Learning-based tone mapping operator for efficient image matching

Aakanksha A Rana, Giuseppe Valenzise, Frédéric Dufaux

Learning-based tone mapping operator for efficient image matching

Aakanksha Rana, *Student Member, IEEE,* Giuseppe Valenzise, *Member, IEEE,* Frédéric Dufaux, *Fellow, IEEE,*

***Abstract*—In this paper, we propose a new framework to optimally tone map the high dynamic range (HDR) content for image matching under drastic illumination variations. Since tone mapping operators (TMO) have traditionally been used for displaying HDR scenes, their design is suboptimal when used for computer vision tasks such as image matching. We address this sub-optimality by proposing a two-step framework, consisting of: a) a luminance-invariant guidance model based on a Support Vector Regressor (SVR) to optimally adapt the tone mapping function for image matching; and b) an energy maximization model to generate appropriate training samples for learning the SVR. At each step, we collectively address both stages of keypoint detection and descriptor extraction in the feature matching framework. By locally altering the intrinsic characteristics of the tone mapping function, the learned guidance model facilitates the extraction of local invariant features in the presence of illumination variations. We demonstrate that the proposed TMO significantly outperforms perceptually-driven state-of-the-art TMOs on a dataset of HDR scenes characterized by challenging lighting variations, such as day/night transitions.**

***Index Terms*—High dynamic range, tone mapping operator, image matching, stochastic gradient descent, machine learning.**

I. Introduction

From acquisition to display, several significant developments have been made in High Dynamic Range (HDR) imagery in the last couple of decades [3]. It has been successfully applied to several multimedia technologies, including video coding [4], inverse tone mapping [5], saliency detection [6], data hiding [7], and quality assessment [8], but also to fields such as automotive or spacecrafts imaging [9]. HDR enables to capture a wider dynamic range and color gamut, encapsulating a vast amount of information. In computer vision applications, for instance, the performance of existing algorithms degrades substantially with drastic lighting variations [10] when the scenes are captured using traditional low dynamic range (LDR) images/videos. In such scenarios, a high contrast-preserving technology like HDR can be of potential interest as it enables to draw on subtle, yet discriminating details present both in the extremely dark and bright areas of a scene [11,12], which would otherwise get lost.

In this paper, we study how HDR can be employed to solve one fundamental problem of *image matching* [13]. The latter lies at the basis of many high-level multimedia applications such as image/video search [14,15], classification [16] and localization [17]. Generally, it relies on the matching of distinctive features that are extracted from key (interesting) image locations and are invariant to geometric (scaling, rotation, etc.) and photometric variations [13,18]. Such algorithms have been designed and tuned for LDR images, which are represented using gamma-corrected 8-bit integer representation and approximately linear to human perception.

HDR images, in contrast to LDR, consist of real-valued pixels which are proportional to the physical luminance of the scene and are expressed in cd/m^2. As a consequence, HDR linear values can vary up to $10^5 cd/m^2$ on a sunny day [3], and are inappropriate when used with LDR-optimized features extraction pipelines [12,19]. Our previous studies [11,19] on keypoint detection and image matching demonstrate experimentally that using HDR linear values significantly biases the localization of keypoints towards the extremely bright areas. In such scenarios, a simple solution introduced in recent studies [1,2,11,12,19,20] is to convert HDR into an adequate LDR representation using a Tone Mapping Operator (TMO) [9], and then using conventional features extraction pipelines to perform image matching.

Classical TMOs have been designed to convert a HDR content into a suitable 8-bit LDR representation for display purposes [5,21,22]. For instance, a popular technique involves the compression of the estimated luminance (e.g., using edge preserving filters such as bilateral [23] or in the gradient domain [24]) from the HDR scenes in order to produce a visually pleasing tone mapped output. Generally speaking, existing TMOs are oriented towards preserving human-vision attributes such as brightness and perceptual contrast [3,9,25].

Differently to human visual perception, image matching is a task for machines. Its goal is not to yield a satisfying quality of experience, but rather to extract unique signatures from image locations which can be matched when the same scene is captured under different transformations. In contrast to perceptual attributes [26], such signatures are specifically designed for invariance to geometric and photometric changes. As a result, existing perceptually motivated TMOs might be sub-optimal for image matching techniques.

Several recent studies emphasize the necessity [11,12,20] and explain the requisites [27] for designing TMOs which are optimal for individual tasks such as keypoint detection. We made the first contribution in this direction and designed a detector-optimal TMO [1] controlled by a guidance model which is *learned* to understand the keypoint's locally extremal and covariant characteristics. Similarly, we also introduced a descriptor-optimal TMO where the guidance model is mainly

Aakanksha Rana is with the Laboratoire Traitement et Communication de l'Information (LTCI), Télécom ParisTech, Université Paris Saclay

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

Part of this work has been presented at the IEEE ICME 2017 [1] and IEEE ICIP 2017 [2]

trained to facilitate the invariant densely-sampled descriptor extraction [2]. However, both TMOs in [1,2] only handle one aspect at a time, namely, keypoint detection or descriptor extraction. This is inefficient in practice for the image matching task, e.g., a poor detector degrades descriptor matching [13]. Therefore, the main contribution of this paper, compared to our previous work, is to optimize TMO for the *full* features extraction chain. To our knowledge, this is the first work targeting this problem on HDR content.

Notice that optimizing a TMO considering keypoint detection and description concurrently is not trivial, as the corresponding design objectives are generally different and somehow contrasting. For instance, an optimal TMO for detection aims to produce covariant feature points, while TMOs optimal for description should guarantee some form of invariance to transformations over a local neighborhood. In addition, optimal detection requires an accurate localization of keypoint position, while optimal description is a patch-level process. In our previous work [19], we have showed that TMOs that are optimal for detection are not necessarily so when the full matching chain is considered.

In this paper, we address this problem and design an optimal tone mapping operator (OpTMO) to enhance the detection and matching of features extracted from HDR scenes captured under complex real-world illumination transitions. To this end, we initially introduce a tone mapping function which can be locally modulated by spatially varying (pixel-wise) its parameters as a function of the HDR content characteristics. Afterwards, we propose a *guidance model* to map HDR-based local characteristics features (detection and description-based) to a low-dimensional TMO parameter space, by means of a support vector regressor (SVR) [28]. In order to train this SVR-based guidance model, we further address the problem of a missing standard dataset. To this end, we compute the ground-truth parameter maps on a dataset of HDR scenes captured under drastic illumination variations. Specifically, we obtain per pixel ground-truth TMO parameters by solving an optimization problem using a stochastic gradient descent (SGD) [29] approach, which simultaneously ensures: 1) stable keypoint detection; and 2) keypoint description robust to illumination changes. Since these two objectives are, in general, non differentiable, we also propose a proxy cost function which enables to compute the required derivatives and obtain an optimal solution.

We formulate the proposed optimization framework to optimize tone mapping with respect to a popular corner detector and a gradient-based descriptor. Nevertheless, the very same design principles can be used with other detectors/descriptors. In this paper, however, the selection of detector, descriptor and the SVR-based regressor model has been motivated by state-of-the-art baselines [1,2]. We compare our proposed model with state-of-the-art TMOs using different features extraction schemes. We evaluate the performance of OpTMO at both detection and description levels. The results show consistent gains in term of overall matching scores [30] and mean average precision [13] across different illumination conditions. In addition, our results show that the choice of detector/descriptor is not critical, i.e., the obtained tone mapped images lead to improved matching performance even if a different detection/description approach is used.

In a nutshell,

- We propose a novel, locally adaptive, image-matching-optimal TMO which is guided by the SVR based predictor model. The proposed model collectively addresses the detection and description stages of the features extraction pipelines.
- We introduce an efficient method to generate appropriate training samples for learning the prediction model. Additionally, we propose a differentiable surrogate objective function which builds on the detection and description level characteristics simultaneously.
- We evaluate our proposed TMO against the state-of-the-art methodologies. Furthermore, we show an applicative scenario of object localization.

The paper is organized as follows. In Section II, we provide a brief overview of the background information. In Section III, we detail our proposed approach. We present experimental results and analysis in Section IV. Finally, conclusions are drawn in Section V, along with future research directions.

II. Background and related work

A. HDR Imagery for Computer Vision

The literature of HDR imaging applied to computer vision problems is not very vast. It is only recently that HDR imaging has been considered in computer vision applications such as keypoint detection [11,12,31], image matching [19], video surveillance [32,33] and photogrammetric applications [34]. Suma et al. [20] presented the added value of using HDR imagery and evaluated the performance of different TMOs in the context of photogrammetric applications. Rerabek et al. [35] considered the impact of HDR content on privacy protection. In [36], Korshunov et al. evaluated TMOs for face recognition applications. [31] investigated the enhanced number of local invariant features on detailed architectural scenes in HDR over LDR images. In [37], an interesting scenario of enhanced people detection and tracking in indoor HDR scenes is presented.

One commonality amongst all these studies is the use of existing perception-based TMOs. These techniques have been directly used to convert HDR images to LDR. In [11,19], we observe that the performance of such operators varies with the content as TMOs are scene-dependent [3]. In general, in those studies, no single TMO has been found to be the best for any of the considered computer vision tasks. In [27], we further investigated this problem and studied strategies for designing an optimal TMO for keypoint detection task. Our results confirmed that optimizing TMO parameters with respect to task-specific measures can improve features extraction performance.

B. Tone Mapping Operators

Tone mapping operators enable to compress the dynamic range of an HDR image to LDR, and have been mainly developed to display HDR pictures on conventional LDR

displays [21]. TMOs are broadly classified into *global* approaches, where a compression function is applied globally to all the image pixels [21,38], and *local* techniques, where a tone-mapped pixel depends on the values of neighboring pixels [22,23]. In general, global TMOs such as DragoTMO [38], which maps the HDR content based on adaptive logarithmic scaling, preserve the overall perceived contrast of the original scene. Conversely, local TMOs, such as ChiuTMO [39] and BilateralTMO (BTMO) [23], are better in conveying local structure by normalizing the estimated luminance component using filters (*e.g.*, Gaussian or the edge preserving bilateral). Other popular local TMOs includes ReinhardTMO and MantiukTMO [21,22], which yield high visual quality output with appealing brightness and contrast.

Traditionally, the performance of these TMOs have been widely studied from a perceptual point of view [25,40,41], generally for display applications. However, in this paper, we discuss TMO for the fundamental computer vision problem of image matching where the input is traditionally assumed to be an LDR image.

Fine tuning of parameters to enhance the perceived visual quality of tone mapped image has been previously explored in the TMO literature [3,42]. Mostly, such parameters were tuned either by a trial-test or grid-search based approach to yield favorable outputs for a wide variety of scenes [21–23,39]. Although some works even propose to automate the parameters selection [43], the tuned values are applied globally over the scene. However, these methods cannot be used for optimizing a TMO for tasks such as local features extraction where the parameters needs to be adapted spatially to maximize the desired local responses such as extremal cornerness response for keypoint detection or detailed gradient-level information for descriptors. This is especially important to cope with local illumination changes. Note that most existing features extraction algorithms fail in practice under drastic non-affine transformations such as day/night change [44].

C. Local Invariant Feature Extraction

Feature extraction algorithms play a critical role in several computer vision pipelines. Essentially, these algorithms comprise two stages, *i.e.*, keypoint detection and descriptor extraction. Keypoint detection methods look for covariant salient locations in a scene that can be repeatedly detected when the latter is undergoing drastic geometric and photometric transformations [18,45,46]. Later, descriptor extraction algorithms are applied to extract discriminative invariant signatures from these selected keypoint locations [13,47–51]. In this paper, we build an illumination invariant model that provides pixel-wise optimal parameter maps for the full features extraction chain.

a) Keypoint Detection: the literature on keypoint detection algorithms has been extensively explored in the past decades. Keypoint detection algorithms, in general, have been categorized in corner and blob detectors [13]. In this paper, we consider the most popular and widely used keypoint detection schemes: Harris [46], FAST [52], BRISK [53], SURF [54] and SIFT [18]. These methods are computationally fast and are widely used for real time applications such as object localization and tracking.

b) Descriptor Extraction: descriptor extraction algorithms have gone hand-in-hand with keypoint detection and have been thoroughly studied (see, e.g., [13]). In this paper, we consider the following features extraction schemes as previously used in [19,20]: BRISK [53] and FREAK [55] (corner based), SIFT [18] and SURF [54] (blob based). BRISK [53] is a computationally efficient scheme made up of a fast multi-scale detector and a binary descriptor. Its detection module is an extension of corner-based detectors such as FAST or Harris. The BRISK descriptor is a binary string computed by brightness comparisons on circular sampling patterns around the detected regions. We also consider FREAK [55] which is composed of a Harris corner detector and a binary descriptor. Similar to BRISK, FREAK also uses a concentric rings arrangement, but the sampling grid is non-uniform as inner circular rings have exponentially more points. The third extraction scheme is SIFT [18] which is a classical algorithm consisting of a blob keypoint detector (based on difference of Gaussians) and a gradient-based descriptor. The SIFT descriptor is a 128-dimensional histogram formed by concatenation of the image gradients computed on 4×4 grid spatial neighborhood around the detected keypoint. Lastly, we use SURF [54] features extraction scheme which is composed of a computationally efficient blob type detector mainly based on the Hessian matrix approximation, along with a descriptor computed as the sum of the Haar wavelet response, around the point of interest.

D. Learning models for TMO and image matching

In the case of image matching, a key problem in learning a tone mapping is generating a proper training set. In [1], we were the first to address this problem and designed an optimization model which generated parameter maps optimal for a keypoint detection task. However, the generalization of such optimization model for an image matching task is not straightforward as each stage targets different objective. Specifically, there are two major challenges: (1) designing a differentiable objective function encompassing both key stages of the features extraction pipeline, (2) acknowledging the keypoint localization dependency of the description stage.

Similar problems can be observed in designing learning-based features extraction pipelines. In [44], authors propose an end-to-end features extraction system by learning each stage individually with their respective similarity-based objectives. However, they train the models in a *sequential* manner, *i.e.*, no combined/collective objective is designed addressing each stage. This approach is appropriate to learn the features extraction pipeline. Conversely, our target is to obtain optimal TMO parameters while maximizing the efficiency of such pipelines. Hence, a similar paradigm of sequential objectives cannot be directly applied in our problem. Furthermore, the detection and description stages accuracy measures, namely repeatability rate (RR) [45] and mean average precision (mAP) [13], are two non-differentiable entities and hence, cannot be directly employed as objective functions.

Therefore, in this paper, we, firstly, employ an alternate approach to use a proxy differential objective function to

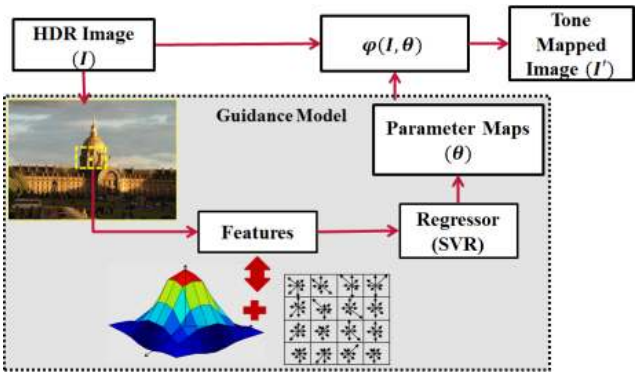


Fig. 1: Optimal Tone Mapping Design. The tone mapping function is modulated by the SVR-based guidance model, which predicts optimal parameter maps using the characteristic features.

mimic the desired behaviors. Secondly, we define an objective function with two weighted terms subjected to maximize the stability of detection response maps while enhancing the similarity of the descriptors. Additionally, we also propose to address the keypoint localization by applying a constraint on the description term.

III. Optimal Tone Mapping for Image Matching

Fig. 1 outlines the general framework of our proposed optimal TMO for image matching. It primarily consists of a tone mapping function φ which maps the linear-valued HDR content of an image I to an output LDR I'. Secondly, the framework consists of a guidance model where a learned SVR predicts an optimal parameter map $\boldsymbol{\theta}$ based on local HDR content characteristics. In the following subsections, we discuss the design of our proposed framework and how to generate training data for the SVR.

A. Optimal Tone Mapping Design

Let φ be a tone mapping function which maps the linear-valued HDR content of an image I to an output LDR I'. In general, for each image pixel x, the TMO operates as:

$$I'(x) = \varphi(I(x), \boldsymbol{\theta}), \tag{1}$$

where $I(x) \in \Re$, $I'(x) \in [0, 255]$. The $\boldsymbol{\theta}$ represents a set of parameters, given as $\boldsymbol{\theta} = \{\theta_1, \theta_2, .., \theta_h\}$ where h is the number of parameters.

The parameter count h typically varies depending on the TMO [21,22,38,39]. One example is ChiuTMO [39], where $\boldsymbol{\theta}$ contains one parameter only, *i.e.*, the variance of the Gaussian kernel which control the estimation of global lighting component from the scene. Other examples include sharpening constant in ReinhardTMO [21] and range and spatial variance in bilateral filtering based TMO [27]. Conventionally, each of these parameters is just a scalar value and is often tuned globally by cumbersome trial-test procedures to produce visually pleasing output images.

In this paper, we not only propose to exploit the potential of local tuning of these parameters for image matching problem but also to automate them by proposing a learning mechanism. We assume the function φ as an extension of existing tone mapping functions which can be modulated spatially by adapting their parameter $\boldsymbol{\theta}$ locally (pixelwise). We will define these adaptive parameters as parameter maps, as shown in Fig. 2.

The basic idea of this work is to facilitate the local adaption of the function φ at sparse keypoint locations so as to further ease their identification, and also to preserve the unique gradient-based local signature in the surrounding of a region, so as to aid the extraction of invariant descriptors.

To automate the prediction of these optimal set of parameter maps, we propose to learn a model by employing SVR [28], which minimizes the non-linear problem of predicting $\boldsymbol{\theta}$ by linearly separating the input samples in a high-dimensional space by using kernel mapping. The SVR model is learned while complying with the following three desired constraints: (1) to distinguish and localize a keypoint from its neighborhood locations; (2) to preserve local gradient orientation patterns around the keypoint; and (3) to bring invariance (as much as possible) to non-affine lighting variations in physical world scenes.

B. Generation of Training Set

In this section, we address the problem of generating an adequate ground truth for training the SVR-based model. We aim to find such ground truth parameter maps $\boldsymbol{\theta}$, which result in efficient image matching (*i.e.*, mAP score) for a scene which undergoes drastic lighting variations, as shown in Fig. 2. In this section, we, therefore, formulate an objective function f, which we minimize over the $\boldsymbol{\theta}$ to yield the optimal parameter maps. The proposed total energy f represents the difference in the image matching pairs. We quantify this difference in terms of both keypoint detection and descriptor extraction stages, depicted as 'Detection Response' and 'Description' in Fig. 2. Finally, we propose to optimize the objective using the SGD based optimization method to obtain the optimal $\boldsymbol{\theta}$.

In the following, we first discuss the formulation of the objective function f. Then, we detail the considerations with respect to image matching components in view of designing the objective function f. Finally, we detail the SGD-based method to optimize *the objective* to obtain the optimal $\boldsymbol{\theta}$.

1) Objective Function: We aim to optimize $\boldsymbol{\theta}$ to tone map an image for the full features extraction pipeline. Therefore, the objective function should consolidate each stage of the features extraction pipeline *i.e.*, to locate and extract the features. Henceforth, we introduce two energy terms dedicated to keypoint localization (E_{det}) and descriptor extraction (E_{des}), respectively and define combined objective function as:

$$\underset{\theta}{\text{minimize}} \quad f(\theta) = E_{det}(\theta) + E_{des}(\theta), \tag{2}$$

where each energy term is computed over a scene consisting of N HDR images with lighting variations as shown in Fig. 2 (a). We denote $P = \{(1,2),(2,3),\dots\}$ the set of $K = \binom{N}{2}$ pair combinations of N images. The E_{det} term aims to ensure the covariance of the corner response maps. Conversely, the E_{des} term helps in retaining the invariance of the discriminative patterns around the *key* locations in the image pairs when

undergoing drastic transformations. Both terms are detailed as follows:

2) E_{det}: To ensure efficient matching, we observe that it is important to enforce the similarity in detection response maps [27]. This is mainly because highly similar response maps increase the probability of detection of keypoints at similar locations and thereby enhance the probability of correct matches.

We define the detection similarity term E_{det}, by summing the penalty computed from each pair in the set K, as:

$$E_{det} = \frac{\lambda_{det}}{K} \sum_{\{i,j\} \in P} \mathcal{C}_1(\mathcal{R}_i(\theta), \mathcal{R}_j(\theta)). \tag{3}$$

For each sample pair $\{i, j\} \in P$, we penalize the response maps dissimilarity by a logistic cost function given as:

$$\mathcal{C}_1(i, j) = \log(1 + \exp(\epsilon_c - \langle \mathcal{R}_i \cdot \mathcal{R}_j \rangle)), \tag{4}$$

where ϵ_c is the penalty control factor, $\mathcal{R}_i$ and $\mathcal{R}_j$ are the response maps corresponding to the images i, j and $\langle \cdot \rangle$ denotes the scalar product. The selection of $\mathcal{R}$ is detailed later in this section.

Inspired by the max-margin formulations applied to retrieval [15] or classification tasks [56], we use the logistic function as the penalty in our detection term. It is a smooth differential operator and ideally penalizes less if there is high similarity and vice-versa. Note that the term E_{det} is somewhat similar to the one we proposed in the detector optimal TMO in [1]. But, in this paper, we include an additional factor λ_{det} which weights the penalization corresponding to detection.

a) Selection of $\mathcal{R}$: From handcrafted [45] to deep-learning [44] era, the concept of corner-like keypoint detection methods has gained popularity for low-latency vision tasks due to high speed, less computational complexity and competitive accuracy. By definition, corners exhibit low correlation with neighboring pixels in all directions. The most basic and widely adopted corner detectors [46,57,58] localize the extrema primarily by computing the per pixel gradient autocorrelation matrix, given as:

$$\boldsymbol{M} = \begin{bmatrix} I_x^2 & I_{xy} \\ I_{yx} & I_y^2 \end{bmatrix}, \tag{5}$$

where each component represents the directional derivative. Thereafter, different methods are proposed in the literature to localize the extrema "keypoints" [45]. In this paper, we use [46] which describes the response for each pixel $\mathbf{x}$ without directly computing the eigenvectors of $\boldsymbol{M}$ as:

$$\mathcal{R}(\mathbf{x}) = det\{\boldsymbol{M}(\mathbf{x})\} - k \cdot tr\{\boldsymbol{M}(\mathbf{x})\}^2, \tag{6}$$

where k is tuned empirically.

Similar to the baseline [1], we employ the detector response in Eq. (6), mainly because it is based on the popular structural matrix $\boldsymbol{M}$, which is simpler to differentiate than alternative approaches, thus aiding in backpropagation. Note that alternate detection methods could also be used, but our choice has been made entirely based on the computation complexity and ease of use in backpropagation.

3) E_{des}: The energy term E_{des} aims to penalize the dissimilarity of the descriptors extracted from the tone mapped images. Previously in [2], we proposed a densely sampled patch-based method where a model is learned to predict global parametric values for an individual patch. Hence, not only the method optimized $\boldsymbol{\theta}$ for a patch *globally*, but it also lacked the consideration of keypoint localization. In contrast, the image matching pipeline additionally relies on the localization of the descriptors. Hence, in this paper, we argue that it is important to compute the gradient orientation impact per pixel and to focus on its locations prior to designing a descriptor-based penalty function. It not only helps in preserving the salient locations but also avoids any "look-alike" redundant matches [2]. Therefore, we propose to constraint the penalization to the dissimilarity of those descriptors that belong to some potential *keypoint* region. We define E_{des} as:

$$E_{des} = \frac{\lambda_{des}}{K} \sum_{\{i,j\} \in P} \mathcal{C}_2(\mathcal{D}_i(\theta) - \mathcal{D}_j(\theta)), \tag{7}$$

where $\mathcal{C}_2$ is the Euclidean distance and λ_{des} is a weighting factor. To apply the constraint in practice, we compute the descriptor $\mathcal{D}$ after the keypoint localization which is obtained by applying the softargmax operation $\mathcal{S}$ [59] on the resulting response map. In general terms, $\mathcal{S}$ is given as

$$\mathcal{S} = \sum_i \frac{\exp(\beta z_i)}{\sum_j \exp(\beta z_j)} \cdot i \tag{8}$$

where z_i is the pixel location and β is a hyper-parameter for defining the shape parameter. The softargmax operation is a differentiable function to obtain local optima and helps in avoiding the cluttering in response maps. Cluttering refers to a phenomenon when several keypoints are located close to each other [2].

To compute an accurate keypoint localization, we define the final gradient orientation around each pixel location as follows:

$$\mathcal{D} = \begin{cases} h(\nu|p), & \text{if } \mathcal{S}(\mathcal{R}) \geq \Lambda \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where $h(\nu|p)$ is the gradient orientation feature map explained later in Eq. (10) and Λ is the maximum softargmax value in a 16×16 neighborhood window of the considered pixel. It simply means that if the softargmax response score for the considered pixel location is maximum in its neighborhood window, only then the gradient orientation map is taken into account to contribute in the final descriptor-based penalty term in Eq. (7).

a) Selection of h: A common image matching approach relies on the similarity of features extracted from patches corresponding to detected keypoint locations. One widely used descriptor extraction algorithm is the Scale Invariant Feature Transform (SIFT) [18] which is a concatenation of 16 unnormalized cells *i.e.*, $[\mathbf{c}_1, ...\mathbf{c}_{16}]$, where each cell can be compactly defined as [60,61]:

$$h(\nu|p)[\mathbf{c}] = \int \mathcal{G}_\delta(\nu - \angle\nabla p(y)) \mathcal{G}_{\hat{\sigma}}(y - c) \|\nabla p(y)\| d(y), \tag{10}$$

where $\mathbf{c}$ is the center location of the cell in the restricted square patch p of size 16×16. The independent variable ν

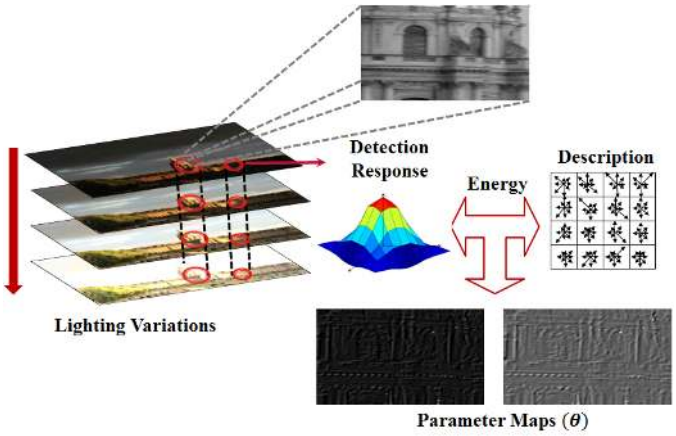


Fig. 2: Generation of Training Set. Ground-truth parameter maps are generated by minimizing the total energy determined from a set of images of the same scene, undergoing lighting variations, using the procedure in Section III-B.

represents the gradient orientation ranging in $[0, 2\pi]$. Moreover, $\mathcal{G}$ represents the Gaussian kernel with standard deviation $\hat{\sigma}$ and an angular dispersion parameter δ. Once histograms are computed, they are normalized and concatenated into a single 128-dimensional descriptor. Finally, the distance between the resulting descriptor can measured using the ℓ_2 metric.

4) Stochastic Gradient Descent Implementation Details: We optimize the objective function in Eq. (2) using Stochastic Gradient Descent (SGD) [29]. It is a fast and robust optimization technique to estimate the incremental gradient descent by its stochastic approximation using a randomly chosen sample from the initial set. To implement the SGD optimization, we follow the backpropagation procedure. We initially build the required partial derivative framework with the objective function given in Eq. (2). It is more formally expressed as

$$\nabla \boldsymbol{\mathcal{C}}_{\{i,j\}}(\boldsymbol{\theta}) = \left\{ \frac{\partial \mathcal{C}_1}{\partial \mathcal{R}_l} \cdot \frac{\partial \mathcal{R}_l}{\partial \varphi_l} \cdot \frac{\partial \varphi_l}{\partial \boldsymbol{\theta}} + \frac{\partial \mathcal{C}_2}{\partial \mathcal{R}_l} \cdot \frac{\partial \mathcal{R}_l}{\partial \varphi_l} \cdot \frac{\partial \varphi_l}{\partial \boldsymbol{\theta}} \right\} \Big|_{l=i,j} \tag{11}$$

Then, following the SGD rule, we iteratively estimate $\boldsymbol{\theta}$ by randomly selecting sample (i, j) from the set P. Finally, we compute the gradient of the objective in Eq. (2), that is:

$$\frac{\partial f}{\partial \theta}\Big|_{\theta_t} = \frac{1}{K} \sum_{\{i,j\} \in P} \frac{\partial \boldsymbol{\mathcal{C}}}{\partial \theta}\Big|_{\theta_t} \tag{12}$$

where

$$\boldsymbol{\mathcal{C}} = \lambda_{det} \cdot \mathcal{C}_1(\mathcal{R}_i, \mathcal{R}_j) + \lambda_{des} \cdot \mathcal{C}_2(\mathcal{D}_i, \mathcal{D}_j), \tag{13}$$

with the single (i, j) selected image pair. Thereafter, at each iteration t, SGD update rule is given as:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \boldsymbol{\gamma}_t \cdot \nabla \boldsymbol{\mathcal{C}}_{\{i,j\}t}(\boldsymbol{\theta}_t), \tag{14}$$

where γ_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 function in Eq. (2) is replaced by the gradient of a randomly chosen sample pair $\{i, j\}$ at time t, *i.e.*,

$$\nabla \boldsymbol{\mathcal{C}}_{\{i_t,j_t\}}(\boldsymbol{\theta}_t) \triangleq \frac{\partial \boldsymbol{\mathcal{C}}(\mathcal{R}_{i_t}, \mathcal{R}_{j_t}, \mathcal{D}_{i_t}, \mathcal{D}_{j_t})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_t}. \tag{15}$$

For SGD-based optimization, we start from a randomly initialized set of $\boldsymbol{\theta}$ which are updated iteratively using the update rule in Eq. (14). In total, the model comprises 3 hyperparameters: $\gamma_0, \lambda_{det}, \lambda_{des}$. To estimate these hyperparameters, we follow the standard approach used in [62] and take a small set of pairs from P and perform a simple cross-validation using the grid search method in the log scale. For the SGD related optimization and convergence proofs along with the asymptotic analysis, we refer the reader to [29].

This proposed mechanism for finding the optimal parameters $\boldsymbol{\theta}$ for a function φ using SGD is generic, *i.e.*, one can easily tune the parameter maps of any TMOs that can be expressed as Eq.(1). In this work, we propose to learn the local spatial and range variance of the bilateral filtering based tone mapping which is described in the following subsection. Note that our proposed OpTMO will be a learned local adaption of bilateral filtering based tone mapping BTMO [23,27].

C. Selected Tone Mapping Operator

Many tone mapping approaches aim at separating scene illumination, which can display large dynamic range variations, from the reflectance of objects, which instead has lower dynamic range characteristics [27,39]. Following this idea, we consider a tone mapping function φ, expressed as: $\varphi = I \cdot L^{-1}$. The illumination component L is estimated by an adaptive version of bilateral filtering [63] and is given as:

$$L(x, \boldsymbol{\theta}) = \frac{1}{W} \cdot \sum_{y \in \Omega} \mathcal{G}_{\theta_1(x)}(\|x-y\|) \cdot \mathcal{G}_{\theta_2(x)}(\|I(x)-I(y)\|) I(y), \tag{16}$$

where $\mathcal{G}$ is a Gaussian kernel. The parameter map vector $\boldsymbol{\theta}$ has two components, θ_1 and θ_2, also known as spatial and range variance. For each pixel location x, y is a pixel in the neighborhood Ω of x. The normalization factor is given as:

$$W = \sum_{y \in \Omega} \mathcal{G}_{\theta_1(x)}(\|x - y\|) \cdot \mathcal{G}_{\theta_2(x)}(\|I(x) - I(y)\|). \tag{17}$$

D. Support Vector Regressor Training

SVR [28] is a learning-based algorithm to estimate the unknown functions which map the input samples into a high dimensional space where the data becomes linearly separable. Consider the sample set of characteristic features $\mathcal{F} = \{\boldsymbol{f_1}, \ldots, \boldsymbol{f_n}\}$ and the corresponding output denoted by $\mathcal{Y} = \{\theta_{k(1)}, \ldots, \theta_{k(n)}\}$ where $k = 1, 2$ in our case. To build our predictor model, we want feature samples which capture distinctive information for both descriptor and detector. To that end, we build our feature sample $\boldsymbol{f_i}$ by concatenating two parts: a) the gradient-based SIFT pattern [18], 64 dimensional feature; and b) the 5×5 grid-based detector response feature [1], 25 dimensional feature. This forms a total dimension of 89. The features $\boldsymbol{f_k}$ are computed from the original HDR linear values, without any processing. This is not contradictory with the need to perform a TMO as, locally, HDR images generally display limited dynamic range [12]. Finally, for each training sample, we get the following input-output corresponding pairs $\{(\boldsymbol{f_1}, \theta_{k(1)}), ..., (\boldsymbol{f_n}, \theta_{k(n)})\}$ and

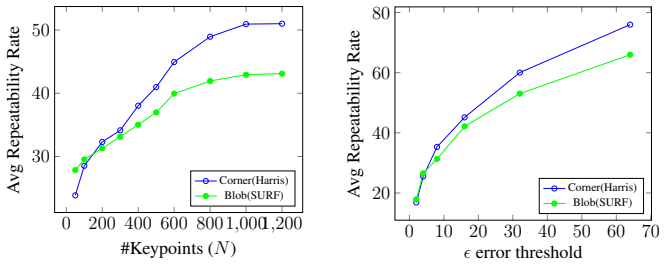


Fig. 4: Repeatability Rates (RR) computed for OpTMO using a corner (Harris) and a blob (SURF) keypoint detector.

formulate our prediction problem using SVR. To fit the desired nonlinear SVR prediction function, the corresponding optimization problem is solved using the dual maximization approach. For further details on SVR, we refer the reader to [28].

IV. Experimental Results and Discussion

A. Dataset

We consider the HDR dataset presented in our previous work [1], which is composed of 8 different HDR scenes as shown in Fig. 3. The Light Room, Project Room and Poster have been used for evaluating HDR for keypoint detection problems [11,64] [1]. 4 outdoor scenes have been captured at famous locations in Paris: Notre-Dame, Louvre, Invalides and Grande Arche, at different hours of the day with a Canon 700D camera. The Camroom indoor scene is shot with a Canon Mark III camera in the presence of powerful 2K Watt reflectors. HDR images have been created by fusing multiple exposure LDR images using the algorithm in [65]. All scenes have been geometrically calibrated for image matching evaluation.

B. Evaluation Metrics

We evaluate the keypoint detection and descriptor extraction performance on the tone mapped images using the standard measures of Repeatability Rate (RR) and Matching Score (MS) respectively, as detailed in [13,30]. For the evaluation of the full image matching, we compute the mean average precision (mAP) scores [13].

RR is a measure of detector efficiency, defined as

$$\frac{r_{ref}(\epsilon)}{\min(n_{ref}, n_{test})}, \tag{18}$$

where r_{ref} is the number of keypoints detected in the reference image which are *repeated* in the test image, and n_{ref} and n_{test} are the number of detected keypoints in the reference and test image, respectively. A keypoint is considered to be *repeated* in the test image if: a) it is detected as a keypoint in the test image, and b) it lies in a circle of radius ϵ centered on the projection of the reference keypoint onto the test image.

MS is defined as the fraction of correct matches to the total number of correspondences in the image pair. To define a match, three different matching strategies have been discussed in [13]. In nearest neighbor (NN) matching, a descriptor A finds its matches B only if A is the nearest neighbor to B and if the distance between them is below a threshold. Nearest neighbor distance ratio (NNDR) extends NN by introducing a threshold to the ratio of the distance descriptors. More precisely, a descriptor finds a good match if the ratio between its distance from the first closest match and its distance from the second closest match is less than a given threshold th. These distances depend on the descriptor type, *i.e.*, Hamming distance metric is used for binary descriptors and Euclidean distance is used for non-binary descriptors. In this paper, we have used NNDR matching strategy to compare the performance of our TMO with other techniques.

To define a correct match, feature location is taken into account. Two descriptors yield a true positive match if they correspond to two keypoints/regions which are repeated [13] in the reference and query images. Similarly, a match is labeled as a false positive if the corresponding keypoints are not repeated.

MS gives only the estimate of correct matches, while in practice, many incorrect matches may occur. Therefore, for completeness, we also evaluate the performance using mAP score. To this end, we generate a Precision-Recall (P-R) curve by varying the matching strategy parameter th from 0 to 1. 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. Once the P-R curves are generated for each scene, we then compute the mAP scores by determining the area under the curves.

C. Evaluation Setup

We test our proposed OpTMO for image matching task on 8 HDR scenes (shown in Fig. 3) at detection and description levels and compare with state-of-the-art TMOs. The HDR dataset is composed of a total of 52 images. For detection and description stage, we formulated a total of 280 test image pairs respectively from the 8 scenes.

We compare the proposed OpTMO with classical perception-based TMOs, including: BTMO [23], ChiuTMO [39], DragoTMO [38], ReinhardTMO [21] and MantiukTMO [22]. We consider these TMOs as they have been previously applied for HDR evaluation studies [20,27] for similar keypoint detection task. In addition, we also consider our previously proposed DetTMO [1] and DesTMO [2], which are optimized methods for detection and description only, respectively.

SVR Training and Implementation: We use the SVR implementation of LibSVM [66] using the Radial Basis Function (RBF) kernel. The optimal values of SVR parameters, the regularization cost and epsilon, are obtained by 10-fold cross validation from the range of $[2^{-5}, 2^{15}]$ and $[2^{-10}, 2^{5}]$, respectively.

To train and validate the SVR model, we build the training set with 5000 sample feature set for each test scene. This training set is drawn from other scenes excluding the corresponding test scene. For instance, to test the Project Room scene, we build the training set by randomly selecting samples from all

[1]Light Room and Project Room dataset can be downloaded from http://webpages.l2s.centralesupelec.fr/perso/giuseppe.valenzise/download.htm

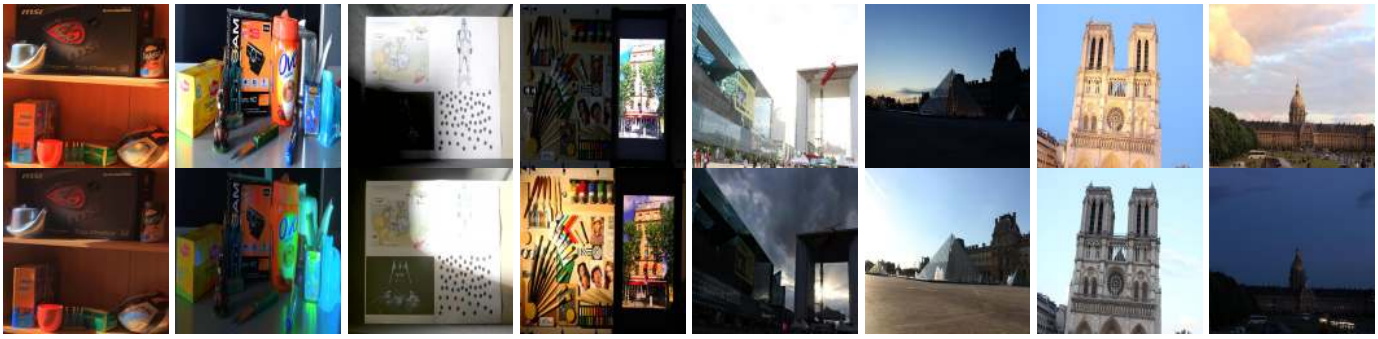


Fig. 3: Sample images from *HDR dataset*, composed of 8 scene from different indoor/outdoor locations, taken with different artificial/natural lighting variations.

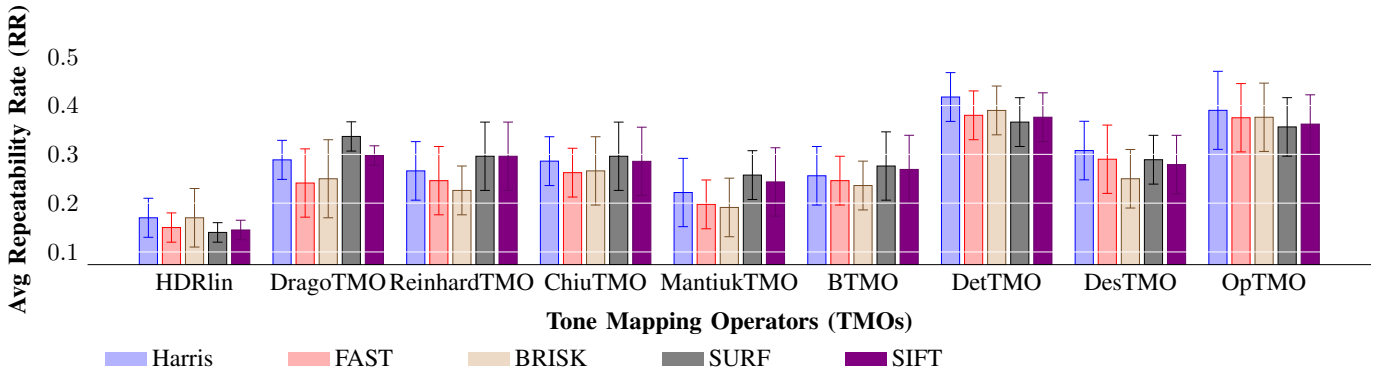


Fig. 5: **Keypoint Detection I**: Average Repeatability Rates (AvgRR) computed on different TMOs using various keypoint detection schemes. The average is calculated over all test scenes.

other 7 scenes. For each training sample, we randomly select a pixel location and compute characteristic features around the selected location, while following the features extraction procedure described in Section III-D. A window of 16×16 is selected around the pixel location for computing the gradient based SIFT pattern part of the feature $\boldsymbol{f_i}$, whereas a 5×5 grid based detector response is used for the second part of $\boldsymbol{f_i}$.

D. Keypoint Detection

We evaluate all the considered TMOs using Harris [46], FAST [52], BRISK [53], SURF [54] and SIFT [18] (as detailed in Section II-C). We selected these detection methods based on state-of-the-art studies in evaluating the performance of TMOs [2,11,64,67] and also due to their popularity in real time applications [68].

The RR is the performance measure as given in Eq. (18). RR [13] is sensitive to the number of detected keypoints and the error rate ϵ. For instance, large variations in the number of keypoints across different scenes might lead to biased average scores. Therefore, we fix the keypoint detection to the strongest N keypoints as suggested in prior TMO evaluation studies [11,12,19]. The impact of N and ϵ over average RR score is shown in Fig. 4. Overall increase in number of keypoints leads to an increase in average RR, but the growth slows down after a certain number, partially due to the detection of cluttered keypoints. On the other hand, increase in the average RR with the increasing ϵ is in coherence with the findings of [45]. In this paper, we choose the values $N = 500$ and $\epsilon = 10$.

Implementation: We use the HDR Toolbox [42] for the implementation of the considered TMOs. Moreover, we use the Matlab's Computer Vision toolbox for Harris, FAST, BRISK and SURF, and Vlfeat for SIFT.

Comparison: We perform a thorough evaluation of our proposed OpTMO quantitatively using the RR measure. In Fig. 5, we initially show the performance of our OpTMO and other state-of-the-art TMOs in terms of RR averaged over all test scenes. For the sake of completeness, we also report the average RR obtained using HDR linear photometric values (HDRLin), without any tone mapping. Our results clearly show that the proposed OpTMO outperforms all the perception-based TMOs. In addition, the significant drop in performance with HDRlin demonstrates that HDR linear values are highly sub-optimal for keypoint detection task, similar to what is found in previous studies [11,19].

In Fig. 6, we expand our experimental test bench for each scene and compare the performance of our OpTMO with the globally optimized BTMO [27] and our previously proposed detector-optimal DetTMO [1]. The per scene gains of OpTMO over BTMO prove that local modulation of parameters significantly improves the keypoint stability. In addition, we

observe that the gain in performance between local and global optimization depends significantly on content characteristics. Especially for indoor scenes, which have been acquired by varying locally the illumination and introducing stark shadows, local parameter tuning enables to obtain important RR gains. We also notice that OpTMO achieves similar (within 2-4% per scene) RR as DetTMO, which is optimized for keypoint detection *only* and thus provides an upper bound in the achievable repeatability.

In order to further confirm these observations, we report a head-to-head comparison of OpTMO versus BTMO and DetTMO, respectively, in Fig. 7, for two different detectors: Harris (corner) and SURF (blob). OpTMO has higher RR whenever a point (representing a specific scene and illumination condition) is above the $45°$ line. As expected, we observe that this is often the case for BTMO, while for DetTMO the two methods have very similar performances. As mentioned above, the loss in keypoint repeatability compared to DetTMO is expected, and is mainly due to two reasons. On one hand, the additional descriptor-level cost term in Eq. (7) changes the objective function with respect to detector repeatability only (as in DetTMO). On the other hand, the use of the softargmax localization measure in Eq. (8) reduces cluttering of keypoints in our OpTMO. This is illustrated on a detail of the "Project-Room" scene in Fig. 8. For instance, cluttered keypoints are detected near the beaver's eyes in DetTMO, whereas OpTMO handles such detections efficiently. Interestingly, the composite objective function in Eq. (2) enables to achieve RR almost as good as DetTMO, but with a significantly improved descriptor matching and thus overall image matching performance, as shown in the next section.

Finally, we observe from Fig. 7 that these conclusions are valid for both Harris and SURF detectors, in spite of the fact that OpTMO is trained with respect to a classical corner response function (Eq. (6)). This demonstrates experimentally that images tone mapped with the proposed approach lead to increased detection performance even when the actual used detector is different from the specific response characteristics captured by the proxy cost function used for training.

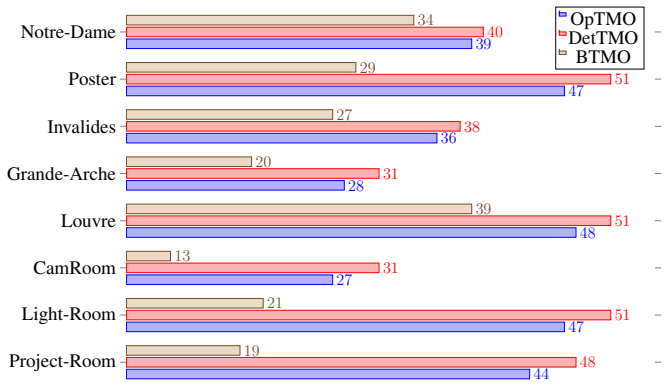
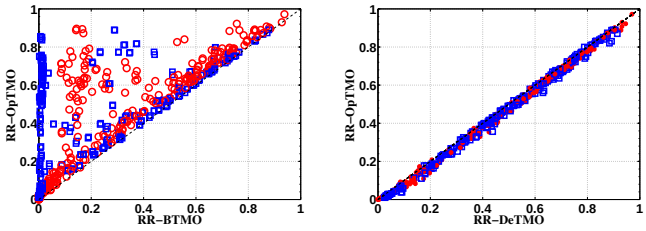
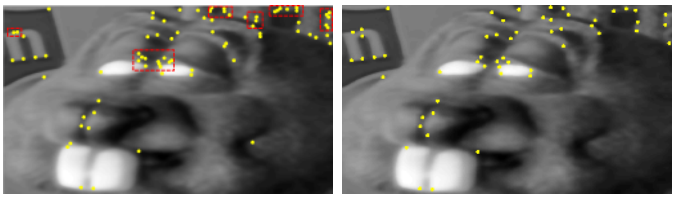


Fig. 6: **Keypoint Detection II**: Average Repeatability Rates (RR) computed using BTMO [27], DetTMO [1] and the proposed OpTMO for each test scene using Harris keypoint detector.



OpTMO vs BTMO OpTMO vs DetTMO

Fig. 7: **Keypoint Detection I**. The head to head comparison between (a) OpTMO vs BTMO and, (b) OpTMO vs DetTMO. Each point represent an image pair with different lighting conditions from the HDR dataset. The points represented using o depict the Harris corner detector and □ represents the SURF blob detector.



DetTMO OpTMO

Fig. 8: **Keypoint Detection IV**. Harris corner keypoints on the DetTMO and proposed OpTMO. The cluttered keypoints in DetTMO are highlighted using the red squares.

E. Descriptor Matching

We perform a thorough evaluation of our proposed OpTMO for descriptor matching using BRISK [53], FREAK [55], SIFT [18] and SURF [54] descriptors. We use the matching score (MS) as performance measure, as described in Section IV-B, considering the NNDR matching criteria with a threshold value $th = 0.5$.

Implementation: We use the Matlab's Computer Vision toolbox for FREAK, BRISK and SURF, and Vlfeat for SIFT, with their default parameter settings.

Comparison: In Fig. 9, we compare the average OpTMO MS with respect to state-of-the-art TMOs. Overall, we attain significant gains in terms of MS using all features extraction methods. With $th = 0.5$ (default value [13,18]), the gains are considerable for gradient-based features schemes such as SIFT and SURF, which is expected by design given the definition of the descriptor signature in Eq. (9).

To further analyze these results quantitatively, in Fig. 10 we report per scene comparison between the competing TMOs that are observed from Fig. 9. We observe that for each scene (indoor or outdoor) our OpTMO outperforms all the other TMOs. As in Section IV-D, we observe considerable gains with respect to traditional BTMO, confirming the potential of local parameter optimization. In comparison to DetTMO, we observe that gains are not as high as what are obtained with BTMO. This can be explained by the higher RR of the DetTMO (Fig. 6) which improves the probability of correct matches. Interestingly, we also observe that in many scenes DetTMO and DesTMO perform equally well, e.g., *Invalides* and *Project-Room* scenes. This is mainly because DesTMO is not optimal for detection, which entails a higher number of

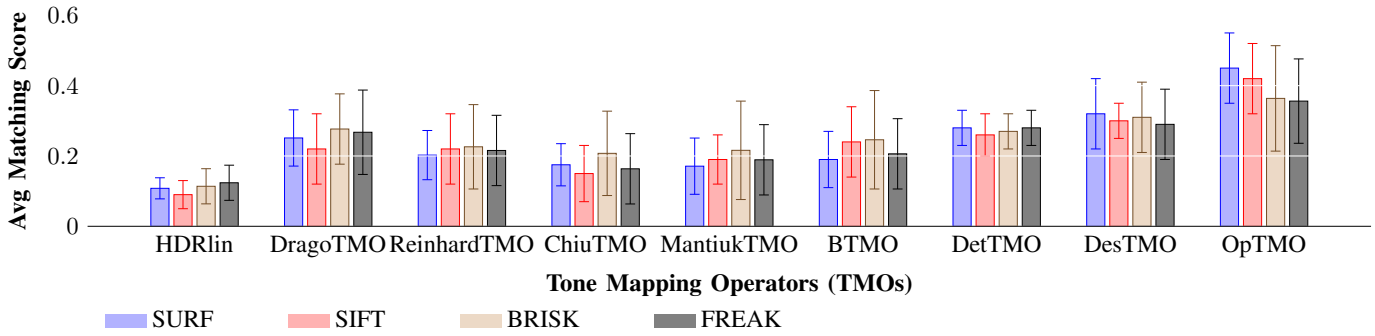


Fig. 9: **Descriptor Matching I** computed on different TMOs using SURF, SIFT, FREAK, BRISK descriptor extraction schemes. The average is calculated over all test scenes.

false matches.

Finally, we show the per image-pair analysis in Fig. 11 to further analyze the behavior of individual test pairs. We observe that our OpTMO improves the MS over DesTMO across the whole dataset (i.e., the gains are not concentrated on specific image pairs). In fact, there is not a single case where there is a significant drop in OpTMO's performance against the descriptor-optimal DesTMO, which again confirms the advantages of simultaneously optimizing the TMO for keypoint detection and description. In addition, the OpTMO produces consistent gains even if a binary descriptor such as FREAK is employed, in spite of the use of a gradient-based cost function in Eq. (9).

Note that MS is sensitive to the choice of th. Therefore, in the following section, we perform a global image matching evaluation using mAP to overcome the impact of the threshold.

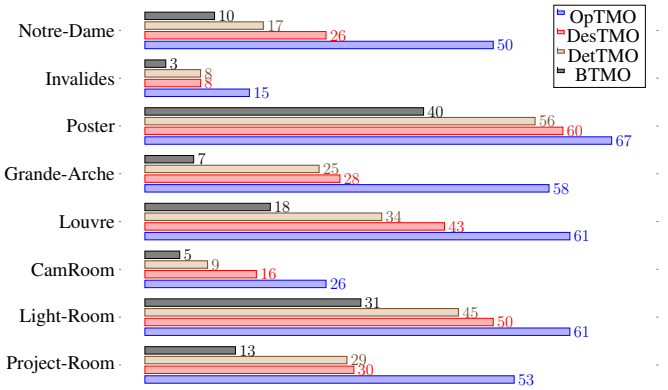


Fig. 10: **Descriptor Matching II**. Matching Score comparison between BTMO [27], OpTMO, DetTMO [1] and DesTMO [2] over all the scenes in the HDR dataset using SURF features extraction scheme.

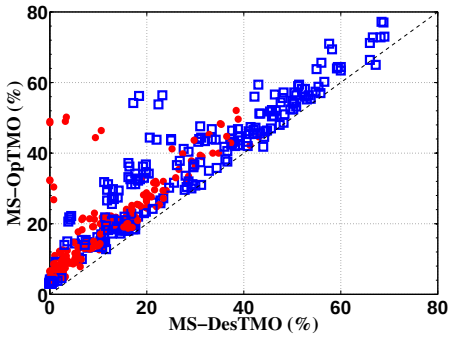


Fig. 11: **Descriptor Matching III**. Matching Score comparison between DesTMO vs OpTMO over all the scenes in HDR dataset. The points represented using o corresponds to FREAK features detection scheme and □ corresponds to SURF scheme.

F. Image Matching

We evaluate the full image matching chain by computing mean average precision (mAP) scores over the complete dataset. We obtain the mAP rates by averaging the area-under-the-curve of PR curves [13]. The results per TMO are reported in Fig. 12. We observe that for every descriptor extraction scheme our proposed model outperforms all the other TMOs. High mAP scores imply that our model obtains more correct matches and reduce the probability of false matches. An illustration of matching results is given in Fig. 13, showing that the proposed full-chain optimal tone mapping improves the matching efficiency in drastic lighting variations. Notice that ReinhardTMO and MantiukTMO provide poor image matching results compared with the proposed approach, although they provide better visually looking images. From Fig. 13, we observe that optimizing only for detector response (DetTMO) might produce a higher number of false matches. On the other hand, optimizing with respect to descriptor matching only (DesTMO) cannot ensure high matching efficiency due to the lower keypoint repeatability. Instead, efficient image matching can only be ensured by optimizing the TMO with respect to the full features extraction chain, as in the proposed OpTMO.

G. Applications

Localization of objects is a high-precision and pivotal task in many computer vision applications, *e.g.*, to find region of interest for fine-grained recognition challenges. For localization, first a homography matrix is computed by finding the best matching correspondences between the target and the test image. Then, the desired object is localized based on the estimated geometric relationship. In Fig. 14 and Fig. 15, we show a similar applicative scenario of localization of

selected objects such as structures in images undergoing both lighting and rotational transformations. We compare the performance of our proposed image-matching optimal TMO and the widely used ReinhardTMO over three scenes, namely *Louvre*, *ProjectRoom* and *Notre Dame*. In Fig. 14, we first find the corresponding matches between the two scenes using the SURF scheme for each TMO. Then, based on those resulting matches, we estimate the homography as proposed in [69]. We observe that our model gives more correct corresponding matches in all three scenes as compared to ReinhardTMO. In challenging outdoor scenes such as *Louvre* where there is a direct impact of sunshine, we observe that ReinhardTMO results in all incorrect matches, mainly concentrated in the brightest regions. In Fig. 15, we overlay the results on the test tone mapped images to show where exactly our desired object should be located based on the obtained correspondences. In *Louvre* and *ProjectRoom* scenes, we observe that tone mapped images using our proposed model result in correct localization of the desired object in the test image, as compared to ReinhardTMO. In the *Notre Dame* scene, the impact of illumination on the target region is smaller, and we are able to find correct overlaying results using both tone mappings.

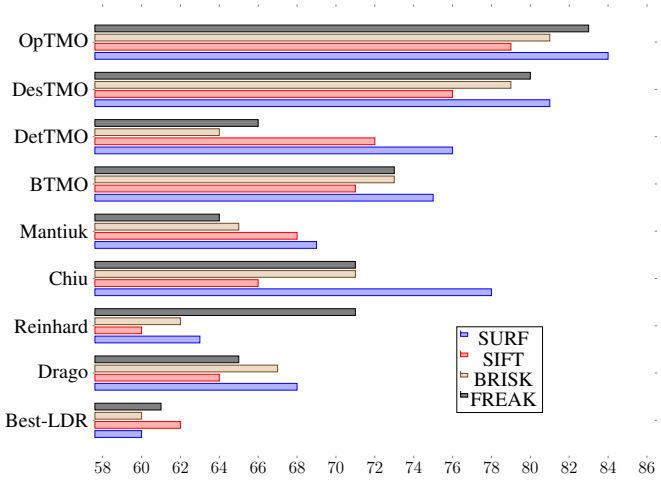


Fig. 12: **Image Matching I**. mAP % scores for the 9 different LDR modalities using 4 feature extraction schemes. Scores are averaged over 8 lighting change datasets.

Computation Time: In Fig 16, we compare the execution time (i.e. to tone map an HDR image) of the most competing state-of-the-art TMOs namely, BTMO, DetTMO, DesTMO and OpTMO. The computational time of our proposed method is not very far from the DesTMO. Note that the current implementation has been carried on a Intel Xeon CPU 4 cores processor, 16 Gb RAM windows 7 machine and has not been parallelized. An efficient parallelized implementation can further sped up the execution.

V. Conclusions

In this paper, we propose a new learning-based adaptive tone mapping framework for efficient image matching under drastic changes of lighting conditions. To this end, we first generate training samples by proposing a bi-objective function capturing both the detection and description stages of the features

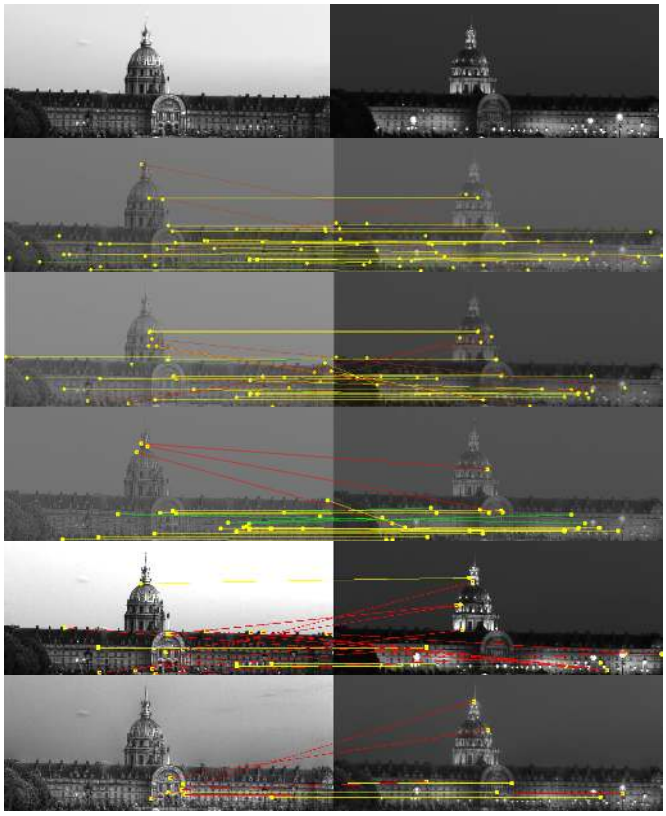


Fig. 13: **Image Matching II**. Day/Night matching using SURF. Row I: 2 HDR images from *Invalides* scene are displayed after log scaling [38]. Correct and incorrect matches are shown with yellow and red lines, respectively. Green lines represent the special case of mismatch due to repetitive structure. Row II: the feature matching using our proposed OpTMO (21 correct and 3 incorrect matches). Row III: using DetTMO (13 correct and 6 incorrect matches). Row IV: using DesTMOusing (11 correct and 3 incorrect matches). Row V using Reinhard TMO (3 correct and 11 incorrect matches). Row VI: using MantiukTMO (3 correct and 4 incorrect matches).

Fig. 16: *Computation time in sec (log scale)*. The time is computed by running all TMOs for an image size (512×512) on a Intel Xeon CPU 4 cores processor, 16 Gb RAM windows 7 machine.

extraction pipeline. Later, we train a Support Vector Regressor using local characteristics to learn a model which predicts spatially varying TMO parameters. We evaluate the proposed OpTMO on a HDR dataset of indoor/outdoor scenes where it outperforms state-of-the-art TMOs across different image matching algorithms. Finally, we demonstrate the performance of our method over other TMOs in a simple localization based application scenario.

Our proposed task-optimal TMO can be applied to different detection/description approaches. Hence, it can be directly fused with any existing local feature based applications such as structure from stereo, scene reconstruction, object tracking, recognition and photogrammetric applications. In the future, instead of learning the tone mapping parameter, we will focus

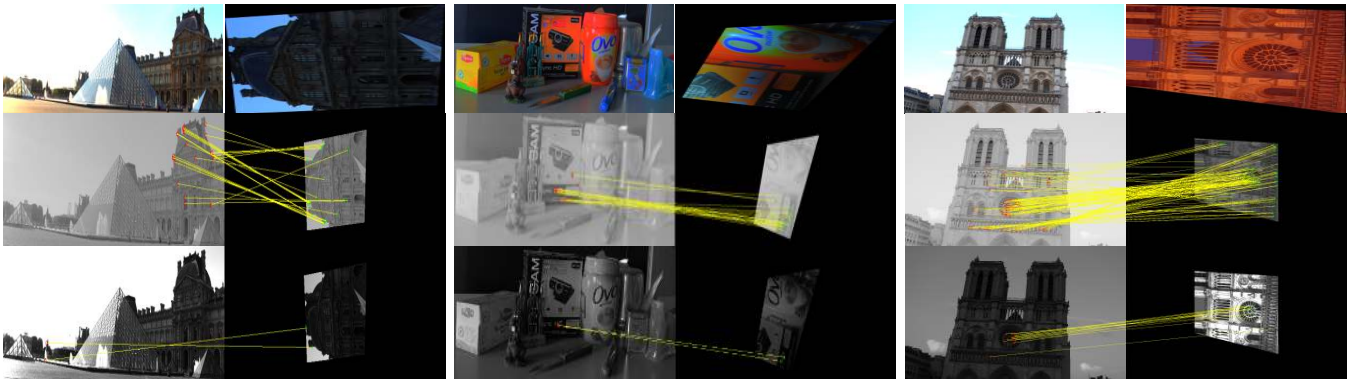


Fig. 14: **(Match & Locate)** Row I: Pair of HDR images from *Louvre*, *ProjectRoom* and *Notredame* scenes, with one reference and other being a selected region undergone lighting change and rotation. Row II: the feature matching using our proposed OpTMO. Row III: using Reinhard TMO.

(a) OpTMO (b) Reinhard TMO (c) OpTMO (d) Reinhard TMO (e) OpTMO (f) Reinhard TMO

Fig. 15: **(Match & Locate)** Results of matching images with different illumination and matched-regions (shown in Figure 14), where the shaded area is the matched image region. Results are shown for OpTMO and ReinhardTMO.

on directly learning a tone mapping function for the image matching problem.

ACKNOWLEDGMENT

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.

REFERENCES

[1] A. Rana, G. Valenzise, and F. Dufaux, "Learning-based Adaptive Tone Mapping for Keypoint Detection," in *IEEE International Conference on Multimedia & Expo (ICME'2017)*, Hong Kong, China, July 2017.

[2] A. Rana, G. Valenzise, and F. Dufaux, "Learning-Based Tone Mapping Operator for Image Matching," in *IEEE International Conference on Image Processing (ICIP'2017)*, Beijing, China, 2017, IEEE.

[3] E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec, *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting (The Morgan Kaufmann Series in Computer Graphics)*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[4] Z. Mai, H. Mansour, P. Nasiopoulos, and R. K. Ward, "Visually favorable tone-mapping with high compression performance in bit-depth scalable video coding," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1503–1518, Nov 2013.

[5] T. H. Wang, C. W. Chiu, W. C. Wu, J. W. Wang, C. Y. Lin, C. T. Chiu, and J. J. Liou, "Pseudo-multiple-exposure-based tone fusion with local region adjustment," *IEEE Transactions on Multimedia*, vol. 17, no. 4, pp. 470–484, April 2015.

[6] Y. Dong, M. T. Pourazad, and P. Nasiopoulos, "Human visual system-based saliency detection for high dynamic range content," *IEEE Transactions on Multimedia*, vol. 18, no. 4, pp. 549–562, April 2016.

[7] Y. T. Lin, C. M. Wang, W. S. Chen, F. P. Lin, and W. Lin, "A novel data hiding algorithm for high dynamic range images," *IEEE Transactions on Multimedia*, vol. 19, no. 1, pp. 196–211, Jan 2017.

[8] H. Hadizadeh and I. V. Bajic, "Full-reference objective quality assessment of tone-mapped images," *IEEE Transactions on Multimedia*, vol. PP, no. 99, pp. 1–1, 2017.

[9] F. Dufaux, P. Le Callet, R. Mantiuk, and M. Mrak, *High Dynamic Range Video: From Acquisition, to Display and Applications*, Academic Press, 2016.

[10] H. Zhou, T. Sattler, and D. W. Jacobs, "Evaluating local features for day-night matching," in *Computer Vision – ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, 2016, pp. 724–736.

[11] A. Rana, G. Valenzise, and F. Dufaux, "Evaluation of feature detection in HDR based imaging under changes in illumination conditions," in *IEEE International Symposium on Multimedia (ISM), Miami,USA, December, 2015*, 2015, pp. 289–294.

[12] P. Bronislav, A. Chalmers, P. Zemcík, L. Hooberman, and M. Cadík, "Evaluation of feature point detection in high dynamic range imagery," *Journal of Visual Communication and Image Representation*, pp. 141 – 160, 2016.

[13] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[14] Z. Liu, H. Li, W. Zhou, R. Hong, and Q. Tian, "Uniting keypoints: Local visual information fusion for large-scale image search," *IEEE Transactions on Multimedia*, vol. 17, no. 4, pp. 538–548, April 2015.

[15] A. Rana, J. Zepeda, and P. Pérez, "Feature learning for the image retrieval task," in *Computer Vision - FSLCV, Asian Conference on Computer Vision (ACCV) 2014 - Singapore, November 1-2, 2014*, 2014, pp. 152–165.

[16] U. L. Altintakan and A. Yazici, "Towards effective image classification using class-specific codebooks and distinctive local features," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 323–332, March 2015.

[17] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[19] A. Rana, G. Valenzise, and F. Dufaux, "An evaluation of HDR image matching under extreme illumination changes," in *The International Conference on Visual Communications and Image Processing (VCIP)*, Chengdu, China, Nov. 2016.

[20] R. Suma, G. Stavropoulou, E. Stathopoulou, L. V. Gool, A. Georgopoulos, and A. Chalmers, "Evaluation of the effectiveness of HDR tone-mapping operators for photogrammetric applications," *Virtual Archaeology Review*, vol. 7, no. 15, 2016.

[21] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Transactions on Graphics*, pp. 267–276, July 2002.

[22] R. Mantiuk, K. Myszkowski, and H. P. Seidel, "A perceptual framework for contrast processing of high dynamic range images," *ACM Transactions on Applied Perception*, vol. 3, no. 3, pp. 286–308, July 2006.

[23] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," in *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, 2002, SIGGRAPH

’02, pp. 257–266.

[24] R. Fattal, D. Lischinski, and M. Werman, “Gradient domain high dynamic range compression,” *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 249–256, July 2002.

[25] A. Chalmers and K. Debattista, “HDR Video Past, Present and Future: A Perspective ,” *Signal Processing: Image Communication*, vol. 54, pp. 49 – 55, 2017.

[26] R. Hong, L. Zhang, and D. Tao, “Unified photo enhancement by discovering aesthetic communities from flickr,” *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1124–1135, March 2016.

[27] A. Rana, G. Valenzise, and F. Dufaux, “Optimizing Tone Mapping Operators for Keypoint Detection under Illumination Changes,” in *2016 IEEE Workshop on Multimedia Signal Processing (MMSP 2016)*, Montréal, Canada, Sept. 2016.

[28] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” *Statistics and Computing*, vol. 14, no. 3, Aug. 2004.

[29] L. Bottou, *Stochastic Gradient Descent Tricks*, pp. 421–436, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[30] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, “A Comparison of Affine Region Detectors,” *International Journal of Computer Vision*, 2005.

[31] G. Kontogianni, E. K. Stathopoulou, A. Georgopoulos, and A. Doulamis, “HDR imaging for feature detection on detailed architectural scenes,” *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 325–330, 2015.

[32] A. Boschetti, N. Adami, R. Leonardi, and M. Okuda, “An optimal video-surveillance approach for HDR videos tone mapping,” in *Signal Processing Conference, 2011 19th European*, Aug 2011, pp. 274–277.

[33] T. Jinno, S. Kuriyama, and M. Okuda, “Tone-mapping for an hdr surveillance system using SIFT features,” in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, Sept 2013, pp. 1–5.

[34] A. Georgopoulos, A. Ntragka, and M. Quintero, “Photogrammetric exploitation of hdr images for cultural heritage documentation,” vol. II-5/W1, 09 2013.

[35] M. Rerabek, L. Yuan, L. Krasula, P. Korshunov, K. Fliegel, and T. Ebrahimi, “Evaluation of privacy in high dynamic range video sequences,” in *Applications of Digital Image Processing XXXVII*, Sept. 2014.

[36] P. Korshunov, M. V. Bernardo, A. M. G. Pinheiro, and T. Ebrahimi, “Impact of tone-mapping algorithms on subjective and objective face recognition in HDR images,” in *International ACM Workshop on Crowdsourcing for Multimedia (CrowdMM)*, 2015.

[37] Georgopoulos A. Agrafiotis P., Stathopoulou E. and Doulamis A., “HDR imaging for enchancing people detection and tracking in indoor environments,” in *In Proceedings of the 10th International Conf. on Computer Vision Theory and Applications (VISIGRAPP)*, 2015.

[38] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, “Adaptive logarithmic mapping for displaying high contrast scenes,” *Computer Graphics Forum*, pp. 419–426, 2003.

[39] K. Chiu, M. Herf, P. Shirley, S. Swamy, C. Wang, and K. Zimmerman, “Spatially nonuniform scaling functions for high contrast images,” in *Proceedings of Graphics Interface ’93*, Toronto, Ontario, Canada, 1993, GI ’93, pp. 245–253.

[40] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen, “Evaluation of tone mapping operators using a high dynamic range display,” *ACM Transactions on Graphics*, pp. 640–648, 2005.

[41] M. Cadik, M. Wimmer, L. Neumann, and A. Artusi, “Evaluation of HDR tone mapping methods using essential perceptual attributes,” *Computers and Graphics*, pp. 330 – 349, 2008.

[42] F. Banterle, A. Artusi, K. Debattista, and A. Chalmers, *Advanced High Dynamic Range Imaging: Theory and Practice*, Natick, MA, USA, 2011.

[43] E. Reinhard, “Parameter estimation for photographic tone reproduction,” *J. Graphics, GPU, & Game Tools*, vol. 7, no. 1, pp. 45–51, 2002.

[44] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, “LIFT: Learned Invariant Feature Transform,” in *Proceedings of the European Conference on Computer Vision*, 2016.

[45] C. Schmid, R. Mohr, and C. Bauckhage, “Evaluation of interest point detectors,” *International Journal of Computer Vision*, pp. 151–172, June 2000.

[46] C. Harris and M. Stephens, “A combined corner and edge detector,” in *In Proc. of Fourth Alvey Vision Conference*, 1988, pp. 147–151.

[47] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: An efficient alternative to sift or surf,” in *Proceedings of the 2011 International Conference on Computer Vision*, Washington, DC, USA, 2011, ICCV ’11, pp. 2564–2571, IEEE Computer Society.

[48] M. Bober, “MPEG-7 visual shape descriptors,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 716–719, Jun 2001.

[49] P. Salembier and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*, John Wiley & Sons, Inc., New York, NY, USA, 2002.

[50] Konstantinos M., Anastasios D. D., Nikolaos D. D., and Marinos I., “In the wild image retrieval and clustering for 3d cultural heritage landmarks reconstruction,” *Multimedia Tools Appl.*, vol. 75, no. 7, pp. 3593–3629, 2016.

[51] J. Hunter, “Combining the cidoc crm and mpeg-7 to describe multimedia in museums,” 2002.

[52] E. Rosten and T. Drummond, “Machine learning for high-speed corner detection,” in *Proceedings of the 9th European Conference on Computer Vision - Volume Part I*, Berlin, Heidelberg, ECCV’06, pp. 430–443, Springer-Verlag.

[53] S. Leutenegger, M. Chli, and R. Y. Siegwart, “BRISK: Binary robust invariant scalable keypoints,” in *Proceedings of the 2011 International Conference on Computer Vision*, Washington, DC, USA, 2011, ICCV ’11, pp. 2548–2555.

[54] H. Bay, T. Tuytelaars, and L. V. Gool, “SURF: Speeded up robust features,” in *9th European Conference on Computer Vision (ECCV)*, pp. 404–417. 2006.

[55] R. Ortiz, “Freak: Fast retina keypoint,” in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, 2012, CVPR ’12, pp. 510–517.

[56] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014.

[57] C. Tomasi and R. Manduchi, “Bilateral filtering for gray and color images,” in *Proceedings of the Sixth International Conference on Computer Vision*, Washington, DC, USA, 1998, ICCV ’98.

[58] W. Förstner, T. Dickscheid, and F. Schindler, “Detecting interpretable and accurate scale-invariant keypoints,” in *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, 2009, pp. 2256–2263.

[59] O. Chapelle and M. Wu, “Gradient descent optimization of smoothed information retrieval metrics,” *Information Retrieval*, vol. 13, no. 3, pp. 216–235, 2010.

[60] J. Dong, N. Karianakis, D. Davis, J. Hernandez, J. Balzer, and S. Soatto, “Multi-view Feature Engineering and Learning,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015.

[61] A. Vedaldi and B. Fulkerson, “VLFeat: An Open and Portable Library of Computer Vision Algorithms,” 2008.

[62] Y. Verdie, K. M. Yi, P. Fua, and V. Lepetit, “TILDE: A temporally invariant learned detector.,” in *CVPR*. 2015, pp. 5279–5288, IEEE Computer Society.

[63] C. Tomasi and R. Manduchi, “Bilateral filtering for gray and color images,” in *Sixth IEEE International Conference on Computer Vision (ICCV)*, 1998.

[64] P. Bronislav, A. Chalmers, and P. Zemcík, “Feature point detection under extreme lighting conditions,” in *Spring Conference on Computer Graphics*, 2012, pp. 156–163.

[65] P. E. Debevec and J. Malik, “Recovering high dynamic range radiance maps from photographs,” in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, New York, USA, 1997, SIGGRAPH, pp. 369–378.

[66] C.C. Chang and C.J. Lin, “Libsvm: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, May 2011.

[67] L. Chermak and N. Aouf, “Enhanced feature detection and matching under extreme illumination conditions with a hdr imaging sensor,” in *IEEE 11th International Conference on Cybernetic Intelligent Systems*, Aug 2012, pp. 64–69.

[68] T. Tuytelaars and K. Mikolajczyk, *Local Invariant Feature Detectors: A Survey*, Now Publishers Inc., Hanover, MA, USA, 2008.

[69] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, second edition, 2004.

Dr. Aakanksha Rana is a post-doctoral research fellow at Trinity College Dublin, Dublin, Ireland, since 2018. She received M.Engg. in multimedia technologies and Ph.D. in signal and image processing from Telecom ParisTech, Paris, France, in 2014 and 2018 respectively. During her masters, she was an intern in the exploratory research group at Technicolor Rennes Research & Innovation Center and AYIN group, INRIA, Sophia Antipolis in 2014 and 2013, respectively. Her areas of research broadly include computer vision, deep learning, computational imaging, and satellite image analysis.

Giuseppe Valenzise is a CNRS researcher at Telecom ParisTech, Paris, France, since October 2012. Previously, he worked as post-doc researcher in the same lab, starting from July 2011. He completed a master degree and a Ph.D. in Information Technology at the Politecnico di Milano in 2007 and 2011, respectively. From January 2009 to July 2009 he was a visiting scholar at the Signal and Image Processing Institute at the University of Southern California. His research interests span different fields of image and video processing, including high-dynamic range imaging, video quality assessment, single and multi-view video coding, video surveillance, image and video forensics, image and video analysis. He is co-author of more than 40 research publications, including award winning papers at ICIP 2011 and ICIP 2014. He has been actively involved in several French and EU-funded research projects. Dr. Valenzise serves as Associate Editor for IEEE Trans. on Circuits and Systems for Video Technology, and Signal Processing: Image communication.

Dr. Frédéric Dufaux (S'93, M'95, SM09, F'16) is a CNRS Research Director at Laboratoire des Signaux et Systémes (L2S, UMR 8506), CNRS - CentraleSupelec - Université Paris-Sud, where he is head of the Telecom and Networking division. He is also Editor-in-Chief of Signal Processing: Image Communication.

Frédéric received his M.Sc. in physics and Ph.D. in electrical engineering from EPFL in 1990 and 1994 respectively. He has over 20 years of experience in research, previously holding positions at EPFL, Emitall Surveillance, Genimedia, Compaq, Digital Equipment, and MIT.

Frédéric is a Fellow of IEEE. He is Chair of the IEEE SPS Multimedia Signal Processing (MMSP) Technical Committee. He is a founding member and Chair of the EURASIP Special Area Team on Visual Information Processing. He has been involved in the standardization of digital video and imaging technologies, participating both in the MPEG and JPEG committees. He is the recipient of two ISO awards for his contributions.

His research interests include image and video coding, 3D video, high dynamic range imaging, visual quality assessment, and video transmission over wireless network. He is author or co-author of 3 books "High Dynamic Range Video","Digital Holographic Data Representation and Compression", "Emerging Technologies for 3D Video"), more than 120 research publications and 17 patents issued or pending.