The Maximal Self-dissimilarity Interest Point Detector

FEDERICO TOMBARI^{1,a)} LUIGI DI STEFANO^{1,b)}

Received: February 6, 2015, Accepted: September 10, 2015, Released: December 16, 2015

Abstract: We propose a novel interest point detector stemming from the intuition that image patches which are highly dissimilar over a relatively large extent of their surroundings hold the property of being repeatable and distinctive. This concept of *contextual self-dissimilarity* reverses the key paradigm of recent successful techniques such as the Local Self-Similarity descriptor and the Non-Local Means filter, which build upon the presence of similar - rather than dissimilar - patches. Moreover, our approach extends to contextual information the local self-dissimilarity notion embedded in established detectors of corner-like interest points, thereby achieving enhanced repeatability, distinctiveness and localization accuracy. As the key principle and machinery of our method are amenable to a variety of data kinds, including multi-channel images and organized 3D measurements, we delineate how to extend the basic formulation in order to deal with range and RGB-D images, such as those provided by consumer depth cameras.

Keywords: interest point, keypoint detection, image matching

1. Introduction

The *self-similarity* of an image patch is a powerful computational tool that has been deployed in numerous and diverse image processing and analysis tasks. It can be defined as the set of dissimilarities of a patch to those located in its surroundings, with dissimilarities usually measured through the Sum of Squared Distances (SSD), although a variety of alternatives does exist [8]. Whenever the task mandates looking for large rather than small minima over such quantities, we will use the term *self-dissimilarity*. Analogous to self-similarity is auto-correlation, which relies on the cross-correlation to compare the given to surrounding patches. An early example of deployment of self-dissimilarity in the computer vision literature is the Moravec operator [22], which detects interest points exhibiting a sufficiently large intensity variation along all directions by computing the minimum SSD between a patch and its 8 adjacent ones. The Harris Corner Detector [10] extends the Moravec operator by proposing Taylor's expansion of the directional intensity variation together with a saliency score which highlights corner-like interest points. Then, Mikolajczyk and Schmid developed the Harris-Laplace operator [20] to achieve scale-invariant detection of corner-like features.

More recently, the self-similarity concept has been used to develop the Local Self Similarity (LSS) region descriptor [24], which leverages on relative positions between nearby similar patches to provide an invariant representation of a pixel's neighborhood. One of the main innovations introduced by this method with respect to previous approaches deploying self-similarities consists in the reference patch being spatially compared with a much larger neighborhood rather than with just its nearest vicin-

ity. The LSS method computes a *self-similarity surface* associated with an image point, which is then quantized into a log-polar grid to build the descriptor. Notably, the inherent traits of self-similarity endow the descriptor with peculiar robustness with respect to diversity of the image acquisition modality [13], [24]. As a further example, [18] exploits the concept of self-similarity to detect interest points associated with symmetrical regions in images. Specifically, auto-correlation based on Normalized Cross-Correlation among image patches is used as a saliency measure to highlight image regions exhibiting symmetries with respect to either a line (mirror symmetries) or a point (rotational symmetries). Interest points are successively detected as extrema of the saliency function over a scale-space.

Though aimed at a different purpose such as denoising, the Non-Local Means (NLM) [6] and BM3D [7] filters exploit the presence of similar patches within an image to estimate the noiseless intensity of each pixel. In [6], this is done by computing the weighted average of measured intensities within a relatively large area surrounding each pixel, with weights proportional to the self-similarity between the patch centered at the given pixel and those around the other ones in the area. Instead, in [7] self-similarity allows for sifting-out sets of image patches grouped together to undergo a more complex computational process referred to as collaborative filtering.

In this paper we propose a novel interest point detector obtained by reverting the classical exploitation of self-similarity so as to highlight those image patches that are most *dissimilar* from nearby ones within a relatively large surrounding area. This concept, which will be referred to in the following as *contextual self-dissimilarity* (CSD), associates a patch's saliency with the absence of similar patches in its surroundings, as exemplified in **Fig. 1** (a). Accordingly, CSD may be thought of as relying on the rarity of a patch, which, interestingly, is identified as the basic saliency cue also in the interest point detector by Kadir and

¹ DISI – University of Bologna, Bologna, Italy

^{a)} federico.tombari@unibo.it

^{b)} luigi.distefano@unibo.it

Fig. 1 Rare patches may be provided by diverse kinds of image structures, such as (a), (b), (c). Rarity is resilient to scale and rotation (d), tone mapping (e) and blur (f).

Brady [14]. However, their work ascertains rarity in a strictly local rather than contextual approach, due to saliency consisting in the entropy of the gray-level distribution within a patch [14]. Also interestingly, a recent study has highlighted a sound link between the salient image structures computed in an uncommitted way by simple weighted raw image values and the computational models of preattentive human visual perception [15].

A similar concept to CSD has been exploited in [9] for the purpose of detecting salient regions to create a visual summary of an image. In particular, the proposed saliency for a patch is directly proportional to the distance in the CIELab space to surrounding most similar patches and inversely proportional to their 2D spatial distance, the latter requirement due to the addressed task calling for spatially close rather than scattered salient pixels. Unlike [9], we aim here at exploiting self-dissimilarities for the task of interest point detection and propose a saliency measure which relies solely upon the CSD measured in the intensity domain. It is worth pointing out that the role of context in determining salient visual features has been highlighted also in Ref. [5]. Another relevant approach which can highlight rare patches for the goal of creating a visual saliency map of the whole image is the work in [12]. Here, however, the authors do not rely on spatial distance but on the log-spectrum of the image.

Unlike several prominent feature detectors bound to fire at specific structures, e.g., Refs. [3], [16], the rarity of a patch permits to highlight a variety of diverse patterns, including distinctive edge traits and smooth textures (see Fig. 1 (b) and (c), respectively). Moreover, inherent to the saliency criterion is the ability of our approach to withstand the major nuisances to be dealt with in the realm of interest point detection, such as scale change, rotation, tone mapping and blur (see again Fig. 1), as well as, peculiarly, to cope effectively with diversity in the image sensing modality, as vouched by the analogy to the work on matching images by self-similarities [13], [24].

It is worth pointing out that our interest point detector is not constrained to operate on the intensity channel only, but, rather, naturally amenable to generalization to any kind of multi-channel images, such as color and range images, as well as the RGB-D images provided by consumer depth cameras, like the Microsoft

Kinect or the Asus Xtion, which are becoming more and more widespread in computer vision research and applications. Therefore, in this paper we also propose the extension of the our approach to such kind of organized 3D data.

We provide an extensive experimental evaluation to demonstrate the effectiveness of the proposed detector in finding excellent interest points in a variety of settings. Experiments address measurement of feature repeatability on the most relevant benchmarks for interest point detection, namely the *Oxford* and *Robot* datasets, as well as assessment of feature distinctiveness through the image retrieval test proposed in *VLBenchmark* framework. Besides, we report experimental results dealing with diverse kinds of data such as multi-modal and RGB-D images.

2. Contextual Self Dissimilarity

The saliency concept used by our interest point detector relies on the computation of a patch's self-similarity over an extended neighborhood, which has already been exploited by popular techniques such as the LSS descriptor [24] and the NLM filter [6]. Unlike these methods, though, we do not aim at detecting highly similar patches within the surroundings of a pixel, but, instead, at determining whether a pixel shows similar patches in its surroundings or not. Thus, the proposed technique relies on a saliency operator, λ , which measures the *Contextual Self-Dissimilarity* (CSD) at a pixel position p , i.e., quantifies how dissimilar is the patch around p to the most similar one in its surroundings:

$$\lambda(p, \rho_w, \rho_a) = \frac{1}{\rho_w^2} \min_{q \in \omega(p, \rho_a), q \neq p} \delta(\omega(p, \rho_w), \omega(q, \rho_w)) \quad (1)$$

As shown by (1), the proposed saliency operator is characterized by two parameters, ρ_w and ρ_a , representing the size of the patches under comparison and the size of the area from which the patches to be compared are drawn, respectively. Moreover, $\omega(p, \rho_w)$ denotes the operator defining a square image neighborhood centered at pixel p and having size equal to ρ_w pixels, while δ denotes the distance between two patches collecting the intensities associated in two equally sized neighborhoods, which in its simplest form can be the squared L_2 distance, or Sum of Squared Distances (SSD):

$$\delta(\omega(p, \rho_w), \omega(q, \rho_w)) = \|I(\omega(p, \rho_w)) - I(\omega(q, \rho_w))\|_2^2 \quad (2)$$

Computing λ at all pixels determines a saliency map whose values are proportional to the rarity of the patch centered at each pixel with respect to the surrounding area. Normalization by the number of pixels involved in the computation of the self-dissimilarity helps rendering the saliency score independent of the patch size ρ_w . Parameter ρ_a establishes the spatial support of the saliency criterion. As a well-known trait in literature [4], certain saliency operators can be defined either locally or globally, depending on a patch's rarity being computed over small local neighborhoods or the whole image. By increasing ρ_a , the λ operator moves gradually from a local toward a contextual or even global saliency criterion. As mentioned in Section 1, we advocate replacing the local self-dissimilarity underpinning all existing interest point detectors rooted in the Moravec operator with

Fig. 2 Results on a subset of the *Oxford* dataset. Top row (a-d): Contextual vs. Local Self-Dissimilarity. Bottom row: repeatability (e-g) and relative execution times (h) for CSD interest points detected with different k values.

a contextual self-dissimilarity notion. To begin substantiating the claim, in the top-row of **Fig. 2**, we report results on a subset of the *Oxford* dataset that show how deployment of a contextual rather than local saliency criterion delivers dramatic improvements in terms of repeatability of detected interest points^{*1}.

The saliency defined in Eq. (1) relies on estimating the minimum distance between the given and neighboring patches by simply picking one sample from observations, which is potentially prone to noise. Indeed, noise on both the central as well as the most dissimilar neighboring patch can induce notable variations in saliency scores, which may hinder repeatability and accurate localization of salient points. On the other hand, most existing operators grounded on self-similarity average out estimates over several samples. In the NLM filter, e.g., the noiseless value to be assigned to each pixel is averaged over all samples. Likewise, in the LSS descriptor the discriminative trait associated to an image point is the union of the locations of similar patches in the neighborhood. A further ingredient which confers robustness to noise to the LSS descriptor is the binning carried out by quantizing into a spatial histogram the locations of most similar patches.

Therefore, we propose to modify Eq. (1) in the way the minimum of the distribution is estimated. Finding the most similar patch among a set of candidates can be interpreted as a 1-Nearest Neighbor (1-NN) search problem. We propose to modify the search task to a k -NN problem (with $k \geq 1$) and, accordingly, to estimate the minimum as the average across the k most similar patches:

$$\lambda^{(k)}(p, \rho_w, \rho_a) = \frac{1}{\rho_w^2 \cdot k} \sum_{i=1}^k \tilde{\delta}^i(\omega(p, \rho_w), \omega(q, \rho_w)) \quad (3)$$

where $\tilde{\delta}^1, \dots, \tilde{\delta}^k$ are the k smallest value of the δ function found within the search area defined by ρ_a . Parameter k thus trades distinctiveness and computational efficiency for repeatability and accurate localization in noisy conditions. Figure 2, bottom row,

highlights the impact of the chosen k on both performance as well as computational efficiency: a higher k yields generally improved repeatability at the expense of a higher computational cost. Although the optimal value may depend on the specific nuisances related to the addressed scenario, we found $k = 4$ to provide generally a good trade-off between performance and speed, and we thus suggest this as default setting in Eq. (3).

2.1 Computational Efficiency

Computing the CSD operator across an image with n pixels implies the operation in Eq. (3) to be repeated as many times as n , this yielding a complexity equal to $O(n \cdot \rho_w^2 \cdot \rho_a^2)$ which may turn out prohibitive for common image sizes. To reduce the computational burden inherent to the saliency operator presented thus far, we have devised an incremental scheme which can decrease the complexity to $O(n \cdot \rho_a^2)$, i.e., so as to render it independent of patch size.

The main intuition relies on the observation that, once the CSD operator has been computed at pixel p , most of the calculations associated with the next position, p' , can be recycled. This is sketched in **Fig. 3** (a), where the patches associated with p and q are depicted in blue, those associated with p' and q' highlighted in red. The figure intuitively shows that the distance between the patches at p' and q' can be computed as:

$$\delta(\omega(p', \rho_w), \omega(q', \rho_w)) = \delta(\omega(p, \rho_w), \omega(q, \rho_w)) - \delta(\alpha(p'), \alpha(q')) + \delta(\beta(p'), \beta(q')) \quad (4)$$

where $\alpha(p')$, $\beta(p')$, $\alpha(q')$, $\beta(q')$ are the vectors collecting the intensities of the left and right vertical sides of the two patches, as highlighted in the Figure. In turn, as illustrated in **Fig. 3** (b), the two distances between the corresponding sides of the patches appearing in Eq. (4) can be computed incrementally from the position just above p' , denoted as p'' and highlighted in green, by adding and subtracting properly the squared differences between the intensities at the four corner positions of the patches, referred

^{*1} Interest point detection is run at multiple scales as described in Section 3

Fig. 3 Efficient computation of the distance between two patches: recursive scheme applied along columns (3 (a)) and along both columns and rows (3 (b)).

to as i, j, u, v . Accordingly, Eq. (4) can be further manipulated so to reach:

$$\begin{aligned} \delta(\omega(p', \rho_w), \omega(q', \rho_w)) &= \delta(\omega(p, \rho_w), \omega(q, \rho_w)) \\ &\quad - \delta(\alpha(p''), \alpha(q'')) + \delta(\beta(p''), \beta(q'')) \\ &\quad - (I(i(p')) - I(i(q')))^2 - (I(j(p')) - I(j(q')))^2 \\ &\quad + (I(u(p')) - I(u(q')))^2 + (I(v(p')) - I(v(q')))^2 \end{aligned}$$

As it can be noticed from the above equation, the distance between the current pair p', q' needs not to be calculated from scratch but, rather, can be obtained incrementally from already available quantities by means of a few elementary operations. This approach, which can be regarded as a particular form of Box Filtering [19], allows calculating all distances between the central patch and those contained in the search area with a limited computational complexity and may be deployed to reduce the complexity of self-similarity-based techniques alike, such as e.g., Refs. [6], [24].

The overall algorithm to compute the saliency operator λ is showcased in Algorithm 1, where, for illustrative purposes, we consider the simplest case of Eq. (1), i.e., $k = 1$. In its practical implementation, δ_α and δ_β are assimilated to the same memory structure having size $w \cdot \rho_a^2$ elements, which is initialized by explicitly computing the column-wise squared difference within all search areas on the first image row. The δ_ω data structure is instead as large as ρ_a^2 elements. Thus, the overhead in memory footprint required by incremental computation turns out as small as $(w + 1) \cdot \rho_a^2$, which is favorably counterbalanced by a speed-up of about one order of magnitude with respect to the standard implementation.

3. Detection of Interest Points

Given its definition, the CSD operator yields a high score only when the current patch is highly dissimilar from all surrounding

Algorithm 1 Incremental computation of the λ operator

```

for  $p \in$  first row do
    for  $q \in \omega(p, \rho_a), q \neq p$  do
         $\delta_\alpha(p, q) = \delta(\alpha(p), \alpha(q))$ 
         $\delta_\beta(p, q) = \delta(\beta(p), \beta(q))$ 
    end for
end for
for  $p \in$  all other rows do
     $\delta_{min} = \inf$ 
    for  $q \in \omega(p, \rho_a), q \neq p$  do
        if  $p$  is the first pixel of the row then
             $\delta_\omega(q) = \delta(\omega(p, \rho_w), \omega(q, \rho_w))$ 
        else
             $\delta_\alpha(p, q) \leftarrow \delta(u(p), u(q)) - \delta(i(p), i(q))$ 
             $\delta_\beta(p, q) \leftarrow \delta(v(p), v(q)) - \delta(j(p), j(q))$ 
             $\delta_\omega(q) \leftarrow \delta_\beta(p, q) - \delta_\alpha(p, q)$ 
        end if
        if  $\delta_\omega(q) < \delta_{min}$  then
             $\delta_{min} = \delta_\omega(q)$ 
        end if
    end for
     $\lambda(p) = \frac{1}{\rho_a^2} \cdot \delta_{min}$ 
end for
    
```

ones. This trait can be exploited to develop an interest point detector whereby interest points are given by the centers of those patches featuring a distinctive structure with respect to their surroundings, whatever such a structure may be. It is worth observing that, with the proposed approach, the self-similarity surface around interest points tends inherently to exhibit a sharp peak rather than a plateau, which is a desirable property as far as precise localization of extracted features is concerned. Indeed, given that the patch centered at an interest point must be highly dissimilar also to adjacent patches, it is unlikely for nearby points to exhibit a similar saliency as that of the interest point. Another benefit of relying on CSD to detect interest points concerns its

Fig. 4 Qualitative comparison between the interest points provided by MSD (green dots) and the Harris-Laplace detector [20] (red dots) on 3 image regions from the *Oxford* dataset. For clarity of visual comparison, only features of approximately the same medium-size scale are displayed for both methods.

potential effectiveness in presence of strong photometric distortions as well as multi-modal data, as intuitable by the affinity to the work related to the LSS descriptor [24]. Moreover, intuition suggest the approach to be robust to nuisances such as viewpoint variations and blur, given that the property of a patch to be somehow unique within its surroundings is likely to hold even though the scene is seen from a (moderately) different vantage point and under some degree of blur.

However, ρ_w and ρ_a would set the scale of the structures of interest firing the detector. To endow the detector with scale invariance, as well as to associate a characteristic scale to extracted features, we build a simple image pyramid $I(l)$ comprising L levels, starting from level 1 (original image resolution) and rescaling, at each level l , the image of a factor f^l with respect to the base level. Denoting as w and h , respectively, the number of image columns and rows, once the scale factor f and the parameters ρ_a, ρ_w are chosen, the number of pyramid levels L can be automatically determined according to:

$$L = \left\lceil \log_f \left(\frac{\min(w, h)}{(\rho_w + \rho_a) \cdot 2 + 1} \right) \right\rceil \quad (5)$$

based on the constraint that the top level of the pyramid cannot be smaller than the area required to compute the saliency on one single point:

$$\frac{\min(w, h)}{f^L} > (\rho_w + \rho_a) \cdot 2 + 1 \quad (6)$$

Once the saliency in Eq. (3) is computed at each point within the several layers of the image pyramid, for each level l the set of interest points, $\tilde{P}_l = \{\tilde{p}_1, \dots, \tilde{p}_n\} \in I(l)$, is extracted by means of a Non-Maxima Suppression (NMS) procedure. Specifically, an interest point $\tilde{p} \in I(l)$ is detected if it yields a saliency higher than all other saliency values within a window of size ρ_v :

$$\tilde{p} \in I(l) \text{ s.t. } \max_{p \in \omega(\tilde{p}, \rho_v), p \neq \tilde{p}} \lambda^{(k)}(p, \rho_w, \rho_a) < \lambda^{(k)}(\tilde{p}, \rho_w, \rho_a) \quad (7)$$

As the features detected through the NMS stage are local maxima of the CSD operator, hereinafter our proposal will be also referred to as *Maximal Self-Dissimilarity* interest point detector (MSD). Afterwards, weak local maxima may be further pruned based on a saliency threshold τ_δ , which in our experiments is set to $\tau_\delta = 250$.

The search for local maxima throughout the image pyramid allows associating a characteristic scale to each detected interest point; given an interest point $\tilde{p}$ detected at coordinates (i_l, j_l) and pyramid level l , its associated i, j coordinates into the original

image and characteristic scale size (or diameter) s are given by:

$$i(\tilde{p}) = i_l \cdot f^l \quad (8)$$

$$j(\tilde{p}) = j_l \cdot f^l \quad (9)$$

$$s(\tilde{p}) = (\rho_w \cdot 2 + 1) \cdot f^l \quad (10)$$

For the purpose of successive feature description, a canonical orientation may also be associated to each interest point $\tilde{p}$ by accumulating into a histogram the angles between the interest point and the centers of the k most similar patches within $\omega(\tilde{p}, \rho_a)$ weighted by their dissimilarity, so as to then choose the direction corresponding to the highest bin in the histogram.

As already pointed-out, assessment of saliency based on the self-dissimilarity of a patch underpins both MSD as well as established detectors of corner-like structures, such as Moravec [22], Harris [10] and, more recently, the Harris-Laplace and Harris-affine detectors [20], [21], the key difference consisting in our proposal advocating assessment to occur across a larger surrounding area referred to as context rather than locally. It is also worth pointing out that, accordingly, our approach cannot deploy Taylor expansion of the dissimilarity function, as it is indeed the case of Harris-style detectors, due to Taylor expansion providing a correct approximation only locally, i.e., within a small neighborhood of the pixel under evaluation.

To further highlight the differences between the two approaches, in **Fig. 4** we compare qualitatively the interest points extracted by MSD to those provided by the Harris-Laplace (*harlap*) scale-invariant corner detector [20]. One of the most noticeable differences between the two approaches concerns *harlap* tending to yield multiple nearby responses around the most salient (and corner-like) structures, while this is not the case of MSD, as nearby corner-like structures tend to be similar and thus inhibit each other due to the requirement for interest points to be salient within the context. This is a favorable property as implies dealing with inherently fewer distinctive interest points in the successive feature matching stage. It can also be observed how, again due to the use of context, MSD features tend to be scattered over a more ample image area and in a more uniform way. Moreover, and unlike *harlap*, MSD can detect also a variety of salient structures quite different from corner-like ones, such as blob-like features, edge fragments and smoothly-textured distinctive patches.

As a final remark, the choice of parameters ρ_a, ρ_w is key to the performance of the proposed detector. In particular, too small a patch does not contain enough information to render the self-

Fig. 5 Repeatability on the 8 sets of images of the *Oxford* dataset. The x axis denotes the level of difficulty of the considered nuisance.

dissimilarity concept meaningful and effective due to dissimilarity tending to appear quite often small. Alike, this is the case of too big a patch, with dissimilarity getting now always high. Given the chosen patch size, as context is enlarged the detector tends to sift-out increasingly distinctive features, but this hinders both the quantity of extracted interest points, as it implies a high probability of finding similar structures around, as well as their repeatability, the latter issue occurring in cluttered scenes due to the likely inclusion into the context of similar patches belonging to nearby objects. Therefore, we have run several experiments to carefully select the key parameters of our method and found quite an effective trade-off pair to consist in $\rho_w = 7$, $\rho_a = 11$.

4. MSD for Range and RGB-D Images

As the CSD principle is not peculiar to the intensity channel, interest point detection based on MSD can be naturally extended

to other types of data, including multi-channel images (e.g., color images) as well as data structures encoding 3D information. As for the latter type of data, the range and RGB-D images acquired by consumer depth cameras, such as the Microsoft Kinect and the Asus Xtion, assume nowadays particular relevance as these very low-cost devices can provide 3D data at VGA-resolution in real-time.

These premises lead us to consider *organized* data structures, i.e., depth measurements taken across a rectangular pixel grid, and calibrated cameras, so that each depth value can be turned into an x, y, z triplet in the 3D space. Accordingly, we define a range image, R , to be an image that takes a triplet of real numbers at each pixel position p , such numbers representing the coordinates of the 3D point associated with p :

$$R(p) = (x(p), y(p), z(p)) \quad (11)$$

Fig. 6 Number of correct correspondences on the *Oxford* dataset. The x axis denotes the level of difficulty of the considered nuisance.

Following a similar notation as that adopted to define the CSD operator in Eq. (2), we introduce here the concept of a patch of size ρ_w , so that the patch $R(\omega(p, \rho_w))$ would represent the set of the 3D coordinates associated with the pixels belonging to a square neighborhood of size ρ_w around p .

We now introduce an operator, dR , which takes two pixels, p and q , and computes the squared Euclidean distance between their associated 3D points:

$$dR(p, q) = \|R(p) - R(q)\|_2^2 \quad (12)$$

Then, dR can be extended to equally sized patches so to define vector $dR(\omega(p, \rho_w), \omega(q, \rho_w))$, which collects the distances between the 3D points associated with corresponding pixels in two neighborhoods of size ρ_w around p and q . Given the above definitions, the CSD operator can be generalized quite seamlessly to the case of range images by defining the distance between two

patches to be squared Euclidean norm of the vector encoding the pixel-wise 3D distances:

$$\delta_R(\omega(p, \rho_w), \omega(q, \rho_w)) = \|dR(\omega(p, \rho_w), \omega(q, \rho_w))\|_2^2 \quad (13)$$

Then, following straightforwardly the formulation for intensity images given in Eq. (1), the computation of the saliency at each pixel position, p , deals with minimization of δ_R across the context area provided by $\omega(p, \rho_a)$. Alike, features are detected by first running an NMS process and then pruning weak local maxima by a threshold on the 3D saliency score denoted as τ_ρ . The thus defined extension to range images of the MSD algorithm will be referred to hereinafter as R-MSD. Then, following straightforwardly the formulation for intensity images given in Eq. (1), the computation of the saliency at each pixel position, p , deals with minimization of δ_R across the context area provided by $\omega(p, \rho_a)$. Alike, features are detected by first running an NMS process and

Fig. 7 Comparison between MSD and the 4 variants of the proposal in [18] on the Oxford dataset.

then pruning weak local maxima by a threshold on the saliency score. The thus defined extension to range images of the MSD algorithm will be referred to hereinafter as R-MSD.

RGB-D images may be seen as range images featuring also a color, i.e. RGB, triplet at each pixel position p . As such, a very simple approach to extend MSD would consist in packing together color and 3D measurements at each pixel into one single structure so as to, by analogy with Eqs. (12) and (13), be able to rely on distances and norms computed on higher dimensional vectors. In practice, though, just mixing together blindly photometric and 3D informations appears too simplistic and hardly effective an approach. Moreover, we argue that it is uncommon to find scene structures that are likely to turn out salient only due to the mixed combination of appearance and shape, whilst they had not been so based on either of the two cues. More often than not, rather, a scene structure may be salient because of either its

3D shape, e.g., an evenly colored sharp bulge, or its appearance, such as a richly textured planar surface.

Based on the above considerations, we pursue detection of interest points by two independent flows, one dealing with appearance the other with 3D information. More precisely, our extension to RGB-D images deploys the standard MSD algorithm to detect interest points in the intensity channel and the previously introduced formulation for range images to achieve detection of salient 3D structures. As such, it will be referred to as RI-MSD in the remainder of this paper. The two sets of detections brought in by the two independent flows are then simply gathered together to form the final set, which means that a pixel in an RGB-D image is found to be an interest point if it has been detected by either MSD or R-MSD or both.

Finally, it is important to point out that the extension to 3D data requires particular care so to deal properly with invalid mea-

Fig. 8 Robot dataset: repeatability of interest point detectors along 4 camera paths.

measurements. Indeed, 3D sensors providing organized data can not yield a fully dense range image due to the presence of invalid measurements caused by a variety of reasons, such as, usually, specularities, transparent or non-Lambertian structures, as well as the limited sensing range of the device. This means that the range image, R , contains both 3D measurements and invalid points. To address this, we build a range image pyramid $R(l)$ where each invalid point is propagated from the highest resolution level (corresponding to the input range image) to coarser levels. At the same time, when computing the CSD saliency score at each pyramid level and at each point p , we discard invalid data as well as we accumulate, for valid ones, the number of invalid neighbors within the search area defined by ρ_a . We thus discard also those positions whose number of valid neighbors is less than a certain minimum number (set equal to k in our experiments), as the estimation of the CSD operator would not be reliable enough in those cases due to the number of neighbors being too small.

5. Experimental Results

We provide here an extensive experimental validation of the MSD detector, whereby we compare its performance with respect to the state of the art in interest point detection. Firstly, we address interest points repeatability, based on the classical *Oxford* benchmark (Section 5.1) as well as the more recent *Robot* dataset (Section 5.2) and a collection of image pairs acquired by different modalities (5.3). Moreover, we evaluate the distinctiveness of detected features through an image retrieval experiment based

on matching SIFT descriptors on the *Oxford Buildings* dataset (Section 5.4). Finally, in Section 5.5 we report the experimental findings concerning the extension of MSD to range and RGB-D images proposed in Section 4. As anticipated, in all kind of experiments and with all datasets, we ran MSD with the same set of parameters, i.e., $\rho_w = 7, \rho_a = \rho_v = 11, \tau_\delta = 250, f = 1.25$.

From the computational point of view, the incremental scheme described in Section 2.1 enables a quite efficient implementation even without deployment of the parallel multimedia-oriented instructions available in modern CPUs or advanced code optimizations. Indeed, with the parameter settings used throughout experiments, our implementation takes averagely 600 ms for image size 640×480 and 150 ms for image size 256×256 on a Intel i7 processor. We have made available an open source implementation of the MSD detector for research and education purposes^{*2}.

5.1 Evaluation on the *Oxford* Dataset

The *Oxford* dataset provides the standard benchmark for evaluating interest point – aka keypoint – detectors [21]. The dataset includes 8 planar scenes and addresses 5 nuisance factors: scale and rotation changes, viewpoint changes, decreasing illumination, blur and JPEG compression. Performance is measured according to two indicators: repeatability and quantity of correct correspondences, which account for, respectively, the relative and the absolute number of repeatable keypoints detected between the

^{*2} github.com/fedassa/msdDetector

Fig. 9 Robot dataset: repeatability of interest point detectors with the light direction changing from back to front at 7 camera positions denoted as 1, 12, 20, 25, 64, 65, and 95

first - *reference* - image of a scene and each of the other five - *corrupted* - images. Based on the *Oxford* benchmark, MSD has been compared to state-of-the-art detectors including Difference-of-Gaussian (DoG) [16], Harris-Affine, Harris-Laplace, Hessian-Affine, Hessian-Laplace [20], [21], MSER [17], FastHessian [3], and the recently introduced Wade algorithm [23]. All methods were tested using the binaries provided by the authors of [21], except for FastHessian, for which the original SURF code^{*3} was deployed, and Wade, for which we used the binaries provided by the authors^{*4}.

Figure 5 reports the performance of the evaluated detectors in terms of repeatability on the 8 image sets of the *Oxford* dataset, with each plot related to one image set. By looking at chart 5 (c) we can see that MSD delivers the highest repeatability with respect to all other detectors in case of illumination changes. As vouched by charts 5 (d), 5 (e), MSD is also quite effective in withstanding viewpoint variations: it yields overall the best invariance on *Wall* and provides the best performance between similarity rather than affine-invariant detectors on the tougher *Graf* set. It is also worth pointing out that, on *Graf*, up to 30° in-depth rotation MSD features are significantly more repeatable than those provided by affine-invariant detectors such as MSER, Hessian-

Affine and Harris-Affine. MSD is also remarkably robust to blur: charts 5 (g) and 5 (h) show that its repeatability is inferior to that of Wade only, while also providing some moderate advantages at low blur levels on the *Trees* dataset. These experimental findings seem to substantiate the conjectured inherent effectiveness of the CSD operator to highlight patches remaining quite unique within their context under illumination variations, blur and moderate viewpoint changes. As far as the other nuisances addressed by the *Oxford* dataset are concerned, charts 5 (a), 5 (b) show that MSD yields overall satisfactory scale invariance, turning out the second-best method in *Boat* and performing slightly worse than the best methods in *Bark*. Resilience to JPEG compression appears to be good alike, MSD ranking among the best methods in image set *ubc*. Considering again the comparison with established methods whose roots can be traced back to the self-dissimilarity concept, we wish to point out how MSD provides substantially better performance than the Harris-Laplace detector throughout all the experiments related to the *Oxford* dataset.

Figure 6 deals with the quantity of correct correspondences yielded by the considered detectors, a performance index also referred to as *absolute repeatability*. As witnessed by the charts, MSD delivers excellent performance in providing a high quantity of repeatable features, ranking overall among the best methods together with Wade and DoG.

^{*3} <http://www.vision.ee.ethz.ch/~surf/>

^{*4} <http://vision.deis.unibo.it/ssalti/Wave>

Fig. 10 Robot dataset: repeatability of interest point detectors with the light direction changing from right to left at 7 camera positions denoted as 1, 12, 20, 25, 64, 65, and 95.

Eventually, we have compared MSD to the proposal in Ref. [18], which detects interest points driven by the concept of patch self-similarity. Accordingly, MSD is compared to the 4 variants of the detector proposed in Ref. [18] and results are displayed in Fig. 7 for better clarity. As vouched by the charts, overall our proposal outperforms neatly all the variants proposed in Ref. [18], the margin appearing particularly substantial when it comes to nuisances such as illumination and view-point changes.

5.2 Evaluation on the Robot Dataset

The DTU Robot dataset [1] contains 60 scenes of planar and non-planar objects belonging to diverse categories and captured along four different paths by means of a camera mounted on a robotic arm. Nuisances are represented by scale and viewpoint changes as well as relighting.

To compute our results we considered the same subset of scenes as in Ref. [1] and, to obtain the charts shown in Figs. 8, 9 and 10, we plotted our curves dealing with MSD and Wade together with those reported in [1], whose data was kindly provided by their authors. Similarly to Fig. 9 in [1], in Fig. 8 we show the Recall Rate (the same performance index as the repeatability in Figs. 5 and 7) averaged across the 60 scenes for 4 camera paths that include either viewpoint changes (Fig. 8 (a)) or increasing scale variations (Fig. 8 (b)), as well as both nuisances

altogether (Fig. 8 (b), 8 (c)) due to images had been taken from different distances than the reference position and from varying viewing directions. In addition, likewise Ref. [1], we provide results addressing the resilience of detectors to relighting scenes from a different direction than in the reference image. In particular, Fig. 9 reports the mean Recall Rates across the 60 scenes for 7 camera positions and the lighting direction changing gradually from back to front (see also Ref. [1], Fig. 11), while Fig. 10 deals with the same camera positions and the light source moving gradually from left to right (as in Ref. [1], Fig. 12).

The results in Fig. 8 demonstrate that MSD outperforms all the other considered feature detectors quite neatly as far as robustness to the significant scale and viewpoint variations set forth by the Robot dataset are concerned. Its ability to withstand changes of the lighting direction turns out remarkable too, Figs. 9 and 10 showing how with both kinds of relighting MSD falls within the top performers almost at each camera position and consistently delivers the highest repeatability in 3 out of the 7 positions.

5.3 Evaluation on Multi-modal Images

MSD has also been compared to the other detectors on a dataset containing 4 image pairs acquired with different modalities, kindly provided by the authors of Ref. [11]. This dataset includes an optical-infrared pair (“square”), a multi-temporal (day-night)

Fig. 11 The 4 considered multi-modal image pairs together with the features detected by MSD on the “remote” pair (rightmost column).

Fig. 12 Repeatability and number of correct correspondences on the 4 multi-modal image pairs.

pair (“building”) and two SAR remote sensing pairs (“satellite” and “remote”). The dataset is shown in Fig. 11, together with qualitative results dealing with the interest points extracted by MSD on image pair “remote”. Results are reported in terms of both repeatability and quantity (Fig. 12). Repeatability results (left chart) demonstrate that MSD yields remarkable performance on multi-modal images, so as to turn out, in particular, the best method in 3 out of the 4 pairs. As such, it provides the highest average repeatability. Moreover, MSD provides the largest quantity of repeatable features in 3 out of the 4 pairs, and just slightly less than the largest in the remaining pair (right chart). Accordingly, it turns out neatly the best method also in terms of average quantity of repeatable features on the considered multi-modal dataset.

5.4 Evaluation by the Retrieval Test in *VLBenchmarks*

In most computer vision tasks dealing with local invariant features, the interest points detected in a set of images need to be told apart to establish as many correct point-to-point correspondences as possible. Besides repeatability, thus, it is quite a desirable detector’s property to be fired by rather distinctive image structures that, as such, may be described and matched effectively across diverse views of the subject matter under observation.

Therefore, to assess how effective MSD features may turn out

Fig. 13 Mean Average Precision on the *Oxford Buildings* dataset in the retrieval test provided with the *VLBenchmarks* framework.

when injected into a typical application pipeline, we have considered the retrieval test provided within the *VLBenchmarks*^{*5} framework, whereby keypoints detected by different algorithms can be feed into an image search engine based on matching SIFT [16] descriptors computed at the locations and scales determined by the feature detectors. In this experiment we have used the *Oxford Buildings*^{*6} dataset and, according to *VLBenchmarks*,

^{*5} www.vlfeat.org/benchmarks/#vlbenchmarks.ref2

^{*6} www.robots.ox.ac.uk/~vgg/data/oxbuildings

Fig. 14 Sample frames and number of correct correspondences yielded by RI-MSD, R-MSD and MSD for 3 sequences of the *RGB-D SLAM* “Freiburg3” dataset.

have evaluated performance through the popular Mean Average Precision (MAP) indicator. **Figure 13** reports the MAP scores yielded by the considered state-of-the-art interest point detectors. As witnessed by the chart, when described by SIFT, the image structures detected by MSD provide the top retrieval performance on *Oxford Buildings* together with *DoG*, i.e. SIFT’s own, tightly coupled, feature detector, and *Harris Affine*.

5.5 Evaluation of the RGB-D Detector

To conclude the experimental section, we present here an evaluation of the proposed extension of MSD to RGB-D images (Section 4). In particular, we have compared the three variants of the MSD detector described in this paper: R-MSD, that detects interest points on range images; RI-MSD, that combines together the features found on the range and intensity channels, and the MSD formulation dealing with the intensity channel only. The three variants are tested using two different threshold values for the 3D saliency score, i.e. $\tau_\rho = 0.25$ and $\tau_\rho = 4$, while the same default setting as in all previous experiment is used for the threshold concerning intensity saliency ($\tau_\delta = 250$).

The chosen dataset is the *RGB-D SLAM Dataset*^{*7} [25], which provides several RGB-D sequences together with the associated ground-truth camera poses. In particular, we have considered three sequences of the *Freiburg3* subset, namely *Teddy*, *Structure Texture Far* and *Sitting Static*, with the first sequence dealing with a 3D reconstruction scenario (a single object in the center of the scene), the last two addressing typical SLAM applications. A sample frame from each of the three sequences is shown above the corresponding chart in **Fig. 14**. The charts report the average number of correct correspondences per frame, that is the average quantity of repeatable interest points found between two succes-

sive frames of the sequence. As the task of detecting repeatable interest points gets more difficult as the views are farer apart from each other, we have run the experiments by applying an increasing temporal sub-sampling step to the original sequences, as denoted by the x-axis of the charts. Moreover, as we now deal with arbitrary — rather than planar — 3D scenes and calibrated cameras, to establish whether two interest points detected in a pair of frames do correspond one to another we project both in the same 3D reference frame according to the ground truth camera motion and check whether their distance is below a threshold (i.e. 2 cm).

A first finding regards intensity being likely a much richer cue to provide MSD-type interest points than 3D shape in typical indoor scenarios, as vouched by the green curve lying substantially above both the purple and cyan curves for all the difficulty levels in the two charts depicted in **Fig. 14** (b) and (c). This might not be the case should the scene include mostly evenly colored surfaces together with objects featuring rather complex 3D shapes, as it is the case of the *Teddy* puppet of the sequence in **Fig. 14** (a), though the advantage provided by 3D shape tends to decrease as the degree of difficulty of the task increases. More relevantly, leveraging on both intensity and 3D shape according to the strategy proposed in Section 4 is significantly beneficial with RGB-D images as it allows for extracting a substantially higher quantity of repeatable MSD features than relying on either of the two cues alone, as shown by the red and blue curves lying neatly above the others in all the three charts of **Fig. 14**.

6. Conclusion and Future Work

The MSD detector is fired by image patches that look very dissimilar from their surroundings, whatever the structure of such patches may be (e.g., corners, edges, blobs, textures...). Despite its simplicity, such an approach inherently conveys remarkable invariance to nuisances such as illumination changes, viewpoint

^{*7} vision.in.tum.de/data/datasets/rgbd-dataset

variations and blur. Likewise, it enables detection of repeatable features across multi-modal image pairs, as required, e.g., by remote sensing and medical imaging applications. Due to computer vision ever increasingly leveraging richer data than intensity measurements only, another favorable trait of our proposal deals with its quite seamless extendability to organized multi-dimensional data. Accordingly, we have shown how repeatable interest points may be gathered by generalizing the MSD formulation to work with the RGB-D images provided by widespread low-cost sensors such as the Microsoft Kinect and the Asus Xtion. A direction for future investigation deals with the use of approximate k-NN techniques for dense patch matching, such as [2], to possibly further ameliorate the efficiency of the detector. In addition, pairing the MSD detector with an appropriate descriptor is another topic we plan to investigate next, LSS [24] likely representing a suitable starting point.

References

- [1] Aanæs, H., Dahl, A.L. and Steenstrup Pedersen, K.: Interesting Interest Points, *Int. J. Comput. Vision*, Vol.97, No.1, pp.18–35 (2012).
- [2] Barnes, C., Shechtman, E., Goldman, D.B. and Finkelstein, A.: The Generalized PatchMatch Correspondence Algorithm, *Proc. European Conference on Computer Vision (ECCV)* (2010).
- [3] Bay, H., Ess, A., Tuytelaars, T. and Van Gool, L.: Speeded-Up Robust Features (SURF), *Comput. Vis. Image Underst.*, Vol.110, No.3, pp.346–359 (2008).
- [4] Borji, A. and Itti, L.: Exploiting Local and Global Patch Rarities for Saliency Detection, *Proc. Conf. Computer Vision and Pattern Recognition (CVPR'12)* (2012).
- [5] Bruce, N. and Kornprobst, P.: On the role of context in probabilistic models of visual saliency, *Proc. Int. Conf. Image Processing (ICIP'10)* (2010).
- [6] Buades, A., Coll, B. and Morel, J.: A review of image denoising methods, with a new one, *Multiscale Modeling and Simulation*, Vol.4, No.2, pp.490–530 (2006).
- [7] Dabov, K., Foi, A., Katkovnik, V. and Egiazarian, K.: Image denoising by sparse 3D transform-domain collaborative filtering, *IEEE Trans. Image Processing*, Vol.16, No.8 (2007).
- [8] Duin, R. and Pekalska, E.: *The dissimilarity representation for pattern recognition, a tutorial* (2010).
- [9] Goferman, S., Zelnik-Manor, L. and Tal, A.: Saliency Detection: A Spectral Residual Approach, *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'07)* (2007).
- [10] Harris, C. and Stephens, M.: A Combined Corner and Edge Detector, *Proc. Alvey Vis. Conf.*, pp.147–151 (1988).
- [11] Hel-Or, Y., Hel-Or, H. and David, E.: Fast Template Matching in Non-Linear Tone-Mapped Images, *Proc. Int. Conf. Computer Vision (ICCV)* (2011).
- [12] Hou, X. and Zhang, L.: Context-aware saliency detection, *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'10)* (2010).
- [13] Huang, J., You, S. and Zhao, J.: Multimodal image matching using self similarity, *Proc. Workshop on Applied Imagery Pattern Recognition (AIPR)* (2011).
- [14] Kadir, T. and Brady, M.: Saliency, Scale and Image Description, *International Journal of Computer Vision*, Vol.45, No.2, pp.83–105 (2000).
- [15] Loog, M. and Lauze, F.: The improbability of Harris Interest Points, *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, Vol.32, No.6, pp.1141–1147 (2010).
- [16] Lowe, D.G.: Distinctive Image Features from Scale-Invariant Key-points, *Int. J. Comput. Vis.*, Vol.60, No.2, pp.91–110 (2004).
- [17] Matas, J., Chum, O., Urban, M. and Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions, *Proc. British Machine Vision Conference, BMVC'02*, Vol.1, pp.384–393 (2002).
- [18] Maver, J.: Self-Similarity and Points of Interest, *Trans. Pattern Analysis and Machine Intelligence (PAMI)*, Vol.32, No.7, pp.1211–1226 (2010).
- [19] Mc Donnell, M.: Box-Filtering Techniques, *Computer Graphics and Image Processing*, Vol.17, pp.65–70 (1981).
- [20] Mikolajczyk, K. and Schmid, C.: Scale and Affine Invariant Interest Point Detectors, *Int. J. Comput. Vis.*, Vol.60, No.1, pp.63–86 (2004).
- [21] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T. and Gool, L.V.: A Comparison of Affine Region Detectors, *Int. J. Comput. Vis.*, Vol.65, No.1–2, pp.43–72 (2005).
- [22] Moravec, H.: Towards Automatic Visual Obstacle Avoidance, *Proc. Int. Joint Conf. Artificial Intelligence* (1977).
- [23] Salti, S., Lanza, A. and Stefano, L.D.: Keypoints from Symmetries by Wave Propagation, *Proc. Int. Conf. Computer Vision and Pattern Recognition* (2013).
- [24] Shechtman, E. and Irani, M.: Matching Local Self-Similarities across Images and Videos, *Proc. Conf. Computer Vision and Pattern Recognition (CVPR'07)* (2007).
- [25] Sturm, J., Engelhard, N., Endres, F., Burgard, W. and Cremers, D.: A Benchmark for the Evaluation of RGB-D SLAM Systems, *Proc. International Conference on Intelligent Robot Systems (IROS)* (2012).

Federico Tomba is Assistant Professor at the University of Bologna since 2013, after obtaining from the same institution a Ph.D in 2009. Since 2014, he is also Guest Professor at the Computer Aided Medical Procedures (CAMP) Chair at the Technical University of Munich (TUM). His research activity concerns computer vision

and robot perception, and it encompasses co-authoring more than 60 refereed papers. He is a member of the Open Perception foundation and a developer for the Point Cloud Library, where in 2014 he served as administrator and mentor for PCL in the Google Summer of Code. In 2015 he was the recipient a the Google Research Faculty Award. He is member of IEEE and IAPR-GIRPR.

Luigi Di Stefano received his Ph.D in Electronic Engineering and Computer Science from the University of Bologna in 1994. He is currently associate professor at the Department of Computer Science and Engineering, University of Bologna. His research interests include computer vision, image processing and computer architecture.

He is the author of more than 150 papers and five patents. He is a member of the IEEE Computer Society and the IAPR-IC. He is a member of the Scientific Council of T3LAB, a public institution devoted to technology transfer, and of the Board of Directors of Datalogic.

(Communicated by *Hideo Saito*)