% Comments in black

% Response in red

% Added text in green

Reviewer #1 (Comments for the Author (Required)):

Authors assessed computational luminance constancy with AMA algorithm, with naturalistic images generated by computer graphics tools. It was interesting approach. However, some of critical information to understand the approach seemed to be missing or less comprehensive. It would be great if authors could address those issues.

Thank you for the careful reading and helpful comments. Please see below for how we have clarified in response.

Luminance constancy was mentioned as "constitutive component of ... general color constancy". However, the definition was not formally given. Authors could have provided the background, any of their specific definition, logic, concept and any assumptions in more details in Introduction.

Thanks for the suggestion. In the original submission, we provided a definition of the computational problem of luminance constancy in the last paragraph of the introduction (Line 55). We have now added a parallel definition of the more general color constancy case at the start of the second paragraph of the introduction (Line 26):

“The computational problem of color constancy may be framed as how to obtain stable descriptions of the spectral surface reflectance functions of the objects in a scene.”

(p.3, para 4) "We define the computational problem of luminance constancy as that of estimating the light reflectance value (LRV) of a target object's surface reflectance function. Estimating the LRV from a surface reflectance function proceeds in two steps. First, one computes the luminance of the light that would be reflected from the surface under a reference illuminant. Second, one normalizes the result by the luminance of the reference illuminant itself."

A problem of conventional color constancy is that we do not know the surface reflectances. However, authors' approach seemed that they assume that the reflectances are already available to viewers, independent of illuminants. Those assumptions, if any, and the links to the computations, estimation of the task-optimal receptive fields with cone-excitations and their normalizations, could have been more clearly explained.

Correct. Human and computational observers do not generally have information about surface reflectance when viewing novel scenes. Our work only makes use of groundtruth information about surface reflectance in the construction and design of our computational observer. When we test the observer’s ability to estimate LRV on images of novel scenes, it does not have any more information about the surface reflectance than would in principle be available to a human observer viewing the same scenes. When evaluating performance, the observer never has direct access to any quantity other than the cone-responses.

The passage quoted by the reviewer describes how LRV is defined, rather than how our computational observer estimates LRV. To prevent confusion, we have changed the third sentence in the paragraph from: “Estimating the LRV from a surface reflectance function proceeds in two steps.” to (Line 57):

“Defining the LRV from a surface reflectance function requires two steps.”

Without those explanations, it is difficult to follow the computations and their results.

e.g. "...datasets to determine how well target object LRV can be estimated from cone excitations and from normalized cone contrasts. Studying both representations allows us to understand how early contrast coding and normalization affect luminance constancy. We applied accuracy maximization analysis (AMA) to learn the optimal receptive fields for estimating LRV, and evaluated the performance obtained when the responses of these receptive fields are optimally decoded."

We hope our response to the previous comment addresses the issue.

Authors introduced the concept of the light reflectance values (LRV) as a "specific problem of luminance constancy, as constitutive component of the more general color constancy problem". However, those "problems" were not well identified in Introduction.

Please see above.

The relationship of the LRV in a physical world could be clarified.

We are not sure we understand what the reviewer is after here. We have defined LRV in the last paragraph of the introduction.

Figure 1 could have improved and to be used to explain the LRV and "object-extrinsic factors".

In Introduction, the property of LRV seemed to be part of physical properties. However, the LRV was one of the parameter in the computation, as if it is one of the internal properties (within visual system).

The LRV is not a parameter in the estimation process. It is the physical property that the computational observer is tasked with estimating. JDB: SHOULD WE HAVE A FIGURE THAT SCHEMATIZES HOW TO COMPUTE LRV?

One of the authors has publications about the illumination geometry and its importance. Mutual reflections, shadow, specularity and multiple illuminations are also important in color constancy. Authors could have commented how these properties were considered in the present model.

Thanks for the suggestion. We now specifically indicate that (Line 87):

“All surfaces in the scene model were matte and did not have specularities.”

We did, however, examine the impact of secondary reflections. They had a minor effect. Please see the last paragraph of the results section.

(p.14, last para) "the secondary reflections have minimal effect on LRV estimation: the estimates without secondary reflections were similar to those with reflections."

Does this mean that the computational luminance constancy with AMA cannot address the mutual reflection or the mutual reflection has no effect on the constancy?

We have reworded the paragraph to increase clarity. It now reads (Line 276):

“Our rendering software allows us to compare the effect of background surface reflectances on target object LRV with and without simulation of secondary reflections of light from one object onto another. These secondary reflections were included in the dataset from which we report our primary results. When we turn off this feature of the rendering, we find (data not shown) that LRV estimation performance is essentially unchanged. Estimates with and without secondary reflections are very similar. This result suggests that the primary source of the estimation error in Condition 3 is caused by image-to-image variation in the reflectance of the background objects.”

We hope this addresses the issue.

Performance of luminance constancy was discussed briefly in Discussion with RMSE. The definition of the relative RMSE and how it could evaluate the luminance constancy was not given in the main text.

We have added the definition of “relative RMSE” at the end of the METHODS section. It now reads (Line 209):

“**Quantifying estimation performance**

We quantified the performance of AMA and the baseline methods at estimating LRV through relative root mean squared error (relative RMSE). Relative RMSE is the square root of the mean of the squared difference between the estimated and true LRV divided by the true LRV. The mean is taken over all stimuli in the test set.”

As for technical matters, there seemed to be any restrictions in using the AMA. Such disadvantages of the computations adopted in the present study could be identified, as well as the advantages. The experimental setting up and parameters could have been explained in more details. An essential point could be what the task or optimization criterion in the AMA was.

Sorry for the confusion. The task was to estimate LRV. This was indicated multiple times throughout the original submission. Information about the AMA cost function was included in the original submission in the second paragraph of the Methods subsection titled: “Learning optimal receptive fields”. We wrote: (Line 178) “In our implementation of AMA, we used both the Kullback-Leibler divergence cost function (corresponding to the maximum a posteriori estimator) and the mean squared error cost function (corresponding to the posterior mean estimator) and assumed that receptive field responses were corrupted by scaled Gaussian noise (i.e. Poisson-like noise with a fano factor of 1.3 (Geisler & Albrecht, 1997)). Training with both cost functions yielded similar estimation performance; the results reported here are for the Kullback-Leibler divergence cost function.”

The spatial resolution of the image data and area seemed to be very small: e.g. "the target by cropping the rendered images to 1 x 1 degrees of visual angle around the target object (51 x 51 pixels)". Authors could provide justifications whether these sizes are large enough to evaluate the effect of LRV, interaction of the geometry of the object surfaces and multiple illuminations.

The choice of a 1x1º analysis area was informed by data on the size of receptive fields in early visual cortex (Gattass et al, 1981; Gattass et al, 1988). Our thinking was that the selectivity of neurons in early visual cortex are better aligned with the perceptually relevant luminance (L+M), red-green (L-M), and blue-yellow (L+M-S) directions of color space (Horwitz & Hass, 2012), than neurons earlier in the visual pathway (e.g. LGN relay cells or retinal ganglion cells).

VS: The cropped area is shown in figure 9b. The target covers about 1/4 th of the area (655 pixels out of 51\*51 pixels). It has several other objects, at multiple 3D distances.

Justification of “size big enough for studying interaction”??

It is unclear how 51 pixels correspond to 1 degree in visual angle. Thus, the parameters for the simulations were not fully explained; thus, corresponding physical size, distance, direction of light sources, intensity,

JB: Not sure about the best response here.

Definition of "naturalistic" should be given. For example, a sphere or Xylophone in the air is not seen in everyday life. The simulated images were based on the indoor structure, but authors applied outdoor illuminants. The simulations of the illuminants were based on the Granada natural illuminants. Thus, despite using the "natural" dataset, those were decomposed and fitted with linear combination of Gaussians. This may sounds as if "natural" data was transformed to "unnatural".

JB: Not sure about the best response here. Maybe just go with Vijay’s? =]

Some terms and acronym should be defined and explained at the first appearance.

e.g. LRV and the definition of the "relative RMSE" were given in the Fig 14 caption.

The LRV acronym is defined in the abstract (Line 6) and the introduction (Line 55).

LRV is defined in the introduction (Line 56).

We have added the definition of “relative RMSE” at the end of the METHODS section. It now reads (Line 209):

Quantifying estimation performance

We quantified the performance of AMA and the baseline methods at estimating LRV through relative root mean squared error (relative RMSE). Relative RMSE is the square root of the mean of the squared difference between the estimated and true LRV divided by the true LRV. The mean is taken over all stimuli in the test set.

[Methods]

(p.5, second from the last para) "The LRV values were equally spaced between 0.2 and 0.6. For each LRV value, we generated a different relative target object surface reflectance for each scene."

The range of LRV was [0.2 0.6]. What was the meaning of this range? What is the meaning of the LRV 0 and 1?

The range [0.2 0.6] was chosen because the LRV of most (>90%) of the generated surface spectra fell within this range. We have now mentioned this in the text (Line 100):

More than 90\% of the surface reflectance spectra (generated as described below) fell within this range.

The meaning of LRV 0 and 1 is now provided in the text (Line 61):

An LRV of 0.0 means that none of the light from the reference illuminant is reflected from the surface. An LRV of 1.0 means that all of the light from the reference illuminant is reflected from the surface.

(p.5, last para) "The Library base scene contains 2 area lights. We inserted one additional spherical light source into the scene. The position and size of the inserted object, the inserted light source, and the viewpoint on the scene were held fixed across all scenes. "

What was the rationale to use the two area lights?

How these multiple lights were manipulated in the Conditions 2 and 3?

We have added the following to explain the manipulation of the light spectra for Condition 2 and 3 (Line 103).

In condition 2 and 3, the overall intensity of the three light source illumination spectra were equal, while their relative shape varied. The overall intensity varied from scene to scene.

(p.4, para 1) "The package builds..."

It would be useful for readers if authors could inform the system requirements and any practical restriction you may be aware of.

The github repository provides such details.

[Baseline methods]

Why was the 3 x 3 pixels region used? Was it center of the 51 x 51 pixels?

L:M:S ratio was 6:3:1. Does this mean that it was possible that no S-cone was included depending on the area?

We used a demosaiced version of the cone-responses, so all 3 types of cones were present in the analysis. This is mentioned in Line 154.

Figure 10

(b) What is the meaning of the negative values on x-axis?

(c) What was the spatial dimension of the RF?

Were the computations of the RF independent across L, M, and S?

Did they have the same spatial size?

(b) This is the RF response compared to a baseline

(c) Same as the cropped region.

No.

Yes.

[Typos]

p.2, last sentence: "these factors (?, ?; Brainard...)"

We have corrected this.

Figure 10 (a): no "filled region" in the panel.

For this case, the filled region is too small to be visible. This is mentioned in the caption.

” The standard deviations are too small to be visible.”

Reviewer #2 (Comments for the Author (Required)):

The authors investigate how differences in object relative reflectance spectrum (i.e. color but not albedo), illumination spectrum and reflectance spectrum of the background affects performance of an optimal decoder in predicting a measure of the surface reflectance (light reflectance value - LRV, which is the reflected luminance under a reference illuminant normalized by the luminance of the illuminant itself - thus, being conceptually similar to albedo).

Specifically, as a first step they generate a large set of rendered naturalistic images systematic varying LRV. In addition to LRV changes, reflectance spectra of the target images, of the illumination and of the background surfaces was varied, according to three conditions. 1) The relative reflectance spectrum of the surfaces was varied while keeping the illumination spectrum and the background reflectance constant, 2) the reflectance and the illumination spectra were varied while keeping the background constant, and 3) all the three factors varied. From condition 1 to 3 estimating LRV from the pixel images is a harder problem because of the additional confounding variations.

As a second step, they used a model of the early visual system to mimic the optical blurring of the eye and the spatial sampling of the three classes of cones. The simulated cone excitations in response to the pixels at the corresponding sampled positions were transformed into images by demosaicing via linear interpolation. Then, the three L M S excitation images were normalized to equate the response magnitude across classes. Cone contrast images were computed from the normalized cone excitations, and both excitation and contrast images were separately used in the analyses.

As a third step, the authors used accuracy maximization analysis (AMA) to determine a set of linear filters (weighting functions applied to the L M S images) chosen to best classify LRV. The AMA searched over the space of linear filters to find the ones that minimize a given cost function. These linear filters are the optimal receptive fields for decoding LRV.

As a final step, they tested how well the responses of these optimal receptive fields can be decoded to estimate LRV. As a baseline, they used predictions form a linear regression fit of LRV as a function of the cone excitation and contrast from a central region of the target images. The receptive fields and the regression coefficients were estimated on the 90% of the images and tested on the remaining 10%.

In condition 1 performance of both AMA and linear regression was close to perfect, based on cone excitation. This is not surprising because only LRV is changing, yielding to a monotonic (linear) increase of cone excitation. In fact, receptive fields are characterized by random weights in the background regions and high weights corresponding to the target regions. This is true for the L and M images, but receptive field applied to S excitation images present a random distribution of small weights, indicating poor contribution of S cones. This is interesting because cone excitation were normalized before the analyses.

In condition 2, based on cone excitation AMA performance was rather poor, reflecting the additional complexity introduced by changing illumination spectrum (thus affecting luminance of the target object). Regression performance was as bad as guessing the mean LRV of the training set. When based on cone contrast, AMA performance was again nearly perfect, and regression improved, presumably because background luminance information is implicit in the images because of the normalization procedure. The shape of the receptive fields was not reported.

In condition 3, performance was only evaluated based on cone contrast images. Both AMA and linear regression performed worse than in condition 2 (AMA performed better than linear regression), reflecting the increased complexity, however they provided information about LRV. The shape of the optimal receptive fields revealed systematic contribution of cone excitations from background object locations, with positive and negative relatively high weights from different background regions, presumably because of the correspondence to background objects.

I think the approach presented in the manuscript is interesting because 1) simulations through physically accurate renderings allows to generate databases large enough for statistical learning, and 2) AMA allows to assess what information is relevant, by investing how manipulation affects performance, but also by looking at the structure of the receptive fields. Also, the structure of the receptive fields, will depend on the geometry of the scene (since it was held fixed), thus it can tell where relevant information is (e.g. in the background objects).

Thank you for this concise summary and positive evaluation of our work.

In only recommend to improve clarity. In particular, I think confusion is made between definitions of constancy. I think the use of "luminance constancy" is at least misleading if not a contradiction in terms. The authors refer to a normalized luminance measure (LRV), which is close to albedo. I think this needs to be made clear from the title. Also, I do not understand why LRV is chosen rather than albedo, since also albedo can be computed based on luminance and to my knowledge it is a more common magnitude in perception research and computer graphics. I think this choice needs to be commented. If as I think, LRV conceptually corresponds to albedo, I recommend changing "luminance constancy" with "lightness constancy", since lightness is commonly referred to as the perceptual correlate of surface albedo.

This is a reasonable point, and the question of terminology is one we grappled with as we wrote the initial draft. In the literature, "lightness constancy" generally refers to studies where the stimuli are restricted by be achromatic. Similarly, "albedo" is a concept that applies in the case where there is no spectral variation in the stimuli. These conditions do not apply to our work – we consider full spectral variation in the stimuli. Our restriction to a special case occurs later in the development, as we only attempt to estimate a scalar valued function of object surface reflectance. Thus, we feel it is not appropriate for us to use the lightness/albedo terminology. Because measure, the LRV, is the luminance of the light that would be reflected from an object under a standardized reference illuminant, we adopted the terminology "luminance constancy."

That said, we agree that we did not sufficiently explain our thinking. We have now added a footnote to make these points explicitly where we introduce luminance constancy. The footnote says (Page 2 bottom):

“We distinguish luminance constancy and LRV from lightness constancy and albedo, respectively, as the latter terms refer to achromatic stimuli.”

Minor comments:

Reference to previous work might be extended, especially concerning research in perception. In fact, although in my understanding, the presented work is about lightness constancy, there is no definition of lightness and it is not clear what are the factors involved in lightness constancy. For a definition of lightness and brightness I recommend referring to "Lightness Perception and Lightness Illusions"- Adelson, 2000. For the factors contributing to lightness constancy I suggest "Seeing black and white" - Gilchrist, 2006.

Also, there is a certain body of work on which scenes aspects potential cues for lightness (e.g "Cues to an Equivalent Lighting Model" Boyaci, Doerschner & Maloney, 2006; "Illumination estimation in three-dimensional scenes with and without specular cues" - Snyder, Doerschner & Maloney). Specular reflections are one of those cues. However, there are human and simulation studies reporting that specular highlight are discounted in lightness judgments and that specular reflections potentially impair lightness discrimination (e.g. "Lightness constancy in the presence of specular highlights" - Todd, Normal & Mingolla, 2004; "Lightness perception for matte and glossy complex shapes", Toscani, Valsecchi & Gegenfurtner, 2017; "The effect of gloss on perceived lightness" - Beck, 1964 ).

We have now added citations to the bodies of work listed above, so as to better connect to the literature. [David to look for where to do this.]

I think that the approach presented in the manuscript might help investigating the role of specular highlights for an ideal observers. In fact, with a fixed geometry of the scene and the illumination (as it was in the reported simulations) the distribution of the weights in the receptive fields is informative about the role of the elements in the scene. Given the interest that specular reflection received by color and lightness constancy investigations, I would add this in the "Future Directions" section.

Also, I suggest stating that the rendered scenes were matte in the "Images of Virtual vs. Real Scenes" section, as a limitation of the simulation given that specular reflections might interact with lightness constancy, as discussed in the literature.

Good point. We have adopted this suggestion. We have added the following to the "Images of Virtual vs. Real Scenes" section (Line 333):

“Similarly, while the scenes rendered for this work were matte, we could introduce specularity and study interaction of its effect on luminance constancy.”

‘’Future work” (Line 351)

These include variation in non-spectral properties of objects and lighting in the scene, including object texture, material, specularity, and shape as well as lighting geometry.

Classical ("Lightness and retinex theory", Land & McCann, 1971) but also recent theories of lightness constancy ("A cortical edge-integration model of object-based lightness computation that explains effects of spatial context and individual differences" - Rudd, 2014) propose that visual system spatially integrates the luminance steps corresponding to reflectance edges (as given object boundaries). By looking at the shape of the receptive field in condition 3, it seems that rather large positive weights are flanked by negative weights corresponding to borders between the objects in the background, suggesting edge related computations. I suppose one of the potentiality of the approach is to reveal such local computations, thus if the authors find my speculation sensible, I would add it in the discussion, showing how the proposed approach has the power to reveal lightness constancy computations as proposed in the literature.

This is an interesting connection, which we now make. A full test would require variation of the geometry of the scene and an understanding of what RFs are optimal for that case, a manipulation that represents an interesting direction for future studies. [David to have a go at this.]

The idea of generating large datasets of rendered surfaces in order to investigate classification of an ideal observers (ROC and linear classification) on their material properties is not new ("Optimal sampling of visual information for lightness judgments" - Toscani, Valsecchi & Gegenfurtner 2013; "Lightness perception for matte and glossy complex shapes" - Toscani, Valsecchi & Gegenfurtner 2017; "Statistical correlates of perceived gloss in natural images" -Wiebel, Toscani & Gegenfurtner, 2015).

However, to my knowledge this is the first time that reflectance spectra are taken into account, as opposed to grayscale images, as the toolbox presented in the paper allows. I would stress the novelty respect to previous work.

Let's take a look at these papers and then discuss them appropriately. [David and Vijay to have a look at these and figure out what to say about them.]

Optimal sampling of visual information for lightness judgments: Uses real objects and rendered images to study relationship between eye fixations and brightness of the object. Concludes that the eye fixations are correlated with brighter parts of the objects. Rendered images are using RTB. No mention of a databases.

Lightness perception for matte and glossy complex shapes:

<http://graphics.stanford.edu/data/3Dscanrep/>

http://visionair.ge.imati.cnr.it/ontologies/shapes/

The distribution of surface albedos in natural environments is approximated by a specific beta distribution ("The distribution of reflectances within the visual environment", Attewell & Baddeley, 2006) and the discernible colors present in nature only cover a specific portion of theoretical solid of visible colors ("The number of discernible colors in natural scenes" Linhares, Pinto & Nascimento, 2008). For the simulation presented in the manuscript, reflectance spectra are sampled from a statistical model approximating a largely variable set of colors, as the Munsell chips is supposed to represent the space of visible colors rather than resembling the occurrence of colors in the word. I suppose this gives an upper limit to the limitation in performance due to the increasing complexity with conditions, and results might change considering the natural distribution of reflectance spectra.

There are databases providing a large collection of reflectance spectra or reflected spectra from isolated surfaces under a known illuminant, although they do not span color spaces as well as the munsell system. In fact, they focus on leaves fruits and vegetables ("Fruits, foliage and the evolution of primate colour vision" - Regan, Julliot, Simmen, Vienot, Charles-Dominique & Mollon, 2011; "Hyperspectral database of fruits and vegetables" - Ennis, Schiller, Toscani & Gegenfurtner, 2018).

This seems like a good point to discuss. I am not sure these data are available, but we might have a look at the papers and then see.

I found only one typo at the end of page 2: "(?, ?; Brainard and Freeman, 1997)", probably due to the reference manager.

Fixed. [Carefully check submission for any such typos, search on "?" in PDF, etc.]

I would find interesting to have the shape of the receptive fields reported also for the analysis about the scenes in condition 2.

We have added these to the appendix.