Interest Points via Maximal Self-Dissimilarities

Federico Tombari, Luigi Di Stefano

DISI - University of Bologna, Italy
{federico.tombari,luigi.distefano}@unibo.it
www.vision.deis.unibo.it

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.

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 distances of a patch to those located in its surroundings, with distances usually measured through the Sum of Squared Distances (SSD). Whenever the task mandates looking for large rather than small minima over such distances, 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 [1], 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 [2] 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 [3] 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 [4], which leverages on relative positions between nearby similar patches to provide invariant representations 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 vicinity. The LSS method computes a *self-similarity surface* associated with an image point, which is then quantized 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 [4, 5]. As a further example, [6] exploits the concept of self-similarity to detect interest points associated with symmetrical regions in images. Specifically, autocorrelation 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) [7] and BM3D [8] filters exploit the presence of similar patches within an image to estimate the noiseless intensity of each pixel. In [7], 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 [8] 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. 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 Brady [9]. 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 [9].

A peculiar trait with respect to several prominent feature detectors like [10–12] is that CSD endows our approach with the ability to withstand significant, possibly non-linear, tone mappings, such as e.g. due to light changes, as well as to cope effectively with diversity in the image sensing modality. A similar concept to CSD has been exploited in [13] 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 [13], 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.

Experiments demonstrate the effectiveness of the proposed detector in finding repeatable interest points. In particular, evaluation on the standard *Oxford* dataset as well as on the more recent *Robot* dataset vouches that our method attains state-of-the-art invariance with respect to illumination changes and remarkable performance with most other nuisances, such as blur, viewpoint changes and compression. Furthermore, we show the peculiar effectiveness of the proposed approach on a dataset of images acquired by different modalities.

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[4] and the NLM filter [7]. 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) of a point p, i.e. how much the patch around p is dissimilar from the most similar one in its surroundings:

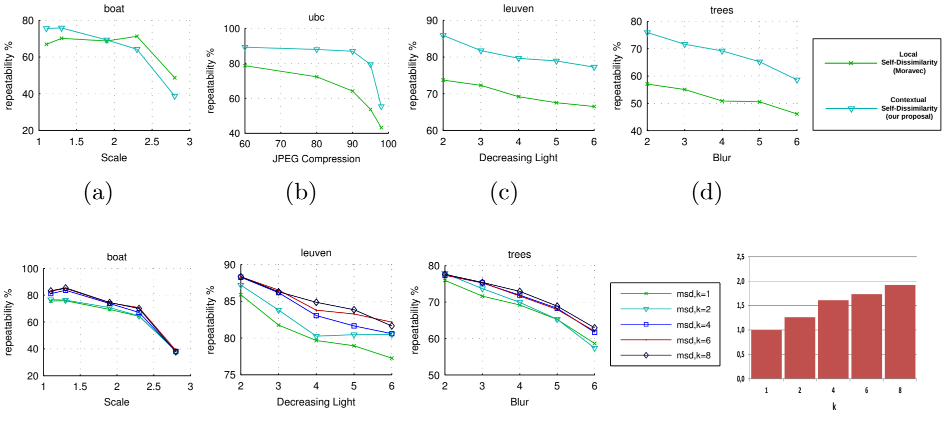
$$\lambda\left(p,\rho_w,\rho_a\right)=\frac{1}{\rho_w^2}\min_{q\in\omega\left(p,\rho_a\right),q\neq p}\delta\Big(\omega\left(p,\rho_w\right),\omega\left(q,\rho_w\right)\Big) \tag{1}$$

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

$$\delta\Big(\omega\left(p,\rho_w\right),\omega\left(q,\rho_w\right)\Big)=\parallel I\left(\omega\left(p,\rho_w\right)\right)-I\left(\omega\left(q,\rho_w\right)\right)\parallel_2^2 \tag{2}$$

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 means of 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 [14], 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 Sec. 1, we advocate replacing the local self-dissimilarity underpinning all the popular interest point detectors rooted in the Moravec operator with a contextual self-dissimilarity notion. To begin substantiating the claim, in the top-row of Fig. 1, 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 the interest points[1].

[1] Interest point detection is run at multiple scales as described in Sec.3



(a) (b) (c) (d)

(e) (f) (g) (h)

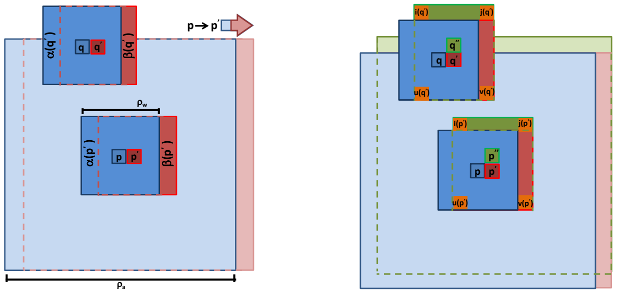
Fig. 1: Results on a subset of the *Oxford* dataset. Top row (a-d): Contextual vs. Local Self-Dissimilarity. Bottom row: repeatability (e-g) and relative execution times (h) for CSD interest points detected with different k values.

The saliency defined in (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 operation which confers robustness to noise to the LSS descriptor is the binning operation carried out by quantizing into a spatial histogram the locations of most similar patches.

Therefore, we propose to modify (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:

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

where $\tilde{\delta}^1, \cdots, \tilde{\delta}^k$ 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. Fig. 1,



(a) (b)

Fig. 2: Efficient computation of the distance between two patches: recursive scheme applied along columns (2a) and along both columns and rows (2b).

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 (3).

2.1 Computational efficiency

Computing the CSD operator over an image with n pixels implies the operation in (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 on 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. 2a, 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:

$$\delta\Big(\omega\left(p', \rho_w\right), \omega\left(q', \rho_w\right)\Big) = \delta\Big(\omega\left(p, \rho_w\right), \omega\left(q, \rho_w\right)\Big) + \\ -\delta\Big(\alpha(p'), \alpha(q')\Big) + \delta\Big(\beta(p'), \beta(q')\Big) \qquad (4)$$

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. 2b, the two distances between the corresponding sides of the patches appearing in (4) can be computed incrementally from the position just above p', denoted as p'' and highlighted in green, by adding and

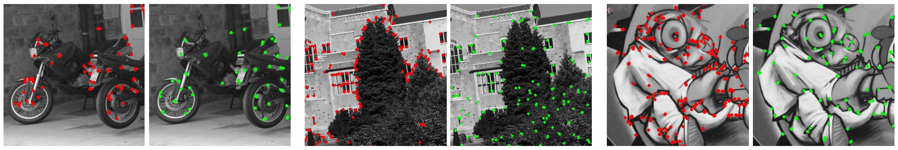


Fig. 3: Qualitative comparison between the interest points provided by MSD (green dots) and the Harris-Laplace detector [3] (red dots) on 3 image regions from the *Oxford* dataset. For clarity of visual comparison, only features of approximately the same medium-size scale are displayed for both methods.

subtracting properly the squared differences between the intensities at the four corner positions of the patches, referred to as i, j, u, v. Accordingly, equation (4) can be further manipulated so to reach:

$$\begin{aligned}\delta\Big(\omega\left(p',\rho_w\right),\omega\left(q',\rho_w\right)\Big)&=\delta\Big(\omega\left(p,\rho_w\right),\omega\left(q,\rho_w\right)\Big)-\delta\Big(\alpha(p''),\alpha(q'')\Big)+\\ +\delta\Big(\beta(p''),\beta(q'')\Big)&-\Big(I\left(i\left(p'\right)\right)-I\left(i\left(q'\right)\right)\Big)^2-\Big(I\left(j\left(p'\right)\right)-I\left(j\left(q'\right)\right)\Big)^2+\\ &+\Big(I\left(u\left(p'\right)\right)-I\left(u\left(q'\right)\right)\Big)^2+\Big(I\left(v\left(p'\right)\right)-I\left(v\left(q'\right)\right)\Big)^2 \quad (5)\end{aligned}$$

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 can instead be achieved 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 [15], allows calculating all distances between the central patch and those contained in the search area with a limited computational complexity and could be usefully deployed to reduce the complexity of self-similarity-based techniques too, such as [4, 7].

The overall algorithm to compute the saliency operator λ is showcased in Alg. 1, where for illustrative purposes only we consider the simplest case of equation (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 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

Algorithm 1 Incremental computation of the λ operator

```
for $p \in$ first row do
  for $\mathbf{q} \in \omega(p, \rho_a), \mathbf{q} \neq \mathbf{p}$ do
    $\delta_\alpha(p,q) = \delta\Big(\alpha(p), \alpha(q)\Big)$
    $\delta_\beta(p,q) = \delta\Big(\beta(p), \beta(q)\Big)$
  end for
end for
for $p \in$ all other rows do
  $\delta_{min} = inf$
  for $\mathbf{q} \in \omega(p, \rho_a), \mathbf{q} \neq \mathbf{p}$ do
    if p is the first pixel of the row then
      $\delta_\omega(q) = \delta\Big(\omega\left(p, \rho_w\right), \omega\left(q, \rho_w\right)\Big)$
    else
      $\delta_\alpha(p,q) \;\; += \;\; \delta\Big(u\left(p\right), u\left(q\right)\Big) - \delta\Big(i\left(p\right), i\left(q\right)\Big)$
      $\delta_\beta(p,q) \;\; += \;\; \delta\Big(v\left(p\right), v\left(q\right)\Big) - \delta\Big((j\left(p\right), j\left(q\right)\Big)$
      $\delta_\omega(q) \;\; += \;\; \delta_\beta(p,q) - \delta_\alpha(p,q)$
    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
```

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 potential effectiveness in presence of strong photometric distortions as well as multi-modal data, as vouched by the work related to the LSS descriptor[4]. 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:

$$L = \left\lfloor log_f \Big(\frac{min\,(w,h)}{(\rho_w + \rho_a) \cdot 2 + 1} \Big) \right\rfloor \tag{6}$$

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:

$$\frac{min\,(w,h)}{f^L} > (\rho_w + \rho_a) \cdot 2 + 1 \tag{7}$$

Once the saliency in (3) is computed at each point within the several layers of the image pyramid, for each level l the set of interest points, $\tilde{P}_l = \{\tilde{p}_1, \cdots, \tilde{p}_n\} \in I(l)$, 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_ν:

$$\tilde{p} \in I(l) s.t. \max_{p \in \omega \left(\tilde{p}, \rho_\nu \right), p \neq \tilde{p}} \lambda^{(k)} \big(p, \rho_w, \rho_a \big) < \lambda^{(k)} \big(\tilde{p}, \rho_w, \rho_a \big) \tag{8}$$

As the features detected through the NMS stage are local maxima of the CSD operator, our proposal will be hereinafter 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:

$$i(\tilde{p}) = i_l \cdot f^l \qquad j(\tilde{p}) = j_l \cdot f^l \qquad s(\tilde{p}) = (\rho_w \cdot 2 + 1) \cdot f^l \tag{9}$$

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 $\omega\big(\tilde{p}, \rho_a\big)$ 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 [1], Harris [2] and, more recently, the Harris-Laplace and Harris-affine detectors[3, 16], 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. 3 we compare qualitatively the interest points extracted by MSD to those provided by the Harris-Laplace (*harlap*) scale-invariant corner detector [3]. 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-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 $\rho_w = 7$, $\rho_a = 11$.

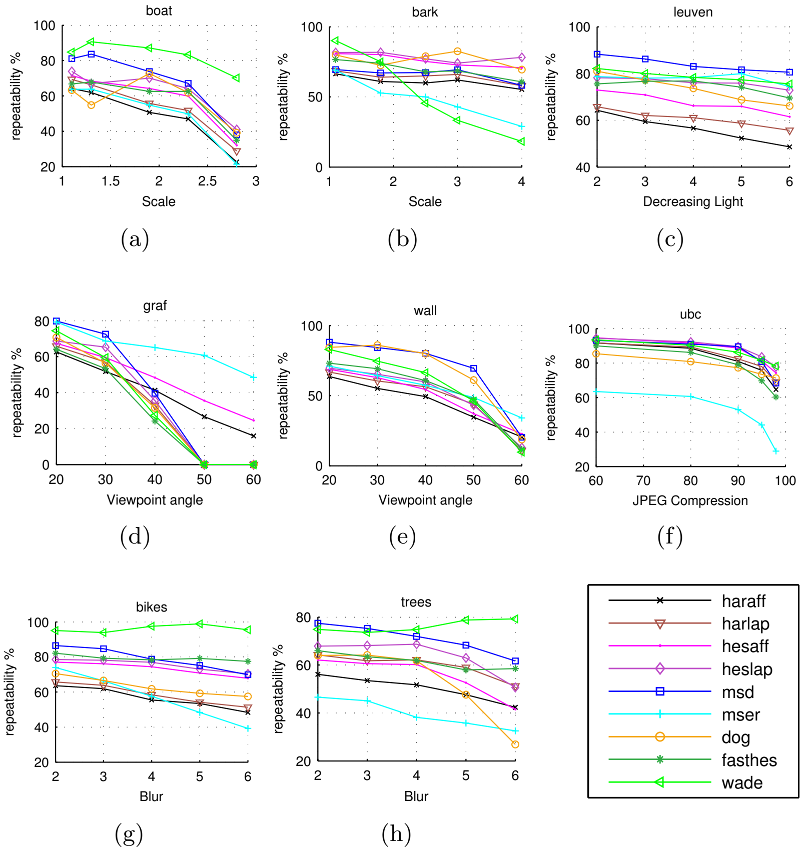
4 Experimental results

To assess its performance, we compare here the proposed MSD algorithm to the state of the art in interest point detection. We consider first the standard *Oxford* benchmark dataset (4.1), then the more recent *Robot* dataset (4.2) and finally an additional dataset made out of image pairs acquired by different modalities (4.3). As anticipated, in all experiments we have ran MSD with the same set of parameters, i.e. $\rho_w = 7, \rho_a = \rho_\nu = 11, \tau_\delta = 250, f = 1.25$.

From the computational point of view, the incremental scheme outlined in Sec. 2.1 enables a quite efficient implementation even without advanced optimizations or deployment of the parallel multimedia-oriented instructions available in modern CPUs. Indeed, with the parameter settings used in the experiments, our implementation takes averagely $600ms$ for image size 640×480 and $150ms$ for image size 256×256 on a Intel i7 processor.

4.1 Evaluation on the *Oxford* dataset

MSD has been tested on the *Oxford* dataset, a benchmark for keypoint detection evaluation introduced in [16]. The dataset includes 8 planar scenes and 5



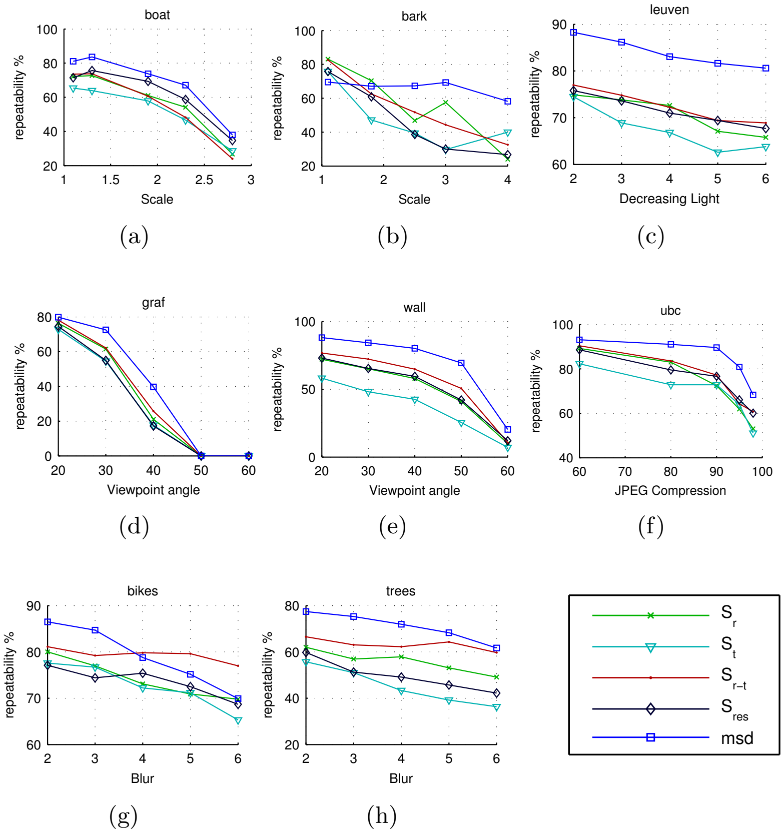
(a) (b) (c)

(d) (e) (f)

(g) (h)

Fig. 4: Repeatability on the 8 sets of images of the Oxford dataset. The x axis denotes the level of difficulty of the considered nuisance.

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 first - *reference* - image of a scene and each of the other five - *distorted* - images. Our proposal has been compared with state-of-the-art detectors including Difference-of-Gaussian (DoG)[10], Harris-Affine, Harris-Laplace, Hessian-Affine, Hessian-Laplace [3, 16], MSER [11], FastHessian [12], and the recently introduced Wade algorithm [17]. All methods were tested using the binaries provided by the authors of [16], except for FastHessian, for



(a) (b) (c)

(d) (e) (f)

(g) (h)

Fig. 5: Comparison between MSD and the 4 variants of the proposal in [6] on the Oxford dataset.

which the original SURF code[2] was deployed, and Wade, for which the binaries provided by the authors[3] were used.

Figure 4 reports the performance of the evaluated detectors in terms of repeatability on the 8 image sets of the *Oxford* dataset, with each plot in both figures related to one image set. By looking at chart 4c we can see that MSD delivers the highest repeatability with respect to all other detectors in case of illumination changes. As vouched by charts 4d, 4e, 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*, MSD features are significantly more repeatable up to 30° in-depth rotation than

[2] http://www.vision.ee.ethz.ch/ surf/

[3] http://vision.deis.unibo.it/ssalti/Wave

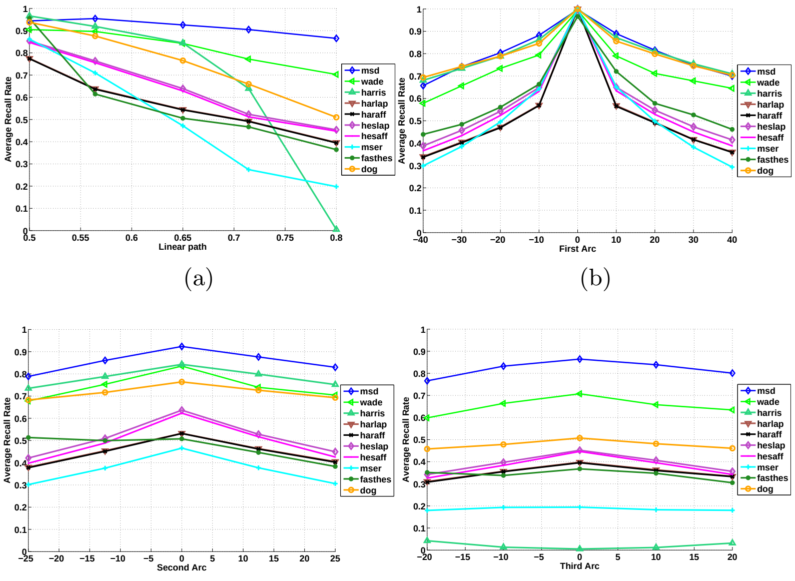
those provided by affine-invariant detectors such as MSER, Hessian-Affine and Harris-Affine. MSD is also remarkably robust to blur: charts 4g and 4h show that its repeatability is surpassed only by Wade, while also providing some moderate advantage at low blur levels on the *Trees* dataset. These experimental 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 4a, 4b 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. Due to lack of space, we include the results dealing with the quantity of correct correspondences together with examples of detected features in the supplementary material. Yet, we wish to highlight here that also in terms of number of repeatable features MSD provides excellent performance, ranking among the best methods on this dataset together with Wade and Dog.

In addition to previous results, we have compared our method to the proposal in [6], which detects interest points driven by the concept of patch self-similarity (for better clarity, the results are displayed in a distinct figure, i.e. 5). As for this experiment, MSD is compared on the *Oxford* dataset to the 4 variants of the detector tested in [6]: as vouched by the charts, overall our proposal outperforms neatly all the variants proposed in [6], the margin appearing particularly substantial when it comes to nuisances such as illumination and view-point changes.

4.2 Evaluation on the *Robot* dataset

We have also evaluated MSD on the more recently introduced DTU *Robot* dataset [18]. This dataset contains 60 scenes of planar and non-planar objects, from different categories captured along four different paths by means of a robotic arm. As for this dataset, nuisances are represented mostly by scale and viewpoint changes as well as relighting. Due to space constraint, we could not include results on the whole dataset. Thus, as MSD already showed state-of-the-art performance with respect to illumination changes on the *Oxford* dataset, we have focused the evaluation on the scene subsets covering increasing scale variations (i.e., *linear path*) and different viewpoint changes (i.e., *first arc*, *second arc* and *third arc*, these last two also including scale variations since they were acquired at different distances from the reference image).

Results shown in Figures 6a-6d report the Average Recall Rate (analogous of the Repeatability) at increasing scale variations (Figure 6a) and different viewpoint angles (Figures 6b-6d). To plot these charts, we added the MSD and Wade curves to those shown in [18] (whose data was kindly provided by their authors). These results show that MSD keypoints yield outstanding repeatability even



(a) (b)

(c) (d)

Fig. 6: Comparison of interest point detectors over the *Robot* dataset.

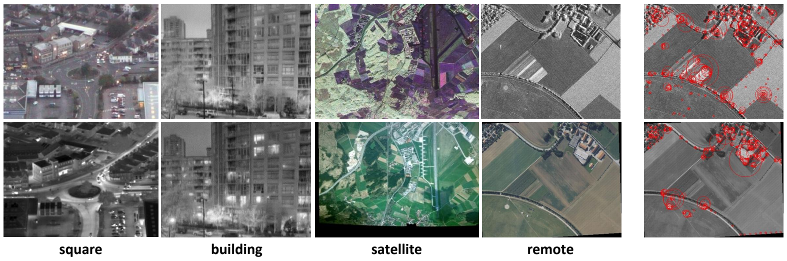


Fig. 7: The 4 considered multi-modal image pairs together with the features detected by MSD on the "remote" pair (rightmost column).

when tested at high scale differences and notable viewpoint changes, remarkably outperforming all state-of-the-art methods on each evaluated scene subset. Also, the higher the scale variation, the higher the gap between MSD and the state of the art: this can be noticed especially in Figure 6a and by considering that scale variations increase moving from the *first arc* through the *third arc*.

4.3 Evaluation on multi-modal images

Finally, MSD has been compared to the other considered detectors on a dataset containing 4 image pairs acquired with different modalities, kindly provided by

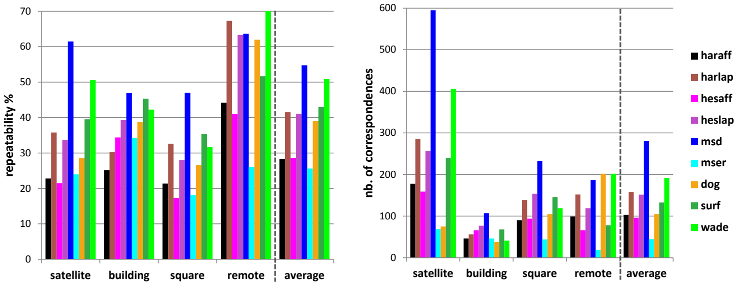


Fig. 8: Comparison of interest point detectors over 4 pairs of images related to different modalities in terms of repeatability and number of correct correspondences.

the authors of [19]. This dataset includes an optical-infrared pair ("square"), a multi-temporal (day-night) pair ("building") and two SAR remote sensing pairs ("satellite" and "remote"). The dataset is shown in Fig. 7, 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. 8). 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 in terms of average quantity of repeatable features on the considered multi-modal dataset.

5 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 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. Peculiarly, the MSD approach generalizes straightforwardly to detect interest points in any kind of multi-channel images, such as color 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. Another direction for future investigation deals with the use of approximate k-NN techniques for dense patch matching, such as [20], to possibly further ameliorate the efficiency of the detector. Finally, pairing the MSD detector with an appropriate descriptor is another topic we plan to investigate next, LSS [4] likely representing a suitable starting point.

References

1. Moravec, H.: Towards automatic visual obstacle avoidance. In: Proc. Int. Joint Conf. on Artificial Intelligence. (1977)
2. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proc. Alvey Vis. Conf. (1988) 147–151
3. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. Int. J. Comput. Vis. **60** (2004) 63–86
4. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: Proc. Conf. on Computer Vision and Pattern Recognition (CVPR'07). (2007)
5. Huang, J., You, S., Zhao, J.: Multimodal image matching using self similarity. In: Proc. Workshop on Applied Imagery Pattern Recognition (AIPR). (2011)
6. Maver, J.: Self-similarity and points of interest. Trans. on Pattern Analysis and Machine Intelligence (PAMI) **32** (2010) 1211–1226
7. Buades, A., Coll, B., Morel, J.: A review of image denoising methods, with a new one. Multiscale Modeling and Simulation **4** (2006) 490–530
8. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3d transform-domain collaborative filtering. IEEE Tran. Image Processing **16** (2007)
9. Kadir, T., Brady, M.: Saliency, scale and image description. International Journal of Computer Vision **45** (2000) 83–105
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60** (2004) 91–110
11. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proc. British Machine Vision Conference. Volume 1 of BMVC'02. (2002) 384–393
12. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Comput. Vis. Image Underst. **110** (2008) 346–359
13. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'10). (2010)
14. Borji, A., Itti, L.: Exploiting local and global patch rarities for saliency detection. In: Proc. Conf. on Computer Vision and Pattern Recognition (CVPR'12). (2012)
15. Mc Donnel, M.: Box-filtering techniques. Computer Graphics and Image Processing **17** (1981) 65–70
16. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. Int. J. Comput. Vis. **65** (2005) 43–72
17. Salti, S., Lanza, A., Stefano, L.D.: Keypoints from symmetries by wave propagation. In: Proc. Int. Conf. on Computer Vision and Pattern Recognition. (2013)
18. Aanæs, H., Dahl, A.L., Steenstrup Pedersen, K.: Interesting interest points. Int. J. Comput. Vision **97** (2012) 18–35
19. Hel-Or, Y., Hel-Or, H., David, E.: Fast template matching in non-linear tone-mapped images. In: Proc. Int. Conf. on Computer Vision (ICCV). (2011)
20. Barnes, C., Shechtman, E., Goldman, D.B., Finkelstein, A.: The generalized PatchMatch correspondence algorithm. In: Proc. European Conference on Computer Vision (ECCV). (2010)