



Research Article

Supervised framework for top-down color interest point detection



Yogeshwar Singh Dadwhal^{1,2}  · Satish Kumar^{1,3} · Harish Kumar Sardana^{1,3}

Received: 20 January 2020 / Accepted: 8 July 2020 / Published online: 22 July 2020

© Springer Nature Switzerland AG 2020

Abstract

Selective attention is one of the dominant properties of the biological visual system to locate regions of interest in the scene. This article presents a local feature detector based on selective attention. Most previous approaches are bottom-up and do not consider prior information for known object categories. They detect feature points using spatial information present in the image. In contrast, this article presents an attention inspired algorithm that encapsulates the second-moment matrix-based detector to identify feature points based on some color information selectively. The intuition is to use skin color as a top-down cue to generate interest points, finding a high similarity to known spectra. It has applications in a variety of real-time applications such as image retrieval, gesture classification, virtual reality, etc. The technique is inspired by the human visual perception to gain cognizance of regions based on selective boosting of colors. The properties of color models are used to form a distinctiveness function to suppress unwanted background clutter. A relationship between interest points and salient colors in the image is computed using partial correlations in color derivative space. The system is evaluated on the MSRA dataset commonly used for saliency detection. The experiments are based on finding distinct regions having an affinity towards skin color. The robustness of the algorithm is tested in a realistic scenario by separate training and testing datasets. Experimental results show a high level of repeatability for different noise variations, image compression, and blur. The simplicity, robustness, and efficiency of the technique to locate color interest points make it appropriate for real-time vision systems.

Keywords Distinctiveness · Color interest points · Feature detector · Saliency

1 Introduction

Computer vision is one of the most active fields of research for providing autonomy to machines and computers. With a lot of advancement in electronic gadgets and surveillance systems, there is an abundance of visual data for generating inferences. It has created a lot of research interest in extracting features from digital images in different applications. Most applications in computer vision use low-level features to extract information from the images. It refers to identifying important locations (key-points) on the image that can efficiently describe the content present in it using a feature descriptor. These features

discriminately identify different objects in the images. However, there can be variations in data on which the system is tested and can hinder its performance. This can also be due to different environmental factors affecting the imaging conditions.

In real applications, an object of interest may be present with background clutter in the form of other objects. These objects may be trees, poles, furniture etc. which may lead to false detections. The presence of clutter in the scene reduces the efficiency of detectors and classifiers to locate objects. Occlusions are also a type of clutter that inhibits the detection of an object. In these situations, interest points within the image having more affinity towards the

✉ Yogeshwar Singh Dadwhal, dadwhal.yogeshwar@gmail.com | ¹Academy of Scientific and Innovative Research (AcSIR), (CSIR-CSIO), Ghaziabad 201002, India. ²Shri Mata Vaishno Devi University, Katra, J&K 182320, India. ³CSIR-Central Scientific Instruments Organisation, Chandigarh 160030, India.



object of interest based on physical properties such as color, spatial characteristics, etc. may be useful. For keypoints detection, the majority of state-of-art point detectors do not rely on the physical properties of the object to be classified. In applications such as victim detection, face detection, human body tracking, action recognition, etc. the presence of skin color can be used as an important cue to locate interest points that may provide us with a better understanding of the scene. This may be achieved by mimicking the biological visual system, which uses heuristics to identify and locate the presence of object categories based on color.

Biological vision systems are useful in finding salient regions based on the prior color and texture information of objects such as pedestrians, vehicles, plants, obstacles, signs, etc. Saliency is defined as the property of an object in a scene due to which it stands out in contrast to its neighborhood. It has been used for various applications such as object detection [1], object segmentation [2], object appearances [3] and interest point detection [4, 5]. In the context of primates, attention is usually top-down and goal-directed (memory-dependent). Based on constant learning from birth, a model to attention develops in the primates' visual cortex to look for known object categories. Over the years, several techniques have evolved using prior information to find saliency [6–8]. Though there is an abundance of bottom-up (memory-free) approaches [4, 9, 10], this paper restricts to learning model for color attention. A handful of researchers have leveraged color information to guide a saliency for an object category [11–13] and the majority of techniques for point matching work on gray level images, using the spatial information to characterize the interest points. To learn visual attention [14], a Bayesian framework combining low-level saliency is proposed, which was later extended to learn object appearances [3]. A knowledge-based saliency model [15] derived from the Bayesian framework is used to learn the appearance of objects.

Similarly, a Bayesian approach is used to find a likelihood function [16] to learn the target class and further detect potential objects. One efficient way to learn weights is to form category-specific dictionaries in a Conditional Random Field (CRF) framework [6, 7]. These methods use graphical models to generate saliency maps of learned categories. Visual similarity has been the key to learning top-down attention between images. A known test image is used to retrieve matching images from the database [17], learning a Fisher kernels based classifier. This classifier helps in separating salient and non-salient regions. This approach seems intuitive; however, it has computationally higher costs due to visual search and matching.

Moreover, the majority of these approaches use the appearance of objects to learn a model that can

often mislead during the testing phase when objects are occluded or undergo certain deformations such as humans. For instance, learning the appearance of a pedestrian does not guarantees its detection if the shape of the human body is deformed. So detecting interest points that are less dependent on the shape of the object may be helpful.

Interest point detection or simply keypoints detection helps in locating regions that provide high information content and are robust to environmental variations such as illumination, blur, compression, viewpoint, scale, etc. Some keypoint detectors are Harris detector [18], Color boosting with Harris detector [19, 20], Shi-Tomasi [21], Difference of Gaussian (DoG) [22], Laplacian of Gaussian (LoG) [23], etc. They are used to make a reference model of an object using local descriptors such as Scale Invariant Feature Transform (SIFT) [22], Speed-Up Robust Features (SURF) [24], Binary robust independent elementary features (BRIEF) [25]. These descriptors are then matched against the test data using a selected similarity score for further classification. These techniques are efficient when there is a limited number of discriminatory invariant keypoints, which is not always the case. So out of all the detected keypoints, there exist many redundant keypoints due to background clutter, which may degrade the performance of discriminatory features for classification.

The contribution of this paper is threefold. It presents an extension of traditional bottom-up models to the human-centric top-down model. Humans, while looking for certain object categories, use prior information to locate objects [26], in contrast, to bottom-up models, e.g., while driving a car a person looks at the road (distinct unique color) to navigate. A novel way to incorporate visual perception of color, inspired distinctiveness function is proposed based on prior color information. And lastly, identifying salient points based on color and spatial statistics to locate closest data points to the learned model, e.g., while looking for disaster victims, the rescuers tend to look for the presence of skin, blood in the debris.

2 Related work

Primitive features for image representation are defined in terms of edges, corners, template region properties, pixel values, boundary, etc. These features are detected at locations having high information content using edge, corner and blob detection by making use of gradients, templates, pixel values etc. Differentiation based methods locate edges by finding local maxima's of gradient values by using Sobel and Prewitt filters. Laplacian of Gaussian [23] uses second-order differentiation to find zero crossings to locate edges. A corner is an intersection of two edges,

and intuitively there is a lot of information content around the corners as it has two different gradient orientations. Harris corner detector [18] is based on finding autocorrelation of gradients on shifting windows over the image. Other detectors, such as Kanade–Lucas–Tomasi (KLT) [27] and Shi-Tomasi [21] detector are based on gradient calculations. These gradients provide illumination invariance to the detector but are highly sensitive to spatial noise. Another Low Complexity Corner (LOCOCO) detector [28, 29] is based on Harris and KLT corner measure. It approximates first-order Gaussian derivative by a box kernel and computes gradient-based integral images followed by non-maximal suppression. Other methods of locating corners use template matching by comparing the intensity of surrounding pixels around a corner. One such method is the Smallest Univalued Segment Assimilarity Nucleus (SUSAN) [30], where intensity differences are calculated by every pixel inside a circular area with the center pixel. All centre pixels having these differences less than a threshold are labelled as corners. Another corner detector is Features from Accelerated Segment Test (FAST) [31] where a point is considered as a corner if some pixels in the circular region are brighter or darker than the centre pixel. The computational cost of these detectors is lower than gradient-based methods. However, the stability of template-based corner detector is less compared to gradient-based methods. The differentiation-based methods focus on local region contrast and hence are sensitive to noise. Statistical learning using multiple cues have shown significant advantages in terms of robustness and efficiency [32]. Many classical methods such as LoG, DoG, Hessian–Laplacian etc. use scale-space feature detection to achieve scale invariance of feature points. Majority of feature descriptors use these feature detectors to locate the point of interest. SIFT [22] uses local extremes of DoG pyramid to detect keypoints and Hessian matrix as a measurement function. SURF [24] uses filters to approximate Hessian matrix to detect interest points. An extension of this algorithm is presented in [33] using centre-surround differencing to locate keypoints. DART [34] is another version of the keypoint detector using weighted triangular responses to approximate determinant of the Hessian matrix. In recent years detectors based on non-linear Partial Differential Equations (PDE) are used to extract features such as KAZE features [35]. Additive operator splitting scheme is used to solve the PDEs to find local extrema by non-linear diffusion filtering, which leads to a high computational cost. Wave-based Detector (WADE) [36] is another interest point detection algorithm based on the wave equation to isolate salient symmetries. Some detectors find regions of interest such as Maximally stable extremal region (MSER) [37] which exploit constancy of image properties to locate regions of interest. An extension of MSER using color information

is Maximally Stable Color Region (MSCR) [38]. It uses color distances by finding Poisson statistics of pixel values by grouping pixels with a similar color. The reader is referred to comprehensive surveys [39, 40] for more information on feature detectors and descriptors.

All the feature detectors described above define models and algorithms that apply directly to the image. An alternative is to train a model and then apply the model to the image. This approach is called a top-down approach where prior information is fed to the model. The feature detectors proposed so far are generalized detectors. They do not have any pretraining of the objects of interest, and hence redundant keypoints are suppressed by manual tuning. The use of color attention in images is proposed to locate points of interest. Color attention (CA) [11] is used to modulate the shape words in a Bag-of-words approach in an image during histogram construction. The bag-of-words framework has its roots in natural language processing; an object is represented in the form of multiset objects and is commonly used to handle occlusions in computer vision applications [42, 43]. Multiple cues from shape and color are used herein in separate stages. Color is used to guide attention, and further shape features are modulated within the image where the probability of occurrence of the object is more. Color attention is dependent on the occurrence frequency of color within a patch-category, and hence different attention values are assigned to different colors. In another approach [12], class-specific discriminative colors define an object. The attention maps do not include objects having discriminative colors. The shortcomings of this approach were handled using object patches [13] and dividing these patches into strong and weak patches. Lastly, the attention maps modulate the weights of Bag-Of-Words for image representation. The location of an object in the scene can be determined by statistics of low-level features [26] using guided attention. These techniques are based on patches and cannot handle a certain amount of the noise and blur variations in the image. To deal with these challenges, the technique presented in this article works at the pixel level and generates a map which is further used to locate interest points.

Keypoints are detected using some saliency measure to enable a sparse search for locating objects in clutter. Due to its ease and robustness, the Color Harris detector is frequently used for extracting color features [4, 10]. A color Gaussian pyramid is used in [5] to make these points scale-invariant. These interest points are detected using color information for texture classification. Frequently occurring points in the pyramid and color channels are filtered as interest points. The properties of color descriptors [44], show invariance to some categories of objects and are robust to light variations. In [45], color information based shadow, shading illumination and specularities invariant

interest points are detected using Lambertian and specular reflection. The method uses fixed scales to match images in varying illumination. Harris second moment matrix is used to locate points. Later, color distinctiveness was used to locate interest points in the image. Boosting of color using local differential structure in images was studied in [19] to locate salient points. This preliminary work based on color boosting hypothesis was inspired by the information theory, stating that the information contained in color derivatives increases if the probability of occurrence of a descriptor is small, which forms the basis of color saliency boosting function. This approach provides a salient point detector robust to varying illumination and shadowing effects in quasi-invariant space. On similar grounds, the approach was extended by introducing a scale selection strategy to detect sparse interest points based on color discriminative and invariant properties [20]. Conventionally, color-based keypoint detectors follow a new bottom-up paradigm to find attention maps followed by feature descriptors and vocabulary construction. The generation of attention maps is independent of the object categories to be detected. This paper proposes a method to compute a model based on prior color information to generate an attention map and further locate keypoints in the image, making it a more application-specific approach.

3 Proposed method

The proposed method is divided into two stages; In the first stage, a skin color model in the form of attention vectors is learned, referred to as a training phase in this article. The second stage finds distinct color interest points in the image and referred to as the testing phase. Figure 1 shows an overview of the proposed algorithm. In the training phase, a set of color data points are used to

generate inference to compute an affinity model. In this stage, the data points are transformed into another space, having less residual variance making it more suitable to understand the context of the scene based on color. The testing stage calculates the skin affinity map by applying the model to the input color image. It helps in generating advanced features which can be further utilized for various applications. Color Harris energy points are detected to check their invariance to various disturbances as in [19].

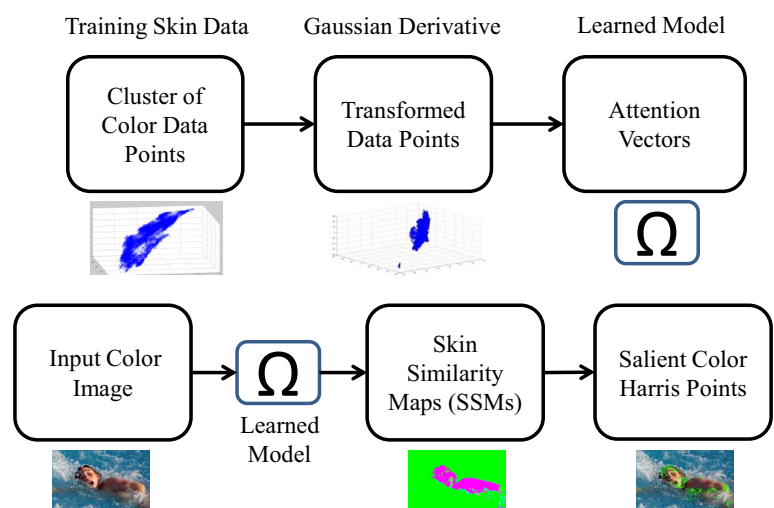
3.1 Gaussian derivative transformation

Learning a skin affinity model includes preparing the data points for their transformation into another space. The data points consist of skin color pixels generated from different skin tones and varying illumination conditions. The training data, skin pixels are reshaped in the form of a matrix for an easy transformation. Incorporating color to find skin similarities in the image leads to a computationally efficient model based on differential color structure.

It is observed that the structure of skin color in Red–Green–Blue (RGB) space is not sufficient to find a solution to locate pixels with an affinity towards the skin. It is due to multiple clusters of skin color in RGB space, which makes it difficult to find an efficient boundary. Hence, the data is transformed into differential vector space to explore the hidden structure within the skin data, proposed in this section.

Skin Similarity Maps (SSM) are proposed to handle skin tone and illumination variations. This work aims to find a suitable transformation that would attenuate these variations by exploring the hidden differential structure in RGB channels. A simple addition of differential color channels results in the cancellation of opposing vectors [46]. This will lead to an unwanted addition of noise and removal of corner and T junctions in the image structure.

Fig. 1 Overview of the proposed algorithm



A combination of differential structure in color images [19] was extracted by incorporating principles of color statistics. The RGB values are transformed into the Gaussian color model space by finding spatial derivatives computed with spatial Gaussian derivative operator convolution at scale σ .

the other end of the diagonal represents lighter skin tones. The information contained in a color cloud depends on the probability of the derivatives. Differential color channels are known to be correlated due to the physical nature of the color model. The first-order derivative of the color

Algorithm: Proposed color interest point detector

Input: Training data δ_λ (Skin pixel map), test image $I_{\{r,g,b\}}(x,y)$

1. $\sigma \leftarrow 1.0$
2. $x \leftarrow -3 : 1 : 3$
3. $Gauss \leftarrow 1/(\sqrt{2 \times \pi i}) \times \sigma \times \exp(-(x^2)/(-2 \times \sigma \times \sigma))$
4. for (R,G,B channels of δ_λ compute gradient in x-direction as $[R_x; G_x; B_x]$):
5. $G_x \leftarrow -(x/\sigma^2) \times Gauss$
6. $G_x \leftarrow G_x/(\text{sum}(\text{sum}(x \times G_x)))$
7. $G_y \leftarrow Gauss/\text{sum}(Gauss)$
8. $\delta_{\lambda_x} \leftarrow G_x \otimes \delta_\lambda$
9. $\delta_{\lambda_x} \leftarrow G'_y \otimes \delta_{\lambda_x}$
10. for (R,G,B channels of δ_λ compute gradient in y-direction as $[R_y; G_y; B_y]$):
11. $G_x \leftarrow Gauss/\text{sum}(Gauss)$
12. $G_y \leftarrow -(x/\sigma^2) \times Gauss$
13. $G_y \leftarrow G_y/(\text{sum}(\text{sum}(x \times G_y)))$
14. $\delta_{\lambda_y} \leftarrow G_x \otimes \delta_\lambda$
15. $\delta_{\lambda_y} \leftarrow G'_y \otimes \delta_{\lambda_y}$
16. Transform into a 1-jet descriptor
17. $G_{\{r,g,b\}\lambda}^{x,y} \leftarrow [R_{xy}; G_{xy}; B_{xy}]'$
18. Calculate the partial correlation within all set of variables.
19. $S_\rho \leftarrow \text{covariance}(G_{\{r,g,b\}\lambda}^{x,y})$ using eq 3
20. $\Theta \leftarrow S_\rho^{-1}$
21. For a test image I , multiply it with vectors obtained as in eq 4.
22. $SSM_{\{r',g',b'\}}(x,y) \leftarrow \Theta * I$
23. Compute gradients for $SSM_{\{r',g',b'\}}(x,y)$ for all pixels :
24. $(r_x, g_x, b_x) \leftarrow \text{repeat steps 4 to 9 for } SSM(R, G, B) \text{ instead of } \delta_\lambda$
25. $(r_y, g_y, b_y) \leftarrow \text{repeat steps 10 to 15 for } SSM(R, G, B) \text{ instead of } \delta_\lambda$
26. For each pixel compute:
27. $L(x, y, \sigma_D) \leftarrow \text{eq 5 by looping over neighbors } x, y$
28. $M(p, \sigma_I, \sigma_D) \leftarrow \text{eq 6}$
29. Repeat steps 26 to 28 for different scales.
30. Compute the eigenvalues using:
31. $E(p, \sigma_I, \sigma_D) \leftarrow \text{eq 7}$
32. At each scale find local maxima using a 3×3 window to generate interest points:
33. **Interest points at different scales** $\leftarrow E(p, \sigma_I, \sigma_D) > \text{threshold}$

Output: Interest points in scale space

The data points in the RGB color model have maximum variance along the diagonal of the RGB cube, which represents grayscale intensity value as it has equal amounts of red, green and blue components. All data points form a cloud around this diagonal. Closer, the data points to the origin represent darker skin tones and data points closer to

cloud yields a remarkably compact and straightforward structure that resembles an ellipsoid [19, 20]. The vectors in \mathfrak{R}^3 can be computed to approximate this structure. These vectors represent the skin color model for generating SSMs.

The training data in the form of skin-colored pixels are described in the form of an image patch having a specific chromatic characteristic in RGB space given by $\delta_\lambda \in \{R_\lambda, G_\lambda, B_\lambda\}$, where λ specifies the small band of skin color tones for training the model. The derivative vectors for scale σ in the neighborhood are given by

$$g = \left(\delta_{r\lambda_x}, \delta_{g\lambda_x}, \delta_{b\lambda_x}, \delta_{r\lambda_y}, \delta_{g\lambda_y}, \delta_{b\lambda_y} \right)^T \tag{1}$$

where $\delta_{r\lambda_x}$ is the red channel Gaussian derivative along x direction for an image patch consisting of skin colored pixels. Equation (1) is a 1-jet descriptor and its components are obtained by Eq. (2) using Gaussian filters in x and y-direction. The Gaussian derivative filter is used to capture the physical structure of color, using a spatial convolution kernel G_σ with scale $\sigma = 1$ to extract information from a limited set of color bands centered at λ .

$$G_{\{r,g,b\}\lambda}^{x,y} = G_\sigma^{x,y} \otimes \delta_{i\lambda}, \quad i \in \{r, g, b\} \tag{2}$$

Here the symbol \otimes represents the 2-D convolution operator. These color features are extracted by convolving the training set of pixels with a Gaussian operator in horizontal and vertical directions. It helps in finding a differential color structure for different spatio-chromatic levels, which can be leveraged to learn a skin similarity function. The RGB channels in $\delta_\lambda \in \{R_\lambda, G_\lambda, B_\lambda\}$ and vectors in the transformed Gaussian space are correlated due to the physical properties of colors. Equation (2) shows the transformation of color data points $\delta_{i\lambda}$ into Gaussian derivative space where $\delta_{i\lambda}$ is the data centered at λ (skin color), This transformation present a dynamic solution to skin color structure. In order to find the vectors that represent the information of skin color for selectively color boosting, the inverse of the covariance matrix is computed for this transformation.

3.2 Skin similarity maps

Vector weights of the model can be obtained by computing a covariance matrix S_ρ , which is real, symmetric and qualifies for spectral decomposition. Matrix S_ρ can be decomposed into eigenvector matrix U and eigenvalue Λ .

$$S_\rho = \begin{pmatrix} \overline{R_x R_x} + \overline{R_y R_y} & \overline{R_x G_x} + \overline{R_y G_y} & \overline{R_x B_x} + \overline{R_y B_y} \\ \overline{G_x R_x} + \overline{G_y R_y} & \overline{G_x G_x} + \overline{G_y G_y} & \overline{G_x B_x} + \overline{G_y B_y} \\ \overline{B_x R_x} + \overline{B_y R_y} & \overline{B_x G_x} + \overline{B_y G_y} & \overline{B_x B_x} + \overline{B_y B_y} \end{pmatrix} \tag{3}$$

The matrix elements of the covariance matrix S_ρ in equation [3] are computed by $\overline{R_x R_x} = \sum \sum R_x R_x$. It represents the information of unconditional correlation between all pairs set of variables, which can be computed by finding the partial correlation values within all sets of variables. For

skin training data, it can be assumed that these variables are neither conditionally independent nor equal to zero.

The selective boosting of skin colors for a test color image $I_{\{r,g,b\}}(x, y) = (R, G, B)^T$ can be computed using these vector weights. The aim is to find a map which selectively boosts skin hue and discards the rest of the hues in the image. The desired color transformation or the skin similarity map (SSM) is obtained by:

$$SSM_{\{r',g',b'\}}(x, y) = \Theta \left(G_{\{r,g,b\}\lambda}^{x,y} \right) * I_{\{r,g,b\}}(x, y) \tag{4}$$

In the above equation, Θ is the transformation to compute the precision matrix which represents the information content of training data, calculated by eigen decomposition $U\Lambda^{-1}U^T$ of equation [3]. This matrix consists of weights having an affinity towards skin color. Both $I_{\{r,g,b\}}(x, y)$ and $SSM_{\{r',g',b'\}}(x, y)$ are in R^3 . The skin similarity map separates the entire image into different weights depending on the affinity with the trained model. Here the input image and the training data has three channels, and the computations are simple to calculate. This approach lays the foundation of finding interest points in image cubes having many spectral bands where the interpretation of objects of interest is non-trivial. The methodology can thus be extended to hyperspectral images having high spectral dimensions.

3.3 Color interest points

The proposed skin attention map is used to locate salient interest points in the image using the second-order moment based detector. Harris and Stephens [18] proposed a second-moment matrix-based detector, sometimes referred to as an auto-correlation matrix describing local structure in images. It was further extended to color images by computing local derivative using Gaussian kernels of scale σ_D further averaged in the neighborhood of the point by a Gaussian of scale σ_I [10].

$$L(x, y, \sigma_D) = \begin{bmatrix} (r_x^2 + g_x^2 + b_x^2) & (r_x r_y + g_x g_y + b_x b_y) \\ (r_x r_y + g_x g_y + b_x b_y) & (r_y^2 + g_y^2 + b_y^2) \end{bmatrix} \tag{5}$$

$$M(p, \sigma_I, \sigma_D) = G(\sigma_I) \otimes L(x, y, \sigma_D) \tag{6}$$

where r_x is the result of the convolution of the red component of an image with the first derivative of the Gaussian kernel of scale σ_D in the x-direction, subscript to each color channel refers to the differentiation with respect to the parameter, \otimes denotes 2D convolution of the gaussian kernel of size $\sigma_I = 1$. All components of the second moment matrix are computed by finding gradients for color channel, following the multiplication and summation of these



Fig. 2 Images from the MSRA salient object database [41]

gradients as in [47]. A corner measurement based on Eigenvalues of M is then computed by:

$$E(p, \sigma_I, \sigma_D) = \det(M(p, \sigma_I, \sigma_D)) - k \cdot \text{trace}^2(M(p, \sigma_I, \sigma_D)) \tag{7}$$

The value of k is set to be 0.04 [18], representing the slope of the border between the corner and edge. In Eq. (7), the interest points are extracted from the attention map generated in Eq. (4) based on the computation of the Harris corner measure at different scales. A robust interest point detector will detect the same point even in the noisy version of the image. The experiments evaluate the suitability of various transformations for interest point detection that generate a minimal residual variance.

The algorithm was implemented in MATLAB on a system powered with Intel i7 4.0 GHz processor and 8 GB RAM running on Windows 10 platform. The algorithm took 0.4321 s to compute interest points on a color image of size $300 \times 400 \times 3$. The computational complexity of top-down interest point detector is of square order $O(n^2)$.

4 Experimental results

In order to demonstrate the robustness and limitations of the proposed detector, experiments were performed on MSRA dataset (Fig. 2). A qualitative comparison between the proposed method and two state-of-art color interest point detectors and quantitative results for the proposed color attention maps to extract color feature points with six feature detectors is presented. The experiments are used to check the robustness of the interest points to different image variations. However, some of the feature

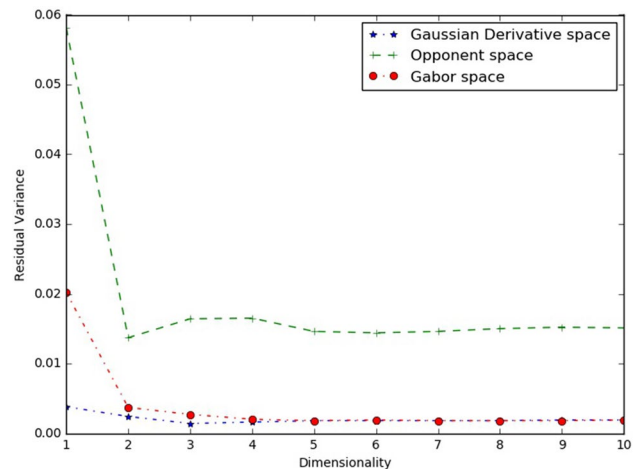


Fig. 3 Residual variance data points after Gaussian derivative, opponent and Gabor transformation

detectors are improvised to compare them using color images.

The training data consist of RGB values taken from the Skin Segmentation dataset [48]. A total of 50859-pixel samples are used from the dataset for learning a transformation model, which is used to calculate a pseudo map. This dataset has skin pixel values from face images of diverse age (young, middle and old), gender and race (white, black and Asian) from the Face Recognition Technology (FERET) database and Park Aging Mind Laboratory (PAL) database. The training data consists of different skin tones from various age groups and illumination, making it generic to the color type skin. The training data is

reshaped in the form of a rectangular image having all the pixels put together.

The partial derivatives for all three channels are computed for the training data. For each channel, the derivative is calculated along the horizontal and vertical direction to extract the frequency filtered transformation of the training data. Other transformations available are opponent space transformation [19, 20] and Gabor space transformation [49, 50] using different Gabor wavelets. These wavelets are computed on four different scales at obtained by a factor of 2 and six orientations (0°, 45°, 90°, and 135°) [50].

The isomap residual variances of these transformations are computed (Fig. 3) to find the dimension for which minimum variance is obtained [51]. The relationship between $G_{\{r\}\lambda}$, $G_{\{g\}\lambda}$, and $G_{\{b\}\lambda}$ vectors and their interdependence for a known chroma help in nonlinear feature integration. For varying isomap dimensionality, it is found that

projecting the color data points in Gaussian derivative space has minimum residual variance in comparison to Gabor and opponent space transformations.

The residual variance in all the three cases decays approximately linearly at the dimensionality of 2. At this point, it forms a knee corresponding to the correct dimension to find interdependencies for learning an affinity model. Beyond this knee point, the variances for all transformations remain almost steady. For this reason, a partial correlation for the training data gives an approximate solution to learn a model.

The Gaussian derivative transformation has a minimal residual variance of 0.0024 for a second-order dimensionality as compared to 0.0037 for Gabor transformation and 0.0137 for opponent space transformation. It lays the foundation of the proposed technique using Gaussian derivative transformation instead of using Gabor and opponent space transformations.

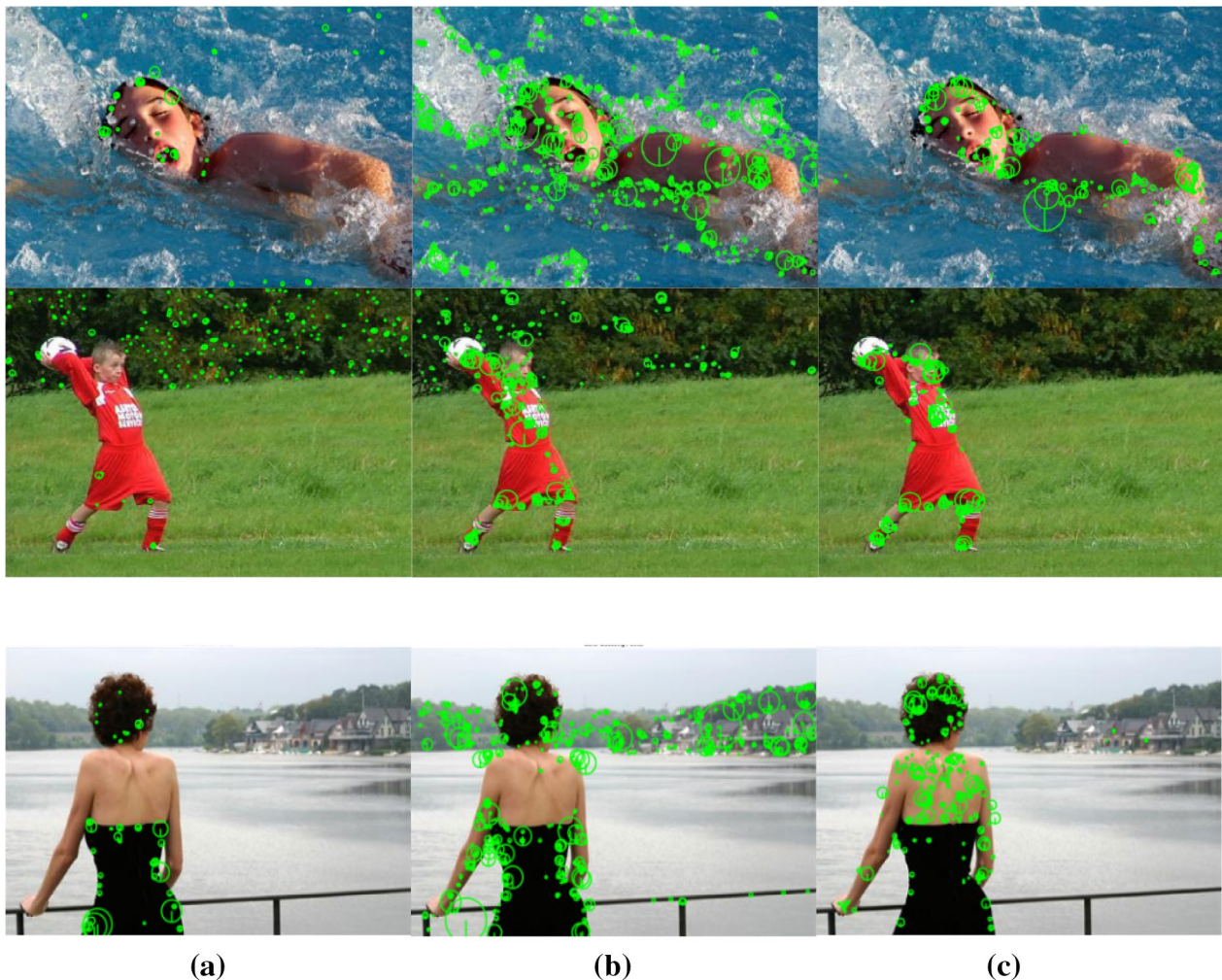


Fig. 4 Interest points using **a** Laplacian of Gaussian, **b** color boosting, **c** proposed technique plotted on images from MSRA salient object database [41]

The proposed method of detecting color interest points is compared to the state-of-art methods for locating color-based interest points. The approach is tested on the MSRA dataset [41] used for salient object detection. A subset of 1088 randomly selected images consisting of humans is derived from the MSRA dataset to evaluate the proposed approach. These images have a variety of variations in terms of human postures, age, skin tones, illumination effects, etc. making it challenging and a suitable choice for the experiment. Figure 2 shows a few images from the derived dataset.

The proposed approach is compared to state-of-art feature point detectors based on chromatic characteristics. The RGB Color Harris points are computed by calculating the corneriness measure in the image computed using equation [7]. The second corner detector is based on the color statistics and local differential structure in images commonly known as color boosted image features. Color boosted features are detected using a saliency boosting function estimated over every location in the image. Color boosting extracts color based interest points by finding rare colors in the image, subsequently detecting Harris points [19, 20]. The third interest point detector for comparison is the Laplacian of Gaussian [23] detector. It is an extension of the luminance-based Laplacian of Gaussian over multiple color channels combining scale-normalized individual channels. These methods were selected for comparison as they rely on color properties of the image to locate interest points and are best suited in this context. In this article, these detectors are termed as Top-Down Color Attention (Proposed), Laplacian Of Gaussian [23], and Color Boosting [19, 20], SIFT [22], SURF [24], MSER [37] and SUSAN [30].

The article illustrates the performance of different color based detectors in Fig. 4. It shows interest points based on RGB gradient-based color Harris detector, color boosted Harris points and top-down skin color attention. The locations of these interest points and their scales are represented by circles indicated in green. The RGB gradient-based color Harris detector generated a random and lesser number of color interest points in the images, Fig. 4a. It detects background features due to shadow and shading. Whereas significant improvement can be seen in case of color boosted Harris points proposed in [19]. The point detector focuses on the color information content and generates distinctive regions Fig. 4b. This method locates interest points based on the selective color boosting of the image. The substantial gain in information content helps in outperforming the RGB gradient-based detector. It is observed that these points are detected throughout the image and not localized to the object of interest. The proposed approach using top-down attention cues Fig. 4c shows a considerable improvement in the detection

of color interest points. As the detector has an affinity towards skin pixels from the learned model, the majority of the interest points are localized near the skin region in the image. This helps in making the detector more robust and informative.

The Harris threshold is set to $10e-9$, and the Laplace threshold is set to 0.03. These values are used for all the detectors in the experiments to generate keypoints. The results depict that the proposed approach is most suitable to detect interest points in relevant regions in the image. The guided color attention in the model generates photometric robustness of the interest points. This helps in reducing clutter in the image and derives more accurate interest points in the image.

Further, the quantitative comparison and performance of the proposed interest point detector is presented. An evaluation framework [52] proposed by Mikolajczyk and Schmid is a good metric to evaluate keypoint detectors [39, 40]. This helps in successful classification under real-time scenarios. The robustness of detected points is calculated by the repeatability score between two different versions of the same image. This work considers blur, noise variations, and compression variations in the derived MSRA dataset. Each set contains a reference image and some corrupted version of the reference image. In this article, few algorithms were implemented using vFeat [53] repository to carry out comparison and evaluation of color based keypoints detection.

While comparing images for classification and matching tasks, it is observed that there are large differences in the training and test data. The interest points are evaluated for repeatability of detected points under different variations such as noise, blur and image compression. A robust interest point detector shows invariance to locations of identified points to these changes. The original image and the corrupted image are matched to find corresponding regions.

Repeatability score computed for a given set of images is the ratio of one to one match correspondences of the interest points to the minimum number of interest points in the reference image. For instance, any keypoint A from a source image is said to be repeated, if there exists a keypoint B in the test image within the vicinity of A having a limited overlap error. The image points that do not show correspondences corrupt the repeatability score. So, a larger number of these correspondences signifies a good repeatability score. The amount of overlap of features is used to compute the amount of correspondence percentage in the form of overlap error. The repeatability score is calculated over the overlap error of the detected feature point in both the images. For a high overlap error, a good repeatability score is obtained. As the error threshold is reduced, the repeatability score tends to decrease. The

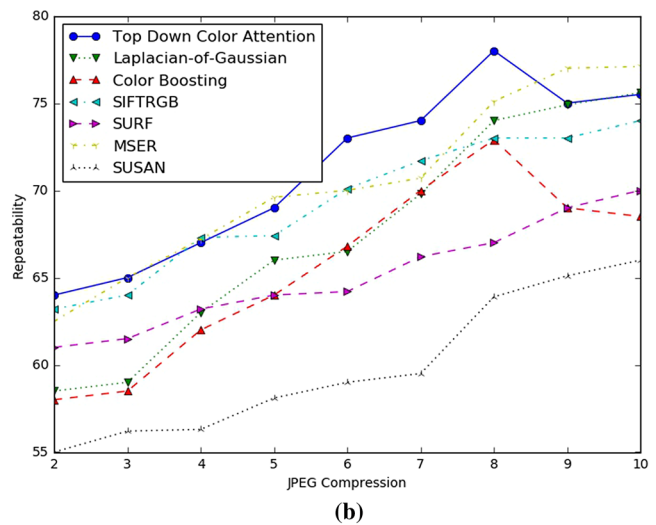
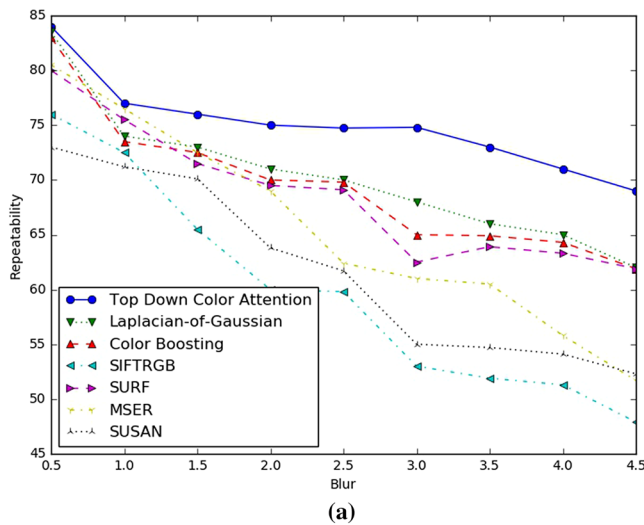


Fig. 5 Mean Repeatability Scores on MSRA dataset. **a** This figure shows the performance of different detectors with variations in image blur introduced by varying the standard deviation of Gauss-

ian low pass filter. **b** In this figure the X-axis denotes increasing image quality index introduced using the JPEG compression

experiment assumes, a feature to have correspondence if the overlap error is less than 40%. The overlap of corresponding regions is defined as

$$\epsilon = 1 - \frac{I_{B_i} \cap I_{T_i}}{I_{B_i} \cup I_{T_i}} \tag{8}$$

where ϵ is the overlap error for the corresponding region i in the base image I_B and I_T . To introduce blur variations in the dataset, the reference image transformed and passed through a Gaussian low pass filter of 5×5 with varying standard deviation of 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4 and 4.5. Similarly, some noise variations were added to evaluate the performance of different detectors. Gaussian noise is added to the reference images with a varying variance of 10^{-1} , 5×10^{-2} , 10^{-2} , 5×10^{-3} , 10^{-3} , 5×10^{-4} , 10^{-4} , 5×10^{-5} , 10^{-5} ; speckle and salt & pepper noise is added with densities ranging from 0.1 to 0.7 in the dataset images. Joint Photographic Experts Group (JPEG) compression is carried out by lowering the number of Discrete Cosine Transform (DCT) coefficients used for reconstructing the image. Furthermore, these compressions are scaled on a quality index of 2, 3, 4, 5, 6, 7, 8, 9, and 10 for understanding. More the value of quality index better is the quality of the image. The experiment was conducted on a derived MSRA dataset containing 1088 images with 500 features per image having the maximum Harris energy values. The comparison results show the proposed repeatability scores for different variations of input images for Top-Down Color Attention interest point detector, Laplacian of Gaussian

interest point detector [23], and Color boosted interest point detector [19, 20],.

The repeatability rates with corresponding parameters for different variations are presented: Image blur Fig. 5a, JPEG compression Fig. 5b, Gaussian Noise, speckle noise and salt and pepper noise in Fig. 6. These plots show the mean repeatability scores of all the images from the derived MSRA dataset. It can be seen that for blur variations, JPEG Compression the proposed Top-Down Color shows a higher repeatability percentage as compared to Laplacian of Gaussian and Color-Boosting, SIFT RGB, SURF, MSER and SUSAN detectors. At certain noise levels some detectors perform well in comparison to the proposed detector. This is due to the change in gradient around the edges and corners whose strength increases abruptly at points of occurrence of noise particles. Similarly, is the case for different noise models, the proposed approach has competitive repeatability score in comparison to other techniques except for salt & pepper noise where the proposed detector is affected by the high frequency noise in the image. In color images the salt & pepper noise generates random color gradients due to which the repeatability rates are lower. However, median filtering of such images will improve the detector performance. It is intuitive that as the quality of the image is increased, the percentage of points matching for a fixed overlap error increase. The overall results demonstrate that the proposed top-down skin distinctiveness approach performs effectively in comparison to other detectors.

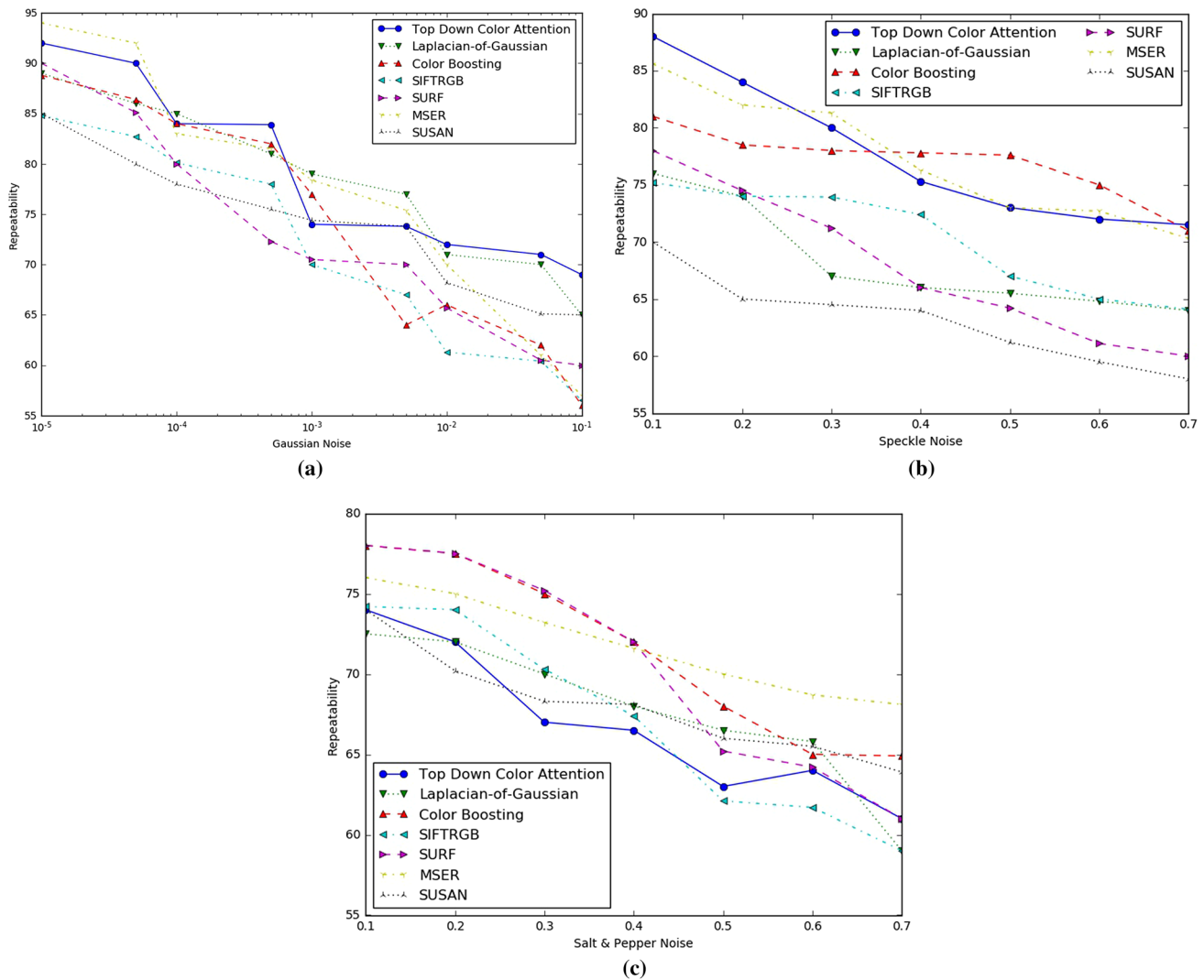


Fig. 6 **a** Mean Repeatability Scores for varying Gaussian Noise, **b** Mean Repeatability Scores for varying density of speckle Noise and **c** Mean Repeatability Scores for the varying density of salt and pepper Noise

Table 1 Performance evaluation using Mean Repeatability Scores at different varying overlap error threshold for different variations in image

Image variations	Overlap error					
	10%	20%	30%	40%	50%	60%
Blur	26.728	43.614	61.461	74.729	84.311	88.802
JPEG compression	8.4828	28.769	49.422	63.958	74.034	78.480
Gaussian noise	23.741	43.465	63.325	77.263	87.226	91.678
Speckle noise	24.231	38.254	71.572	80.005	85.018	89.108
Salt and pepper noise	15.254	31.290	67.270	72.258	78.251	80.212

The behavior of the proposed feature detector varies for different overlap error thresholds; To illustrate the potential of the proposed detector, Table 1 shows the variations in the repeatability rates for Blur, JPEG compression, Gaussian noise, speckle noise and salt & pepper noise for different error overlaps. From the experiment, it is observed that the

proposed detector is more robust to noise and Blur variations as compared to JPEG compression. As the error overlap threshold is increased, there is a considerable increase in repeatability scores. At 60% error overlap, a noisy image has a repeatability of 91.67% for Gaussian noise, 89.108% for speckle noise and 80.212% for salt and pepper noise. In

case of blur variations a repeatability of 88.80% is obtained, whereas the compressed image shows repeatability of 78.48% for detected color interest points. This paper compares the state of art detectors on single view images, for applications related to object detection and classification. The base algorithm can further be extended to tracking applications using point matching. The performance of the color features detector has scope of improvement in terms of noise handling capacity. As already stated 2-D filtering can enhance the performance of the system.

5 Conclusions

This article proposes the use of color and spatial information for detecting interest points in the image. It can be seen in Fig. 4 that color based interest point detectors yield better keypoints on the image. The proposed method generates attention maps based on color affinity. These vectors are generated by transforming testing data into the derivative space and finding vectors in the direction of maximum variation. The testing data is obtained by category-specific color information, skin color pixels in this case, followed by a transformation that generates vector weights for the color category. Further, these vectors are used to compute color saliency maps for the test images, which are then used to locate interest points.

Experiments are conducted on a derived MSRA dataset consisting of 1088 images. Evaluation for the robustness of interest points was carried out using the repeatability metric tested for a variety of variations, including three noise variations, blur, and JPEG compression. From the experimental results, it can be derived that the detector shows invariance to Gaussian noise, blur and image quality variations. However, the detector is confused by the high-frequency components introduced due to salt & pepper noise. The results further reveal that the top-down approach yields much informative and precise number of keypoints on the image. For applications, which include motion blur and transmission losses, the proposed color feature detector is recommended. The analysis leads to define a process to generate an affinity model towards a hue; skin color is taken up in this article. It was further used to locate interest points on the image to generate relevant information. Qualitative results show the effectiveness of using a learning-centric model for locating interest points. The technique was evaluated using different training and testing data to simulate real-time scenario where unknown images will be fed into the detector.

The evaluation show that the proposed technique shows a comparable invariance to noise variations in comparison to other state-of-the-art color-based interest point detectors. The technique is one of its kind using a

top down approach to train a model for locating interest points in the image. The article also presented the behavior of the proposed technique to a varied amount of overlap error for computing repeatability scores.

As a future research direction, the existing methodology can be extended to videos where temporal may be used to generate visual features for reducing clutter in challenging environments. That way, different objects present within different temporal frames can be matched for tracking purposes. It would also be interesting to integrate the system into a robot for visual odometry where different feature points can be tracked for their relative locations due to motion. The proposed feature detector can also be used for hyperspectral images where it is challenging to visualize different spectrums of known categories.

Acknowledgements This work was funded by Council of Scientific and Industrial Research (CSIR) (ESC0113 and PSC0202 Activity No. 2.3.1) and was carried out at CSIR- Central Scientific Instruments Organisation. The authors would like to acknowledge the Academy of Scientific and Innovative Research (AcSIR) for supporting this research work.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Pan J, Li X, Li X, Pang Y (2016) Incrementally detecting moving objects in video with sparsity and connectivity. *Cogn Comput* 8(3):420–428
- Tu Z, Abel A, Zhang L, Luo B, Hussain A (2016) A new spatio-temporal saliency-based video object segmentation. *Cogn Comput* 8(4):629–647
- Ehinger KA, Hidalgo-Sotelo B, Torralba A, Oliva A (2009) Modeling search for people in 900 scenes: a combined source model of eye guidance. *Vis Cogn* 17(6–7):945–978
- Gabriel P, Hayet J-B, Piater J, Verly J (2005) Object tracking using color interest points. In: IEEE conference on advanced video and signal based surveillance, 2005. IEEE, pp 159–164
- Da Rugna J, Konik H (2001) Color interest points detector for visual information retrieval. In: Electronic imaging 2002. International Society for Optics and Photonics, pp 139–146
- Khan N, Tappen MF (2013) Discriminative dictionary learning with spatial priors. In: 2013 IEEE international conference on image processing. IEEE, pp 166–170
- Yang J, Yang M-H (2012) Top-down visual saliency via joint CRF and dictionary learning. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 2296–2303
- Kocak A, Cizmeciler K, Erdem A, Erdem E (2014) Top down saliency estimation via superpixel-based discriminative dictionaries. In: Proceedings of the British machine vision conference. BMVA Press. <http://dx.doi.org/10.5244/C.28.73>
- Zhao J, Sun S, Liu X, Sun J, Yang A (2014) A novel biologically inspired visual saliency model. *Cogn Comput* 6(4):841–848
- Montesinos P, Gouet V, Deriche R (1998) Differential invariants for color images. In: Proceedings fourteenth international conference on pattern recognition, 1998. IEEE, pp 838–840

11. Khan FS, Van De Weijer J, Vanrell M (2009) Top-down color attention for object recognition. In: 2009 IEEE 12th international conference on computer vision. IEEE, pp 979–986
12. Zhu J, Yu J, Wang C, Li F-Z (2014) Colour combination attention for object recognition. *IET Image Proc* 8(9):539–547
13. Zhu J, Yu J, Wang C, Li F-Z (2015) Object recognition via contextual color attention. *J Vis Commun Image Represent* 27:44–56
14. Torralba A, Oliva A, Castelano MS, Henderson JM (2006) Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychol Rev* 113(4):766
15. Kanan C, Tong MH, Zhang L, Cottrell GW (2009) SUN: top-down saliency using natural statistics. *Vis Cogn* 17(6–7):979–1003
16. Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW (2008) SUN: a Bayesian framework for saliency using natural statistics. *J Vis* 8(7):32
17. Marchesotti L, Cifarelli C, Csurka G (2009) A framework for visual saliency detection with applications to image thumbnailing. In: 2009 IEEE 12th international conference on computer vision. IEEE, pp 2232–2239
18. Harris C, Stephens M (1998) A combined corner and edge detector. In: *Alvey vision conference, 1988*. Citeseer, p 50
19. Van de Weijer J, Gevers T, Bagdanov AD (2006) Boosting color saliency in image feature detection. *IEEE Trans Pattern Anal Mach Intell* 28(1):150–156
20. Stottinger J, Hanbury A, Sebe N, Gevers T (2012) Sparse color interest points for image retrieval and object categorization. *IEEE Trans Image Process* 21(5):2681–2692
21. Shi J, Tomasi C (1994) Good features to track. In: 1994 IEEE computer society conference on computer vision and pattern recognition, 1994. *Proceedings CVPR'94*. IEEE, pp 593–600
22. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vis* 60(2):91–110
23. Vigo DAR, Khan FS, Van De Weijer J, Gevers T (2010) The impact of color on bag-of-words based object recognition. In: 2010 20th international conference on pattern recognition (ICPR). IEEE, pp 1549–1553
24. Bay H, Ess A, Tuytelaars T, Van Gool L (2008) Speeded-up robust features (SURF). *Comput Vis Image Underst* 110(3):346–359
25. Calonder M, Lepetit V, Strecha C, Fua P (2010) Brief: binary robust independent elementary features. In: *European conference on computer vision, 2010*. Springer, pp 778–792
26. Oliva A, Torralba A, Castelano MS, Henderson JM (2003) Top-down control of visual attention in object detection. In: 2003 international conference on image processing, 2003. *ICIP 2003*. *Proceedings*, vol 251. IEEE, pp I-253–256
27. Tomasi C, Kanade T (1991) Detection and tracking of point features. *Carnegie Mellon University, Pittsburgh*
28. Mainali P, Yang Q, Lafruit G, Van Gool L et al (2011) Robust low complexity corner detector. *IEEE Trans Circuits Syst Video Technol* 21(4):435–445
29. Mainali P, Yang Q, Lafruit G, Lauwereins R, Van Gool L (2010) Lococo: low complexity corner detector. In: 2010 IEEE international conference on acoustics, speech and signal processing. IEEE, pp 810–813
30. Smith SM, Brady JM (1997) SUSAN—a new approach to low level image processing. *Int J Comput Vis* 23(1):45–78
31. Rosten E, Drummond T (2006) Machine learning for high-speed corner detection. In: *European conference on computer vision, 2006*. Springer, Berlin, pp 430–443
32. Martin DR, Fowlkes CC et al (2004) Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans Pattern Anal Mach Intell* 26(5):530–549
33. Agrawal M, Konolige K, Blas MR (2008) Censur: center surround extremas for realtime feature detection and matching. In: *European conference on computer vision, 2008*. Springer, pp 102–115
34. Marimon D, Bonnin A, Adamek T, Gimeno R (2010) DARTs: efficient scale-space extraction of DAISY keypoints. In: 2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, pp 2416–2423
35. Alcantarilla PF, Bartoli A, Davison AJ (2012) KAZE features. In: *European conference on computer vision, 2012*. Springer, pp 214–227
36. Salti S, Lanza A, Di Stefano L (2013) Keypoints from symmetries by wave propagation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition, 2013*, pp 2898–2905
37. Matas J, Chum O, Urban M, Pajdla TJI et al (2004) Robust wide-baseline stereo from maximally stable extremal regions. *Image Vis Comput* 22(10):761–767
38. Forssén P-E (2007) Maximally stable colour regions for recognition and matching. In: 2007 IEEE conference on computer vision and pattern recognition. IEEE, pp 1–8
39. Salahat E, Qasaimeh M (2017) Recent advances in features extraction and description algorithms: a comprehensive survey. In: 2017 IEEE international conference on industrial technology (ICIT). IEEE, pp 1059–1063
40. Awad AI, Hassaballah M (2016) Image feature detectors and descriptors. In: *Studies in computational intelligence*. Springer, Cham. <https://doi.org/10.1007/978-3-319-28854-3>
41. Liu T, Yuan Z, Sun J, Wang J, Zheng N, Tang X, Shum H-Y (2011) Learning to detect a salient object. *IEEE Trans Pattern Anal Mach Intell* 33(2):353–367
42. Zeng F, Ji Y et al (2017) Contextual bag-of-words for robust visual tracking. *IEEE Trans Image Process* 27(3):1433–1447
43. Weinland D, Özuysal M, Fua P (2010) Making action recognition robust to occlusions and viewpoint changes. In: *European conference on computer vision, 2010*. Springer, pp 635–648
44. Van De Sande K, Gevers T, Snoek C (2010) Evaluating color descriptors for object and scene recognition. *IEEE Trans Pattern Anal Mach Intell* 32(9):1582–1596
45. Faillie F (2005) Stable interest point detection under illumination changes using colour invariants. In: William Clocksin AF, Torr P (eds) *The British machine vision conference*. British Library Cataloging in Publication Data, Oxford
46. Di Zenzo S (1986) A note on the gradient of a multi-image. *Comput Vis Graph Image Process* 33(1):116–125
47. Van de Weijer J, Gevers T, Geusebroek J-M (2005) Edge and corner detection by photometric quasi-invariants. *IEEE Trans Pattern Anal Mach Intell* 27(4):625–630
48. Bhatt R, Dhall AJUMLR (2010) Skin segmentation dataset. UCI machine learning repository. IIT Delhi, Delhi
49. Dadwhal YS, Kumar S, Sardana H (2020) Data-driven skin detection in cluttered search & rescue environments. *IEEE Sens J* 20(7):3697–3708. <https://doi.org/10.1109/JSEN.2019.2959787>
50. Lee TS (1996) Image representation using 2D Gabor wavelets. *IEEE Trans Pattern Anal Mach Intell* 18(10):959–971
51. Tenenbaum JB, De Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290(5500):2319–2323
52. Mikolajczyk K, Tuytelaars T, Schmid C, Zisserman A, Matas J, Schaffalitzky F, Kadir T, Van Gool L (2005) A comparison of affine region detectors. *Int J Comput Vis* 65(1–2):43–72
53. Vedaldi A, Fulkerson B (2010) VLFeat: an open and portable library of computer vision algorithms. In: *Proceedings of the 18th ACM international conference on multimedia, 2010*. ACM, pp 1469–1472

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.