improvement of 4.39 AP points over the importance-agnostic baseline. On the other hand, the performance improvement when employing traditional test settings is minor. A similar observation holds for the VGG-IG model. Importance-guided training specifically targets visual challenges and appearance patterns of objects of higher importance, thereby significantly improving the detection performance for such objects. This experiment demonstrates the feasibility of employing the collected importance annotations for evaluating vision tasks, as traditional and importance metrics are shown to be complementary.

VI. CONCLUDING REMARKS

This paper analyzed the task of video-based importance prediction using a variety of multi-modal spatio-temporal cues. To that end, a human-centric object-level importance annotations dataset was collected for KITTI videos. Two types of models were used to perform in-depth analysis into what types of cues make an object important. In addition to subtle object-level cues, such as object occlusion state and orientation, temporal dynamics of a proposed set of multi-modal situation attributes were shown to be crucial for object importance classification. Moreover, we demonstrated how the collected annotations can be useful when evaluating vision tasks, in particular for object detection performance evaluation. The experimental analysis uncovered the role of dataset bias and motivated training cost-sensitive object detection models which are better at detecting objects of higher relevance to the driving task. In the future, we would like to further study spatio-temporal visual modeling for the purpose of importance prediction [20], [22], analyze generalization and importance in other contextual settings, such as U.S. highways [27], and research the usefulness of the importance annotations for other vision tasks on KITTI. Another future direction is in developing subject-specific scene perception models, as motivated by the demonstrated relationship between subject personal attributes (e.g. driving experience) and importance perception.

VII. ACKNOWLEDGMENTS

We thank the subjects who participated in the study, our CVRR colleagues for helpful discussions, and the reviewers for their constructive comments. We are grateful for the support of our associated industry partners.

REFERENCES