Optical Engineering 53(4), 043101 (April 2014)

Perceptual evaluation of color transformed multispectral imagery

Alexander Toet,[a,b,*] Michael J. de Jong,[a] Maarten A. Hogervorst,[a] and Ignace T. C. Hooge[b]
[a]Netherlands Organization for Applied Scientific Research (TNO), Kampweg 5, 3769 DE Soesterberg, The Netherlands
[b]Utrecht University, Helmholtz Institute, Experimental Psychology, Utrecht, The Netherlands

Abstract. Color remapping can give multispectral imagery a realistic appearance. We assessed the practical value of this technique in two observer experiments using monochrome intensified (II) and long-wave infrared (IR) imagery, and color daylight (REF) and fused multispectral (CF) imagery. First, we investigated the amount of detail observers perceive in a short timespan. REF and CF imagery yielded the highest precision and recall measures, while II and IR imagery yielded significantly lower values. This suggests that observers have more difficulty in extracting information from monochrome than from color imagery. Next, we measured eye fixations during free image exploration. Although the overall fixation behavior was similar across image modalities, the order in which certain details were fixated varied. Persons and vehicles were typically fixated first in REF, CF, and IR imagery, while they were fixated later in II imagery. In some cases, color remapping II imagery and fusion with IR imagery restored the fixation order of these image details. We conclude that color remapping can yield enhanced scene perception compared to conventional monochrome nighttime imagery, and may be deployed to tune multispectral image representations such that the resulting fixation behavior resembles the fixation behavior corresponding to daylight color imagery. [DOI: 10.1117/1.OE.53.4.043101]

Keywords: image fusion; color fusion; false color; color mapping; night vision; scene gist; gaze behavior.

Paper 140027P received Jan. 7, 2014; revised manuscript received Feb. 14, 2014; accepted for publication Mar. 4, 2014; published online Apr. 1, 2014.

1 Introduction

We recently introduced a real-time color transform that enables full color rendering of multispectral nighttime images, giving them a realistic and stable color appearance.[1] Although the resulting images have the intuitive appearance of daylight images, it has not yet been established whether human visual perception actually benefits from this color transform. The current study was performed to assess whether this new color transform can indeed enhance human visual scene recognition and understanding.

1.1 *Color Night Vision*

Night vision cameras are a vital source of information for a wide range of critical military and law enforcement applications such as surveillance, reconnaissance, intelligence gathering, and security.[2] Common nighttime imaging systems, cameras are low-light-level cameras, which amplify reflected visible to near-infrared (NIR) light, and thermal or long-wave infrared (LWIR) cameras, which convert the thermal energy into a visible image. Currently, monochrome display of night vision imagery is still the standard. However, monochrome night vision imagery often does not look natural, provides lower feature contrast, and tends to induce visual illusions and fatigue.[3,4] Moreover, the absence of color severely impairs scene recognition.[5] Intuitive color representations of night vision imagery may alleviate these problems.[6]

The increasing availability of multiband night vision systems has led to a growing interest in the color display of multispectral imagery.[6–10] The underlying assumption is that mapping the different spectral bands to a given color space can increase the dynamic range of a sensor system,[11] and can enhance the feature contrast and reduce the visual clutter, resulting in better human visual scene recognition, object detection, and depth perception. It has indeed been observed that appropriately designed false color rendering of nighttime multispectral imagery can significantly improve the observer performance and reaction times in tasks that involve scene segmentation and classification.[12–17] However, most color mappings do not achieve color constancy or produce color images with an unnatural appearance, thereby seriously degrading observer performance.[15,16,18] We recently introduced a simple color remapping technique that can give multispectral nighttime imagery an intuitive and stable color appearance.[1] This lookup table-based method (which will be described briefly in Sec. 2) is computationally efficient and can easily be deployed in real time.[19]

1.2 *Benefits of Color Imagery*

In principle, color imagery has several benefits over monochrome imagery for human inspection. Although the human eye can only distinguish about 100 shades of gray at any instant, it can discriminate several thousands of colors. By improving the feature contrast and reducing the visual clutter color may help the visual system to parse (complex) images both faster and more efficiently, achieving superior segmentation into separate, identifiable objects, and thereby aiding the semantic "tagging" of visual objects.[20] Color imagery may therefore yield a more complete and accurate mental

*Address all correspondence to: Alexander Toet, E-mail: lex.toet@tno.nl

representation of the perceived scene, resulting in better situational awareness. Scene understanding and recognition, reaction time, and object identification are indeed faster and more accurate with realistic and diagnostically (and also—though to a lesser extent—nondiagnostically[21]) colored imagery than with monochrome imagery.[20,22–25] Color also contributes to ultra-rapid scene categorization or gist perception[26–29] and drives overt visual attention.[30] It appears that color facilitates the processing of color diagnostic objects at the (higher) semantic level of visual processing,[25] while it facilitates the processing of noncolor diagnostic objects at the (lower) level of structural description.[23,31] Moreover, observers are able to selectively attend to task-relevant color targets and to ignore nontargets with a task-irrelevant color.[32–34] Hence, simply mapping multiple spectral bands into a three-dimensional (false) color space may already serve to increase the dynamic range of a sensor system,[11] and may thus provide immediate benefits like improved detection probability, reduced false alarm rates, reduced search times, and increased capability to detect camouflaged targets and to discriminate targets from decoys.[35,36]

1.3 *Color and Scene Gist*

Viewers can rapidly extract semantic information from a real-world scene within a single eye fixation, an ability known as "scene gist recognition."[26,27,37–41] An appreciable amount of meaningful information can be gleaned even from an extremely brief (e.g., 20 ms) glance at a scene, while asymptotic levels are already reached after about 100 ms.[26,37,40] The extracted information can be at different levels of abstraction,[42] ranging from a general description of the nature (e.g., urban, rural) or the emotional valence (pleasant, unpleasant) of the scene to an inventory of specific objects therein (e.g., a red car, a tree, a parking lot) and their spatial layout (a red car behind a tree on a parking lot).

Color contributes to gist perception either directly by providing an additional classification cue or indirectly by facilitating the initial segmentation of a scene.[26] As a result, diagnostically colored images are processed faster and more accurately than their grayscale version, followed by their nondiagnostically colored version.[27–29] These findings and the fact that the contribution of color to gist perception is robust for image degradation[26,29] suggest that the addition of color may also enhance the gist perception of multispectral imagery. Thus, gist perception may be a useful paradigm to quantify and optimize the effectiveness of color transforms for human visual perception of multispectral imagery.

In this study "scene gist recognition" is operationalized as the amount of meaningful (nameable) detail perceived during a brief image presentation (i.e., in a single glance). Color daylight photographs typically provide scene representations in which most details are clearly distinguishable and recognizable. Monochrome intensified (visual and near-infrared) imagery often has lower feature contrast and a somewhat unnatural appearance.[3] LWIR typically even has a less natural appearance since the thermal contrast in a scene is not strictly related to its visual contrast. Based on these considerations, we hypothesize (H1) that the gist of a scene will be optimally conveyed by daylight color photographs, followed in decreasing order by color fused imagery, intensified imagery, and infrared imagery. Hence, we expect that the amount of detail that is correctly reported after a brief image presentation will be largest for daylight color imagery, followed by color fused imagery, intensified imagery, and infrared imagery.

1.4 *Current Study*

In this study, we performed two observer experiments to investigate whether human visual scene perception indeed benefits from a realistic color representation of multispectral nighttime imagery. In the first experiment, observers were asked to give a detailed description of briefly presented night vision images in a free-recall paradigm. The night vision modalities tested were intensified visual (II), longwave infrared (IR), color fused multispectral (CF; produced with our new color mapping algorithm[1]), and digital daytime color photographs (REF). As stated before, we hypothesized that observers would be able to extract information from imagery in the REF and CF categories more accurately and with less effort than from imagery in the II or IR categories. We expected (H1) that imagery in the REF category would yield the most complete and most accurate image descriptions, followed by imagery in the CF, II, and IR categories (in that order). In the second experiment, we registered the fixation behavior of observers who freely explored imagery from each of the four categories investigated. Our second hypothesis was (H2) that fixation behavior for REF and CF imagery would be similar, since realistically colored multispectral nighttime imagery (CF) should be equally informative (and therefore drive fixation behavior in a similar way) as daytime color imagery (REF). Our third hypothesis was (H3) that fixation behavior for both REF and CF imagery would be different from fixation behavior for II and IR imagery, since observers probably have more difficulty in extracting information from grayscale nighttime imagery than from either colorized multispectral imagery or daytime photographs.

Until now only a few studies investigated to what extent scene gist recognition with colored (fused) multispectral nighttime imagery differs from performance with monochrome and unimodal sensor imagery.[43–45] Some studies investigated how well a second image matched a briefly presented first one for different combinations of (single band, fused, and colorized) image modalities.[44,45] In other studies, subjects were asked to report whether a scene contained an exemplar of a predetermined set of object categories.[43,46] The results of these studies suggest that fused and colorized multispectral imagery can indeed facilitate more accurate scene gist recognition. However, none of these studies investigated how detailed and accurate scene perception actually was for each of the tested image modalities. In contrast to those previous studies, the free-recall paradigm used to test gist perception in the current study involved no cuing of scene or target properties and merely involved the registration of unbiased responses from participants who watched brief image presentations.

Another way to study the potential benefits of color transformed representations of multispectral nighttime imagery is by registering human eye fixation behavior. The underlying assumption is that images displaying different information will induce different gaze strategies. The idea is that an observer obtains the gist of a given scene during an initial fixation, while low-level visual cues such as color,

brightness, and contours guide the observer's eye movements during further inspection of the scene. In this view, the observer's fixation behavior characterizes the level of informativeness of an image. It has, for instance, been suggested that low contrast (e.g., monochrome II and IR) imagery will induce longer saccade lengths and shorter fixation durations,[47] while more informative (e.g., REF or CF) imagery should induce more and longer lasting fixations on the part of the observer due to the higher saliency of informative details.[36] However, an increase in the number of uninformative details like clutter or noise may also increase the number of fixations and their duration. Moreover, eye movement data merely illustrate where and when an observer fixates a scene, but do not tell why more or longer fixations are needed in some cases (either because there is simply more detail to explore, or because the detail is harder to resolve?) and what information was actually extracted from the scene. Also, visual scan behavior is highly task dependent. As a result, the findings from studies that characterize the quality of (fused and/or colorized) multispectral imagery for human target detection and tracking through human fixation behavior are not unequivocal.[36,47,48] In contrast to these previous studies, we studied eye movements during free inspection of static images (to minimize the influence of task) without any *a priori* assumption about the nature of the effect of image informativeness. Thus, we hypothesize that differences in informativeness will induce different fixation behavior (H2 and H3), without specifying the exact nature of this effect.

The rest of this paper is as follows. In Sec. 2 (for the sake of completeness), we briefly present the color fusion algorithm. Then, we will present the method, results, and discussion of a scene gist experiment and an eye movement study, respectively, in Secs. 3 and 4. Next, we will discuss the overall findings of this study in Secs. 5 and 6; and finally, we will present our conclusions in Sec. 7.

2 Color Remapping

The color remapping technique that we recently developed is based on the assumption that there is a fixed relationship between false color tuples and natural color triplets for bands near the visual spectrum.[1] This allows its implementation as a simple color table swapping operation. For bands that are not correlated with the visual spectrum (e.g., LWIR), this assumption evidently does not apply. In that case, the color remapping can, for instance, be deployed to enhance the detectability of targets through contrast enhancement and color highlighting.[49]

Color remapping is achieved by mapping the multiband sensor signal to an indexed false color image and swapping its color table with that of a regular daylight color image of a similar scene (see Fig. 1). Different (e.g., urban, rural, maritime, or desert) environments may require specific color tables. However, in practice, we found that an entire environment is well represented by a single color table, as long as the environmental characteristics do not change too drastically.[50] Thus, only a limited number of color tables is required in practice. These tables need to be constructed only once, before the system is deployed.

For a given environment, the lookup-based color table transformation can be derived as follows. First, take a multispectral image of a scene that is typical for the intended operating theater and transform this image to an indexed color representation. Second, take an indexed representation of a regular color photograph of a similar (but not necessarily the same) scene. Then, there are two options:[6] either transfer the first-order statistics of the lookup table of the color photograph to the lookup table of the false color multispectral image,[51] or establish a direct mapping between corresponding entries in both tables.

When both a multispectral and a daylight color image are available of the same scene, a realistic color mapping can be obtained by establishing a direct relationship between the values of corresponding pixels in both images.[1] When there is only an indexed daylight color image available representing an environment similar to the one in which the multispectral sensor suite will be deployed, a mapping can still be established by transferring the color statistics of the daylight image to the multispectral image.[51] Although the first approach yields more specific colors, both approaches produce intuitively correct and stable color representations. Note that the statistical approach can even be used with imagery from sources like artificial terrain databases or Google Earth[52] (for demonstrations see

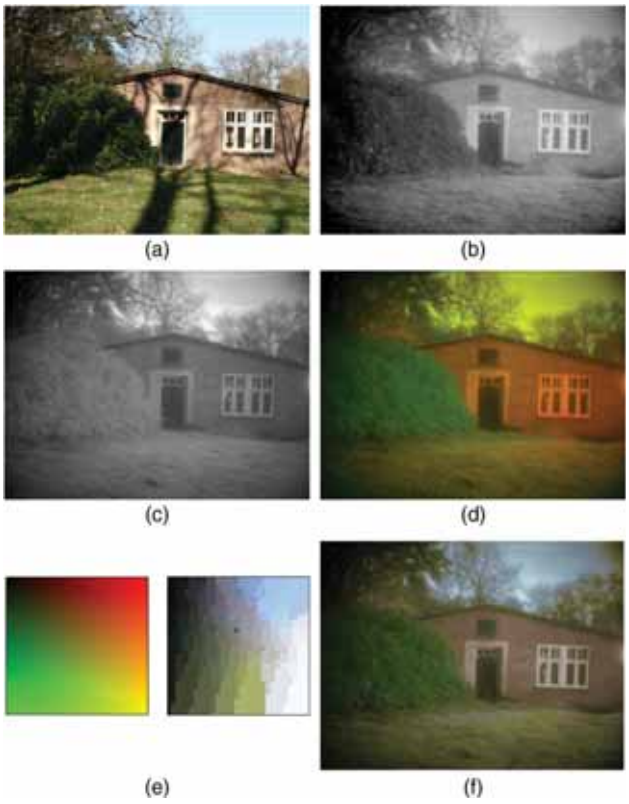

Fig. 1 Example of color remapping Gecko (dual band visual+near-infrared) imagery. (a) Daylight color reference image. Visible (b) and near-infrared (NIR) (c) nighttime images of the same scene provided by the Gecko system. (d) Intermediate RG false color representation of (b) and (c), obtained by assigning (b) to the green and (c) to the red channel of an RGB color image (the blue channel is set to zero). (e) Color mapping derived from corresponding pixel pairs in (a) and (d). (f) Output of the Gecko system, obtained by applying the mapping scheme in (d) to the intermediate two band false color image in (e). The list of key elements that served as ground truth for this particular scene consisted of the following four elements: {shrubs, trees, building, and grass}.

http://www.scivee.tv/node/29094 and http://www.scivee.tv/node/29095). The specificity of the lookup table-based color remapping also allows us to selectively enhance and emphasize certain details (e.g., camouflaged targets) in a given scene.[1,19,53]

For the sake of completeness, we will briefly describe our color transformation here, using the example shown in Fig. 1 (an extensive description is presented elsewhere[1]). Figure 1(a) depicts the full color daytime reference image, which is, in this case, a color photograph taken with a standard digital camera. Figures 1(b) and 1(c) show a visible and NIR image of the same scene, respectively. Figure 1(f) shows the result of applying daytime colors to the two-band nighttime sensor image using our new color mapping technique. The color transfer method works as follows. First, the multiband sensor image is transformed to a false-color image by taking the individual visual and NIR bands [Figs. 1(b) and 1(c), respectively] as input to the R and G channels, referred to as the RGB image [Fig. 1(d)]. In practice, any other combination of two channels can also be used (one could just as well use the combinations of R & B or B & R). Mapping the two bands to a false color RGB image allows us to use standard image conversion techniques, such as indexing.[54] In the next step, the resulting false color (RGB image), Fig. 1(d) is converted to an indexed image. Each pixel in such an image contains a single index. The index refers to an RGB value in a color lookup table (the number of entries can be chosen by the user). In the present example of a sensor image consisting of two bands [R and G; Fig. 1(d)], the color lookup table contains various combinations of R and G values (the B values are set to zero when the sensor or sensor pair provides only two bands). For each index representing a given R, G combination (a given false color), the corresponding realistic color equivalent is obtained by locating the pixels in the target image with this particular index and finding the corresponding pixels in the (realistic color) reference image [Fig. 1(a)]. First, the RGB values are converted to perceptually decorrelated $l\alpha\beta$ values.[55] Next, the average $l\alpha\beta$ vector is calculated over this ensemble of pixels. This assures that the computed average color reflects the perceptual average color. Averaging automatically takes the distribution of the pixels into account: colors that appear more frequently are attributed a greater weight. For instance, let us assume that we would like to derive the realistic color associated with color index i. In that case, we locate all pixels in the (indexed) false color multiband target image with color index i. We then collect all corresponding pixels (i.e., pixels with the same image coordinates) in the reference daytime color image, convert these to $l\alpha\beta$, and calculate the average $l\alpha\beta$ value of this set. Next, we transform the resulting average $l\alpha\beta$ value back to RGB. Finally, we assign this RGB value to index i of the new color lookup table. These steps are successively carried out for all color indices. This process yields a new color lookup table containing the realistic colors associated with the various multiband combinations in the false color (RGB) color lookup table. Replacing the RGB color lookup table [left side of Fig. 1(e)] by the realistic color lookup table [right side of Fig. 1(e)] yields an image with a realistic color appearance, in which the colors are optimized for this particular sample set [Fig. 1(f)].

3 Experiment I: Scene Gist Perception

In the first experiment, we investigated the amount of information observers can extract from a brief presentation of a scene (the gist of the scene) registered in different sensor modalities (REF, COL, II, IR). To enable the quantification and comparison of scene gist recognition, a scoring method is adopted that verifies the amount of extracted information against an inventory of the ground truth. As stated before, we hypothesized (H1) that the gist of a scene would be optimally conveyed by daylight color photographs, followed in decreasing order by color fused imagery, intensified imagery, and infrared imagery.

3.1 *Stimuli*

The stimuli were night vision images and daytime color photographs (called reference images, further indicated as REF) of 28 different urban and semirural scenes. The night vision modalities used in this study were standard (grayscale) intensified imagery (II), (grayscale) long-wave (8 to 12 μm) infrared imagery (IR), and color transformed fused multiband imagery (CF). The imagery used in this study was registered with the Gecko[56] system (providing II and CF imagery, where the CF imagery is obtained as color transformed fused visual and near-infrared imagery; see Fig. 1) and the Triclobs[57] system (providing II, IR, and CF imagery, where CF imagery is obtained as color transformed fused visual, near–infrared, and long-wave infrared imagery; see Fig. 2), respectively. CF imagery was obtained by applying the color mapping algorithm described in Sec. 2 to an intermediate false color fused image that was constructed by mapping the visual part of the II signals to the green channel, the NIR band of the II signals to the red channel, and the IR signals (when available) to the blue channel of an RGB color image. For Gecko imagery, the blue channel was set to zero since this system includes no IR sensor. A total of 14 scenes were represented in all four image modalities (i.e., there were

Fig. 2 Example of color remapping Triclobs imagery. TheTriclobs triband system splits incoming intensified imagery (II) into a visual and NIR band, fuses these signals with a long-wave infrared (IR) signal, and applies color remapping to the intermediate false color RGB image to obtain a realistic color setting (CF) that approaches the color distribution of a regular daylight photograph. The list of key elements that served as ground truth for this particular urban scene consisted of the following nine elements: {road, buildings, vehicles, trees, road sign, sand, grass, lamppost, and wall}.

14 different scenes represented by a full set {REF, CF, II, and IR}). For four scenes, there was no corresponding reference image available (i.e., there were four scenes represented by a partial set[58]), and for 10 scenes, there was no corresponding IR image available (i.e., there were 10 scenes represented by a partial set {REF, CF, and II}). Hence, there were 24 REF images, 28 II images, 18 IR images, and 28 CF images (a total of 98 images). When no exactly matching reference image was available, the lookup table transform used in the color remapping algorithm was derived from a scene that was highly similar to the given scene. The images were divided into four sets, such that each set contained only images of the same modality. All images had a resolution of 640×480 pixels.

3.2 *Ground Truth*

An exhaustive list of key elements was constructed for each of the 28 individual scenes. This list served as a ground truth inventory and enabled the scoring of the precision and recall measures for each individual image. The key elements were determined and described by the three experts who were familiar with the scenes (the authors), according to a procedure described elsewhere.[37] The key elements were described at the highest object level (e.g., house, tree, vehicle, road). A key element was scored as perceived by a participant when it was named in response or if any of its details were named (e.g., window, branch, pavement).

Since the visibility of scene elements can differ significantly between different image modalities, a single ground truth list was constructed for each individual scene, using all the available imagery (i.e., REF, CF, II, or IR) for that particular scene. Figures 1, 2, 3, 4, 5, 6, and 7 show some examples of typical scenes represented in the different image modalities that were investigated in this study, together with the corresponding list of key elements that served as the ground truth for these scenes.

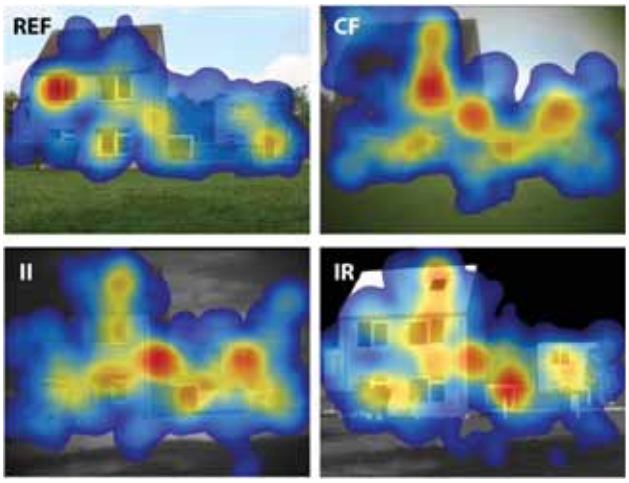

Fig. 3 Fixation distribution (for all observers) plotted as heat maps and superimposed on the corresponding scene. The list of key elements that served as ground truth for this particular scene consisted of the following nine elements: {houses, garage, chair, person, tree, fence, shrubs, grass, and road}.

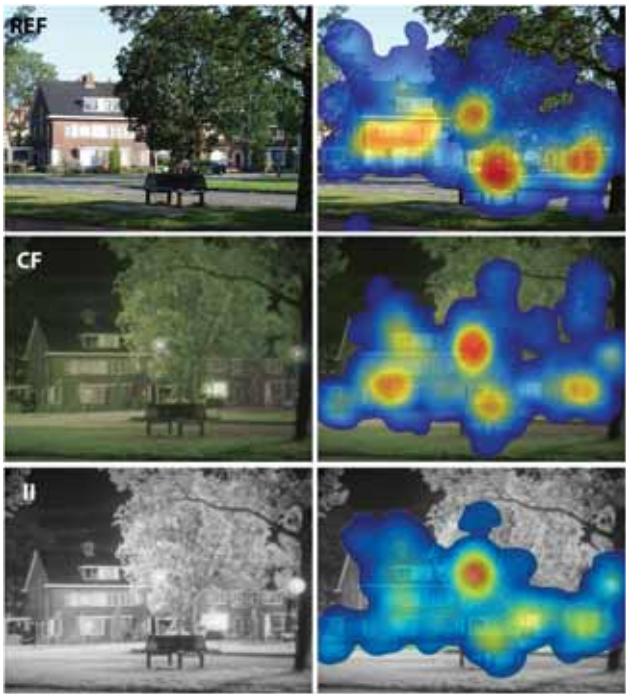

Fig. 4 After color remapping (the visual and NIR bands of) the II signal the corresponding heat map becomes more similar to the heat map of the REF image. The list of key elements that served as ground truth for this particular scene consisted of the following eight elements: {grass, trees, street bench, road sign, buildings, vehicles, bicycle path, and lamppost}.

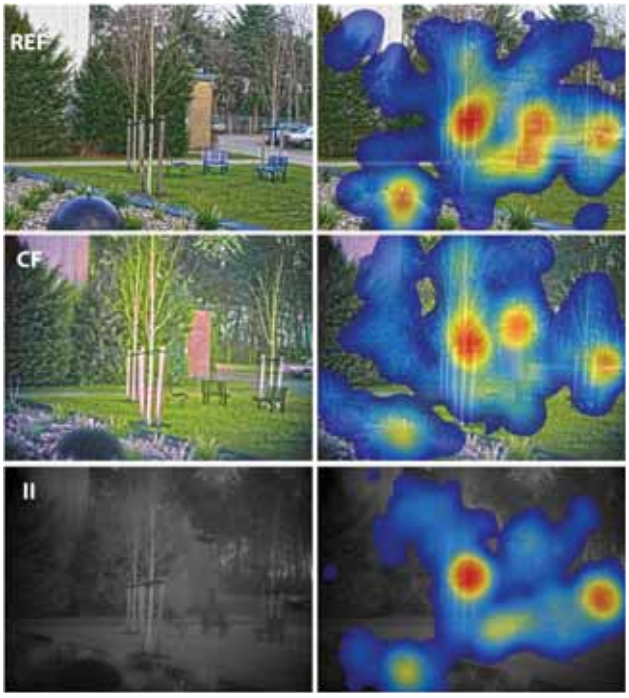

Fig. 5 As Fig. 4, the list of key elements that served as ground truth for this particular scene consisted of the following 11 elements: {grass, trees, ball, benches, vehicle, building, fence, gravel, plants, road, and props}.

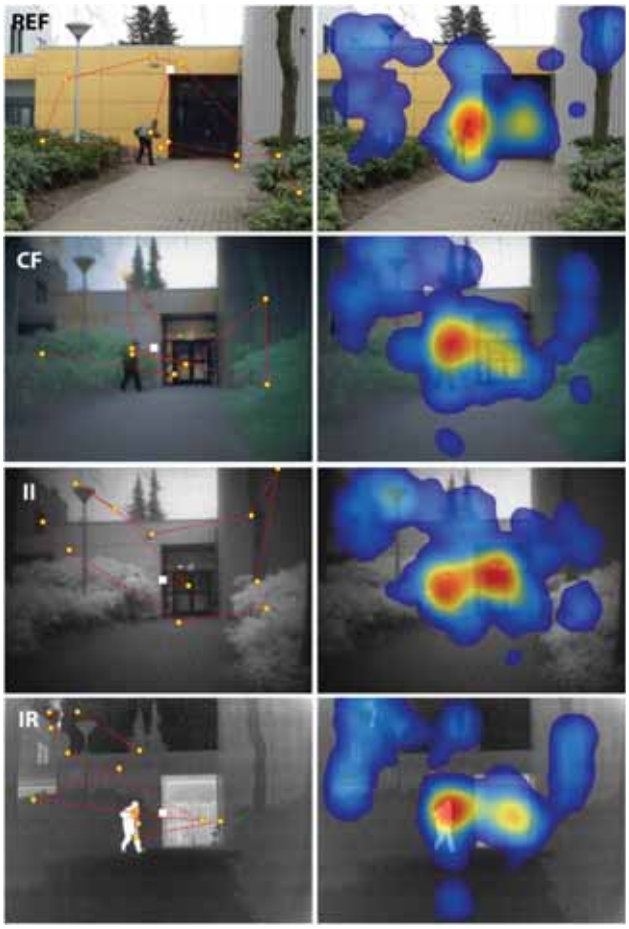

Fig. 6 Example of scanpaths (left column; for individual observers) and relative heat maps (right column; for all observers) for a scene with a person represented in each of the four different image modalities (REF, CF, II, and IR) tested in this study. The white square in the center represents the start of the scanpath (i.e., the first fixation). The list of key elements that served as ground truth for this particular scene consisted of the following six elements: {road, building, lamppost, person, shrubs, and trees}.

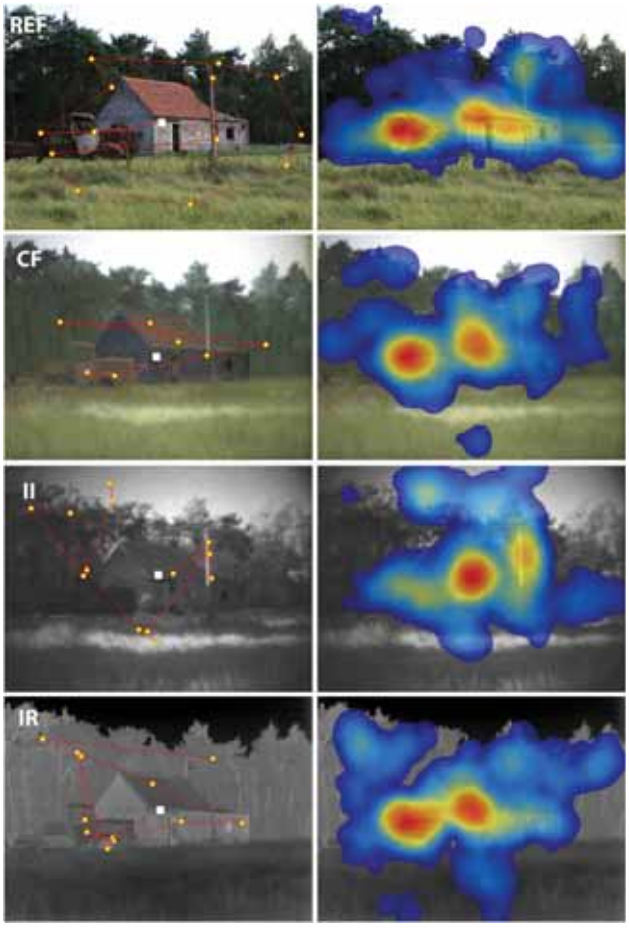

Fig. 7 As Fig. 6, for a scene with a vehicle. The list of key elements that served as ground truth for this particular scene consisted of the following six elements: {grass, trees, building, pillar, vehicle, and barbed wire}.

3.3 *Quantifying Gist Perception*

Since there is currently no generally accepted method to quantify scene gist perception, we will adopt the F_1 accuracy score[59] as our evaluation criterion. The F_1 score has become a standard metric in the field of information retrieval and pattern recognition to quantify search and classification performance. The F_1 score is defined as the harmonic mean of precision P and recall R, where precision is defined as the fraction of retrieved instances that are actually relevant, while recall is defined as the fraction of relevant instances that are actually retrieved. Let t_p, f_p, and f_n denote the number of true positives, false positives, and false negatives, respectively. Precision and recall are then given by

$$P = \frac{t_p}{t_p + f_p}, \tag{1}$$

$$R = \frac{t_p}{t_p + f_n}, \tag{2}$$

and the F_1 score is then given by

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}. \tag{3}$$

Note that P, R, and F_1 are all bounded between 0 and 1. In a precision-recall graph, higher F_1 scores correspond to points closer to $(P, R) = (1, 1)$. For a given scene, a larger F_1 value implies that both a larger number of relevant items are extracted and a larger number of the extracted items are actually relevant. Hence, the F_1 score is highly suitable to represent human observer performance in a scene gist perception task.

3.4 *Participants*

A total of 81 participants (40 males and 41 females, mean age = 31.2, SD = 12.3) performed the experiment. The participants were randomly assigned to one of the four groups based on stimulus type. As a result, 20 participants viewed 24 REF images, 20 participants viewed 28 CF images, 19 participants viewed 28 II images, and 22 participants viewed 18 IR images. All participants had self-reported normal or corrected to normal vision and no color deficiencies. Potential participants were randomly selected from the TNO participant database and sent an invitation via e-mail. When they indicated that they were

interested in participating in the study, they received instructions to download and run an application containing the study. All instructions, stimuli, and questionnaires were embedded in this application. Upon completion of the experiment, participants emailed their results (a text file generated automatically by the application) to the researchers. The instructions included a notice that users could end their participation in the experiment at any time without notifying the experimenter, and that all results would be processed anonymously. Participants were paid 10 Euro after completing the experiment.

3.5 *Procedure and Task*

The application started by presenting written information about the experimental procedure and the user instructions. The information stated that a number of test images would be presented in the course of the experiment, and the observer was asked to provide an accurate and detailed description of the content of each test image immediately following its presentation. To illustrate the purpose of the experiment and to familiarize, the participants with the type of images that would be presented, two example images (of the same modality but not part of the test set) were first shown, together with a written description of their content. To encourage the participants to provide a detailed account of their percept, the written descriptions accompanying the example images were quite extensive. The participants were further informed that each stimulus presentation interval would consist of a brief (500 ms) presentation of a test image preceded for 2 s by a black fixation cross in the middle of a white screen, and followed by a briefly (500 ms) presented random noise image. They were instructed to fixate the cross for 2 s until the test image appeared, and to type in a free-format accurate and detailed description of the test image in a text box that would appear immediately following the noise image. Finally, they were informed that the entire experiment would take about 30 min, and that they were free to withdraw from the experiment at any time. After reading the introduction, the participants could start the actual experiment by pointing and clicking on a button on the screen labeled "Next." The test images were shown on a white background. The observers were given unlimited time to enter their description of the test image in the text box. After typing in their stimulus description they could start the presentation of the next stimulus by pointing and clicking on a button on the screen labeled "Next." Figure 8 shows a graphic representation of the presentation procedure. At the end of the experiment, the application presented three demographic questions (gender, age, and education), thanked the observers for their participation in the experiment, and informed them that they would receive payment after returning the results file by email to the experimenters.

Note that a presentation time of 500 ms is commonly accepted as sufficient for perceiving a natural scene and most of its contents.[37,60–62]

3.6 *Results*

Each image description provided by the participants was evaluated by computing a precision and recall value. This was done by scoring the elements mentioned in the description returned by the participants against the corresponding ground truth inventory (the list of key elements) for the corresponding scene (according to a procedure described elsewhere[37]).

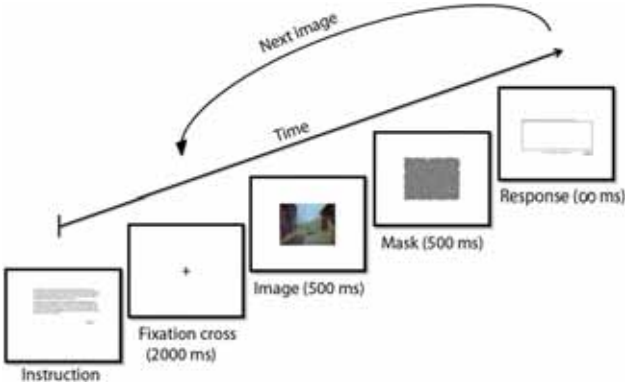

Fig. 8 Schematic representation of the experimental procedure. After reading the instructions participants started the trial sequence by pressing the space bar on the computer keyboard. Each trial consisted of the presentation of a fixation mark (2000 ms), a test image (500 ms), and a mask image (500 ms), followed by a response box (unlimited time).

A between factors analysis of variance (ANOVA) revealed a significant difference between the F_1 scores for the different types of imagery tested [$F(3, 77) = 34.03$, $p < 0.001$, $\eta_p^2 = 0.57$]. Bonferroni *post hoc* tests revealed that participants in the color fused (CF) image condition had a significantly higher F_1 values then participants in the intensified image (II) and infrared image (IR) conditions (both p values <0.001). Participants in the daylight color image (REF) condition did not differ significantly on the F_1 value from participants in the CF image condition ($p = 0.629$). Figure 9 shows the mean (over all observers and all scenes) F_1 scores for each of the four images modalities tested. The higher F_1 score in the CF condition means that participants were able to correctly identify more objects and had better scene recognition, suggesting that the gist of the scene was conveyed better by CF than by either II or IR imagery. The individual precision and recall measures can further clarify this result (Fig. 10). Participants in the CF condition scored significantly higher on both precision and recall than participants in the II and IR conditions. Note that participants in the CF condition yielded similar precision but higher recall scores as participants in the REF condition. This implies that CF imagery (1) clearly conveys the identity

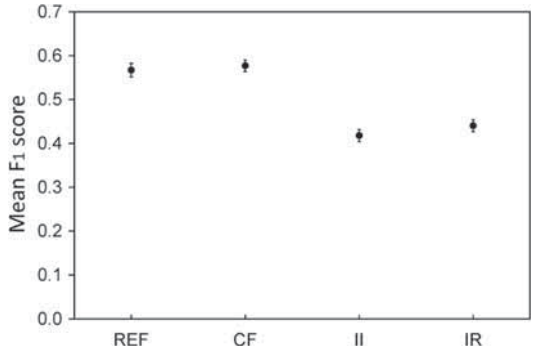

Fig. 9 Mean (over all observers and all scenes) F_1 scores for the four images modalities (CF, REF, II, and IR) investigated in this study. The error bars represent the standard error of the mean.

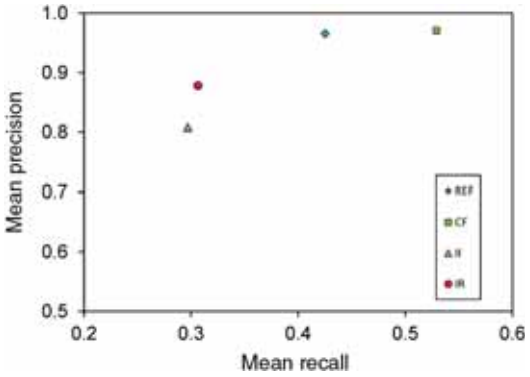

Fig. 10 Mean (over all observers and all scenes) precision and recall scores for the four images modalities (CF, REF, II, and IR) investigated in this study.

of relevant scene elements (leading to a similar precision for CF as for REF), while (2) CF imagery represents relevant scene elements as more distinct (i.e., with higher saliency). This is most likely a direct result of the increased informativeness of the CF imagers due to the fusion of multiple spectral bands. Participants in the intensified image and infrared condition on the other hand were less precise and made more mistakes in what they thought they perceived, resulting in a higher false positive rate. Persons and terrain features are almost always correctly recalled in both REF and CF imagery, while participants in the II condition often failed to detect persons.

To investigate if these results hold across scenes and observers, we also plotted the mean performance over all observers for each individual scene (Fig. 11) and the mean performance over all scenes for each individual observer (Fig. 12) in terms of precision and recall measures. Figure 11 shows that the mean precision and recall performance over all observers is typically higher and similar for REF and CF imagery, whereas it is typically lower for II and IR imagery. This means that the overall observer performance is consistent for the different image modalities. Figure 12 shows that the mean precision and recall performance over all images in a given modality is typically higher and similar for REF and CF imagery for most observers, whereas it is typically lower for II and IR imagery. This means that the performance for a given image modality is consistent across individual observers.

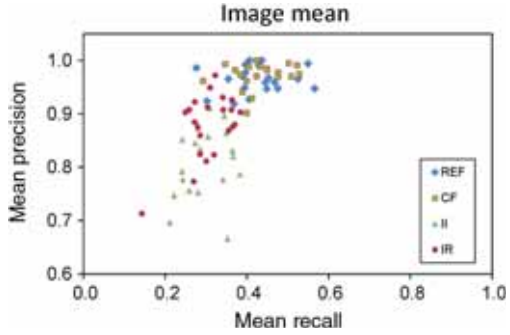

Fig. 11 Precision-recall values for the four image modalities investigated in this study (REF, II, IR, CF, and REF). Each data point represents the mean performance over all observers for a given image.

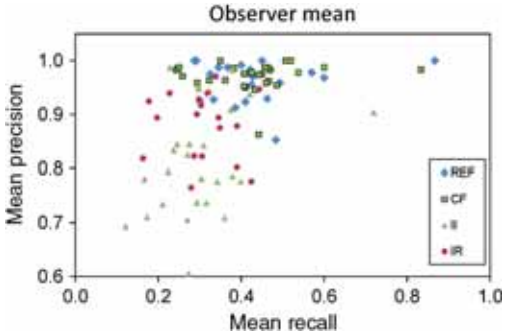

Fig. 12 Mean precision-recall values for the four image modalities investigated in this study (REF, II, IR, CF, and REF). Each data point represents the mean performance over all images in a given modality for a given observer.

3.7 *Discussion*

The present results confirm our hypothesis H1 that the gist of a scene is conveyed better (i.e., to a larger extent and more accurately) with color fused and the daylight color images then with infrared and intensified image images. Participants in the CF and REF condition recalled more details and were more precise than the participants in the IR and II conditions. This performance is consistent across scenes and observers. The fact there is no significant difference between gist perception with CF and REF images suggests that our multispectral color rendering conveys the gist of the scenes used in this study just as well as regular daylight color photographs.

4 Experiment II: Gaze Behavior

4.1 *Stimuli*

The stimuli used in this experiment were the same four sets of REF, CF, II, and IR images, respectively (98 images in total) that were also used in Experiment I (see Sec. 3.1). To enhance the resolution of the eye tracking data, the size of the original images (640×480 pixels) was enlarged with a factor 2 (1280×960 pixels) for display using bicubic interpolation and subsampling.

4.2 *Apparatus and Setup*

MATLAB (www.mathworks.com) with the Psychophysics Toolbox 3 (psychtoolbox.org) was used to display the images on a monitor (refresh rate: 60 Hz) and to record eye movement data. During the experiment, the participants were comfortably seated in front of the monitor. A chinrest was used to prevent head movement and to ensure that the participants remained at a fixed viewing distance of 56 cm from the monitor. An easyGaze eye tracker (designinteractive.net) was used to register eye fixations with a sampling frequency of 52 Hz. The declared system accuracy by the manufacturer is 0.5 deg, which, at a viewing distance of 56 cm, is equivalent to about 18 pixels. QuickGLANCE 6.5.0.3 software (www.eyetechaac.com) was used to calibrate and measure eye movements.

4.3 *Participants*

A total of 100 participants (39 males and 61 females, mean age = 21.4, SD = 2.5) performed the experiment.

All participants were students from Utrecht University (Utrecht, The Netherlands) and had self-reported normal or corrected to normal vision and no color deficiencies. The participants were randomly divided into four groups of 25 persons each. Each group viewed all images from only one of the four image modalities (i.e., either REF, CF, II or IR).

4.4 *Procedure and Task*

The experiment started with a nine-point calibration procedure to calibrate the eye tracker for each individual participant. Participants were then instructed (i) to fixate a cross that appeared in the center of the screen prior to each stimulus presentation, (ii) to start the next stimulus presentation by pressing the spacebar button on the computer keyboard, and (iii) to freely inspect the image that was presented between 0.5 and 1 s after pressing the spacebar and for at most 5 s (maximal presentation duration). To speed up the experiment, participants were able to continue to the next image (by pressing the spacebar again) when they had completely inspected the image. This procedure was repeated until the participant had inspected all images from a given set.

4.5 *Results*

Fixation duration, number of fixations, and number of fixations per second were analyzed for all scenes and image modalities using R software (www.r-project.org). There was no significant difference between the image modalities for the mean fixation duration [$F(3,96) = 0.520$, $p = 0.669$; see Fig. 13], the mean number of fixations [$F(3,96) = 0.150$, $p = 0.930$; see Fig. 14], or the mean number of fixations per second [$F(3,96) = 1.427$, $p = 0.240$; Fig. 15]. Hence, these data provide no support for both our hypotheses H2 (similar fixation behavior for REF and CF imagery) and H3 (fixation behavior for both REF and CF imagery would be different from fixation behavior for II and IR imagery). These results indicate that participants looked at the same number of elements for the same amount of time at roughly the same speed in all image modalities. This does, however, not mean that participants also looked at the same type of elements.

To investigate the type of details that were actually inspected by the observers, the fixation distributions (computed from the data of all observers) were visualized as heat maps and superimposed on the corresponding images

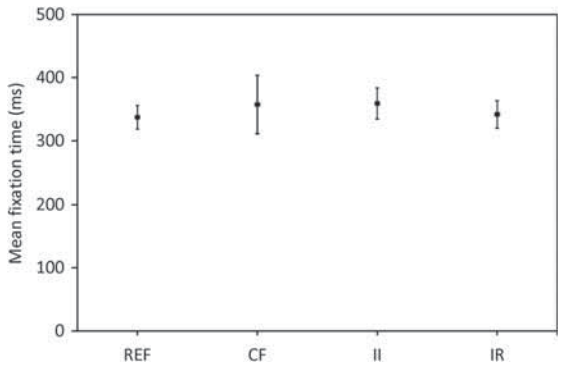

Fig. 13 Mean fixation time for each of the four tested image modalities. Error bars represent the 95% confidence intervals.

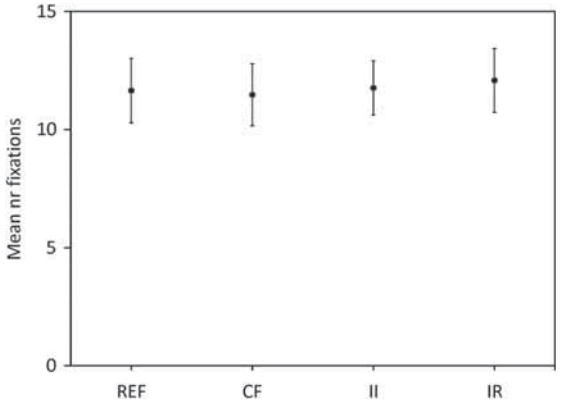

Fig. 14 Mean number of fixations for each of the four tested image modalities. Error bars represent the 95% confidence intervals.

using MATLAB (e.g., Fig. 3). On first sight, these heat maps appear quite similar for the different image modalities. On closer inspection, it appears that color remapping in some cases increases the similarity of human fixation behavior to that on REF imagery. For instance, Figs. 4 and 5 show that after color remapping of (the visual and NIR bands of) the II signal the heat maps of the REF images appear more similar to the heat maps of the CF images than to the heat maps of the II images.

It is difficult to accurately establish where the participants fixated due to the limited resolution of the fixation registration device and the inherent behavioral variation between observers. A solution for this problem is to define extended regions of interest (ROIs) and to compare viewing behavior for the same ROIs across different image modalities.[36] For each scene, we therefore defined several (between 1 and 6, median number of 4) rectangular ROIs, based on the heat maps and the list of key elements resulting from the gist experiment (see Sec. 3.2). These ROIs represented the scene elements that were both (1) fixated by most observers in this experiment and (2) that were also named most frequently by participants in Experiment I (e.g., Fig. 16). In addition, we also included ROIs representing scene areas that yielded notable differences in heat maps (based on visual inspection) across image modalities. The mean overall gaze duration was calculated for each ROI and analyzed using

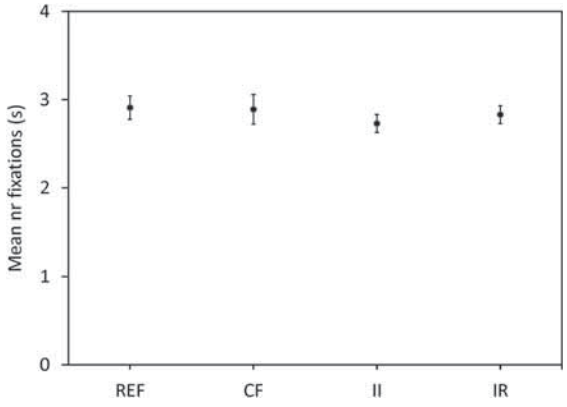

Fig. 15 Mean number of fixations per second or each of the four tested image modalities. Error bars represent the 95% confidence intervals.

Fig. 16 ROIs (outlined by red rectangles) for some of the scenes used in this study.

a between factors ANOVA. Since our main interest is whether the addition of color to multispectral imagery can enhance scene processing up to the level of regular daylight color imagery, we only analyzed fixations in the ROIs for the REF, CF, and II image modalities. The IR modality was not considered here since its addition does not contribute to natural color rendition. Mean overall gaze duration was not significantly different between the ROIs in the REF, CF, and II image modalities [$F(2,72) = 3.02,\ p = 0.055$ see Fig. 17]. Since our procedure to define the ROIs may have been biased, we repeated our analysis for arbitrary ROIs defined as the cells of a regular rectangular image grid, for grid sizes of 4×4, 6×6, and 8×8 pixels. Again, we found no significant difference between the mean overall gaze duration across the different image modalities.

To get a better notion of where—and in what order—participants fixated the scenes, we also compared individual scan paths between image modalities. It appears that the initial fixations in the REF, CF, and IR conditions are typically toward persons and vehicles (when present in the scene), whereas these details are typically inspected at a much later stage in the scan process in the II condition. This suggests that persons and vehicles were more salient in the REF, CF, and IR conditions than in the II condition.

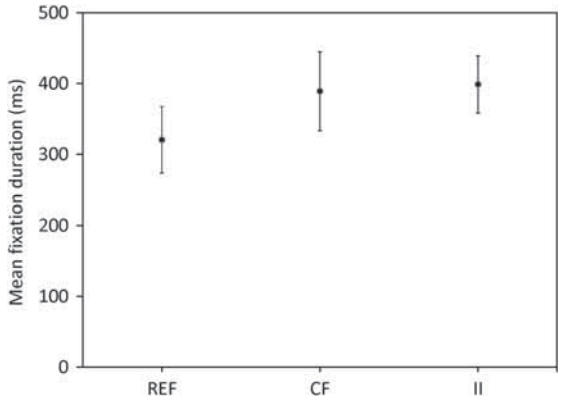

Fig. 17 Mean fixation time over all ROIs for each of the four tested image modalities. Error bars represent the 95% confidence intervals.

Figures 6 and 7 show some examples in which fusion of the visual and NIR bands of the II signal with an additional IR band helps to restore the fixation order of image details by boosting the saliency of details in the resulting CF image to their level of saliency in the REF image. The details that are fixated early and most frequently in the REF, CF, and IR images of the scene shown in Fig. 6 are the person and the doorway, respectively. In the II image, the person is also frequently fixated but only at a much later stage in the scan process. Out of 25 participants, the first fixation was at the person for 25 participants in the REF and IR conditions, for 21 participants in the CF condition, and only for four observers in the II condition. The person is quite salient (represented with high contrast) in the IR image, while he/she is hardly visible (has low contrast) in the II image. Fusion of the IR band with the visual and NIR bands of the II signal, followed by color remapping, boosts the contrast of the person in the resulting CF image so that it becomes equivalent to his contrast in the REF image, resulting in a similar fixation behavior as shown by the corresponding heat maps. A similar example is shown in Fig. 7 where the details that are fixated early and most frequently in the REF, CF, and IR images of the scene are, respectively, the vehicle, the door, and the flagpole. In the II image, the flagpole is frequently inspected because it is depicted at high contrast and therefore quite salient. Out of 25 participants, the first fixation was at the vehicle for 25 participants in the REF condition, for 22 participants in the CF condition, for 23 participants in the IR condition, and only for seven participants in the II condition. The vehicle is represented at high contrast in the IR image, while it has low contrast in the II image. After fusion, its contrast in the CF image is similar to that in the REF image, resulting in a similar fixation behavior as shown in the corresponding heat maps.

4.6 *Discussion*

The results of Experiment II indicate that participants looked at approximately the same number of elements for about the same amount of time at roughly the same speed in all four image modalities tested in this study (REF, CF, II, and IR). Although the overall fixation behavior was similar across image modalities, the fixation order showed some variations. In some cases, persons and vehicles were fixated first in REF, CF, and IR imagery, while these items were fixated at a later stage in the scanning of II imagery, indicating that persons and vehicles were less salient in II imagery. It appears that color remapping an II signal and fusion with an IR band may both serve to restore the fixation order of image details by boosting their saliency levels in the resulting CF image to the levels in the REF image.

5 General Discussion

From Experiment I, we found no significant difference between the perception of scene gist with CF and REF images. This suggests that our multispectral color rendering conveys the gist of the scenes used in this study just as well as regular daylight color photographs. Also, participants in the IR and II conditions recalled significantly less details and were less precise than participants in the CF and REF condition.

Experiment II showed that participants looked at the same amount of elements for the same amount of time at roughly

the same speed in all image modalities tested in this study. For most scenes, human fixation behavior did not differ between image modalities. In natural viewing conditions, eye movements are guided both (1) by the low-level stimulus features (e.g., a scene's luminance, color, edge distribution[63–65]) and (2) by subjectively or semantically informative regions (e.g., the viewer's task and goals and scene familiarity[63,66]). However, it has been found that even in neutral viewing tasks (watching images with neutral content in the absence of a particular task) cognitive factors override the low-level factors in fixation selection: regions with high semantic importance attract fixations regardless of their saliency.[67] As a result, fixation selection is robust for low-level image manipulations such as contrast modulation and color to grayscale conversion.[68] This may explain our current finding that fixation behavior is highly similar across different image modalities.

6 Future Research

Although the overall fixation behavior was similar across image modalities, the order of the fixated elements appears to differ between image modalities for some scenes. For instance, persons and vehicles were sometimes fixated earlier in REF, CF, and IR imagery than in II imagery. This is probably because these elements are often represented at low contrast in II imagery, while they are easily visible in REF and CF imagery and depicted at high contrast in IR imagery. The color transform method essentially establishes a statistical relationship between the daytime (reference) colors and the multiband output. Since there is typically no inherent correlation between the thermal channel and a daytime color image, the resulting colors (in the CF images) are determined by the visual and NIR input channels. The presence of a thermal band only serves to enhance the intensity contrast of hot targets (e.g., persons and vehicles) but does not change the overall color distribution. Thus, the saliency of vehicles and persons may improve due to enhanced luminance contrast when adding a thermal band. Color remapping an II signal and fusion with an IR band may therefore both serve to restore the fixation order of image details by boosting their saliency levels in the resulting CF image to the levels in the REF image. The limited number of images available for this study did not allow us to investigate this issue any further. We plan to investigate this phenomenon in a follow up study, by (1) registering a larger number of scenes both with and without hot targets (persons and vehicles) present and by (2) comparing human performance both for two- (visual and NIR) and three- (visual, NIR, and IR) band CF imagery.

7 Conclusions

Collecting and evaluating free-recall responses from participants watching brief image presentations appears to be an effective and efficient way to assess human scene gist perception. Using this technique, we established that the gist of a scene is conveyed better (i.e., to a larger extent and more accurately) with color transformed multispectral imagery than with monochrome infrared or intensified imagery. The order in which scene details are fixated may provide information on their (cognitive or task-relevant) saliency. We found that color remapping and fusion of multiple spectral bands can affect the fixation order of image details by altering their saliency levels in the resulting color fused image. In a future study, we plan to investigate if this finding can be used (1) to tune multispectral color fusion such that the resulting fixation behavior resembles the fixation behavior corresponding to daylight color imagery, or (2) to enhance the detectability and recognition of features of interest (e.g., camouflage breaking).

Acknowledgments

Effort sponsored by the Air Force Office of Scientific Research, Air Force Material Command, USAF, under Grant No. FA8655-11-1-3015. The U.S. government is authorized to reproduce and distribute reprints for governmental purpose notwithstanding any copyright notation thereon.

References

1. M. A. Hogervorst and A. Toet, "Fast natural color mapping for night-time imagery," *Inf. Fusion* **11**(2), 69–77 (2010).
2. R. S. Blum and Z. Liu, *Multi-Sensor Image Fusion and its Applications*, CRC Press, Taylor & Francis Group, Boca Raton, Florida (2006).
3. F. L. Kooi and A. Toet, "What's crucial in night vision goggle simulation?," *Proc. SPIE* **5802**, 37–46 (2005).
4. W. E. Berkley, "Night vision goggle illusions and visual training," in *Visual Problems in Night Operations, AGARD-LS-187*, Advisory Group for Aerospace Research & Development, pp. 9-1–9-6, North Atlantic Treaty Organization, Neuilly-sur-Seine Cedex, France (1992).
5. G. W. Stuart and P. K. Hughes, "Towards an understanding of the effect of night vision display imagery on scene recognition," *Ergon. Open J.* **2**, 150–158 (2009).
6. A. Toet and M. A. Hogervorst, "Progress in color night vision," *Opt. Eng.* **51**(1), 010901 (2012).
7. J. J. Christinal and T. J. Jebaseeli, "A survey on color image fusion for multi sensor night vision image," *J. Adv. Res. Comput. Eng. Technol.* **1**(9), 151–155 (2012).
8. Y. Zheng, W. Dong, and E. P. Blasch, "Qualitative and quantitative comparisons of multispectral night vision colorization techniques," *Opt. Eng.* **51**(8), 087004 (2012).
9. Y. Zheng, "An overview of night vision colorization techniques using multispectral images: from color fusion to color mapping," in *Proc. of the 2012 Int. Conf. on Audio, Language and Image Processing (ICALIP)*, pp. 134–143, IEEE, Piscataway, New Jersey (2012).
10. Y. Zheng, "An exploration of color fusion with multispectral images for night vision enhancement," in *Image Fusion and its Applications*, Y. Zheng, Ed., InTech Open, Rijeka, Croatia (2011).
11. R. G. Driggers et al., "Target detection threshold in noisy color imagery," *Proc. SPIE* **4372**, 162–169 (2001).
12. E. A. Essock et al., "Perceptual ability with real-world nighttime scenes: image-intensified, infrared, and fused-color imagery," *Hum. Factors* **41**(3), 438–452 (1999).
13. M. J. Sinai et al., "Psychophysical comparisons of single- and dual-band fused imagery," *Proc. SPIE* **3691**, 176–183 (1999).
14. A. Toet et al., "Fusion of visible and thermal imagery improves situational awareness," *Proc. SPIE* **3088**, 177–188 (1997).
15. A. Toet and J. K. IJspeert, "Perceptual evaluation of different image fusion schemes," *Proc. SPIE* **4380**, 427–435 (2001).
16. J. T. Varga, *Evaluation of operator performance using true color and artificial color in natural scene perception*, AD-A363036, Naval Postgraduate School, Monterey, California (1999).
17. B. L. White, *Evaluation of the impact of multispectral image fusion on human performance in global scene processing*, Naval Postgraduate School, Monterey, California (1998).
18. W. K. Krebs et al., "Beyond third generation: a sensor-fusion targeting FLIR pod for the F/A-18," *Proc. SPIE* **3376**, 129–140 (1998).
19. A. Toet and M. A. Hogervorst, "Real-time full color multiband night vision," in *Vision Sensors and Edge Detection, Real-Time Full Color Multiband Night Vision*, F. Gallegos-Funes, Ed., pp. 105–142, INTECHopen, Rijeka, Croatia (2010).
20. F. A. Wichmann, L. T. Sharpe, and K. R. Gegenfurtner, "The contributions of color to recognition memory for natural scenes," *J. Exp. Psychol.* **28**(3), 509–520 (2002).
21. I. Bramão et al., "The role of color information on object recognition: a review and meta-analysis," *Acta Psychol.* **138**(1), 244–253 (2011).
22. M. T. Sampson, *An assessment of the impact of fused monochrome and fused color night vision displays on reaction time and accuracy in target detection*, AD-A321226, Naval Postgraduate School, Monterey, California (1996).
23. I. Spence et al., "How color enhances visual memory for natural scenes," *Psychol. Sci.* **17**(1), 1–6 (2006).

24. K. R. Gegenfurtner and J. Rieger, "Sensory and cognitive contributions of color to the recognition of natural scenes," *Curr. Biol.* **10**(13), 805–808 (2000).
25. J. W. Tanaka and L. M. Presnell, "Color diagnosticity in object recognition," *Percept. Psychophys.* **61**(6), 1140–1153 (1999).
26. M. S. Castelhano and J. M. Henderson, "The influence of color on the perception of scene gist," *J. Exp. Psychol.* **34**(3), 660–675 (2008).
27. G. A. Rousselet, O. R. Joubert, and M. Fabre-Thorpe, "How long to get the 'gist' of real-world natural scenes?," *Visual Cognit.* **12**(6), 852–877 (2005).
28. V. Goffaux et al., "Diagnostic colours contribute to the early stages of scene categorization: behavioural and neurophysiological evidence," *Visual Cognit.* **12**(6), 878–892 (2005).
29. A. Oliva and P. G. Schyns, "Diagnostic colors mediate scene recognition," *Cognit. Psychol.* **41**(2), 176–210 (2000).
30. H.-P. Frey, C. Honey, and P. König, "What's color got to do with it? The influence of color on visual attention in different categories," *J. Vision* **8**(14), 6 (2008).
31. I. Bramão et al., "The influence of color information on the recognition of color diagnostic and noncolor diagnostic objects," *J. Gen. Psychol.* **138**(1), 49–65 (2010).
32. U. Ansorge, G. Horstmann, and E. Carbone, "Top-down contingent capture by color: evidence from RT distribution analyses in a manual choice reaction task," *Acta Psychol.* **120**(3), 243–266 (2005).
33. B. F. Green and L. K. Anderson, "Colour coding in a visual search task," *J. Exp. Psychol.* **51**(1), 19–24 (1956).
34. C. L. Folk and R. Remington, "Selectivity in distraction by irrelevant featural singletons: evidence for two forms of attentional capture," *J. Exp. Psychol.* **24**(3), 847–858 (1998).
35. S. Horn et al., "Monolithic multispectral FPA," in *Proc. of the Int. Military Sensing Symposium*, Paris, France, pp. 1–18, NATO Science and Technology Organization, Neuilly-sur-Seine, France (2002).
36. J. Lanir, M. Maltz, and S. R. Rotman, "Comparing multispectral image fusion methods for a target detection task," *Opt. Eng.* **46**(6), 066402 (2007).
37. L. Fei-Fei et al., "What do we perceive in a glance of a real-world scene?," *J. Vision* **7**(1–10), 1–29 (2007).
38. A. M. Larson and L. C. Loschky, "The contributions of central versus peripheral vision to scene gist recognition," *J. Vision* **9**(10), 6 (2009).
39. A. Oliva and A. Torralba, "Building the gist of a scene: the role of global image features in recognition," *Prog. Brain Res.* **155**(Part B), 23–36 (2006).
40. M. R. Greene and A. Oliva, "The briefest of glances: the time course of natural scene understanding," *Psychol. Sci.* **20**(4), 464–472 (2009).
41. A. Oliva, "Gist of a scene," in *Neurobiology of Attention*, L. Itti et al., Eds., pp. 251–256, Academic Press, San Diego, California (2005).
42. A. M. Larson et al., "The spatiotemporal dynamics of scene gist recognition," *J. Exp. Psychol.* (2013).
43. A. Toet and E. M. Franken, "Perceptual evaluation of different image fusion schemes," *Displays* **24**(1), 25–37 (2003).
44. W. K. Krebs and M. J. Sinai, "Psychophysical assessments of image-sensor fused imagery," *Hum. Factors* **44**(2), 257–271 (2002).
45. M. J. Sinai, J. S. McCarley, and W. K. Krebs, "Scene recognition with infra-red, low-light, and sensor-fused imagery," in *Proc. of the IRIS Specialty Groups on Passive Sensors*, pp. 1–9, IRIS, Monterey, California (1999).
46. E. A. Essock et al., "Human perceptual performance with nonliteral imagery: region recognition and texture-based segmentation," *J. Exp. Psychol.* **10**(2), 97–110 (2004).
47. W. K. Krebs, D. A. Scribner, and J. S. McCarley, "Comparing behavioral receiver operating characteristic curves to multidimensional matched filters," *Opt. Eng.* **40**(9), 1818–1826 (2001).
48. T. D. Dixon et al., "Task-based scanpath assessment of multi-sensor video fusion in complex scenarios," *Inf. Fusion* **11**(1), 51–65 (2010).
49. M. A. Hogervorst and A. Toet, "Evaluation of a color fused dual-band NVG," *Proc. SPIE* **7345**, 734502 (2009).
50. M. A. Hogervorst and A. Toet, "Presenting nighttime imagery in daytime colours," in *Proc. of the 11th Int. Conf. on Information Fusion*, pp. 706–713, International Society of Information Fusion, Cologne, Germany (2008).
51. A. Toet, "Natural colour mapping for multiband nightvision imagery," *Inf. Fusion* **4**(3), 155–166 (2003).
52. A. Toet et al., "Augmenting full color fused multiband night vision imagery with synthetic imagery for enhanced situational awareness," *Int. J. Image Data Fusion* **2**(4), 287–308 (2011).
53. M. A. Hogervorst et al., "Colour-the-INSight: combining a direct view rifle sight with fused intensified and thermal imagery," *Proc. SPIE* **8407**, 1–10 (2012).
54. P. Heckbert, "Color image quantization for frame buffer display," *Comput. Graph.* **16**(3), 297–307 (1982).
55. D. L. Ruderman, T. W. Cronin, and C.-C. Chiao, "Statistics of cone responses to natural images: implications for visual coding," *J. Opt. Soc. Am. A* **15**(8), 2036–2045 (1998).
56. A. Toet and M. A. Hogervorst, "Portable real-time color night vision," *Proc. SPIE* **6974**, 697402 (2008).
57. A. Toet and M. A. Hogervorst, "TRICLOBS portable triband lowlight color observation system," *Proc. SPIE* **7345**, 734503 (2009).
58. J. G. Ackenhusen, "Infrared/hyperspectral methods (paper II)," in *Alternatives for Landmine Detection*, J. MacDonald et al., Eds., pp. 111–125, Rand Corporation, Santa Monica, California (2003).
59. C. J. van Rijsbergen, *Information Retrieval*, 2nd ed., Butterworth-Heinemann, Newton, Massachusetts (1979).
60. I. Biederman, R. C. Teitelbaum, and R. J. Mezzanotte, "Scene perception: a failure to find a benefit from prior expectancy or familiarity," *J. Exp. Psychol.* **9**(3), 411–429 (1983).
61. M. C. Potter, "Short-term conceptual memory for pictures," *J. Exp. Psychol.* **2**(5), 509–522 (1976).
62. S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature* **381**(6582), 520–522 (1996).
63. J. M. Henderson et al., "Visual saliency does not account for eye movements during visual search in real-world scenes," in *Eye Movements: A Window on the Mind and Brain*, R. P. G. van Gompel et al., Eds., pp. 537–562, Elsevier, Amsterdam, The Netherlands (2007).
64. D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Res.* **42**(1), 107–123 (2002).
65. B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," *Vision Res.* **45**(5), 643–659 (2005).
66. M. S. Castelhano, M. L. Mack, and J. M. Henderson, "Viewing task influences eye movement control during active scene perception," *J. Vision* **9**(3), 1–15 (2009).
67. M. Nyström and K. Holmquist, "Semantic override of low-level features in image viewing—both initially and overall," *J. Eye Mov. Res.* **2**(2), 1–11 (2008).
68. G. Harding and M. Bloj, "Real and predicted influence of image manipulations on eye movements during scene recognition," *J. Vision* **10**(2), 1–17 (2010).

Alexander Toet is currently a senior research scientist at TNO (Soesterberg, The Netherlands), where he investigates multimodal image fusion, image quality, computational models of human visual search and detection, the quantification of visual target distinctness and cross-modal perceptual interactions between the visual, auditory, olfactory, and tactile senses. He is a fellow of SPIE and a senior member of IEEE.

Michael J. de Jong received his MSc in applied cognitive psychology from the Utrecht University in The Netherlands, in 2013. His research interests include visual information processing and human fixation behavior.

Maarten A. Hogervorst is employed at TNO as a research scientist. His current research interests include visual information processing, electro-optical system performance, search and target acquisition modeling, image quality assessment, image enhancement, information fusion, color imaging, EEG, and human factors of camera surveillance systems. He is a senior member of SPIE.

Ignace T. C. Hooge is an associate professor in experimental psychology at the Utrecht University, The Netherlands. His research topics involve oculomotor control, visual search, and applied topics with eye tracking.