LEARNING-BASED TONE MAPPING OPERATOR FOR IMAGE MATCHING

Aakanksha Rana
LTCI, Télécom ParisTech
Université Paris Saclay

Giuseppe Valenzise, Frederic Dufaux
Laboratoire des Signaux et Systèmes (L2S, UMR 8506)
CNRS - CentraleSupelec - Université Paris-Sud

ABSTRACT

In this paper, we propose a new framework to optimally tone-map a high dynamic range (HDR) content for image matching under drastic illumination variations. This task is of fundamental importance for many computer vision applications. To design such a framework, we build a luminance invariant guidance model using a Support Vector Regressor (SVR) and learn it to facilitate the extraction of invariant descriptors from scenes subject to wide variety of appearance changes such as day/night transition. To this end, we initially generate appropriate training samples using a simple similarity-maximization mechanism. We then employ the learned model to predict optimal modulation maps that help to locally alter the intrinsic characteristics (such as shape, size) of the tone mapping function. We evaluate the proposed model performance in terms of matching score and mean average precision rate using state-of-the-art descriptor extraction schemes. We demonstrate that our tone mapping framework significantly outperforms the existing perceptually-driven state-of-the-art TMOs on the benchmark datasets.

Index Terms— Descriptor, Image matching, High dynamic range, Tone mapping operator.

1. INTRODUCTION

High Dynamic Range (HDR) [1,2] imaging captures high contrast information from the very dark and bright regions of a scene. As a result, it has brought potential interest in solving illumination-related challenges in computer vision problems such as image matching [3, 4] where performance of the algorithms degrades substantially with drastic lighting variations.

Image matching algorithms [3] look for distinctive feature descriptors that are capable of describing the detected regions and remain invariant under different transformations such as geometrical or lighting variations. Traditionally, such algorithms have been designed and optimized for low dynamic range (LDR) imagery which is represented using gamma-corrected 8-bit integer representation and is approximately linear to human perception. On the other hand, HDR images consist of real-valued pixels which are proportional to the physical luminance of the scene and are expressed in cd/m². As a consequence, HDR linear values are inappropriate when used with LDR-optimized descriptor extraction designs. In such a scenario, a simple solution opted by recent studies [5–9] is to convert HDR into an LDR representation using a Tone Mapping Operator (TMO) [1].

Tracing roots from computer graphics, TMOs have been designed to map HDR content in a suitable 8-bit LDR representation for display purposes [10,11]. For instance, TMOs such as [12,13]

The work presented in this document was supported by BPIFrance and Région Ile de France, in the framework of the FUI 18 Plein Phare project.
similarity-maximization approach to generate appropriate training samples. To this end, we define an objective function aiming to maximize the similarities of descriptors if they are extracted from images with lighting variations but from the same location. We carry out the optimization using stochastic gradient descent (SGD) [17] by deriving the required partial derivative architecture. We finally present the comparison of our approach with state-of-the-art TMOs using different descriptor extraction schemes. Our results show consistent gains in terms of overall matching scores [18] and mean Average Precision (mAP) [3] rate across different illumination conditions with respect to popular tone mapping approaches proposed in the literature.

The paper is organized as follows. In Section 2, we provide the details of our proposed approach. We present the experimental results and analysis in Section 3. Finally, the conclusions are drawn in Section 4, along with future research directions.

2. PROPOSED TONE MAPPING MODEL

2.1. Model Overview

Fig. 1 outlines the framework of our proposed algorithm. It primarily consists of a tone mapping function \( \varphi \) which maps the linear-valued HDR content of an image \( I \) to an output LDR \( I' \). More specifically, it is expressed as

\[
I'(x) = \varphi(I(x), \theta),
\]

where \( I \in \mathbb{R}^{m \times n} \), \( I' \) is of size \( m \times n \) with pixel values in the \([0, 255]\) range, and \( \theta \) represents a vector of modulation maps, \( \theta = \{ \theta_1, \theta_2 \} \), where \( \theta_i \) is of size \( m \times n \). Secondly, the framework consists of a guidance model where an SVR predicts the optimal values of these modulation maps \( \theta \) by using the densely extracted local features from the HDR content. To this end, initially, the HDR image is densely sampled into patches of size \( s \times s \) and from each such patch a SIFT feature \( f \) is extracted. Then, these features are fed to the regressor which in turn predicts parameter values for modulation map \( \theta_1, \theta_2 \). Note that the regressor output for each feature is applied over the \( s \times s \) in these modulation maps, corresponding to exact location of the sampled patch from which the feature is extracted. Such patch level tuned vector parameters \( \theta_1, \theta_2 \) are later used by \( \varphi \) to obtain the tone mapped image \( I' \).

2.2. Tone Mapping Function

Inspired by illumination normalization TMOs [12,13,19], our tone mapping function \( \varphi \) in Eq. (1) is expressed as: \( \varphi = I \cdot L^{-\psi} \), where the illumination component \( L \) is estimated by a variant of bilateral filtering [20] and is given as:

\[
L(x, \theta) = \frac{1}{W} \sum_{y \in \Omega} \mathcal{G}_{\theta_1(x)}(||x - y||) \cdot \mathcal{G}_{\theta_2(x)}(||I(x) - I(y)||) I(y),
\]

where \( \mathcal{G} \) is a Gaussian kernel and for each pixel location \( x \), the pixel \( y \) is in the neighborhood set \( \Omega \). The normalization factor \( W \) is equal to \( \sum_{y \in \Omega} \mathcal{G}_{\theta_1(x)}(||x - y||) \cdot \mathcal{G}_{\theta_2(x)}(||I(x) - I(y)||) \). Here, the modulation vector \( \theta \) has two components: \( \theta_1 \) and \( \theta_2 \). They are often globally referred to as spatial and range variance respectively and control the behavior of function \( \varphi \). For example, if \( \theta_2 \) is predicted higher at a patch location, its corresponding Gaussian kernel widens and flattens behaving like a Gaussian blur [20], and finally, a blurred luminance \( L \) is estimated. In such condition, the final tone mapped pixels, which are obtained by normalizing the estimated \( L \) for the corresponding patch, will preserve the structures such as gradients.

Notice that we opted for bilateral filtering because its proposed formulation facilitates the integration of the core concept of local modulation. However, any other tone mapping function with parametric formulations such as [10,12] could be used.

2.3. Guidance Model based on SVR

Suppose we are given a training set \( \{(f_1, o_1), \ldots, (f_n, o_n)\} \), where \( f_i \) is the feature sample and \( o_i \) represents its corresponding observation (scalar or vector), \( i = 1 \ldots n \). A classical linear regressor would solve the problem of fitting a prediction function as: \( r(f_i) = (\omega^T f_i + b) \), where \( \omega, b \) are estimated by minimizing the mean square error. However, such function is often incapable of separating the non-linearly sampled data, like our case where \( f_i \) is the SIFT feature with size 128, and \( o(i) = \theta_k(i), \) where \( k = 1, 2 \). Therefore, with such given inputs, we use the non-linear SVR [21] which maps the input vector \( f_i \) into high dimensional space using the kernel \( \psi \) where data becomes linearly separable and is given as \( r(f_i) = (\omega^T \psi(f_i) + b) \). To fit the desired non-linear SVR prediction function, the following optimization problem is solved:

\[
\min_{\omega, b, \xi, \xi^*} \frac{1}{2} ||\omega||^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)
\]

subject to:

\[
\theta_k(i) - (\omega^T \psi(f_i) + b) \leq \chi + \xi_i, \\
(\omega^T \psi(f_i) + b) - \theta_k(i) \leq \chi + \xi_i^*, \\
\xi_i, \xi_i^* \geq 0, i = 1 \ldots n
\]

where \( \xi, \xi^* \) are the slack variables, \( C \) represents the cost which is imposed for samples that exceed the error \( \chi \). For further understanding of the non-linear SVR optimization problem, we refer the reader to [21].

2.4. Generation of Samples

To train the SVR, we need to find appropriate training features and their corresponding supervised observations \( \theta_1, \theta_2 \) as shown in Fig. 2. To this end, we propose a two step solution. First, we identify key locations in a scene, where we can extract meaningful descriptor features. Second, we build a model to find the optimal \( \theta_1 \)
and \( \theta_k \) that maximize the similarity between those descriptors which are captured from the same key locations of the scene.

To identify key locations, we first detect keypoints independently in each log-scaled HDR image of the scene using the DoG [14] detector. We then iteratively check, for each detected keypoint, whether it is found at about the same location in other images of the same scene, taken under different illumination conditions. If it is detected in the majority of these images, we call it a key location. As we just want to collect ‘meaningful’ key locations with majority occurrence under lighting variations, any other format could also be used instead of log-HDR.

From each key location, we use SIFT [14] as training feature, extracted from linear HDR content. More specifically, it is given as concatenation of 16 unnormalized cells i.e., \([x_1,...x_{16}]\) where each cell can be compactly defined as [22,23]:

\[
h(\Theta)p[x] = \int \mathcal{G}_\sigma(\Theta - \mathcal{L}\nabla p(y))\mathcal{G}_\sigma(y - x)\|\nabla p(y)\|d(y)
\]

where \( x \) is the center location of the cell in the restricted square patch \( p \) of size 16 \( \times \) 16. The independent variable \( \Theta \) represents the gradient orientation ranging from \( 0, 2\pi \). Moreover, \( \mathcal{G} \) represents the Gaussian kernel with standard deviation \( \sigma \) and an angular dispersion parameter \( \delta \).

**Similarity model:** We assume a scene \( S \) consisting of \( n \) HDR images with lighting variations as shown in Fig. 2. We consider \( P = \{(1, 2), (2, 3),...\} \) to be the set of \( K = \binom{N}{2} \) pair combinations of \( N \) descriptors extracted from a key location. Our aim is to minimize the following objective function:

\[
F(\Theta) = \frac{1}{K} \sum_{(i,j) \in P} \Phi(h_i(\Theta), h_j(\Theta)).
\]

We define function \( \Phi \) using the logistic penalty, \( \Phi(h_i, h_j) = \log(1 + \exp(-h_i^T h_j)) \). We optimize the objective function in Eq. (4) using a robust optimization technique, Stochastic Gradient Descent (SGD) [17]. To estimate \( \Theta \) maps at each iteration \( t \), SGD update rule is given as: \( \Theta_{t+1} = \Theta_t - \gamma_t \cdot \nabla \Phi_{(i,j)}(\Theta_t) \), where \( \gamma_t \) is a learning rate that can be made to decay with \( t \) as \( \gamma_t = \gamma_0/(t + 1) \) and the gradient for the objective in Eq. (4) is replaced (as detailed in [17]) with the gradient of a randomly chosen sample pair \( \{i, j\} \) at time \( t \), i.e., \( \nabla \Phi_{(i,j)}(\Theta_t) \triangleq \left. \frac{\partial \Phi_{(i,j)}}{\partial \theta_k} \right|_{\Theta_t} \).

**3. RESULTS AND DISCUSSION**

**3.1. Experimental Setup**

We build the test setup for image matching using the HDR luminance dataset shown in Fig. 3 which consists of 4 indoor and 4 outdoor scenes as detailed in [8]. We compare our proposed DoTMO with the classical perception based TMOs: BTMO [19], ChiuTMO [12], DragoTMO [24], ReinhardTMO [10] and MantiukTMO [11].

The BTMO in [19] and ChiuTMO [12] are also based on normalizing the estimated luminance \( L \) but use global parametric settings. DragoTMO [24] maps the HDR content based on adaptive logarithmic scaling. ReinhardTMO [10] and MantiukTMO [11] are well known tone mapping techniques for high visual quality outputs with appealing brightness and contrast. We considered these TMOs as they have been previously applied for HDR evaluation studies [7,19] for the related task of feature detection.

To effectively evaluate the impact of descriptor extraction scheme, we selected the strongest 500 keypoints using the DoG detector [14] for each tone mapped image. Then, we use four popular and widely used descriptor schemes SURF [25] and SIFT [14], FREAK[26] and BRISK [27] (binary descriptors).

**Metrics:** We evaluated the descriptor performance using the standard measures of Matching Score and mAP rates as detailed in [3,18]. Matching Score is defined as the fraction of correct matches in the minimum of total number of correspondences in the image pair. mAP is calculated as the mean of the area under the precision-recall (PR) curves where recall is defined as the fraction of true positives over total correspondences and precision is given as the ratio of true positives to the total number of matches.

To define a match, we use the standard nearest neighbor distance ratio (NNDR) matching strategy. According to NNDR, a descriptor finds a good match if the ratio between its distance from first closest match and its distance from second closest match is less than a given threshold \( th \). Hamming and Euclidean distances are used for binary (BRISK) and non-binary (SIFT, SURF) descriptors, respectively. Two descriptors yield a true positive match if they correspond to two keypoints/regions which are indeed repeated [3] in the reference and query images. Similarly, a match is labeled as a false positive if the corresponding keypoints are not repeated. A PR curve is generated by varying the NNDR threshold.

**3.1.1. Training and Implementation details**

For each test scene, we build the training set with 5000 training samples and use it to train and validate the SVR model. Given a test scene from our dataset (Fig. 3), the training set is drawn from the other 7 scenes. For each training sample, we compute the SIFT feature on a patch size of 16 \( \times \) 16.

**Implementation:** We use the SVR implementation of LibSVM [28].
3.2. Evaluation Results

We perform a thorough evaluation of our proposed DoTMO quantitatively using the matching score and mAP. We initially show in Fig. 4 the performance of our method over all test scenes using the SURF descriptor, where we compare our algorithm with BTMO [19] and the best exposure LDR. Our results clearly show that predicted local modulation of the bilateral filtering helps in preserving the invariance of the local gradient and hence, boosts the average number of correct matches in both the indoor and outdoor scenes. However, we observe small gains in outdoor scenes such as Invalides. This can be explained by strong lighting transitions and is partially due to increased false matches due to repetitive structures in the images as shown in Fig. 7. Note that, we use threshold $th = 0.2$ to avoid ambiguous matches and to improve the readability of descriptor matching in Fig. 7.

**Comparison with popular TMOs.** We evaluate the performance of our method across different descriptor extraction schemes including both gradient based and binary descriptors. In terms of average matching score, we observe that by using every extraction scheme our DoTMO overall yields a higher number of correct matches, as shown in Fig. 5. Furthermore, in Fig. 6, we compute the mAP rates by averaging the area-under-the-curve of PR curves of the complete dataset. We observe that for every descriptor extraction scheme our proposed model outperforms all the other TMOs. Additionally, we compare our proposed TMO with popular and visually pleasing Reinhard TMO [10] and MantiukTMO [11] in Fig. 7, where we show that our method produces a higher number of correct matches in difficult day/night matching.

4. CONCLUSIONS

We propose a novel TMO approach to improve the descriptor discriminability under drastic changes of lighting conditions. To this end, we train a SVR using SIFT features to learn a model which spatially modulates the pixel-wise adaptive TMO. Further, we introduce a simple and effective method for generating the training set to learn the SVR for the given problem. We evaluate our model on our proposed HDR benchmark dataset of indoor/outdoor scenes. Our model achieves significantly better matching score and mean average precision than state-of-the-art TMOs on the HDR dataset and across different descriptor extraction algorithms. In the future, we plan to extend our model to combine detector and descriptor, and explore its usability for real-time problems such as object matching.
References


