Seeing Signs of Danger: Attention-Accelerated Hazmat Label Detection

Mahmoud A. Mohamed, Jan Tünnermann, Bärbel Mertsching

Abstract— Rescue robots and similar vehicles must recognize various visual objects. Some are of particular interest and must be reliably recognized, for example, hazard signs. Hazmat labels and other intentionally placed signs of danger are typically attached to walls, containers, or vehicles, in locations where they attract attention. These backgrounds typically are of relatively simple structure (though not guaranteed to be plain) while the labels have saturated colors and high contrasts. We provide a new dataset that contains such images and a novel hazmat detection method. It includes an attentional preselection, which explicts the salient design and placement of the labels to locate them, followed by a SIFT-based classification that determines the concrete label type. The results show substantial speed improvements and accuracy gains over the traditional method without an attention stage.

I. INTRODUCTION

Future fully autonomous rescue robots could assist emergency responders by exploring and mapping inaccessible areas, for instance in collapsed structures. Areas with human victims are of highest priority in rescue missions. Typical environments contain many elements originally designed for humans but which are highly relevant for the technical rescue units. For instance, labels with specific shapes, colors, and symbols indicate hazardous materials. Rescue robots must recognize such “hazmat” labels to avoid, for example, triggering explosions, exposing human rescue workers or victims to chemical dangers, and to enter such information into maps of the area.

The detection of hazmat labels is also part of rescue robot test scenarios, such as the RoboCup Rescue League competition (see, e.g., [1]). It belongs to the readiness tasks, which are to be performed before every simulated mission, to test the robot’s ability to detect such signs and its visual capabilities in general. The labels also occur within some exploration tasks.

In the present work, we present a novel procedure for fast hazmat label recognition, which is facilitated by an attentional preselection. Selective visual attention is a mechanism in human visual perception to select relevant among irrelevant stimuli. The rationale behind including a technical model of attention in the procedure is that the labels are designed to attract human attention, using salient shapes and colors. The attentional preselection can massively reduce the search space by excluding non-salient areas of the scene. Candidate patches can be forwarded to object recognition algorithms to confirm the candidates and classify the particular hazmat type. We evaluate the procedure with a novel dataset, which we also make available for testing future approaches.

The detection of hazmat labels is a challenging problem because of several aspects, such as shape variation, partial occlusion, illumination change, inner class variation, and variety of classes. Existing approaches to hazard sign detection can be categorized into three different classes [2]: detection based on shape elements, such as edges and contours [3], based on color using histogram or template matching [4], [5], and based on saliency maps [6]. Several methods have been used to recognize hazmat labels based on geometric constraints [7], [8], boosted cascade features [9], and statistical moments [10]. Recent approaches [11] use deep learning with convolutional neural networks.

These approaches are often tailored toward the detection of hazmat labels, limiting the possibility to extend them toward other objects of interest in rescue scenarios. Therefore, we aim at using the scale-invariant feature transform (SIFT) approach [12] for recognizing hazmat labels, which has been successfully applied in many object recognition applications. To compensate for the fact that SIFT is rather slow compared to other methods, we propose an attention-based preselection that limits the search space by excluding image areas that contain no salient objects.

II. ATTENTION-ACCELERATED HAZMAT LABEL RECOGNITION

The general procedure proposed in this paper is the following: The input image first undergoes an attention analysis, based on a saliency calculation which highlights hazmat labels and similar attention-grabbing objects. A tight image patch containing the most salient location is then forwarded to the hazmat label detection, which either confirms the candidate or requests a larger, more inclusive patch from the attentional stage (see Figure 1).

Fig. 1. General procedure for the attention-accelerated hazmat detection.

Because hazmat labels typically are distinctively colored, and to avoid the influence light and shadows, the $a^*$ and $b^*$ channels of the CIELAB color space are used in the saliency calculations. These channels encode the chromatic information. Moreover, because the details of the labels are
typically symbols, line drawings, and text in black or white, the L*-channel is used for the final recognition with SIFT.

A. Attentional Preselection

To facilitate the detection, an artificial visual attention model is employed to exploit the attention-grabbing design and placement of hazmat labels. There are numerous models available in the literature (see [13] for a survey), many of which could be apt for this task. However, one method that is particularly interesting is the spectral residual saliency approach [14]. It is a frequency-domain procedure, which can be expected to capture the difference in spatial frequencies between the plainly colored hazmat labels and the typical structures of surfaces to which they are attached. To compensate the drawback that this method produces blurry priority maps (cf. [15]), we combine it with a superpixel over-segmentation using the CIELAB color space. Superpixels and other segmentation methods are widely used to produce contour boundaries within the image. We set the parameters number of segments to 30 and the compactness to 40. This produces superpixels that capture hazmat labels as one or two segments under the typical viewing conditions we tested.

The saliency $S_{\psi_i}$ of every superpixel $\psi_i$ is then determined as

$$S_{\psi_i} = \sum_{(x,y) \in \psi_i} \frac{S_a((x,y)) \cdot w_a + S_b((x,y)) \cdot w_b}{w_a + w_b} \frac{1}{|\psi_i|},$$

Hence, the weighted average of the a* and b* spectral residuals is aggregated for each pixel $(x,y)$ within a superpixel and normalized by the superpixel size $\psi_i$. The weights $w_a$ and $w_b$ control the strength of the green–red and blue–yellow CIELAB channels. We set $w_a = 1/\text{var}(L_a(x,y))$, $w_b = 1/\text{var}(L_b(x,y))$, where $\text{var}(\cdot)$ is variance and $L_a(\cdot)$ and $L_b(\cdot)$ are Laplacians of the a* and b* saliency maps. Effectively, this promotes channels with clearer peaks and little noise.

Then, the most salient superpixel is forward to the hazmat label detection. If this does not lead to a detection, the next larger neighborhood of superpixels is forwarded (see Figure 1). Note that we include an additional margin (10% of the selected patch size) when forwarding the extractions and that we do not mask out the background, both because it may help SIFT to match features at the object boundaries.

B. Classification of Hazmat Labels

The classification algorithm contains training and testing phases. In the training phase, we used standard images of hazmat labels\footnote{Note that there is a trade-off: If superpixels are as large or larger as hazmat labels, the procedure in Figure 1 tends to return a full hazmat label in the first selection, if it is sufficiently salient. However, with such large superpixels, the neighborhood (which is determined on the basis of superpixels) increases quickly in subsequent cycles if the label was not found on the first attempt, increasing the computational load quickly. Smaller superpixels may lead to fewer first-attempt detections but allow for more gradual increases of the search area.}. We applied the SIFT algorithm to extract keypoints in each image template and built a SIFT descriptor for each of them. The descriptors were saved in a codebook which can then be used to recognize whether image contain hazmat labels and of what type they are. Figure II-B shows the results after applying SIFT keypoint extraction on different hazmat templates.

In the testing phase, SIFT keypoints are extracted and SIFT descriptors are constructed for the test image. To find the best match with the codebook descriptors, the k-nearest neighbors (KNN) algorithm is used with the Euclidean distance between two descriptors. Figure II-B shows the results after extracting SIFT keypoints from an image, while figure 9 shows the results after finding the best match using the KNN.

\footnote{Obtained from: http://ian-albert.com/hazmat_placards/}
Fig. 2. Examples of labels with SIFT keypoints. The size of the circles indicates the characteristic scale and the oriented line segments indicate the keypoint orientation.

Fig. 3. Keypoints detected on one of the test images.

III. THE DATASET

To evaluate our work, and to encourage the community to test their approaches, we provide a novel high-quality hazmat label test dataset with this publication. Overall, the dataset includes 600 high-resolution (5184 × 3456) images with manually created pixel-accurate ground truth maps. Every image includes one of eight types of hazmat labels. Photographs were taken in front of three different typical backgrounds at a distance of roughly 150 cm and under different lighting conditions (see Table I). Moreover, five different azimuths (−45°, −30°, 0°, 30°, 45°), each with five different image position (center, top left, lower left, top right, lower right), are included.

A hand-held digital single-lens reflex camera was used on automatic settings (no flash). The focus was adjusted to target the image center, even when the objects were placed off-center, to mimic their potential occurrence in peripheral image regions. The test images with the “bricks” background were taken outdoors, are very sharp, but contain harsh light and strong shadows. The images with “OSB” and “woodchip wallpaper” backgrounds were taken indoors under weaker incandescent light and are therefore slightly blurry to varying degrees. Note that harsh light and blurry images are both image deficits found under real conditions, rendering the dataset realistic. Example images are shown in Figure 4, further details are provided in Table I.

For all 600 test images, pixel-accurate binary ground truth masks were created. The dataset is available for download⁴.

IV. RESULTS

We evaluated the procedure presented in this paper on the dataset described above. Three main questions are addressed: (1) Is the attentional preselection based on spectral residual saliency and superpixels generating useful candidates that include the hazmat labels, (2) does the preselection lead to processing speed improvements in the overall procedure, and (3) what is the overall performance of the hazmat label detection with attention and SIFT?

A. Attention-based Preselection

Three versions of the saliency calculation are assessed here. The raw spectral residual “SR” obtained for a grayscale version of the input (cf. [14]), the combined spectral residual of the a* and b* CIELAB channels “SR-a*b*”, and the version enhanced by superpixels, “SR-a*b*-SP”.

Figure 5 illustrates the characteristics of these different approaches for an “explosive” hazmat label from the most difficult class with harsh sunlight. The grayscale spectral residual saliency in Figure 5B is heavily affected by the light conditions. The CIELAB version in Figure 5C is indeed less prone to this and assigns high activity to the location of the label but is blurry, not containing the target contours. The superpixel version in Figure 5D improves on this and accurately renders the shape of the target salient.

To assess the attentional-preselection quantitatively, we use the $F_\beta$-score classification metric (with $\beta^2 = 0.3$; see [13]). First, the saliency threshold that optimizes the $F_\beta$-score over the whole dataset is determined. Then $F_\beta$-scores for the different background classes are calculated. This is done for the three different types of saliency described above. Note that the exact shape of the hazmat labels is considered. The score is high if the saliency has a high precision (all pixels that are selected belong to the hazmat label) and when it leads to a high recall (all pixels of the target are selected).

As can be seen in Figure 6, the proposed “SR-a*b*-SP” saliency calculation has a high performance close to an $F_\beta$-score of 0.9 (1.0 being the highest possible). The two other versions are less successful. The figure demonstrates also how the performance (of more or less all methods) decreases with increasing structuredness of the background. Most notably, “SR” fails for the brick wall background. The crucial factor may not be the bricks, though, but the

<table>
<thead>
<tr>
<th>Background</th>
<th>Bricks, coarse high-contrast</th>
<th>OSB, fine low-contrast</th>
<th>Woodchip wallpaper, very fine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Images &amp; masks</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Hazmat labels</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Light</td>
<td>Harsh sunlight</td>
<td>Incandescent light</td>
<td>Indirect daylight</td>
</tr>
</tbody>
</table>

⁴https://osf.io/b5dap/
Fig. 4. Three exemplary test images from the dataset. A Simple target “oxidizer” on relatively homogeneous woodchip wallpaper background B “Non-flammable gas” label on OSB wood wall. C “Spontaneously combustible” label on sunny outdoor brick wall.

Fig. 5. A Input image with harsh light. B Saliency from the spectral residual [14] of the input in grayscale, “SR”. C Combined spectral residuals of the a* and b* channels of the input in CIELAB color space, “SR-a*b*”. D Proposed saliency map enhanced with superpixels, “SR-a*b*-SP”.

harsh light and shadows, the consequences of which are also illustrated in Figure 5.

Overall, the “SR-a*b*-SP” approach shows promising results, producing many accurate candidates with the target well within the first attended superpixel (as in Figure 5), or at least in the second iteration, when the direct neighborhood is also included.

B. Overall Hazmat Recognition Performance

The evaluation of the overall procedure addresses two main questions: (1) Does the attentional preselection shorten the duration of recognizing a hazmat label and (2) if it does, can the high accuracy of SIFT be maintained? The speedup is expected because the time-consuming feature matching has only to be performed on a small part of the original image if the attentional selections target hazmat labels.

The attention stage, as evaluated in the previous section, takes roughly 150 ms, which is rather long considering that it may have to be invoked multiple times until a label is detected (see Figure 1). To mitigate this drawback, we run this stage with a low resolution of 92 × 64 pixels. At this resolution, the calculations take only approximately 15 ms, but the shape of the target area is still sufficiently well represented to produce (after upsampling to the input image resolution of 960 × 640 pixels) good candidates for the final SIFT-based recognition step.

The computation times we report in the following consist of the attention invocations, the unsuccessful recognition attempts before the target is found, and the final successful one. Typically one or two attempts are required. In case the target is not found even if the whole image was considered, the run is counted as a failure. Failures are rare and not considered in the runtime estimates. However, they are reflected in the accuracy scores. Recognizing a hazmat label as the wrong type happens even less often (only 4 times for class “flammable liquid” and 9 times for class “non-flammable gas”). Such false detections were treated as failures.

We first turn to the computation time estimations. Figure 7A–C shows the results summarized for the different backgrounds and labels4. The proposed method with attentional preselection is substantially faster, requiring only between 10 % and 70 % (depending on label type and background) of the processing time of the traditional SIFT approach. Moreover, the figure illustrates how computation time increases with the complexity of the background, which is due to the increasing number of keypoints in more complex backgrounds.

Concerning the accuracy (Figure 8A–C), the accuracies are improved by up to 50 percentage points. This is an interesting finding because the approach was intended to reduce the computation time, not necessarily to improve

4 Measured on an Intel Core i7-4770 PC at 3.40 GHz with 16 GB RAM.
SIFT’s accuracy which already is rather high. Most likely these unexpected accuracy gains result from the fact that the attentional preselection excludes many areas in which otherwise keypoints would be created. These increase the time in finding matches but also the risk of false positive detections. Hence, excluding such areas early not only speeds up the process but also increases the accuracy.

Figure 9 shows examples of the classification for different classes of hazmat labels and where the labels are localized in the image via the locations of matching SIFT keypoints.

Note that our method is conservative in the sense that it will eventually analyze the whole image if no matches are found on the preselected patches. However, the greatest gains in speed and accuracy result from cases where the attentional preselection provides the correct target in the first or second attempt. In our evaluation, the average number of attempts until target detection is 1.2 and does not vary much between the three different background types. Interestingly, the highest averages of attempts are not found for labels that have two parts in different colors, even though only one of these parts typically is included in the first selection (see two rightmost labels in Figure 2). The highest average (of 1.5 attempts) is found for the “non-flammable gas” label. It also has a relatively weak overall accuracy (see Figure 8), which indicates that the SIFT-based classification and not the attentional preselection leads to the weakness. Possibly, the simple structure of the symbol on the label provides no sufficiently distinct keypoint arrangements (see Figure 2, second label). Moreover, the false positives are classified as a flammable liquid. This originates from the word flammable in both classes. SIFT detects many keypoints on that word, leading to misclassifications.

In comparison to other approaches such as [11] and [2], the fast and accurate hazmat detection reported above could be achieved without a large number of training images. A single image from each class is sufficient. Moreover, our approach
is not limited to special hardware. By contrast, deep learning approaches typically require GPUs to achieve sufficiently fast computations.

V. CONCLUSIONS

In the present paper, we showed that SIFT can be applied in robust hazmat label recognition, but that the computation times are rather long. A novel attentional preselection can substantially speed up the process and also improve the accuracy in recognizing hazmat labels. The use of a combination of the spectral residual approach and a superpixel segmentation proved highly useful. The spectral residual saliency successfully localizes hazmat labels and the superpixels accurately capture their shape. In some classes, the overall accuracy is somewhat weaker than in the others (e.g., the “non-flammable gas” hazmat label; see Figure 8). Importantly, this does not mean that some instances of these labels are never recognized. In real scenarios they may be recognized as the robot changes its location, providing further views (active vision).

Neither the saliency computation nor SIFT are restricted to hazmat labels but are general mechanisms for object detection and recognition. This provides ample possibilities to extend the approach to further relevant static and possibly moving objects (e.g., localized via their optical flow [18]). Selecting candidates via attention is not limited to spectral residual saliency and SLIC superpixels, though we found these choices natural for reasons described in the introduction and successful according to our evaluation. However, future work can explore further saliency and attention approaches (see, e.g., [13] for an overview). In particular, top-down influences on attention (see, e.g., [19]) could guide attention to particular types of targets when the system is extended beyond hazmat labels. Moreover, further object recognition or classification techniques can be explored.

Beyond extending the approach to further objects and possibly improving the computational modules, the test dataset should be extended as well. We plan to include further labels, other relevant objects, more different backgrounds, recordings from varying distances, and further disturbances and defects that can be expected in real-world applications.

REFERENCES


