Post-processing Approach for Radiometric Self-Calibration of Video

Matthias Grundmann[1] Chris McClanahan[1] Sing Bing Kang[2] Irfan Essa[1]
[1]Georgia Institute of Technology, Atlanta, GA, USA [2]Microsoft Research, Redmond, WA, USA
http://www.cc.gatech.edu/cpl/projects/radiometric

Abstract

We present a novel data-driven technique for radiometric self-calibration of video from an unknown camera. Our approach self-calibrates radiometric variations in video, and is applied as a post-process; there is no need to access the camera, and in particular it is applicable to internet videos. This technique builds on empirical evidence that in video the camera response function (CRF) should be regarded time variant, as it changes with scene content and exposure, instead of relying on a single camera response function. We show that a time-varying mixture of responses produces better accuracy and consistently reduces the error in mapping intensity to irradiance when compared to a single response model. Furthermore, our mixture model counteracts the effects of possible nonlinear exposure-dependent intensity perturbations and white-balance changes caused by proprietary camera firmware. We further show how radiometrically calibrated video improves the performance of other video analysis algorithms, enabling a video segmentation algorithm to be invariant to exposure and gain variations over the sequence. We validate our data-driven technique on videos from a variety of cameras and demonstrate the generality of our approach by applying it to internet video.

1. Introduction

Fundamental operations for computational video, like deblurring, stereo matching and tracking have been shown to require radiometric calibration [10, 14, 15] to achieve consistent visual appearance over time. However, most cameras capture videos that first are auto-exposed to optimize the dynamic range at every frame and second are dynamically tone-mapped before encoding. Such auto-exposure and other corrections within the camera result in unreliable output for basic vision algorithms as they mostly rely on consistent appearance over time. To remove the impact of auto-exposure to a frame sequence, the camera would need to undergo radiometric calibration. In practical settings, we usually just have access to the video, *e.g.,* video obtained from the internet, with no further knowledge or access to the capturing camera. In such cases radiometric *self*-calibration is required by simply analyzing the video at hand. In the case of image/photo capture, cameras store metadata information for exposure per frame. At present, we are not aware of any video camera that stores such metadata for each frame or allows access to the uncompressed RAW sensor data for video (current generation high-end RED Cameras store compressed raw video).

Radiometric calibration recovers the camera response function (CRF), which links scene irradiance to observed RGB values given the exposure. While the mapping from irradiance to raw sensor values, known as opto-electric conversion function (OECF), is roughly linear, the subsequent demosaicing, sharpening, white balance, gamut mapping and gamma correction result in the CRF being very camera and scene specific [1, 11, 16, 24]. For competitive reasons, camera manufacturers keep the functionalities of their camera firmware secret and proprietary. At times, the CRF also incorporates some form of in-camera exposure compensation that is dependent on the specified exposure itself. For example, Nikon has a local exposure feature (called active D-lighting [23]) that actually manipulates the shadow and highlight regions; this modifies the radiometric response at the given exposure. Sony has a similar feature called Dynamic Range Optimization [25]. Furthermore, smartphone cameras employ an undisclosed amount of post-processing in software. As a result, it seems very likely that the CRF of video cameras should be regarded *time-varying*, changing with scene content and exposure.

In this paper, we propose a new, data-driven technique for radiometric self-calibration *given only the input video without meta-data*. This allows us to generate a video with consistent color appearance over time, barring loss of information due to low signal or saturation (texture/color transfer is outside the scope of our paper). Based on our empirical observations and validated by a series of experiments, we believe that the CRF should be regarded time-varying. Our technique extracts a mixture of time-varying radiometric response curves to more accurately characterize the mapping between scene irradiance and image brightness. This is in contrast to previous self-calibration techniques that rely on one global CRF.

Our contributions in this paper are as follows:

- We present a radiometric self-calibration post-process approach that works solely from video data, without access

Figure 1: Video recorded with a Canon camcorder in auto-mode (top) and our auto-calibrated result after tone-mapping (bottom). Our algorithm recovers the non-linear mapping of intensity to irradiance, effectively canceling adjustments employed by the camera over time to cover the dynamic range. For example, compare the drastic changes in the lantern's post appearance in the original video to its uniform appearance in our calibrated result. Please see the accompanying video.

to the camera. We show applicability to internet video.

- We use a window of exposures to locally compute the response curves at keyframe exposures and apply a mixture model to interpolate the curves for pixel-to-irradiance mapping. This extends our technique to streaming videos.
- We address the exponential ambiguity (*i.e.,* scene irradiance is up to scale due to lacking ground truth) by using regularization for model parameters and exposure, greatly improving stability in the estimation process.
- We evaluate the effectiveness of our approach over several sequences captured with different camera models. We quantitatively confirm constant irradiance of Lambertian surfaces after calibration.
- We demonstrate improving video segmentation using our technique.

2. Related Work

Radiometric calibration is usually performed using multiple aligned images of the same scene taken at different exposure settings. Assuming the change in exposure is known *a priori*, Debevec and Malik [2] proposed a method for recovering the radiance map and the inverse CRF represented as a non-parametric smooth mapping of irradiance to intensity. Mitsunga and Nayar [21] extended this model by approximating the CRF with a higher-order polynomial. An empirical model of response (EMoR) was introduced by Grossberg and Nayar [7]. After collecting 201 response functions for various film materials and cameras they subsequently projected them into a low-dimensional space using PCA. [7] empirically showed that the unknown CRF can be accurately represented as linear combination of the EMoR basis function. Recently, Lee *et al.*[17] introduced a new solution in this calibrated setting based on rank minimization, while Xiong *et al.*[26] proposed a probabilistic color rendering model leveraging Gaussian process regression applied to matching RAW / intensity images.

The scene dependency of the CRF for images was observed and accounted for by Chakrabarti *et al.*[1]. They also confirmed via experimentation that the OECF mapping of scene irradiance to RAW values is indeed linear. Given matching RAW / intensity images, they fit several color models (independent exponentiation after 3x3 color twist and various polynomials), empirically determining that a 5 degree per-channel polynomial performed best. Diaz and Sturm [3] proposed a model for photo collections that accounts for surface normals by leveraging recovered 3D structure from wide-baseline matches. This model is not generally applicable to unstructured video for which 3D structure can not always be reliably extracted (*e.g.,* dynamic scene or small baseline in case of a rotating camera).

Calibration without correspondence from video via histogram equalization was proposed by Grossberg and Nayer [8], at the expense of restricted camera motion. Lin *et al.* [19] propose a method for recovering the response function from a single image by linearizing edge color distributions via a Bayesian approach. A model free approach to recover the inverse response was presented by Jia and Tang [12] using 2D tensor voting. Litvinoc and Schechner [20] generalize the inverse model by separating it into response function, gain, and optical uniformity. Farbman and Lischinkski [5] proposed a correspondence free method to undo non-linear color changes, however their method does not recover irradiance values nor the calibrated CRF.

Kim and Pollefeys [13] used the inverse EMoR to radiometrically calibrate video sequences from unknown exposure values, by determining the unknown coefficients of the log-inverse response basis functions. We built upon [13]'s use of the EMoR to model the inverse CRF, generalizing it to a time-varying window of CRFs. To handle color, most efforts [2, 8] either estimate the response curve for each channel independently or assume one response curve but different illuminations for each channel to account for white balance changes [13]. We use independent estimation

of each channel and show that this is sufficient to account for white balance adjustment.

Recently, Kim *et al.*[16] proposed a novel in-camera imaging model and performed comprehensive analysis across a wide range of cameras and scenes. In particular, the imaging model describing the mapping of scene irradiance to intensity values is composed of four functions. First, the the white-point of the captured RAW image is adjusted by a diagonal 3x3 matrix, then linearly mapped to sRGB and narrowed to the sRGB gamut via a non-linear function, and finally mapped to pixel intensities by a per-channel CRF. Being a calibrated approach, various training image pairs (RAW/intensity) are taken with a known camera. The CRF is recovered for a subset of matches via [7], while gamut mapping is modeled via radial basis functions. Lin *et al.*[18] recently proposed to replace the last step using lattice regression. If RAW data is not available, aligned raw images of the same scene from a different camera are used as reference. In a video setting this requires aligned frames captured with a raw-capable camera, *e.g.,* RED, making this model not practically applicable to video. Further, [16] assume photographic reproduction mode, *i.e.,* fixed (spatially and intensity invariant) color rendering, as can be achieved with high-end cameras in manual mode. In contrast, in our calibration-free video setting, we must treat the camera pipeline as black box, mapping irradiance to intensity, without making any assumptions about the camera, making our model applicable to even low end mobile phone cameras.

In the next section, we briefly review the topic of radiometric calibration before we introduce our new mixture model of response curves.

3. Radiometric Calibration

Without the availability of raw video data, we regard the camera imaging process as a black box that maps scene irradiance of a point in a scene to three intensity values in RGB. As imaging sensors respond differently to each color [22], we model the color channels separately, which allows for compensation of changes in white balance. Here we briefly review the estimation of the radiometric response function using the empirical model of Grossberg and Nayar [7] with some modifications.

Radiometric response function: The radiometric camera response function R of a camera maps the incoming light (irradiance) to the camera sensor output after color and tone-conversion. The imaging mechanism of the camera is highly non-linear (usually more sensitive to changes in low than high intensity areas), as Grossberg and Nayar [7] showed from their collection of 201 response curves. By applying PCA to the response curves, they obtained the Empirical Model of Response (EMoR), modeling the CRF as linear combination of basis functions. Experiments showed that $5 - 10$ basis functions account for 99% of the model variance. As this greatly increases stability of the estimation by reducing degrees of freedom, we adopt their model using 7 basis functions. We validate this choice of number of basis function in section 4.

Calibration approach: We seek to find scene points of constant radiance across all frames [6, 13]. For a static camera under fixed lighting, this assumption is valid for all points. In case of a moving camera, this assumption only holds for scene points on Lambertian surfaces, even dynamic ones. We use a robust calibration method to account for outliers originating from non-Lambertian (*e.g.,* specular) surfaces (section 3.2). In general, we can track sufficient Lambertian scene points, if this assumption is severely violated, *e.g.,* flickering illumination in a night setting, our method might fail as we show in our supplemental video.

Let a video be represented by frames $(I_1, I_2, ..., I_n)$. Assuming a Lambertian scene point p, the irradiance $L(p)$ of the scene point through the lens is constant. The amount of light reaching the sensor is mostly linear w.r.t. exposure (If raw video values were available, the exact mapping would be given by the OECF, which requires a lab-setting for calibration.) As others, [6, 7, 13], we express this relationship (assuming constant aperture) as

$$L(p) = k_i \cdot L_i(p) = \text{const}, \ \forall i = 1..n, \tag{1}$$

with $L_i(p)$ being the irradiance captured at scene point p in frame I_i and k_i being a linear weight representing the inverse of the exposure value.

This enables us to recover the radiometric response curve R from intensity matches. Let x and y be two pixels in images I_i and I_j, such that x and y capture the same scene point p of a Lambertian surface. Suppose r denotes the inverse of the radiometric response curve R, mapping intensity to irradiance. Then r maps the pixels x intensity $I_i(x)$ to the irradiance of the corresponding scene p that reaches the sensor, *i.e.,* $r(I_i(x)) = L_i(p)$. Using the exposure constraint in eq. (1), the intensities of x and y are related by $r(I_i(x)) \cdot k_i = r(I_j(y)) \cdot k_j$. We linearize this relation by applying the natural logarithm to each side.

$$\log(r(I_i(x))) + K_i = \log(r(I_j(y))) + K_j, \tag{2}$$

where $K_i := \log k_i$. Denoting the log-inverse of the response function R by $l := \log r$ and the change in log-exposure $K_{i,j} = K_i - K_j$, the above constraint becomes

$$l(I_i(x)) - l(I_j(y)) + K_{i,j} = 0. \tag{3}$$

As the right hand side of the above constraint is zero, any recovered solution is only up to scale in the log-domain. This is known as exponential ambiguity [8]. Consequently, without ground truth data, we cannot determine the absolute exposure, but only the change in exposure w.r.t. to an unknown base-exposure. More importantly, if $I_i(x) \sim I_j(y)$

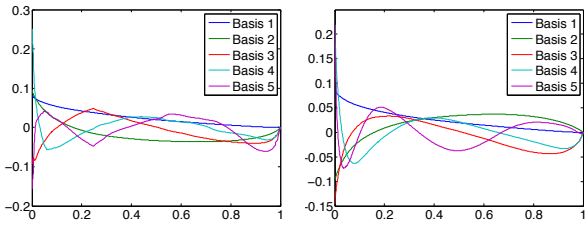


Figure 2: Left: Original PCA model [13] is not C^1 continuous. Right: Our PCA model after removing log-inverse response function with significant changes in direction.

for most pixels x, y, *i.e.,* the video is virtually uniformly exposed, the response function can not be recovered. Accounting for this inherent instability in the solution properly is crucial to us and we address this by apply regularization (section 3.3).

Similar to Kim and Pollefeys [13], we model the log-inverse CRF (which enables a linear relation as described above) using the PCA-based EMoR model [7]. In contrast to [13], we perform some crucial post-processing before applying PCA to the log-inverted response functions. We noticed that some log-inverse response curves are not C^1 continuous, due to the small gradient of many response curves near zero. As PCA is prone to model outliers and noise, we rejected all log-inverse response functions with a local change in gradient larger than 0.01. Figure 2 shows the result of this pre-filtering.

Using the log-inverse EMoR model, we can express the log-inverse response l in eq. (3) as a linear combination of known basis functions $l_0, l_1, .., l_N$ with weights c_n :

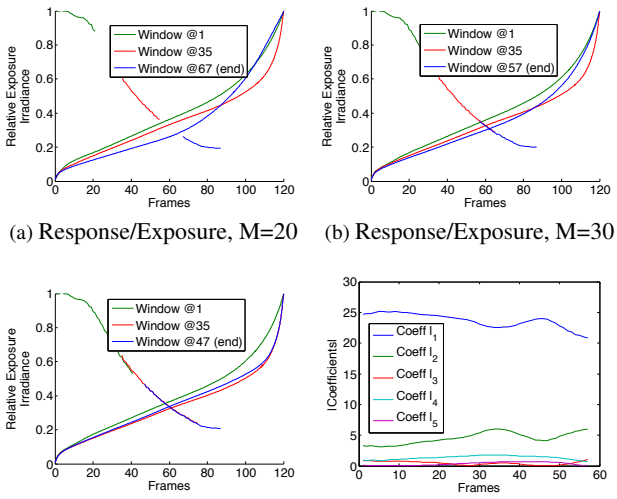
$$l_0(I_i(x)) + \sum_{n=1..N} l_n(I_i(x)) \cdot c_n - l_0(I_j(y)) - \sum_{n=1..N} l_n(I_j(y)) \cdot c_n + K_{i,j} = 0, \quad (4)$$

with l_0 being the mean of the PCA model. The above equation poses an over-constrained least-squares minimization problem, with unknowns c_n and $K_{i,j}$. The solution is again up to scale, and if $I_i(x) \sim I_j(y)$ for most x, y, the solution is numerically unstable. We address this by applying regularization as described in section 3.3.

3.1. Mixture Model of Response Curves

Previous work on self-calibration assumes that the radiometric response function is constant over time for a specific camera, regardless of its settings. However, the CRF is dominated by the scene and image dependent tone mapping function [11, 24] and recent work on radiometric calibration from matching RAW/intensity images has shown the CRF to be scene dependent [1] and non-linearly affected by gamut mapping [16]. Consequently, when recording video in auto-mode it is likely that the camera manufacturer's post-process changes during recording over time, for example, adjusting the gain, which changes the noise level function, or adjusting the exposure, which affects the response function and gamut mapping. Current Canon DSLR models also employ a low-pass filter for dust removal even before the light reaches the CMOS sensor.

To answer the question, if for practical self-calibration of video the CRF for over- and underexposed segments of a video should be regarded time-invariant, we conducted the following experiment: We recorded a static scene (shown in fig. 7) while varying the exposure setting from $+9$ to -9. Note, that by using a static scene we *avoid undue influence* of tracking errors and vignetting.



(a) Response/Exposure, M=20 (b) Response/Exposure, M=30

(c) Response/Exposure, M=40 (d) Coeff. w.r.t. offset, M=30

Figure 3: Time-varying response shown for 3 different windows. (a-c) Inverse CRFs (continuous curves) estimated within a sliding window of size M over increasing frame offsets. Corresponding exposure and inverse CRF are indicated by equal color. Intensity domain is scaled to the number of frames for visualization purposes. Notice, how the CRF varies over time w.r.t. the frame offset. (d) Coefficients for log-inverse response function over frame-offset of sliding window. Change in coefficients is smooth, justifying our mixture model approach.

We estimated the inverse response function over a sliding window of fixed size using the approach described in section 3. The CRF is estimated within each window W_i independently, however to remove undue influence due to exponential ambiguity in eq. (2) we constrain the exposure to be consistent across windows as follows: For two neighboring windows W and W' of fixed size M, starting at adjacent frames I_i and I_{i+1} respectively, we first compute the inverse response function and log-exposure values K_j for window W. Then consistent exposure for window W' is achieved by: (a) Constraining the first $\frac{M}{2}$ differences in log-exposure $K'_{j,j+1}$ for window W' to agree with the already estimated values from the previous window W: $K'_{j,j+1} := K_{j+2} - K_{j+1}$ (window W' is displaced by one frame w.r.t. W). (b) After computing the exposure values K'_j for W', we offset them by the first frame's exposure

K_i in W, therefore aligning them to the same origin. This corresponds to adding a scalar to each side of eq. (2) and represents the fact that we do not know the ground truth irradiance.

The results of this experiment are shown in fig. 3. The recovered response curves and exposure values are shown for various window sizes and frame offsets. Note, that the response function indeed varies across windows, specifically the variation is smooth w.r.t. to the basis function coefficients. Motivated by this *empirical* evidence, that the radiometric curve seems time varying, likely influencing the amount of tonal adjustment and color correction, we propose the *mixture model of response* for videos. Instead of estimating a *single* CRF, we estimate *multiple* CRFs at equidistant key-frames. We chose keyframes 15 frames apart, however we investigate this choice in section 4 and show that a mixture model consistently out-performs a single CRF model.

The coefficients of the response function in-between key-frames are given as weighted linear combination of the coefficients at the key-frames. This is motivated by the evolution of the coefficients for the above sliding window experiment shown in fig. 3d, which empirically, vary smoothly w.r.t. the frame-offset of the sliding window. Specifically, for frame I_i we denote the previous key-frame to the left as $I_{p(i)}$ and the next key-frame to the right as $I_{n(i)}$. We further assume that keyframe spacing $s := n(i) - p(i)$ is constant for all i. Then the mixture model of response is given as direct generalization of eq. (4)

$$\begin{aligned} &l_0(I_i(x)) - l_0(I_j(x)) + \\ &\textstyle\sum_{n=1..N} l_n(I_i(x)) \cdot [w(\alpha) c_n^{p(i)} + (1 - w(\alpha)) c_n^{n(i)}] - \\ &\textstyle\sum_{n=1..N} l_n(I_j(x)) \cdot [w(\beta) c_n^{p(j)} + (1 - w(\beta)) c_n^{n(j)}] + \\ &K_{i,j} = 0, \end{aligned} \tag{5}$$

where $\alpha := \frac{i-p(i)}{s}$ is the normalized distance of frame I_i to the previous keyframe $I_{p(i)}$ (similar $\beta := \frac{j-p(j)}{s}$ for frame I_j) and $w(\alpha)$ a weighting function, satisfying $w(0) = 1$ and $w(1) = 0$. Equation (5) can be optimized within the same linear system approach as eq. (4), as $w(\alpha)$ are fixed scalars for each frame. We chose the cubic-hermite spline as weight, *i.e.,* $w(\alpha) := 2\alpha^3 - 3\alpha^2 + 1$. The recovered response functions at different intervals for our the initial experiment are shown in fig. 7. Finally, by dividing the video into overlapping clips, and constraining the shared models to agree across clips, we enable our approach to be conducive for *streaming* video.

3.2. Tracking Across Multiple Exposures

We use intensity matches from sparse feature tracks, generated using the pyramidal KLT feature tracker in OpenCV. To find features across the whole intensity range of the frame

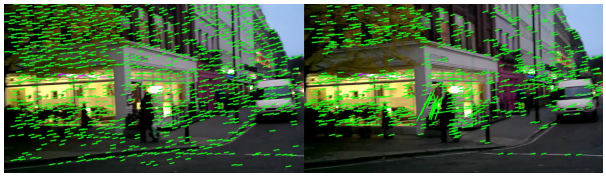


Figure 4: Left: Our grid-based feature extraction and outlier rejection, right: Standard KLT tracks.

we discritize the frame across a grid, using a local threshold for each cell. To reject outliers, we constrain the sparse flow to be locally consistent within each cell, as opposed to enforcing a fundamental matrix constraint, which might discard matches for moving foreground objects. This preprocessing removes spurious matches and inconsistent moving specular reflections, as shown in fig. 4.

If the intensity change between two neighboring frames is small, the solution to eq. (3) becomes less stable. To improve stability, we propose to use *long feature* tracks. For each feature point p_i, we track its corresponding positions $p_{i-1}, p_{i-2}, ...p_{i-N}$ in the last N frames (we use N = 6, as validated in section 4). As the change in log-exposure $K_{i,j}$ is additive in eq. (3), for intensity matches between two adjacent frame pairs (I_i, I_{i-1}) and (I_{i-1}, I_{i-2}), we have:

$$\begin{aligned} l(I_{i-1}) - l(I_i) - K_{i,i-1} = 0 \quad \text{and} \\ l(I_{i-2}) - l(I_{i-1}) - K_{i-1,i-2} = 0, \end{aligned} \tag{6}$$

using eq. (3) scaled by -1. Consequently, for intensity matches between (I_i, I_{i-2}) obtained from long feature tracks, we obtain

$$l(I_{i-2}) - l(I_i) - K_{i,i-1} - K_{i-1,i-2} = 0. \tag{7}$$

Using the EMOR model to write l as linear combination of basis functions, we can derive a similar extension of eq. (4) to multi-frame tracks.

3.3. Stable estimation using regularization

There are several options for removing the exponential ambiguity in eq. (3). One is to fix the difference in log-exposures to a predefined value (*e.g.,* for the first frame pair [13]), which in our experience requires manual adjustment for each video. Further this does not prevent the system in eq. (4) from becoming unstable in the case where a video is uniformly lit.

Instead, we propose the use of a model prior when solving eq. (5). Denoting the solution vector as $w = (c, K)$, where $c = (c_j^i)$ is the vector of all coefficients c^i for all mixtures j and $K = (K_{i,i-1})$ the vector of all changes in log-exposure between adjacent frames, the system in eq. (5) can be written as least squares problem $||A \cdot w - b||$ for appropriate matrix A and vector b. Here, b denotes the log-exposure difference w.r.t. mean l_0 of each intensity match. By computing the mean w_0 and the *inverse* covariance matrix C of

Figure 5: Qualitative outdoor example recorded with a cell phone camera. Original at the top, our calibrated result at the bottom.

the unknowns w, we can use Generalized Tikhonov regularization $||A \cdot w - b|| + \lambda||w - w_0||_C$, which can be solved using normal equations, yielding

$$w = w_0 + (A^T A + \lambda C)^{-1} A^T (b - A w_0). \qquad (8)$$

To compute the mean $w_0 = (c_0, K_0)$, we observe that the mean of the log-inverse response curves is simply obtained when all model coefficients but the DC component are zero, *i.e.,* $c_i = 0 \; \forall i > 0$. The variances of each model parameter are given by the square root of the corresponding singular value from the PCA model. For the prior of K, we compute the mean change and variance in log-intensity for each frame pair, which is equivalent to a gain-change model for adjacent frames under the geometric mean.

Besides effectively removing the exponential ambiguity, our approach has the benefit that if the right hand side b is close to zero (the video is uniformly exposed over time), our regularization reverts to the mean of the EMoR model.

3.4. Irradiance and Tone-mapping

After computing exposure changes and model parameters, we can map a video directly to irradiance values in case of video analysis, or in case of visualization employ tone-mapping.[1] For tone mapping, we follow the approach of [4]: After calibration, we compute the normalized irradiance range across the video. A bilateral filter is applied to each irradiance frame, and the frame is divided by the filtered result to obtain local contrast. Irradiance is compressed and local contrast added back, and if desired, boosted by some power larger than one. We apply conservative boosting of the contrast to highlight our calibration, however if more contrast is desired the power can be increased. Qualitative results shown in this paper are tone mapped, however for quantitative evaluation we only perform normalization to avoid undue influence of tone mapping with our error estimation. As our solution is up to scale, our tone-mapped results can suffer from a noticeable, *but constant*, color tint. To address this issue, we adopt [2] and compute the irradiance value L_c for mean intensity 128 for each color c across frames. Following the gray-world assumption, we compute the mean irradiance L across colors L_c, $c = 1..3$ and bias the log-exposure value of the first frame by $L - L_c$. Consequently, the mean intensity 128 is mapped to L across all color channels.

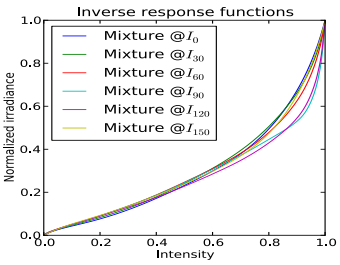
[1] Note that over/underexposed intensities may be mapped to unintended colors in the tone-mapped result, *e.g.,* as shown in the rightmost frame of fig. 1, saturated pixels were mapped to a slightly purple color.

Figure 6: Two examples on YouTube videos (Top: youtu.be/ytv5xBiawmM, Bottom: youtu.be/AyXAw5JtJlQ) Top row: Original frames, Bottom: Our calibrated result.

4. Results

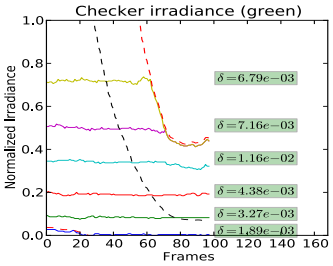
We show several qualitative tone-mapped results after auto-calibration in fig. 1 and fig. 5. We also tested our algorithm on examples we obtained from YouTube, see fig. 6. Please watch the accompanying video for more dynamic scenes and comparison to [5]. For quantitative evaluation and com-

(a) 2 Frames of original sequence, $\sim$ 2 s apart



(b) Recovered CRF

(c) Corresponding calibrated frames



(d) Irradiance of result

Figure 7: Result for static camera shot in aperture priority mode. We vary exposure compensation during recording from +9 to -9. (a) Two frames of the original sequence, $\sim$ 2 seconds apart. (b) The recovered response functions over time via our mixture model. (c) Our radiometrically calibrated result without tone-mapping. (d) The measured irradiance for the top 6 achromatic checkers after calibration and calibration error δ. Dotted red lines denotes over- and underexposure bounds, dotted grey line, the irradiance of 50% intensity. Our mixture model is able to calibrate the sequence with high accuracy (calibrated irradiance is constant within $< 1\%$ error on average). Color chart is *not used for estimation, only for evaluation* and the static sequence is free of undue influences like vignetting and tracking errors.

parison to [13], we measure how well our calibrated results (without tone mapping) respect the constant irradiance of Lambertian scene points (based on eq. (1)). To this end, we used 3 different cameras (Android phone, Nikon DSLR, and Canon camcorder) and recorded a small dataset of 10 sequences of in- and outdoor sequences, each containing a color checker chart. Note that our auto-calibration method is not aware of the presence of the checker, *i.e.,* it is not used to aid or improve the calibration. After auto-calibration, we track the checker through the sequence from its manual specified initial position.

We then measure the calibrated median irradiance (within a frame) for the top 6 achromatic checkers for each frame. We define the calibration error δ as the variance in irradiance for each checker across frames after calibration. Over- and underexposed pixels are excluded from the computation of the variance, specifically those within the immediate vicinity (2%) of the the over- and underexposure bounds (shown in dotted red in fig. 7). The over- and underexposure bounds are computed by mapping an under- and overexposure threshold (5 and 250) to the corresponding irradiance value for each frame. Values outside these envelopes correspond to irradiance values unobserved due to the camera's limited dynamic range. Our error plots also show the locus associated with mean intensity 128 as an indication of the actual exposure change.

We compare the calibration error achieved by our mixture model of response with that of a single model [13] in

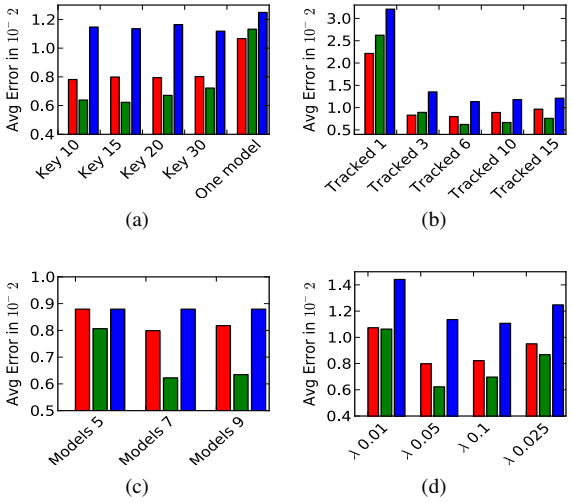


Figure 8: Average calibration error (variance of irradiance after calibration for achromatic checkers) across our dataset for colors RGB. (a) Error for mixture model w.r.t. different keyframe spacing vs. a single model as used by [13]. We chose a key-frame spacing of 15 frames, resulting in an average error reduction of 33%. (b) Including long feature tracks dramatically improves stability. Each frame is tracked w.r.t. to its 6 previous neighbors. (c) Choice of number of basis models. Adding more than 7 models does not improve results. (d) Effect of λ in eq. (8). We chose $\lambda = 0.05$.

fig. 8a. We use our implementation (with our additions of pre-filtering the EMoR model, multiple exposures and regularization) as quantitative results on video for [13] are not available. Our model consistently out-performs the single response model, reducing the calibration error by 33% on average for keyframes placed 15 frames apart. We also investigate the choice of our parameters w.r.t. the calibration error. Including long feature tracks dramatically decreases the error (fig. 8b), we chose to track each frame w.r.t. previous 6 ones for our results. We model the CRF by the first 7 basis functions obtained by applying PCA to the log-inverse EMoR dataset. Including more basis functions does not decrease the error (fig. 8c). Also note, that our regularization prevents over fitting if more models than necessary are used. Finally, fig. 8d motivates our choice of $\lambda = 0.05$.

After demonstrating empirically that the CRF should be regard time-varying in video (see fig. 3 and fig. 7 for original and calibrated result), one might ask how reproducible the change is. To this end we recorded two *different* scenes using the same camera (Canon Vixia HF100), panning to the left while varying manually the exposure compensation from +5 over -8 back to +5. As we do not measure the overall illumination and exposure compensation is adjusted manually, both videos are only qualitatively similar. Sample frames and calibrated results are shown in fig. 9. Independently of calibration, we conducted our window experiment described in section 3.1 to observe how similar the changes in response curves w.r.t. to exposure are across videos for the same camera. We show the response curves for both sequences in fig. 9 for three different window offsets, which demonstrates reproducibility.

Application: Calibrated Video segmentation We evaluate the impact of using our auto calibration method for a subsequent video analysis algorithm. To this end, we apply video segmentation to videos affected by gain change and to their calibrated result. We use the video segmentation approach of [9], and use their website to generate output for both the uncalibrated and calibrated videos. As show in fig. 10 (and in the accompanied video), prior calibration using our method greatly improves temporal consistency. For quantitative evaluation, we measure the percentage of regions that are present across all frames for the static example (fig. 10, left). Before calibration only 47.2 % of regions are present across all frames, after calibration this number is vastly improved (100 %).

5. Concluding Remarks

We have introduced a novel approach for data-driven time-varying radiometric calibration of video. We show using empirical evidence that the camera response should be regarded time varying across frames and propose a need for a mixture of responses, leading to better accuracy and consistently reducing error in mapping intensity to irradiance. We test our approach on several videos from a variety of cameras, dynamic scenes and web video. A major advantage of our approach is that it can be applied to any video from any video camera and does not require any calibration of the cameras. In addition, we demonstrated the benefit of our approach for video segmentation. As of current, our model is based on empirical evidence and the practical limitation that only the rendered intensity values are observable in video. In case uncompressed RAW video becomes ubiquitous, we plan to revisit and expand on our experiments.

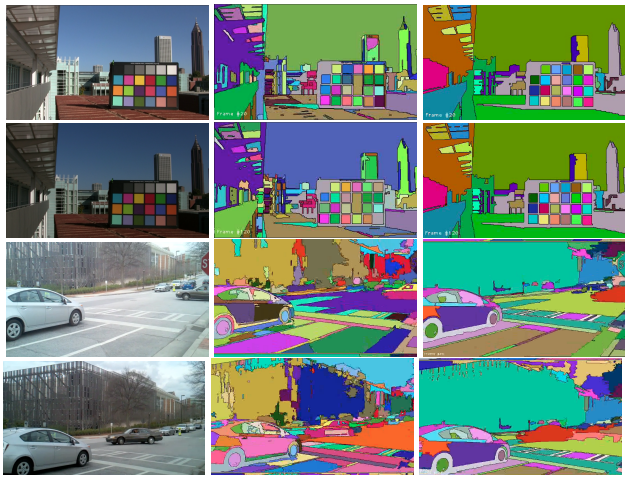


Figure 10: Improving video segmentation by prior auto-calibration. Left: 2x2 frames of the original video. Middle column: Segmented result, heavily affected by gain change. Right: Segmented result after prior auto-calibration, virtually unaffected by the gain change.

As we rely on feature tracks, our algorithm fails if tracking fails, *e.g.,* if the video is severely under- or overexposed, in areas of low texture or with significant motion blur. If lightening changes drastically, *e.g.,* flickering lights during night, our algorithm fails as demonstrated in our video. We currently do not address texture/color transfer to fill-in unobserved information in under- and over-exposed areas. For subsequent video analysis algorithms this is not necessary a limitation, as invalid data should be discarded before the analysis. Our algorithm is efficient, as we can calibrate a 5s video @20 fps in 2 min on a consumer laptop.

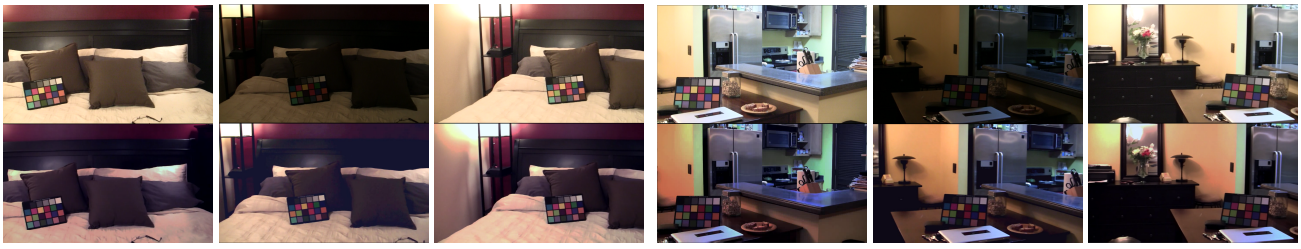
References

[1] A. Chakrabarti, D. Scharstein, and T. Zickler. An empirical camera model for internet color vision. In *BMVC*, 2009. 1, 2, 4

[2] P. E. Debevec and J. Malik. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH*, 1997. 2, 6

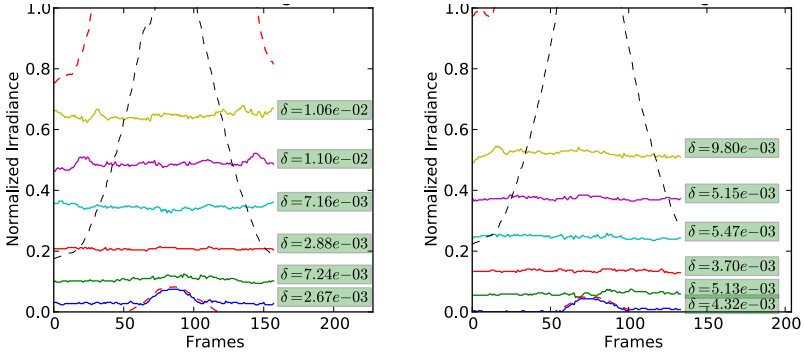
[3] M. Diaz and P. Sturm. Radiometric calibration using photo collections. *ICCP*, 2011. 2

[4] F. Durand and J. Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. *ACM SIGGRAPH*, 2002. 6

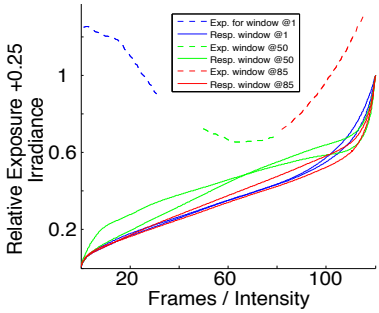


(a) First sequence, camera pan ←

(b) Second sequence, camera pan ←



(c) Error of calibrated result



(d) Windowed response

Figure 9: Moving camera example for 2 sequences (a,b) recorded with Canon Vixia 100. In both sequences camera pans to the left while exposure is changed by varying exposure compensation from +5 over -8 back to +5. (c) Error of top 6 achromatic checkers over frames. Notice that both sequences have similar exposure profiles. (d) Response and exposure independently estimated within a sliding window at three different frame offsets (indicated by color). Results are shown for *both* independently captured sequences within each window.

[5] Z. Farbman and D. Lischinski. Tonal stabilization of video. *ACM SIGGRAPH*, 2011. 2, 6

[6] D. B. Goldman. Vignette and exposure calibration and compensation. *PAMI*, 32, 2010. 3

[7] M. Grossberg and S. Nayar. What is the space of camera response functions? In *CVPR*, 2003. 2, 3, 4

[8] M. D. Grossberg and S. K. Nayar. What can be known about the radiometric response from images? In *ECCV*, 2002. 2, 3

[9] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph based video segmentation. *CVPR*, 2010. 8

[10] Y.-S. Heo, K. M. Lee, and S. U. Lee. Mutual information-based stereo matching combined with sift descriptor in log-chromaticity color space. In *IEEE CVPR*, 2009. 1

[11] J. M. Holm. Pictorial digital image processing incorporating adjustments to compensate for dynamic range differences. *US 6628823*, 2003. 1, 4

[12] J. Jia and C.-K. Tang. Tensor voting for image correction by global and local intensity alignment. *PAMI*, 27(1), 2005. 2

[13] S. Kim and M. Pollefeys. Robust radiometric calibration and vignetting correction. *PAMI*, 30, 2008. 2, 3, 4, 5, 7, 8

[14] S. Kim, Y.-W. Tai, S. J. Kim, M. S. Brown, and Y. Matsushita. Nonlinear camera response functions and image deblurring. In *IEEE CVPR*, 2012. 1

[15] S. J. Kim, D. Gallup, J.-M. Frahm, and M. Pollefeys. Joint radiometric calibration and feature tracking system with an application to stereo. *Comput. Vis. Image Underst.*, 2010. 1

[16] S. J. Kim, H. T. Lin, Z. Lu, S. Süsstrunk, S. Lin, and M. S. Brown. A new in-camera imaging model for color computer vision and its application. *IEEE PAMI*, 2012. 1, 3, 4

[17] J.-Y. Lee, Y. Matsushita, B. Shi, I. S. Kweon, and K. Ikeuchi. Radiometric calibration by rank minimization. *IEEE PAMI*, 2013. 2

[18] H. Lin, Z. Lu, S. Kim, and M. Brown. Nonuniform lattice regression for modeling the camera imaging pipeline. In *ECCV*, 2012. 3

[19] S. Lin, J. Gu, S. Yamazaki, and H.-Y. Shum. Radiometric calibration from a single image. In *CVPR*, 2004. 2

[20] A. Litvinov and Y. Y. Schechner. Addressing radiometric nonidealities: A unified framework. In *CVPR*, 2005. 2

[21] T. Mitsunaga and S. Nayar. Radiometric Self Calibration. In *CVPR*, 1999. 2

[22] J. Nakamura. *Image Sensors and Signal Processing for Digital Still Cameras*. CRC Press, Inc., 2005. 3

[23] Nikon. Active d-lighting. http://tinyurl.com/Active-D-Lighting, 2011. 1

[24] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda. Photographic tone reproduction for digital images. *ACM SIGGRAPH*, 2002. 1, 4

[25] Sony. Dynamic range optimization. http://www.imaging-resource.com/PRODS/AA100/AA100DRO.HTM, 2011. 1

[26] Y. Xiong, K. Saenko, T. Darrell, and T. Zickler. From pixels to physics: Probabilistic color de-rendering. In *IEEE CVPR*, 2012. 2