White Balance Diffusers in Digital Photography

Douglas A. Kerr

Issue 1
February 17, 2008

ABSTRACT

When we photograph an object illuminated by light whose chromaticity does not match the "reference white" chromaticity of the color space used to record the image, then when the "published" image is examined by a viewer, familiar objects will not seem to have their expected chromaticity. To overcome this undesirable effect, we apply color correction (often called "white balance correction") to the captured image. In digital photography, we may actually have the camera do this for us "on the fly". In order for the camera to do so, it must know the actual chromaticity of the incident light—the light that illuminated the subject during its photography.

Although we can measure this with a specialized laboratory instrument, we can also equip the camera temporarily with a special "front end" (often called a white balance diffuser) that will equip it to make the needed measurement itself. There is considerable misunderstanding about the technical principles involved in doing so. In this article we review and explain these principles and show how they pertain to the actual workings of this technique. The article does not discuss the operation or performance of specific available white balance diffusers.

THE NEED FOR COLOR CORRECTION

The color of the light reflected by an object is determined by the interaction between the color of illumination on the object (from the *incident light*) and the object's surface color (where *color* includes both the aspects of *luminance* and *chromaticity*). Thus, the chromaticity of the reflected light, for a given object, varies with the chromaticity of the incident light.[1]

The human visual system ("human eye", for short) has the uncanny ability, when viewing an object "first hand", to discern its reflective

---

[1] More precisely, the spectrum of the reflected light is determined by the interaction of the spectrum of the incident light and the reflective spectrum of the subject. Each of those spectrums defines a chromaticity, but the interaction cannot be precisely determined by considering those chromaticities alone. Nevertheless, in most situations of importance to us, the distinction is not of great consequence (and if it were, we'd have no way to deal with that!). Thus we will continue to think in terms of chromaticity alone.

color in the face of considerable variation in the chromaticity of the illumination[2]. We don't yet know exactly how this works, but it is believed to largely have to do with the eye's interpretation of the light reflected from all the other objects in "the scene", especially when many of these are familiar objects whose reflective color is known from experience. In effect, the eye cleverly makes an estimate of the chromaticity of the incident light and then "backs that out of" the observed chromaticity of the light reflected by the object of interest, leaving an impression of its actual surface chromaticity.

Thus, although the light reaching the eye from a white sheet of paper illuminated by the relatively-yellowish light from the incandescent lighting in the kitchen has quite a different chromaticity from the light reflected by the paper from the relatively-bluish light from an overcast sky, the eye seeing paper directly will in each case consider that this is white paper. This phenomenon is called *color constancy*.

Note that, amazing as this is, it can only take place for a moderate range of chromaticities of the incident light, essentially the range of chromaticity that we are willing to call "white". If the paper is illuminated by a red theatrical light, though, the eye will not recognize the paper as white.

If we take a photograph of the white paper, and if the chromaticity of the incident light differs from the "reference white chromaticity" defined for the color space in which the image is recorded, then when the image is viewed, the impression received by the viewer may not be that the paper is (in terms of its implied surface color) "white", but rather it will perhaps seem "yellowish" or "bluish"—the mechanism of color constancy no longer works as it did for direct viewing. It is believed that this comes from the fact that the eye does not now have available for comparison the entire panorama of the "surround" it would take in during direct viewing of the object.

In any event, this phenomenon (sometimes called a "color cast") is not usually considered a satisfactory result.

COLOR CORRECTION

To dispose of this undesirable effect, we may perform what is often called "color correction" of the image initially captured by the camera. Because this all has to do with variation of the precise chromaticity of illumination that we are still willing to think of broadly as "white", the

---

[2] See appendix A for why I say her "the chromaticity of the illumination" rather than "the chromaticity of the incident light".

process is often spoken of as "white balance" (that is, compensation for "the particular shade of white" in the illumination). We can ask the camera to do this for us (in several ways) before the final image is delivered, or (if we take the "raw" output from the camera) we can conduct the correction during post-processing. In this article, we will assume that the correction is done in the camera.

The chromaticity of the illumination

Essentially, to perform this color correction, the camera needs to know the chromaticity of the illumination of the subject of interest by the ambient light. Conceptually, this would best done by actually measuring the chromaticity of the illumination with an ambient light colorimeter, an instrument that gathers the incident light at a certain location (in a way that mimics the way the incident light would illuminate the surface of a object) and then measures its chromaticity.[3] We would then advise the camera of the instrument's findings.

Measuring with the camera

However, by equipping our camera temporarily with a special "front end", often called a "white balance diffuser", we turn the camera itself into a special-purpose ambient light colorimeter. We use the camera, thus adapted, to measure the chromaticity of the illumination, and ask it to remember that. Then, if we like, when we actually photograph the scene (with the diffuser removed), we ask the camera to make a "white balance correction" based on that remembered value of the chromaticity of the illumination.

The camera participates in both phases of this scheme under what is often called the "custom white balance" feature.

The role of the diffuser

What does the diffuser do for us? Couldn't we just "aim the camera at the light source" and ask it to make a determination of the chromaticity of the arriving light? Not effectively.

In most cases, the light falling on an object comes from many directions, and the components from different directions may not at all have the same chromaticity. Under classical photometric theory, if the surface of the subject is a "perfectly diffusing" surface, each component of the arriving light (that is, the part arriving from a

---

[3] Often the same instrument also will determine the luminance of the incident light, a determination that can be used in "incident light exposure metering" for exposure planning.

particular direction) has an influence on the illuminance of the reflected light that is proportional to the luminous flux density of the light in that component and the cosine of its angle of arrival (the angle between its direction of arrival and a line perpendicular to the surface).[4]

Then, if the different components have different chromaticity, the contribution of each to the chromaticity of the overall illumination is weighted by its contribution to the overall illuminance.

If we merely aim the camera "at the light source", perhaps planning to have it "take the average chromaticity of the light reaching the sensor", then:

- Unless we were using a "fisheye lens", the camera would not take in light from all the relevant directions of arrival into account.

- The averaging process for the light that was embraced would not take into account the "cosine" relationship with regard to its angle of arrival at the subject (whose importance was mentioned just above).

Now suppose we had a translucent disk that would accept light from all directions of arrival (over the hemisphere in front of it, of course); combine all the contributions, weighted according to the "cosine rule"; and turn the resulting collection into a uniform luminous pattern on the rear of the disk.[5] If the camera now observes that "exit glow", and makes a determination of its chromaticity, then we will have a meaningful determination of the chromaticity of the illuminance (as it would affect an object to be photographed). This is the principle of a white balance diffuser.

We typically mount it to the front of our lens. Because the lens probably does not focus at that distance, the exit glow on the rear of the disk will not be imaged "in focus" on the camera sensor. But that is of no consequence. The camera will presumably average the chromaticity across the entire sensor (or some arbitrary central portion of it) anyway, and the severe "blurring" that occurs merely contributes to that process.

---

[4] See Appendix A for a detailed discussion of this concept.

[5] I'll refer to that henceforth as the "exit glow"; "pattern" is not attractive, since we'll be using that term in another sense later.

Technique

The manufacturers of many white balance diffusers suggest that they be used by placing the camera (with the diffuser in place) at the subject location, with the outward face of the diffuser generally pointed toward where the camera will be for taking the actual photograph of the subject. I say "generally" because there are different opinions as to exactly how to aim the rig, among them:

- With the face of the diffuser facing toward the light source (hard to figure out in many cases just what that might be).
- With the face of the diffuser facing toward the camera position for the "shot".
- With the face of the diffuser parallel to the surface of the subject (hard to figure out if the subject isn't "flat").

This technique, especially in its last variation, well matches the concept that, in order to determine the chromaticity of the incident light in a way relevant to its effect on the light that will be reflected by the subject, the instrument should mimic the "acceptance" characteristics of the subject surface itself—that is, should measure the chromaticity of the illumination the subject will receive from the incident light.

Sometimes, it is impossible, impractical, or merely bothersome to take the camera to the subject location to make this measurement. Can we make a meaningful measurement from the camera location?

Well, in some settings we may reasonably believe that the incident light at different locations around the battle zone (and regardless of the orientation of the receiving surface), exhibits a consistent chromaticity. This might well be true outdoors, especially near midday, or in a ballroom illuminated by chandeliers all over the ceiling.

In effect, this would mean that the light incident on the photographer would have essentially the same chromaticity as the light incident on the subject.

If we accept that, then we may consider making the measurement with the camera (equipped temporarily with the diffuser) at the same location it will have during the "shot". Under our assumption about the consistency of the incident light, the direction the diffuser faces doesn't much matter (not down at the ground, obviously) and so we might just as well choose to have the camera aimed in the same direction it will be for the shot (so we don't have to fiddle with the tripod head).

If our assumption about the incident light turns out be correct (which it will in many real settings), this will produce a "good" result.

When this alternate technique is used, it is often described as "pointing the diffuser at the subject." This is of course true if we choose to do it as I described, but not at all relevant to the principle involved. As I suggested above, we may get the same result with the rig aimed in some other direction (somewhat upward, for example).

It is also often said of this alternate technique, "Here we measure not the incident light but rather the light reflected from the subject." This is not true. We are actually measuring the incident light "on the photographer". A fraction (usually small) of that is of course the light reflected from the subject, just as a small part of it is reflected from the photographer's car, just outside the field of view for the shot, or from a person standing nearby, or from every tree in front of the camera.

But it is no more appropriate to think of this as "measuring the light reflected from the subject" than, when a measurement is taken at the subject location, to describe it as "measuring the light reflected by the photographer's assistant" (standing by the empty tripod).

These mischaracterizations of this alternate technique serve to misdirect attention from what the technique really is: measuring the ambient light not at the subject but at the camera location, in the hope that the two will have about the same chromaticity.[6]

In fact, truly measuring the chromaticity of the light reflected by the subject (perhaps with a diffuser with a narrow sensitivity pattern) is of no value in planning color correction. This measurement is wholly dominated by the reflective chromaticity of the subject.

PROPERTIES OF A WHITE BALANCE DIFFUSER

There are several properties of a white balance diffuser which we can readily see would be consequential to their operation. Here are the two biggies:

---

[6] Note, however, that there are those who feel that taking the measurement from the camera position (which they may describe as "measuring the light reflected by the subject") is actually fully valid (not dependent of any assumption of the incident light being the same there as at the subject)—perhaps even more desirable, in some circumstances (they suggest), than taking the measurement at the subject location. I have never received a satisfying conceptual justification from that camp for that outlook.

Sensitivity pattern

If we consider a diffuse reflecting surface, the illuminance provided on the surface from a light "beam" of any given luminous flux density will vary as the cosine of the angle of incidence (the angle that the direction along which the beam arrives makes with a line perpendicular to the surface). This is often described as a "cosine response" (duh!).

If we want the diffuser to essentially mimic that behavior in accepting light for presentation to the camera for measurement, then it should have a cosine sensitivity pattern (where here by "sensitivity" I mean the variation in the luminance of the exit glow presented to the camera from a "beam" of light with a given luminous flux density, as a function of its angle of arrival.

This is not to say that other sensitivity patterns might not be advantageous for subtle reasons not apparent in the basic concepts I describe here. But in any event, the sensitivity pattern of the diffuser is an important property.

A sensitivity pattern is generally presented as a polar plot, with the radius in terms of relative sensitivity (the ratio of exit luminance to entry illuminance). Interestingly enough, the ideal cosine response plot is a circle passing through the origin of the plot (tangent to the vertical axis, if the horizontal axis represents the "aiming axis").

Note that measuring that pattern is not as simple as it might seems. We must in any case state the conditions under which it is measured. In particular, we need to be specific about how the "glow" at the back of the diffuser will be "examined" by the camera. The subtleties of this are beyond the scope of this article.

Chromatic neutrality and spectral uniformity

Clearly we want the chromaticity of the light emitted by the diffuser to closely follow the chromaticity of the incident illumination. After all, we measure the former, whereas what we really want to know is the latter.

It would be desirable for this chromatic neutrality to result from the fact that the transmission of the diffuser is spectrally uniform ("flat"); that is, the ratio of the luminance of the exit glow to the illuminance on the face of the diffuser will be constant with wavelength over the entire visible spectrum. If we have that, then we will have chromatic neutrality.

But spectral uniformity is not necessary for chromatic neutrality (at least in a sense that will well serve our purposes). A diffuser not

having a perfectly uniform spectral response can still present to the camera for measurement an exit glow that has essentially the same chromaticity as the incident light, assuming that the spectrum of the incident light itself isn't too "wild". (And if it is, the color balance procedure probably won't produce the desired result anyway.)

The result of a spectral uniformity test is a plot of relative sensitivity (as defined above) versus wavelength. A problem is that is hard to assign a numeric "score" to any particular curve ("Well, how unflat is it?"). Remember, our real concern is with the preservation of chromaticity, and that is affected in a complicated way by spectral nonuniformity.

Chromatic neutrality can be expressed in different ways. Sometimes it is done by expressing the difference between the chromaticity of the exit glow and the chromaticity of the incident light in terms of displacement on the CIR u-v chromaticity diagram (quoting Δu and Δv). It is, however, difficult to interpret a result in terms of those quantities. A discrepancy of 1 unit in u does not have the same implication on perceived chromaticity as a discrepancy of 1 unit in v. And the significance varies with the starting chromaticity.

Sometimes the test is of the "transmissive color" of the diffuser (analogous to the "reflective color" of a surface), reported in CIE L*a*b* coordinates, only paying attention to the a* and b* values (which would both be zero for a neutral transmissive color). Here again we have a problem. The variables a* and b* are chrominance, not chromaticity, values. The amount of chromaticity discrepancy implied by a one unit discrepancy in a* depends on the value of L*. And again, the perceptual implications of a* being non-zero by one unit are different than for b* being non-zero by one unit.

Sometimes the manufacturer will send through the diffuser light having a certain "standard illuminant" spectrum (and thus a certain chromaticity), typically "illuminant D65" (whose chromaticity is the "reference white" for the sRGB color space) and determine the RGB representation (in the sRGB color space) of the exit glow. To get a consistent "scaling" for R, G, and B, we set the absolute luminance of the test illumination so that the relative luminance represented by R, G, and B corresponds approximately to the "standard exposure" a properly calibrated camera automatic exposure system would produce for a uniform-color scene.

If the test report included in the diffuser box shows, for example, "R,G,B = 121,121,121, then we know that the diffuser exhibits very good neutrality. But if the report shows 120, 121, 119, how bad is that? Again, it is hard to evaluate a repot in these terms.

Perhaps the most useful way to quantify the chrominance discrepancy caused by imperfect neutrality is in terms of MacAdam steps. These relate to the minimum chromaticity difference that can be perceived by an observer, on a statistical distribution basis. For example, 32% of all observers will be not be able to notice a chromaticity difference (in an A-B comparison) of one MacAdam step.[7]

For comparison, industry standards for the chromaticity of various lamps (such as compact fluorescent lamps of a certain "color") generally call for them to be within 4 MacAdam steps of the specified standard chromaticity.

In any event, it is much more difficult to measure transmissive chromatic neutrality that might be thought. Among other things, we must decide at what angle(s) will the "probe" light be allowed to strike the surface, and how will the exit glow be regarded from behind.

In our case, we are concerned with the incident light striking the diffuser at a wide range of angles, and its response may not be the same for all angles. How do we investigate that, and report the results? The subtleties of this are beyond the scope of this article.

THE DILEMMA OF SEVERE MIXED LIGHT

The *bête noir* (!) of color balance is a situation of severe "mixed light", in which the subject is illuminated from different directions by two or more light sources of substantially different chromaticity. We might, for example, have a model illuminated from the left front by sunlight though a window and from the right front by light from an incandescent lamp. How does our concept of white balance measurement deal with this?

To clarify the principle involved, first suppose that our subject is actually perfectly flat (perhaps a mounted vintage newspaper page). For either arriving light "beam", the angle of arrival is constant over the entire subject, and thus the effect of the "cosine" factor for that beam is constant. Thus, the relative contributions to the illuminance on the subject of the two sources is the same over the entire subject. Accordingly, the effective chromaticity of the illumination is the same over the entire subject.

---

[7] One MacAdam step represents the "one sigma" point on the distribution of the sensitivity of observers to chromaticity difference. Thus, 68% of the observers will be able to perceive a one-step difference. However, it is more easily grasped to say that "32% of observers cannot perceive a one-step difference".

If we take an ambient light chromaticity measurement with a diffuser-equipped camera at the subject location, and orient the face of the diffuser parallel to the subject, then the chromaticity of the net illuminance on the diffuser will be the same as for the illuminance on the subject. If the diffuser exhibits essentially a cosine response, then the "reaction" of the diffuser will be the same as that of an actual surface, and the exit glow will have the same chromaticity as that on (any part of) the subject. This will be the chromaticity determined by the camera. If this is used in the color correction process, the correction should be "ideal" (for all parts of the subject).

Thus my earlier suggestion that the "parallel to the subject surface" orientation is the conceptually best one.

Now let's consider a more common subject, perhaps a person's head. What orienbtatioin would "parallel to that surface" be?

Well, before we get frustrated by that, lets look at the photometric situation. The angle of incidence, for the light from any given source, will vary across the face. For a different source, it will also vary, but differently (owing to the different location of that source).

Thus, the relative contributions of the sources will be different at different parts of the face, and the chromaticity of the net effective illuminance will be different at different parts of the face. Now, how should we make the incident light chromaticity measurement?

Well, in fact, we must first recognize that ideal correction of such an image is impossible. If we apply a correction that is appropriate for one side of the face, it won't be for the other. So the issue of "best practice" for the orientation of the diffuser in such a case is almost moot.

Likely, our best bet would be to make the diffuser parallel to the "center of the face" (practically, aim it toward the camera position). Then, we might have ideal color correction for the center of the face, with an error in opposite directions on the two sides of the face.

#

APPENDIX A

INCIDENT LIGHT AND ILLUMINATION

In the body of this article, I speak often of "the chromaticity of the illumination on a surface from ambient light". Couldn't I just as well (and more concisely) say, "The chromaticity of the incident light"? Not really, since there is a distinction between the two, one that is very pivotal to the issues covered here.

I can best explain the difference in an outlook that does not involve chromaticity but rather the "potency" of light (an intentionally non-specific term I use to embrace several related, but distinct, concepts). The potency of an arriving "beam" of light is described by its *luminous flux density*. This is defined as luminous flux[8] per unit of area, where the area is on a plane, traversed by the beam (at the location in its travels where we are interested in its "potency"), at right angles to the beam's direction of travel.

When the beam strikes a surface, we say that it illuminates the surface, and the potency of that illumination is described in terms of its *illuminance*. Illuminance is defined as luminous flux per unit area, where here the area is on the surface of interest.

Thus, for an arriving beam of a certain luminous flux density, the illuminance it provides on a surface declines as the beam arrives more obliquely (that is, away from a line perpendicular to the surface). This does not result from any mysterious principle of physics, but from simple geometry.

Suppose we consider that portion of an arriving beam that lies within a "square tube" 1 cm x 1 cm in dimensions (a cross-sectional area of 1 cm^2). It contains a certain amount of luminous flux. If the beam lands on a surface at an angle of incidence of 0° ("head on"), our little "square tube" will deposit that same amount of luminous flux over an region whose dimensions are also 1 cm x 1 cm (an area of 1 cm^2). Thus, the illuminance will be the same as the luminous flux density.

If, however, that beam lands on a surface at an angle of incidence of 45° (with a flat side of the tube down), then our little tube will deposit that same amount of luminous flux over an region whose

---

[8] Luminous flux is the "stuff" of light. It is wholly analogous to power in an electrical or radio engineering situation, differing only that its measure takes into account the different sensitivity of the eye to different wavelengths.

dimensions are 1 cm x 1.414 cm [9] (an area of 1.414 cm^2). Thus, the illuminance there will be 1/1.414 (0.707) times the luminous flux density of the beam.

In general, then, the illuminance on a surface is the product of the luminous flux density of the arriving beam times the cosine of its angle of incidence.

Now, of course, this doesn't in any way affect the fact that, for a surface illuminated by a single incident light beam, the chromaticity of the illumination will be the chromaticity of the incident light. But now imagine that the surface is illuminated by two incident light beams, landing with different angles of incidence, and having different chromaticities.

The effective chromaticity of the total illumination on the surface (which is what influences the chromaticity of the reflected light) must be reckoned by combining the illumination contributed by the two beams, weighted by their respective illuminance (not their respective luminous flux density). Thus the cosines of the angles of incidence get into the act. In the extreme, a beam arriving at a very oblique angle (almost parallel to the surface has little influence on either the illuminance **or the chromaticity** of the net illumination.

This is why, for example, if we have a cylindrical object illuminated by light from two directions, with the two ambient light "beams" having different chromaticity, the chromaticity of the net illumination will vary around the object (and, if the object has a consistent "reflective color", so will the chromaticity of the reflected light).

#

[9] Visualize a wood stick 1 cm square, with the end cut of at a 45° angle. The dimensions of the cut surface will be 1 cm x 1.414 cm.