Proceedings

EXPERIENCING LIGHT 2009

International Conference on the Effects of Light on Wellbeing

Y. A. W. de Kort, W. A. IJsselsteijn, I. M. L. C. Vogels,
M. P. J. Aarts, A. D. Tenner, & K. C. H. J. Smolders (Eds.)

Keynotes and selected full papers
Eindhoven University of Technology,
Eindhoven, the Netherlands, 26-27 October 2009

Volume Editors

Yvonne de Kort, PhD
Wijnand IJsselsteijn, PhD
Karin Smolders, MSc
Eindhoven University of Technology
IE&IS, Human-Technology Interaction
PO Box 513, 5600 MB Eindhoven, The Netherlands
E-mail: {y.a.w.d.kort, w.a.ijsselsteijn, k.c.h.j.smolders}@tue.nl

Ingrid Vogels, PhD
Visual Experiences Group
Philips Research
High Tech Campus 34, WB 3.029
5656 AE Eindhoven, The Netherlands
E-mail: ingrid.m.vogels@philips.com

Mariëlle Aarts, MSc
Eindhoven University of Technology
Department of Architecture Building and Planning
PO Box 513, VRT 6.34
5600 MB Eindhoven, The Netherlands
E-mail: M.P.J.Aarts@tue.nl

Ariadne Tenner, PhD
Independent consultant
Veldhoven, The Netherlands
E-mail: ariadne.tenner@onsmail.nl

ISBN: 978-90-386-2053-4

Copyright:

Reference specification:
Name Author(s), “Title of the Article”, In: Proceedings of EXPERIENCING LIGHT 2009 International Conference on the Effects of Light on Wellbeing (Eds. Y.A.W. de Kort, W.A. IJsselsteijn, I.M.L.C. Vogels, M.P.J. Aarts, A.D. Tenner, and K.C.H.J. Smolders), 2009, pp. X (startpage) – Y (endpage).

Content-based Adaptation of the Dynamics of Estimated Light Sources

Marc Peters*, Pedro Fonseca*, Lu Wang*, Bas Zoetekouw*, Perry Mevissen#

*Philips Research
Experience Processing Group
High Tech Campus 34, 5656AE Eindhoven

#Philips Consumer Lifestyle
Advanced Technology
High Tech Campus 37, 5656AE Eindhoven

{marc.a.peters, pedro.fonseca, l.wang, bas.zoetekouw, perry.mevissen}@philips.com

ABSTRACT

Lighting is a very important aspect in film-making. Using a technique known as light source estimation, it is possible to estimate the color properties of the light sources used while filming scenes of films or television series. One very important—but often unaddressed—aspect of light source estimation is related to temporal control. In this paper, we propose a novel method for temporal control of the estimated light source of a video scene. After describing the method, we will explain the results of a user study which shows that it is superior when compared with traditional temporal control techniques.

Keywords

Movie lighting, content analysis, light source estimation, temporal control.

INTRODUCTION

Lighting is a very important aspect in film-making. In modern movies and television series, film makers and cinematographers carefully use light to accentuate certain aspects of the story, to change the atmosphere that is conveyed, or to establish a certain mood. For example, candlelight suggests romance and harmony, high contrast lighting achieves accentuated dramatization, and moving light can invoke fear, chaos and madness [17]. Colored light is also used to accentuate certain aspects of the story or to help convey certain emotions. Although there are no specific rules on how to associate colors with emotions, red is often found associated with love or hatred, yellow with happiness and joy, and blue with peace and tranquility [18].

For regular film watchers, these lighting aspects usually have an implicit influence: although they are of crucial importance to help convey the story, most people don't actually realize that they are "manipulated" by lighting changes during a movie. On the other hand, the presence and characteristics of these elements have very often been used in the area of video content analysis; interpretation of cinematographic rules — such as information about the lighting of a scene and the color of the light source that illuminates it — can give important semantic information about that scene or even about the entire movie. For example, Rasheed et al. use scene lighting characteristics, along with other visual features, to automatically classify the genre of a movie [11]. Light source estimation can thus be a very important technique to extract high-level, semantic information about a scene.

One particular application of light source estimation is the creation of a lighting atmosphere which is rendered while users watch a movie on a television screen. If the light source is estimated correctly, the rendered atmosphere will resemble the light settings of the scene in the movie and increase the user's immersion in the content.

The topic of light source estimation has been well studied in the past and several well described techniques are in common use. The techniques range from simple MPEG-7-style dominant color extraction [15] to more advanced systems based on white point extraction. In [1,2], some of these methods are reviewed. As we are mostly interested in the dynamic control of the estimated light source rather than the actual light source estimation itself, a full comparison between the different known methods is beyond the scope of this paper. For simplicity reasons, we will only use a single light source estimation algorithm in this paper to test our dynamic filtering technique. For each content frame, we construct an RGB space in which each the color of each pixel is represented by a point. We then use least squares estimation to determine the best fit of the data from the linear RGB color space into a vector in that space. Because the light source is reflected on objects on the screen, this vector has the property that it reflects precisely the color that is reflected on the different surfaces.

For the application mentioned above, the temporal dynamics of the estimated light source are very important. The actual lighting set-up of a mise-en-scène — i.e. the arrangement of actors and scenery on a setting filmed for a motion picture — filmed for a calm movie scene is usually constant for an entire scene. However, light source estimation techniques are not perfect, and camera action and object movement often cause the estimated light source to vary throughout such a calm scene, even though the actual lighting during recording of the scene didn't change at al. If the raw results of light source estimation are used to render a lighting atmosphere, the atmosphere will be much more dynamic than the actual content as the user sees it on the screen. This effect not only reduces immersion, but it might even distract and annoy the viewer. On the other hand, a very dynamic lighting atmosphere might be

desirable if scene under consideration is equally dynamic, for example when many special visual effects are used. In that case, a dynamic atmosphere will match the content on the screen and will therefore contribute to the viewer's immersion.

This temporal aspect of lighting has not been so much addressed. This is surprising, as dynamics of lighting and light effects are a fundamental part of the cinematographic experience Most existing work on temporal control of estimated lighting focuses on the use of (advanced) low-pass digital filters which "smoothen" the estimated signal over time. In [16] an advanced implementation of such a system is described, which uses substantial time subsampling and employs the spatial detection of special effects on small regions to reset a low-pass filter. Whereas this system is able to react to very localized visual effects, a drawback is that it will often fail to reflect the global dynamic properties of an entire scene. This drawback actually applies to most filters described in literature: however advanced these filters are, unless they take into account the fact that global and local dynamic properties of the content, on a frame by frame basis, are inherently a very important part of the content itself, they will unavoidably decrease the intensity of these elements for the viewer.

Let's look at two extreme cases as an example. During reasonably static, dialog-based scenes — such as those which occur quite often in soap operas and popular comedy television series like *"Friends"* and *"Will and Grace"* — the resulting estimated light source should be rather static throughout the entire scene. However, at the exact frame this scene ends and a completely different scene starts, the light source should reflect the properties of the new scene immediately and should not be slowly smoothed throughout. At the other extreme, consider a scene in a war movie where the battle takes place at night. Due to the dark lighting conditions, any special effect (small such as a gunshot or intense such as an explosion) should be reflected in the estimated light source. In this case, smoothing with a low-pass filter is simply not acceptable.

In this paper, we propose a novel method for temporal control of the estimated light source of a video scene. The goal is to smooth the estimated light source color when the content is static but to allow special, abrupt effects to be instantly reflected without latency in the resulting estimation. This will allow an atmosphere to be rendered while viewers watch a movie on a television. This atmosphere not only reflects the color of the lighting of a scene but also its temporal dynamics. All of this will help increase the immersive experience.

In the next section we will introduce the algorithm. Afterwards we will describe the test content and objective criteria for characterizing the dynamics of that content. We will then describe the user test that we have carried out to evaluate the proposed algorithm. Finally, we will discuss the results and conclude the paper.

TEMPORAL CONTROL ALGORITHM

In the context of this paper, by "temporal control", we mean the process through which the results of light source estimation are modified or filtered to change its temporal characteristics. An example of a very simple temporal control algorithm is a low-pass filter. When applied to the output of a light source estimator, it simply eliminates (or "smoothes") all abrupt color variations given by that estimator. As can be easily imagined, the resulting colors vary slowly in time, regardless of how dynamic the content might be.

For the reasons explained in the introductory section, it is not always desired that the estimated light source follows the changes in the content on a frame-by-frame basis. Particularly for the application described earlier where an atmosphere is rendered in real-time along with a movie being watched by the viewer, the rendered light source should only change significantly when the same happens with the content. With the approach described in this paper we attempt to let the dynamics of each scene dictate how "smooth" the variations of the resulting light source color should be. Very dynamic scenes will lead to fast variations in the light, whereas static, slowly varying scenes will lead to results that are calm and smooth in time.

It should be noted that the temporal control algorithm proposed in this paper does not depend on the light source estimation algorithm used. In fact, it is suitable for any kind of raw color signal input, for example expressing an approximation of the color properties of the dominant light source in a scene. The method described in this section can be used without loss of generality for any such input.

To achieve automatic smoothing as described above, we need to characterize the dynamics of the content in a feature which is simple to use. As physical light sources that illuminate the scene during recording naturally influence both the colors and the luminance of that portion of the motion picture, we will estimate the dynamics based on color- and illumination-based features. From these features, we then compute the dynamics for each frame of the content.

To calculate these features, we make use of a so-called "combined HSV histogram". HSV is a relatively simple three-component color space, characterized by the hue (H), saturation (S) and brightness (value, V) [6]. We use an HSV color space here, rather than the simpler RGB color space, because HSV more accurately describes perceptual color relationships than RGB. On the other hand, it is still computationally simpler to implement than real perceptually uniform color spaces like CIE 1976 La*b*. However, because the HSV color space is not completely perceptually uniform, in particular in the low brightness colors, we need to use a non-trivial distance measure as we will define below.

We define a 256-bin HSV histogram as follows:

- The 256 bins are ordered in a cube of dimensions 16×4×4. The first dimension corresponds to the hue

values (discretized in 16 bins), the second dimension corresponds to the saturation (discretized in 4 bins), and the third dimension corresponds to the value (discretized in 4 bins). We use more bins for the hue component than for the saturation and value components because we want to give more importance to the differences in hue than to differences in saturation or brightness.

- For each video frame, we calculate the HSV values for each of the pixels and fill each bin of the histogram with the number of pixels that have an HSV value in the corresponding range.
- If the hue value is expressed as a number in [0,360) and the saturation and value are numbers in [0,1], the first bin will thus contain the number of pixels that have an hue between 0 and 22.5 (=360/16), a saturation between 0 and 0.25, and a value between 0 and 0.25.

Mathematically, we can express this as follows:

$$\begin{aligned} HSV[h,s,v] = |\{(\mathrm{i},\mathrm{j}): \\ & \frac{360}{16}h \le H_{\mathrm{ij}} < \frac{360}{16}(h+1) \\ & \wedge \frac{1}{4}s \le S_{ij} < \frac{1}{4}(s+1) \\ & \wedge \frac{1}{4}v \le V_{ij} < \frac{1}{4}(v+1) \\ \}| \end{aligned} \quad (1)$$

where i and j correspond to the rows and columns of pixels in each video frame, and H_{ij}, S_{ij}, and V_{ij} are the hue, saturation and value components, respectively, of the pixel at position (i, j).

The histogram is then normalized by dividing each value by the total number of pixels in the video frame, in order to make the feature independent of the dimensions of the video frame.

The HSV histogram is created for each frame in a video sequence. Next we define a distance measure Δ between two of such histograms

$$\Delta = \sum_{h=0}^{15}\sum_{s=0}^{3}\sum_{v=0}^{3}\left|HSV_t[h,s,v] - HSV_{t-1}[h,s,v]\right| \quad (2)$$

as the sum of the absolute differences between corresponding bins in two consecutive frames. As the histograms are normalized, this definition of the distance measure Δ will always yield a value between 0 and 2.

However, large parts of video frame often have very low brightness. Whenever such dark regions are present in two consecutive frames, the distance between the corresponding HSV histograms will also be very small. This is caused by the fact that HSV is not perceptually uniform as explained before. For our application, this is undesirable for two reasons:

1. Dark regions of an image do not convey much information about the light settings of a scene, other than the fact that the light source did not strongly illuminate that area;
2. A small difference between the histograms of two dark video frames might hide the fact that the light settings captured on those two frames are completely different. For example, consider the situation in which a very dark scene is illuminated by a small green light in the first frame, and that that light suddenly changes to red in the next frame. In that case the light condition as determined by the histograms should clearly be very different, even though the majority of the pixels are black or very dark in both frames.

In order to make the distance measure of Equation (2) more robust to such conditions we introduce an alternative distance measure Δ', that takes into account this problem by not counting dark pixels in the frames. We define this alternative distance measure as follows:

$$\Delta' = \frac{\sum_{h=0}^{15}\sum_{s=0}^{3}\sum_{v=1}^{3}\left|HSV_t[h,s,v] - HSV_{t-1}[h,s,v]\right|}{\max\left[\frac{1}{4}, 1 - \min_{t,t-1}\left[\sum_{h,s} HSV[h,s,1]\right]\right]} \quad (3)$$

The numerator counts the bin-to-bin difference between the histograms, but leaves out bins for which the brightness component is low (i.e., only pixels with brightness components in the three highest bins are taken into account). The minimum in the denominator should be taken between the sums of the bins with the lowest brightness of the two consecutive frames. The denominator leaves out the dark parts that are common to both frames. It is bound to a minimum value of 1/4, to avoid extremely large distance values between the histograms of two video frames when both have very large dark areas. The distance measure Δ' of Equation (3) thus emphasizes the difference between the non-dark parts of the images.

Next, for each frame t we compute two values:

1. The first is the **estimated light source** LSE_t (expressed as RGB values) for that frame. The exact nature of the algorithm used to find the light source of the current frame does not matter as long as for each frame we find a color vector (in the linear RGB color space) that reflects some properties of the dominant light source illuminating the scene:
$$LSE_t = (R_{LSE}, G_{LSE}, B_{LSE})_t \quad (4)$$
2. The second value we calculate is the **resulting light source**, LSR_t after temporal filtering:
$$\begin{aligned} LSR_t &= (R_{LSR}, G_{LSR}, B_{LSR})_t \\ &= (1-s)\cdot LSE_t + s\cdot LSR_{t-1} \end{aligned} \quad (5)$$

It is computed as a linear combination of the estimated light source for the current frame and the color that was

calculated from the previous frame. The smoothing factor s is defined as

$$s = \begin{cases} 0.98 & \text{if} \quad \Delta' \le 0.02 \\ 1 - \Delta' & \text{if} \quad 0.02 < \Delta' \le 1 \\ 0 & \text{if} \quad \Delta' > 1 \end{cases} \tag{6}$$

For dynamic scenes the smoothing factor s is small, giving a high weight to the light source estimated for the current frame. For calm, static scenes the smoothing factor is large, leading to a calm and gradual transition between colors because the previous value has a large weight. The minimum smoothing factor is larger than 0 to make sure that the filtered light color will always converge to the estimated light source of the current frame LSE_t if the video freezes or becomes completely static (i.e., if the distance computed between several consecutive frames is 0).

VALIDATION FRAMEWORK

In this section, we present a framework to quantize the dynamic properties of video sequences in order to characterize the properties and the behavior of the proposed temporal control algorithm. We start by describing our test set; this test set is used both for the quantization in this section as well as for the user study of the temporal filtering algorithms that is described in the next section. We then describe which features we extract from the video sequences to characterize their dynamic properties. We continue by analyzing the effects of temporal filtering on the dynamics of estimated light source.

Test content

The test set consists of six video clips of 30 seconds each. They were selected based on the presence of very different lighting conditions which characterize different genres in film and TV series. The clips from the test set can be described as follows:

- ***Walk the Line*** — the first sequence is a clip from the movie "*Walk the Line*"; this particular scene depicts a concert with different illumination sources: the backstage lighting, with a relatively dark and saturated color and the non-saturated, bright illumination of the singer which dominates the scene. The scene has a very high contrast, is very calm and shots are relatively long.
- ***Hellboy*** — the second sequence is a clip from the movie "*Hellboy II: The Golden Army*"; this particular sequence has a filmed part and a computer generated part. Both are highly saturated and the filmed part has the additional particularity of having a very distinctly colored light source in the left and the right side of the screen.
- ***Friends*** — the third sequence is a clip from the episode "*The One with the Kips*" (season 5, episode 5) of the popular soap opera *"Friends"*; like most soap operas, Friends has a flat appearance: there is little contrast, and it is filmed in high-key, i.e., it has an abundance of unsaturated light, and the scenes are free from shadows. The colors are mostly pastel and although there is not a lot of movement on the scene, the shots are relatively short as it mainly consists of dialogues.
- ***Hulk*** — the fourth sequence is a clip from the movie "*The Incredible Hulk*"; this particular clip takes place in a cave, at night, during a thunderstorm. Although the shots are long and dark and the contrast is high, the lightning strikes and the rain add a very dynamic element to the scene lighting.
- ***Wall-E*** — the fifth sequence is a clip from the animation movie "*Wall-E"*; this particular computer generated clip depicts an indoors scene, illuminated by one of the characters (a robot). During the scene the illumination varies drastically, from very well lit, high-key, to very dark and saturated.
- ***Platoon*** — the sixth and last sequence is a clip from the movie "*Platoon*"; this particular scene depicts a combat situation at night; apart from being a very dark scene, contrast is relatively low and gunshots and explosions dominate the scene, making it very dynamic and intense.

Visual feature extraction

The variety of clips and genres will help us explore the behavior of the temporal control algorithm proposed in this paper. As the clips we use are not part of any public domain test set and therefore are not available for free, it is important to characterize them as well as possible. Not only will this offer the reader a better description of the test set used in our study, but it will also help us explain the results of that study later on in this paper.

In this section, we describe this characterization of the dynamics of the video sequences in terms of temporal changes of their visual properties. To quantify these visual properties, we extract a number of video descriptors from the content. These descriptors offer a numerical representation of visual properties which can be analyzed in terms of their dynamic behavior in time.

Video descriptors

In order to characterize the dynamics of each sequence in our test set, we will use three different descriptors: the shot duration, the HSV histogram and the light source color.

The first descriptor we will use is the shot duration. A shot is defined as a sequence of video frames, captured uninterruptedly by a movie camera. It is delimited by a shot transition (commonly called a "shot cut") at its beginning and at its end. The transition between shots can be abrupt, in which case the new shot will start on a frame immediately after the last frame of the previous shot, or gradual, in which case the actual transition lasts for a number of frames (e.g. cross-fade, fade in, fade out).

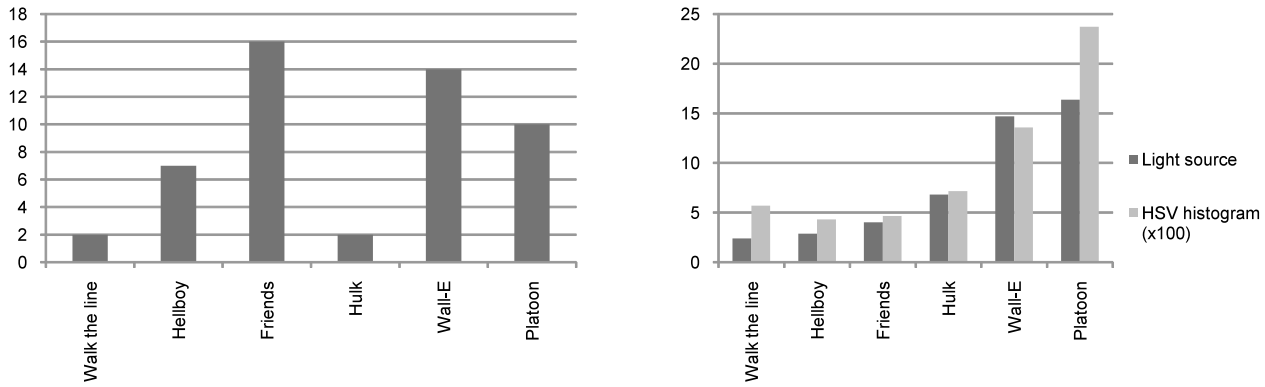


Figure 1 – (left) Number of shots in each of the test sequences and (right) average intra-shot feature differences. Note that the HSV histogram difference values were scaled up so both features can be visually compared.

Different shots often correspond to different points of view within the same scenario; other times, they belong to distinct scenes and thus have completely different visual properties. These shot transitions represent visual discontinuities and should be taken into account when characterizing the video sequences. An important way to characterize the dynamics of film content is by measuring the average shot length. Since most of the shots end in abrupt transitions, shot boundaries are moments when the visual properties (including the lighting of a scene) change drastically from one moment to another. This does not necessarily mean, however, that the sequence is very dynamic. In a soap opera, mainly comprised of dialogues, shots are typically very short and alternate between two or more view points within the same scene. However, the properties of each shot are very similar because the dialogues take place in the same physical location. Furthermore, within each shot, there is little or no change since usually, during dialogues, only the face of the actors is shown in the image. In an action movie, on the other hand, directors usually keep the shots short to indicate action and induce a high tempo. In this case, the visual properties of each shot are very different from each other, offering the viewer very much different visual information within a short period of time.

The second video descriptor we will use is the HSV histogram. This descriptor was introduced in the previous section. It describes the color properties of each frame of the video in terms of the hue, saturation and brightness (value) of all its pixels.

The third and final descriptor we will use is the light source color. It gives an approximation of the color of the light entering a scene. As remarked before, the temporal control algorithm that is the main focus of this paper is independent of the algorithm used to estimate the light source, and a full description of the different methods available is out of scope here. We use a method based on a least-squares fit in the RGB-space representing all pixels in a single video frame.

Characterization of visual properties

The number of shots in each sequence of our test set is illustrated in Figure 1 (left).

To express the temporal behavior of the two remaining visual descriptors, we compute the corresponding average feature differences. The average HSV histogram difference is given by:

$$\overline{\Delta H} = \frac{1}{N-1}\sum_{t=1}^{N-1}\sum_{h,s,v}\left|HSV_t[h,s,v] - HSV_{t-1}[h,s,v]\right| \quad (7)$$

where the histogram $HSV_t[h,s,v]$ was described in the previous section. The outer summation is over all N frames, the first frame being $t=0$, and the inner summation is over all the 16×4×4 bins of the histogram.

The average light source difference is given by:

$$\overline{\Delta L} = \frac{1}{N-1}\sum_{t=1}^{N-1}\Delta L_t \quad (8)$$

with

$$\Delta L_t = \sqrt{(r_t - r_{t-1})^2 + (g_t - g_{t-1})^2 + (b_t - b_{t-1})^2} \quad (9)$$

i.e., the Euclidian distance in RGB space between the light source colors in two consecutive frames, and where r_t, g_t and b_t are the RGB values of the light source computed for frame t.

As was mentioned above, shot cuts represent visual discontinuities. To better assess the dynamic properties of the sequences, while at the same time excluding the influence of the shots boundaries from this process, the computation of the average feature difference will be restricted to all frames that are neither the starting frame of a shot nor the frames that make up a gradual transition. Figure 1 (right) illustrates the intra-shot average feature differences for the sequences in the test set, ordered by increasing average light source differences.

As can be clearly seen, the sequence *Friends* has one of the lowest variations in terms of both visual descriptors, even though the clip contains the highest number of shots in the

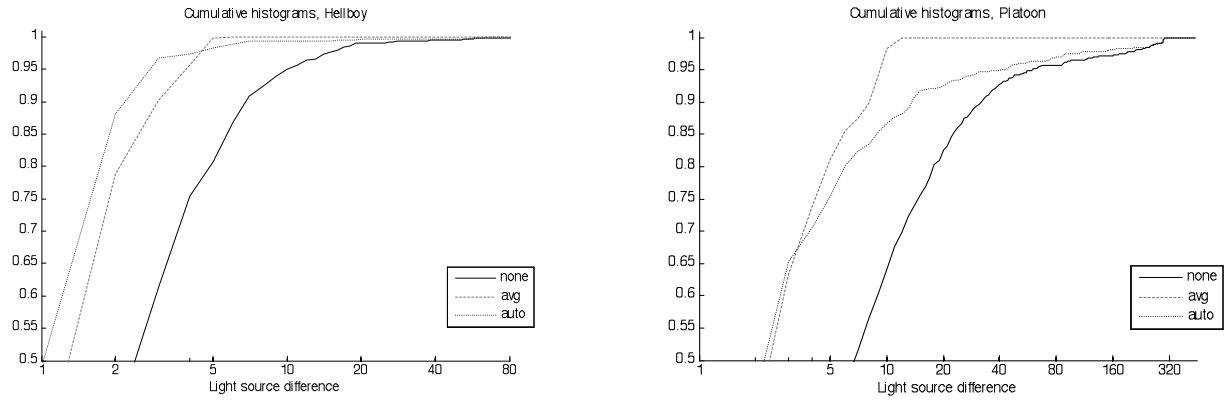


Figure 2 – Cumulative histogram of light source differences, filtered with the three different settings for Hellboy (left) and Platoon (right); note that the horizontal scales are different for the two graphs. The fact that the *avg* setting is in both cases above the unfiltered result, shows that simple averaging always decreases the variations in the light source. The *auto* setting, on the other hand, is first above and later below the *avg* setting: this indicates that small changes are smoothed even more than in the *avg* case, while large changes are kept (see text). Note that the maximum possible light source difference is 442; however, all of the graphs saturate much earlier.

clips in our test set. This is typical of soap operas: short shots are necessary for dialogues but the scene is kept rather static to avoid inducing a notion of action and activity to the viewer. In contrast, *Platoon* is the sequence with the most relative activity, as expressed by the high variation of visual features. This is also expected from an action scene in a war movie: a combination of short shots and high intra-shot visual variation help convey a notion of high activity and intense action.

Characterization of temporal control techniques

In this section we will characterize the properties of three different temporal filtering strategies for the estimated light source estimation. In the next section, these are compared in a user study.

The three strategies (or "settings") that we examine here are the following:

- No smoothing, or ***none***: no filtering is done on the results of light source estimation. Light source is estimated frame by frame without any sort of temporal filtering;
- Low-pass filtering with a windowed average, or ***avg***: a low-pass filter smoothes out abrupt transitions and sudden light source changes, as well as small variations that can occur on certain frames. In particular, the estimated light source of the past 20 frames (i.e., 0.8 seconds in a video with a frame rate of 25 frames per second) is averaged.
- Content-based temporal filtering, or ***auto***: as described in the temporal control algorithm section: small variations that might occur from frame to frame are smoothed out, but abrupt transitions with high amplitude are retained.

In order to visualize the different behavior of each of these temporal filtering mechanisms, we first apply each method to the light source computed for each frame of a sequence. Then, we compute a histogram of light source differences for each setting.

This histogram gives an indication of the type of variations that occur for the filtered light source for each video and helps us compare the different temporal filtering strategies. The light source difference is computed as defined by Equation (8). If r, g and b can have any value in the range $[0,...,255]$ then the maximum light source difference ΔL will be $255 \cdot \sqrt{3} \approx 441.7$. The light source difference will therefore be a value in the range $[0,...,442)$.

The distribution of light source differences ΔL in an entire movie sequence gives important insight into the dynamics of the light source: if all light source differences are small, this means that the light source changes very gradually, whereas a more homogeneous distribution would point towards a case where both small and larger changes are present. To characterize this distribution, we look at the histogram of light source differences, which is computed as follows:

$$D[k] = \frac{1}{N-1} \left| \{ t : k \le \Delta L_t < k+1 \} \right| \qquad (10)$$

where $t = 1,...,N$ for each frame in the sequence except for the first, $t = 0$ and with $k = 0,...,441$. In order to make it easier to visualize, we compute a cumulative histogram based on $D[k]$ as:

$$C[k] = \sum_{i=0}^{k} D[i] \qquad (11)$$

Note that the histogram $D[k]$ is normalized, i.e., the sum of all bins in the histogram will add up to 1. Conversely, the last bin on the cumulative histogram will also be 1, i.e., $C[441] = 1$.

		p		
Sequence	**Setting**	**75%**	**90%**	**99%**
Walk the line	*none*	6	9	54
	avg	2	5	9
	auto	2	3	6
Hellboy	*none*	5	9	52
	avg	3	6	8
	auto	2	3	7
Friends	*none*	5	9	69
	avg	3	5	8
	auto	2	3	62
Hulk	*none*	5	10	264
	avg	2	3	11
	auto	2	3	272
Wall-E	*none*	12	25	102
	avg	5	8	13
	auto	4	7	44
Platoon	*none*	15	35	284
	avg	5	8	11
	auto	5	13	270

Table 1 – Light source difference below which 75%, 90%, or 99% of all light source difference are accounted for, i.e., when compared to the graphs in Figure 2, we look for the light source difference values for which the graphs cross the 75%, 90%, or 99% point, respectively. Note that the maximum possible light source difference is 442, and that almost all of the sequences and settings saturate much earlier than that.

Figure 2 illustrates the cumulative histograms computed for each temporal control setting for two very distinct video clips: *Hellboy* and *Platoon*.

As can be clearly seen for *Hellboy,* – Figure 2 (left) – without any type of filtering (setting *none*), the cumulative histogram saturates much later than for the other two settings. This means that with *avg* and *auto*, most light source variations are simply filtered out.

When comparing this with Figure 2 (right), notice that the horizontal scale is different – light source differences in *Hellboy* are much smaller (for all settings) than for *Platoon*. It can be easily seen that the cumulative histogram for setting *avg* saturates very quickly; this is expected, as this algorithm is nothing more than a low-pass filter which cuts out any large sudden variation.

More interesting is the difference between the settings *none* and *auto*. The latter has a steeper curve for low difference values – this means that small consecutive differences in light source are simply smoothed out. However, after this initial point, the curves of *auto* and *none* are similar for higher difference values. This means that large light source differences are kept. This behavior is characteristic for the the temporal filtering technique proposed in this paper. In scenes with rather static content most variations of estimated light source are smoothed out. In scenes with very dynamic content, on the other hand, variations are sharp and pronounced, reflecting the amount of dynamics of the content on the screen.

In order to characterize the behavior of the temporal filtering techniques for the remaining sequences, we compute the lowest bin in the histograms for which a certain percentage of light source differences are found:

$$T[p] = \min\left[k : \sum_{i=0}^{k} D[i] > p\right] \qquad (12)$$

This measure is computed for the percentages of 75%, 90% and 99%, i.e., $p = 0.75, 0.9, 0.99$. Table 1 lists these values for all sequences, for the three different temporal filtering settings. Compare this to cumulative histograms like those of Figure 2: we look for the light source differences (horizontal scale) for which the curves cross a horizontal line at p=75%, p=90% and p=100%, respectively.

As can be easily seen in Table 1, for sequences with little visual variation (e.g, *Walk the Line*, *Hellboy*), the saturation point for both *avg* and *auto* settings occurs quite early, with 99% of the light source differences occurring already below bin 9 for both settings. This means that 99% of the differences as defined in Equation (8), after each of these temporal control settings were applied, are smaller than a value of 9. On the other hand, for sequences with high visual variation (in particular *Hulk* and *Platoon*), this early saturation stays low for the *avg* setting but is much higher for setting *auto*. This again reflects the characteristic of the setting *auto*, which smoothes out small variations but not large variations in light source.

Based on the analysis done in this section, we expect that the setting *none* will be appropriate for sequences which are very dynamic, because all the variations in detected light source are maintained, but won't be very useful for less dynamics scenes, for which small but potentially disturbing changes in the estimated light source will also be present in the filtered result.

We expect the setting *avg* to be appropriate for sequences which are calm, since most small and large variations are smoothed out. It will probably not work very well for dynamic scenes, particularly those with special effects such as lightning and explosions, as these will be averaged out of the filtered result.

Finally, we expect the setting *auto* to be appropriate for most sequences, both calm and dynamic, because it is able to match the dynamics of the filtered light source to the actual dynamics of the content.

In the next section we will present the results of our user study, performed to test these hypotheses.

USER STUDY

In this section we describe the user study that we have performed in order to evaluate the perceived quality of the temporal filtering method described earlier.

To test the settings, we created a system that plays a movie clip and analyses the light source of the content in real-time. The estimated light source is then filtered in one of the three ways ("settings") described in the previous section. The resulting, filtered light source color is projected into the user's living room using four Philips LivingColors lamps, creating an effect from here on designated as "surround light". In this way, the atmosphere of the movie clip is brought into the user's living room, potentially increasing the user's immersion in the content.

Two basic questions arise:

1. Does this new atmospheric context improve the users' viewing experience?
2. How do the three different settings influence the viewing experience?

Answers to these questions will help us explore the temporal filtering algorithms to further tune and develop them. For this purpose, we need to answer the following research questions:

- Do the surround light settings match the video content?
- Do the surround light settings help increase the level of immersion?
- Do the three different temporal filtering settings help increase the feeling of presence and engagement in different ways?
- Which of the three temporal filtering setting for the surround light system do users prefer?

We expect that the presence of surround light settings will improve the level of presence and engagement for the users, and as explained in the previous section, we expect the *auto* temporal filtering setting to best reflect the dynamics of the video content. Hypotheses are therefore:

1 The level of immersion while watching video with the surround light turned on is higher than without surround light.

2 The level of immersion while watching video using the *auto* temporal filtering setting for the surround light system is higher than when the other two settings (*none* and *avg*) are used.

Based on the research questions and the hypotheses, the surround light settings were evaluated with 25 participants, using a within-subjects design in which participants watched six video clips in four variations: without surround light and with the three different light settings. Additionally, before the six regular clips, an additional movie sequence from the movie "*Shrek*" was used as a training for the participant. A Presence and Engagement questionnaire [10] was used to measure the level of presence and engagement. Additionally, the participants were asked to rank their preference for the three settings.

Participants

For the user test, 25 voluntary participants from Philips Research with ages ranging from 22 to 33 (mean=26.4, sd=3.3), were recruited (11 males and 14 females). The participants were selected not to suffer from color deficiency in red and green hue. Each participant received a 5-euro voucher as a surprise at the end of the experiment.

Material

We developed a surround light system as described before, which can operate using the three different temporal filter settings described in the previous section. The three light settings constitute experimental conditions. In the control condition, no surround light is used.

The six test video clips used were described in the previous section. A seventh 30-second video clip, extracted from the movie "*Shrek*", was used to explain the procedure and to let the participants become familiar with the questionnaires.

Each video clip is presented to each participant four times. The first viewing uses the control condition (i.e., without surround light), and is followed by three experimental conditions (i.e., the three light settings) in a randomized order. The first of the seven video clips shown to the participants (the clip from '*Shrek*') is used as a test video. The test video provides a training opportunity of the experimental setting, and allows the participants to calibrate their rating scales for the following test video clips. The results from the test clip are not used in further analysis of the measurements. The order of remaining six video clips is randomized. The order in which the three light settings and the six video clips are shown to the participants was pre-edited in order to ensure a balanced distribution over the 25 participants. A script program was used to play the video clips and light settings according to the pre-edited order with a single key press by the experiment leader.

In human computer interaction (HCI), immersion [4,13], presence [8,9,13,14,19], engagement [3,9], and flow [5] are often related to the experience of interacting with virtual environments. Various questionnaires [9,10,12,19] were developed to measure immersion or engagement. Most of these questionnaires were developed for interactive virtual environments, whereas our experimental context is passive television watching. This makes some common factors in the aforementioned questionnaires not applicable for our experiment, for example, the 'Control' factor.

Although there is an ongoing scientific debate on the notion of immersion and presence (technology space or subjective experience) [12,19], our goal is to test the participants' subjective experience. The difference between the two notions lies in whether it is measured by objective parameters (e.g., the amount of the virtual space the user can interact with) or by subjective ratings of his/her own experience.

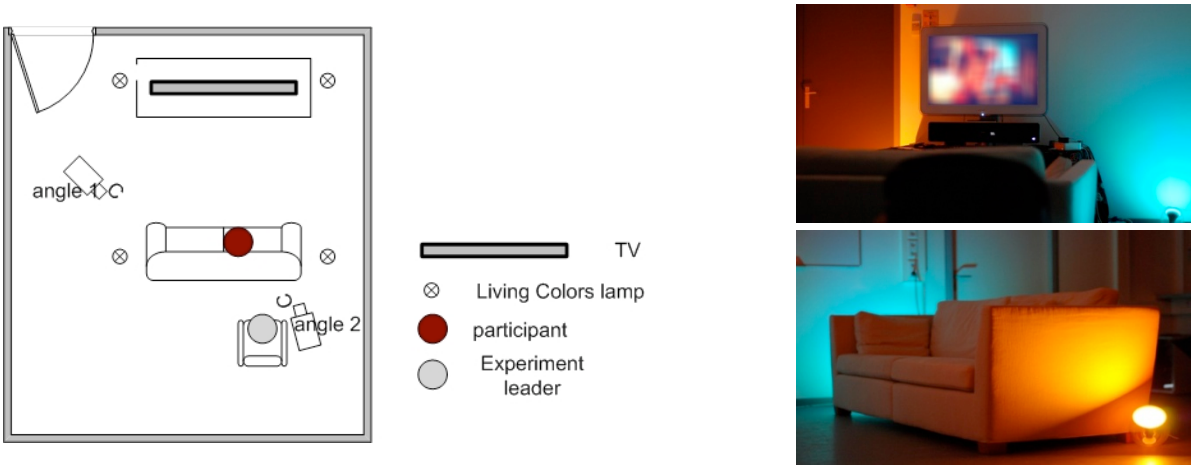


Figure 3 – (left) Map of the lab room used, (right) photo of the setup from angle (top) 2 and (bottom) 1.

For the test, we used a 13-item Presence and Engagement questionnaire[1] [10] which was originally developed for the context of 3D-TV watching, and which measures the subjective experience of the participant. The questionnaire consists of a number of questions which should be answered on a five-point Likert scale, ranging from '*strongly disagree*' to '*strongly agree*'. The question items contribute to two factors: Feeling of Presence (5 items) and Engagement (8 items). However, since the questionnaire was originally developed for 3D-TV, two items (contributing to Feeling of Presence) from the questionnaire are not applicable for our experiment. These two items are '*I had a strong sense that the characters and objects were solid*', and '*I felt I could have reached out and touch things*'. We excluded these from our questionnaires.

Based on the research questions as stated above, we would like to test whether the dynamics of the lights in the living room match the video content. We developed one question item to measure this aspect: '*The surrounding setting matches the video clip*'. This question item was not asked after the control condition, i.e., when no surround light is used.

Procedure

The test started by handing the questionnaire booklet to the participant. The participant was asked to fill in his/her personal information and TV watching experiences. The purpose of this experiment was not explained beforehand in order to avoid biasing participants, thus preventing that they would pay too much attention to the surround light. The experiment leader explained the procedure of the following steps.

The following steps encompassed seven sessions, in which the first session was a training session. Each session consisted of four sub-sessions, where one video was played with one light setting from the four variations. The starting time of each sub-session was manually controlled by the experiment leader with a remote keypad. After each sub-session, the participant had to fill in a questionnaire about the movie–lighting setting combination that he/she had just watched. In addition, he/she was encouraged to write down his/her comments on the same page of the questionnaire. When the participant finished filling in the questionnaire, the experiment leader pressed the keypad to proceed to the next sub-session. At the end of each session, the participant had to rank the light settings based on his/her preference. To make sure the participant could follow the experiment, the experiment leader (only during the first session) asked whether the questions and the procedure were clear. All participants in the experiment understood the questions and the procedure after the first session. The same process was repeated for the remaining sub-sessions.

After the seven sessions, the experiment leader had a short interview with the participant. The conversation was noted down by the experiment leader.

Data analysis

In order to avoid inconsistencies across different raters [7], a within-subject design with two factors was chosen. The two factors are: type of video and type of setting (6×4), where *Walk the Line* is used as the baseline video (as it constitutes the calmest sequence) and the *auto* setting is used as the baseline setting. The questionnaire data on the test video was not used.

On each questionnaire item, we collected 600 (6 video clips × 4 settings) data points from 25 participants. A two-way repeated measures ANOVA was used to analyze data. Main effects were analyzed by multivariate tests.

The interview was conducted in a semi-structured way. Participants were given printed posters of the movies and TV series they had watched on an A4 paper. This helped them remember the video clips they had watched and easily start the conversation. The experiment leader started the conversation by talking about the movies. This helped observe their emotional reaction to the movies and later on the light settings which were not captured in the previous

[1] This questionnaire has not yet been validated.

sessions. During the conversation, the experiment leader asked their general impression about the settings, whether any settings made them uncomfortable or distracting, and if they had any wishes to improve the settings. The experiment leader tried to ask these questions in a spontaneous way, thus not following a particular order. For example, if a participant started talking about annoying settings, the experiment leader followed by asking 'so did any other settings made you annoyed or made you feel uncomfortable?'

Presence and Engagement questionnaire data

The Presence and Engagement questionnaire we used in this experiment was comprised of eleven items, contributing to two factors: Feeling of Presence and Engagement.

On the Feeling of Presence score (range from 3 to 15), which was aggregated from three question items, all main effects: type of video ($F(6,19)=6.35$ $p=0.001$), type of setting ($F(3,22)=78.636$ $p<0.001$), and interaction type of video × type of setting ($F(18,7)=5.174$, $p<0.01$) were significant at $p<0.05$. The significant interaction effect indicates that type of setting had different effects on the Feeling of Presence score depending on which type of video was used. Figure 4 shows the mean scores for all the light settings on each video.

Paired comparisons[2] were performed comparing the three settings to their baseline setting (i.e., no light setting). It revealed that the three settings: *none* setting, *avg* setting and *auto* setting resulted significantly ($p<0.001$) higher ratings on the Feeling of Presence score than the baseline setting. Further, paired comparisons comparing the *none* and *avg* setting to the *auto* setting revealed that the *auto* setting resulted significantly ($p<0.005$) higher rating than the *none* and *avg* setting.

On the Engagement score (range from 8 to 40), which was aggregated from eight question items, significant main effects were found on: type of video ($F(6,19)=7.627$, $p<0.001$) and type of setting ($F(3,22)=16.045$, $p<0.001$). The interaction main effect ($F(18,7)=1.934$, $p=0.147$) was found not significant.

Paired comparisons comparing the three settings to their baseline settings showed that the *avg* setting and the *auto* setting resulted significantly ($p<0.01$) higher rating than the baseline setting, whilst the *none* setting did not result significantly higher rating ($p=0.601$). This can be observed from Figure 5 as well, where the mean score of *none* setting on two video clips (i.e., Walk the Line and Friends) were lower than the baseline setting.

Matching question data

An extra question 'The surrounding setting matches the video clip.' was used with the *none*, *avg* and *auto* setting.

[2] Adjustment for multiple comparisons: Bonferroni was used.

Factor	F	Sig
type of video	$F(6,19) = 6.35$	0.001
type of setting	$F(3,22) = 78.636$	<0.001
type of video × type of setting	$F(18,7) = 5.174$	0.006

Table 2 - Result of main effect analyses on Feeling of Presence score.

Factor	F	Sig
type of video	$F(6,19)= 7.627$	<0.001
type of setting	$F(3,22)= 16.045$	<0.001
type of video × type of setting	$F(18,7)=1.934$	0.147

Table 3 - Result of main effect analyses on Engagement score

Factor	F	Sig
type of video	$F(6,19)= 8.475$	<0.001
type of setting	$F(3,22)= 23.873$	<0.001
type of video × type of setting	$F(18,7)=5.027$	0.003

Table 4 - Result of main effect analyses on the matching score.

On the matching score (range from 1 to 5), all main effects: type of video ($F(6,19)=8.475$, $p<0.001$), type of setting ($F(3,22)=23.873$, $p<0.001$) and interaction ($F(18,7)=5.027$, $p=0.003$), were significant at $p<0.05$.

Paired comparisons revealed that comparing to the *none* setting and the *avg* setting, the *auto* setting had significantly ($p<0.001$) higher score. Figure 6 shows the mean scores of matching, from which one can observe that the *avg* setting had very similar matching score comparing to the *auto* setting on the first three videos with lower amount of dynamics (i.e., *Walk the Line*, *Hellboy* and *Friends*), whilst the effect of the *avg* settings became much less comparing to the *auto* setting on the next three videos with higher amount of dynamics (i.e., *Hulk*, *Wall-E* and *Platoon*).

Preference data

As introduced in the previous section, if difference among the experimental light settings was found, participants were asked to rank the three settings[3] *none* setting, *avg* setting and *auto* setting, to their preference on a ranked order scale where 1 is most preferred and 3 is least preferred. After all light settings corresponding to the video were presented, 150 groups (6 video clips × 25 participants) of orders were collected from the 25 participants, of which 141 groups

[3] The actual setting labels were replaced with the labels 'setting 1', 'setting 2' and 'setting 3', which were randomized over the video clips.

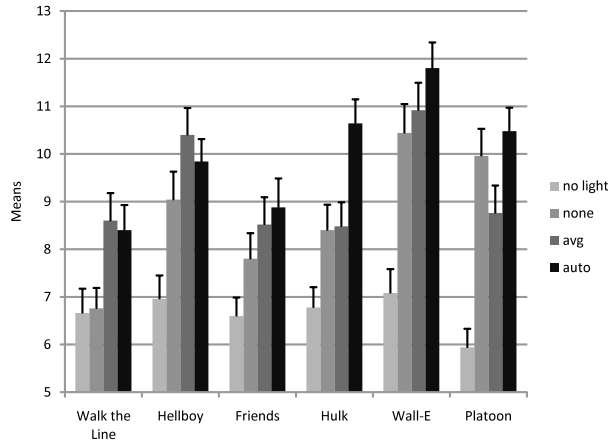


Figure 4 - Mean scores of Feeling of Presence for all the settings on each video clip.

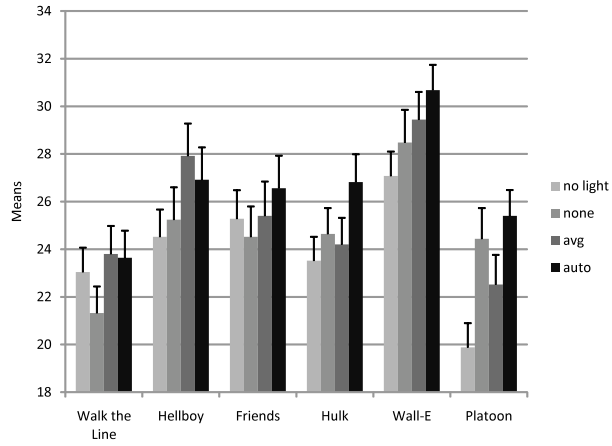


Figure 5 - Mean scores of Engagement for all the settings on each video clip

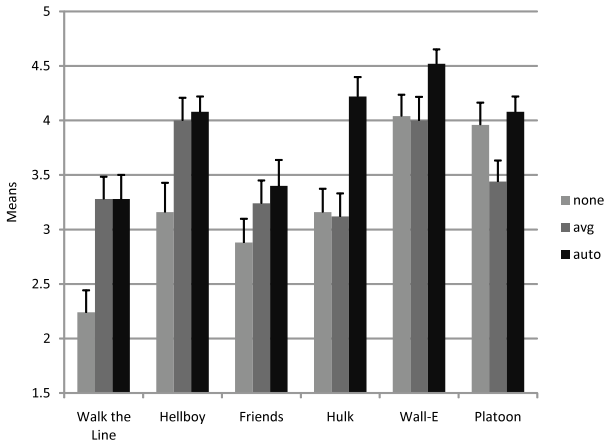


Figure 6 - Mean scores of matching for the three experimental settings on each video clip.

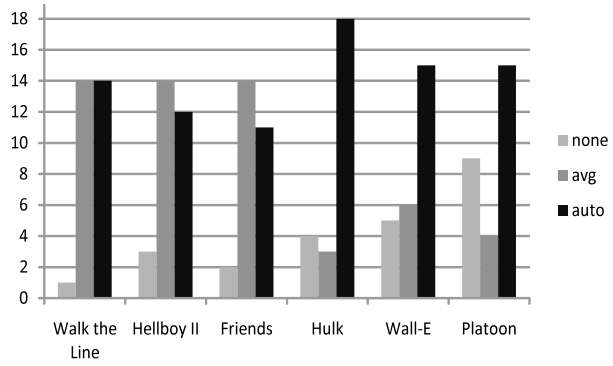
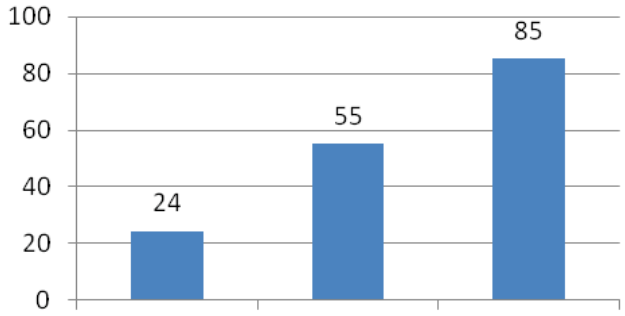


Figure 8 - Number of occurrence in the highest preference for the three experimental settings on each video.

were found different and were indicated with preference order. In general, the *auto* setting got the highest number of highest preference votes from the participants (see Figure 7).



To break down this aggregated result, we categorized it for each video (Figure 8). Figure 8 shows a trend where the *auto* setting is more preferred (when compared with the other two settings) when higher dynamics are present in video clips, whilst preference on the *avg* setting is close to the *auto* setting when lower dynamics are present. This trend is consistent with the results from the matching question item (see Figure 6).

Interview

Most participants liked the idea of surround light settings. They mentioned particularly positive impressions about the settings with cartoon movies, such as *Wall-E*. One participant expressed his feeling as:

Wall-E impressed me the best, because the settings were bright and gave me the feeling of presence in space.

Participants found that the *none* setting is distracting or annoying in most cases, but they accepted it more when it is used with a fighting or action scene, such as *Platoon*. This is confirmed with the ordering of preference shown in Figure 8. A representative comment made by one participant:

Whenever the lights flicker too much, it's very distracting, especially for static moments! However, on Platoon, the flicker at the same time as the bullet was a nice touch!

On calm scenes such as *Walk the Line*, the participants did not express much difference in preference between the *avg*

and the *auto* settings. Recalling Table 1 from an earlier section, this is not surprising since the behavior of these two settings is very similar for this particular video clip.

Discussion

The results of the test suggest that the surround light settings helped increasing the feeling of presence. The *avg* and *auto* settings increased the level of engagement, but the effect from the *none* setting was not significant. This may be explained by the comments made by the participants in the final interview, that is, the *none* setting is distracting in most cases. Moreover, the *auto* setting resulted in higher feeling of presence and engagement comparing to the *none* and the *avg* setting. Similarly, tests showed that the *auto* setting also resulted in a better matching effect than the *none* and the *avg* settings.

From qualitative analysis on the preference ordering, the *auto* setting is in general the most preferred one comparing to the *none* and the *avg* setting. However, the *avg* setting seemed to be an equally preferred setting when the dynamics present in the video clips are low.

CONCLUSIONS

We have described a novel method for temporal filtering of light source colors that are extracted from video content. The resulting dynamics of the detected light source fits much better with the content on the screen than previous methods: dynamic, action scenes or scenes with special effects result in dynamic lighting, whereas slow and static scenes result in calm lighting.

We have tested the new algorithm in a user test in which the lighting conditions from video content were projected into the living room. The results of the test show that the users liked the effect of the proposed new algorithm better than the control condition (no lights) and also better than two other tested algorithms for temporal dynamics. In addition, the user test suggests that the novel method provides increased immersion in the video content when compared to the two other algorithms or to the situation without surround light.

References

1. Barnard, K. *et al*., A Comparison of Computational Color Constancy Algorithms, Part One; Theory and Experiments with Synthetic Data, *IEEE Transactions on Image Processin*g, Vol. 11, No. 9, pp. 972-984, 2002
2. Barnard, K. *et al*., A Comparison of Computational Color Constancy Algorithms; Part Two: Experiments with Image Data, *IEEE Transactions in Image Processing*, Vol. 11. No. 9. pp. 985-996, 2002
3. Bentham, J. *The Theory of Legislation*. Oceana Publications, 1975.
4. Brown, E. and Cairns, P. A grounded investigation of game immersion. *CHI 2004*, ACM Press, 1279-1300.
5. Czikszentmihalyi, M. Flow: The Psychology of Optimal Experience, *Harper Perennial*, 1991.
6. Foley J., van Dam A., Feiner S., Hughes J., Computer Graphics, Principles and Practice, second edition, *Addison-Wesley*, Reading, MA, 1990.
7. Freeman, J., Avons, S.E., Pearson, D.E. and IJsselsteijn, W.A. Effects of sensory information and prior experience on direct subjective ratings of presence, *Presence: Teleoperators and Virtual Environments, 8, 1,* (1999), 1-13.
8. Heeter, C. Being There: The Subjective Experience of Presence. *Teleoperators and Virtual Environments, 1, 2*, (1992), MIT Press, 262-271.
9. Lessiter, J., et al. A Cross-media Presence Questionnaire: The ITC-Sense of Presence Inventory. *Presence, Teleoperators and Virtual Environments*, 10, 3 (2001), 282-297.
10. Rajae-Joordens, R. J. E., Measuring Experiences in Gaming and TV Applications – Investigating the Added Value of a Multi-View Autostereoscopic 3D Display. *Probing Experience – From assessment of User Emotions and Behaviour to Development of Products*. Springer Netherlands, 2008, 77-90.
11. Rasheed, Z. *et al*., On the Use of Computable Features for Film Classification, *IEEE Trans. On Circuits and Systems for Video Technology*, vol. 15, no. 1, Jan 2005.
12. Slater, M. Measuring Presence: A Response to the Witmer and Singer Questionnaire. *Presence*, 8, 5 (1999), 560-566.
13. Slater, M., Usoh, M. and Steed, A. Depth of Presence in Virtual Environments. *Presence, Teleoperators and Virtual Environments*, 3, 2 (1994), MIT Press, 130-144.
14. Slater, M., et al. Taking Steps: The Influence of a Walking Metaphor on Presence in Virtual Reality. *ACM Transactions on Computer Human Interaction (TOCHI) Special Issue on Virtual Reality*, September 1995.
15. Text of ISO/IEC 15938-3/FDIS Information technology – Multimedia content description interface – Part 3 Visual
16. Van den Broek, H.A. Automatic living light effect generation, *Master's thesis, Technische Universiteit Eindhoven*, Dept. of Mathematics and Computer Science, 2004.
17. Van Sijll, J, Cinematic storytelling*, Michael Wiese Productions*, 2005.
18. Wei, C. Dimitrova, N., and Change S., Color-mood analysis of films based on syntactic and psychological models, *IEEE International Conference on Multimedia and Expo (ICME)*, Taipei, Taiwan, June 2004.
19. Witmer, B. G., and Singer, M. J. Measuring presence in virtual environments: A presence questionnaire. *Presence, Teleoperators and Virtual Environments, 7,3* (1998), 225-240.