Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

SPARSE COLOR INTEREST POINTS FOR IMAGE RETRIEVAL AND OBJECT CATEGORIZATION

Julian Stöttinger, Allan Hanbury, Nicu Sebe and Theo Gevers

January 2012

Technical Report # DISI-09-011

Published on IEEE Transactions on Image Processing, Vol. 21, Issue 5, May 2012.

Sparse Color Interest Points for Image Retrieval and Object Categorization

Julian Stöttinger*, Allan Hanbury, Nicu Sebe *Senior Member, IEEE*,
and Theo Gevers *Member, IEEE*

Abstract

Interest point detection is an important research area in the field of image processing and computer vision. In particular, image retrieval and object categorization heavily rely on interest point detection from which local image descriptors are computed for image matching. In general, interest points are based on luminance, and color has been largely ignored. However, the use of color increases the distinctiveness of interest points. The use of color may therefore provide selective search reducing the total number of interest points used for image matching.

This paper proposes color interest points for sparse image representation. To reduce the sensitivity to varying imaging conditions, light invariant interest points are introduced. Color statistics based on occurrence probability lead to color boosted points which are obtained through a saliency-based feature selection. Further, a PCA-based scale selection method is proposed which gives a robust scale estimation per interest point.

From large scale experiments, it is shown that the proposed color interest point detector has a higher repeatability than a luminance-based one. Further, in the context of image retrieval, a reduced and predictable number of color features shows an increase in performance compared to state-of-the-art interest points. Finally, in the context of object recognition, for the Pascal VOC 2007 challenge, our

J. Stöttinger and N. Sebe are with the Department of Information Engineering and Computer Science, University of Trento, via Sommarive 14, 38100 Povo-Trento, Italy. email: julian@disi.unitn.it

A. Hanbury is with the Institute of Software Technology and Interactive Systems, Vienna University of Technology, Favoritenstraße 9-11/188, A-1040 Vienna, Austria.

T. Gevers is with the Informatics Institute, University of Amsterdam, The Netherlands and with the Computer Vision Center (CVC), Universitat Autonoma de Barcelona, Spain.

method gives comparable performance to state-of-the-art methods using only a small fraction of the features, reducing the computing time considerably.

Index Terms

ELI-COL, ARS-IIU, SMR-REP, color invariance, local features, object categorization, image retrieval.

I. INTRODUCTION

Interest point detection is an important research area in the field of image processing and computer vision. In particular, image retrieval and object categorization rely heavily on interest point detection from which local image descriptors are computed for image and object matching [1]. The majority of interest point extraction algorithms are purely intensity based [2], [3], [4]. These methods ignore saliency information contained in the color channels. However, it was shown that the distinctiveness of color-based interest points is larger, and therefore color is important when matching images [5]. Furthermore, color plays an important role in the pre-attentive stage in which features are detected [6], [7] as it is one of the elementary stimulus features [8].

In general, the current trend in object recognition is toward increasing the number of points [9], applying several detectors or combining them [10], [11], or making the interest point distribution as dense as possible [12]. While such a dense sampling approach provides accurate object recognition, they basically shift the task of discarding the non-discriminative points to the classifier. With the explosive growth of image and video datasets, clustering and offline training of features becomes less feasible [13]. By reducing the number of features and working with a predictable number of sparse features, larger image datasets can be processed in less time. Additionally, a stable number of features leads to a more predictable workload for such tasks.

Our aim is to exploit state-of-the-art object classification and to focus on the extraction of distinctive and robust interest points. In fact, the goal is to reduce the number of interest points extracted while still obtaining state-of-the-art image retrieval or object recognition results. Recent work aims to find distinctive features e.g. by performing an evaluation of all features within the dataset or per image class and choosing the most frequent ones [14]. This approach requires an additional calculation step with an inherent demand on memory and processing time dependent on the number of features. Another option is to use color to increase the distinctiveness of interest points [15], [16]. This alternative may therefore

provide selective search for robust features reducing the total number of interest points used for image retrieval.

Therefore, in this paper, we propose color interest points to obtain a sparse image representation. To reduce the sensitivity to imaging conditions, light invariant interest points are proposed. To obtain light invariant points, the quasi-invariant derivatives of the HSI color space are used. For color boosted points, the aim is to exploit color statistics derived from the occurrence probability of colors. In this way, color boosted points are obtained through saliency-based feature selection. Further, a PCA-based scale selection method is proposed which gives robust scale estimation per interest point. The use of color information allows to extract repeatable and scale-invariant interest points. Feature selection takes place at the very first step of the feature extraction and is carried out independently per feature.

Van de Weijer et al. [15], [16] did preliminary work on incorporating color distinctiveness into the design of interest point detectors. Color derivatives were taken to form the basis of a color saliency boosting function to equal the information content and saliency of a given color occurrence. However, our aim is to select interest points based on color discriminative and invariant properties derived from local neighborhoods. Therefore, our focus is on color models that have useful perceptual, saliency and invariant properties to achieve a reduction in the number of interest points. We propose a method of selecting a scale associated with the computed interest points, while maintaining the properties of the color space used, and to steer the characteristic scale by the saliency of the surrounded structure. Opposed to other color interest points used so far [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], the goal here is to enhance an adapted multi-dimensional color Harris corner detector in conjunction with an independent scale selection maintaining the main properties of the chosen color space.

The proposed method includes the following contributions to image retrieval and object categorization:
1. A scale decision strategy is proposed for the multi-channel Harris detector to provide a scale invariant, color interest point detector. So far, the most stable corner detectors are either luminance only or are not scale invariant.
2. The incorporation of perceptual color spaces in local, scale invariant features. The advantages of these color spaces are directly passed on to the representation of the features. Therefore, the instability of luminance based local features due to changing shadowing, reflections, lighting effects and color temperature are implicitly addressed. Invariance to lighting changes and the incorporation of a visual saliency function is achieved by a color transformation and can be passed directly to an image retrieval and object categorization framework.
3. Selection of distinctive features is typically done in the matching stage when the classification system

builds its model [27]. With the proposed method it is possible to perform this choice in the first step of the image matching pipeline making all subsequent operations faster. Moreover, this step is conducted independently for every feature and image (e.g. without considering the global feature space) and is based on the local visual input only (e.g. no spatial inter-relation, ground truth or occurrence frequency of features is used).

4. Runtime of every step of the image matching process decreases with a more sparse representation of local features. Offline procedures like constructing global dictionaries are practically infeasible when the number of features in the training data is extensive [13]. The runtime of online procedures like the quantization of features also depends on the number of features. With the proposed method, the amount of data to be processed is reduced significantly.

5. Higher dimensional data can be processed. The proposed representation of multi-channel information is not limited to a single color space.

The paper is structured as follows. In the next section, related work is discussed. Section III presents the framework for color interest point and scale detection. Experimental results are presented in Section IV.

II. Related Work

In Section II-A, the main steps of image retrieval and object categorization are outlined, see Figure 1. A detailed comparison of interest points is presented in Section II-B.

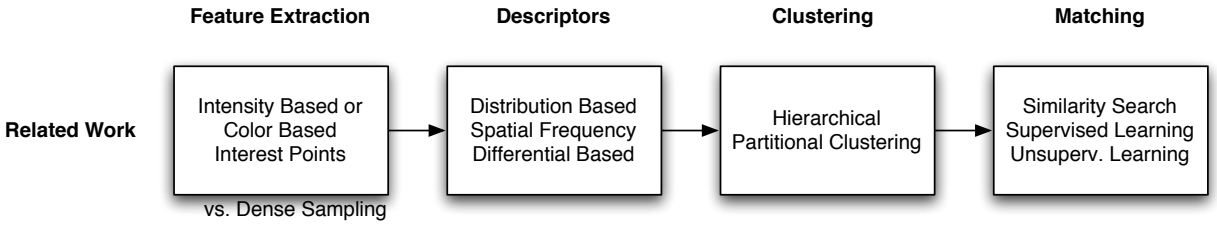


Fig. 1. The main steps of image retrieval and object categorization. (1) Feature extraction is carried out with either global or local features. (2) Descriptors characterize the image information steered by the feature extraction. (3) Clustering is used for signature generation. (4) Matching summarizes the classification of images.

A. Common Pipeline for Image Retrieval and Object Categorization

Feature extraction is carried out with either global or local features. In general, global features lack robustness against occlusions and cluttering (e.g. [28], [29]) and provide a fast and efficient way of image representation. Local features are either intensity-based or color-based interest points. Recently, dense

sampling of local features is used as it provides good performance especially for the bags of words approach and robust learning systems [10], [12].

Descriptors represent the local image information around the interest points. They can be categorized into three classes: They describe the distribution of certain local properties of the image (e.g. SIFT), spatial frequency (e.g. wavelets) or other differentials (e.g. local jets) [30]. For every feature extracted, a local descriptor is computed. A disadvantage is that the runtime increases with their number. Efficient ways to calculate these descriptors exist, e.g. for features with overlapping areas of support, previously calculated results can be used.

Clustering for signature generation, feature generalization or vocabulary estimation assigns the descriptiors into a subset of categories. There are hierarchical and partitional approaches to clustering. Due to the excessive memory and runtime requirements of hierarchical clustering [31], partitional clustering, such as the k-means, is the method of choice in creating feature signatures.

Matching summarizes the classification of images. Image descriptors are compared with previously learnt and stored models. This is computed by a similarity search or by building a model based on supervised or unsupervised learning techniques. Classification approaches need feature selection to discard irrelevant and redundant information [32], [33], [34]. It is shown that a powerful matching step can successfully discard irrelevant information and better performance is gained [12]. Training and clustering are the most time consuming steps in state-of-the-art recognition frameworks. Clustering of a global dictionary takes several days for current benchmark image databases, becoming less feasible for online databases resulting in several billion of features [13]. Therefore, one of our goals is a feature selection using color saliency within the first stage of this scheme. The use of color provides selective search reducing the total number of interest points used for image retrieval. The aim is to use color interest points to obtain sparse image representations. In the next section, an overview is given of the successful approaches to detect interest points.

B. Interest Points

The Harris corner detector [2], which is based on the Moravec corner detector [35], is the first corner detector providing a rotation invariant and isotropic corner measure that is robust to noise and scale changes up to a factor of $\sqrt{2}$.

An extension of the Harris corner detector, the scale invariant Harris Laplacian (also referred to as Harris Laplace, shape-adapted Harris, or multi-scale Harris), is proposed by Mikolajczyk and Schmid [36]. The main idea is to carry out corner and blob detection on different scales. Wherever there is a stable

corner and a stable blob at the same scale, a *characteristic scale* is found. The approach is extended to the Hessian Laplacian detector which computes points that maximize the determinant of the Hessian matrix and uses a similar scale selection.

Scale invariant blob detectors are based on the scale-space theory introduced by Witkin [37] and extended by Lindeberg [38]. The precision of the scale estimation using either Laplacian of Gaussian (LoG), Difference of Gaussian (DoG) or Determinant of Hessian (DoH) [39] methods depends on the choice of the scale sampling rate [40]. Maximally stable extremum regions (MSER) [41] are obtained by a watershed-like algorithm. Connected regions at a certain thresholded range are selected if they remain stable over a set of thresholds. The algorithm is very efficient both in runtime, performance and detection rate and is extended to color in [24]. Following [42], MSER is a robust detector to geometric transformations providing only a few but large blobs. Contrary, for blur and lighting effects it performs significantly worse [30]. Further, it is highly dependent on the contrast of the input image. Unnikrishnan et al. [23] extract scale and illumination invariant blobs through color by an adapted illumination model and a modification of the LoG. This is efficiently approximated by multiplying the LoG functions' output per channel but is of limited robustness. Additionally, it follows the original SIFT's key-points: adding robustness to light temperature change, but maintaining similar performance for geometrical transformations.

The most successful color features are based on the color Harris detector introduced by Montesinos et al. [21] and successfully used for example in [19]. In image retrieval scenarios, they apply the fixed scale detector on gradually downsized images and use all the detections extracted. This leads to multiple ambiguous features, and the inability to match images at different scales. Rugna et al. [22] suggest a method to extract scale-invariant interest points based on color information for texture classification. A color Gaussian pyramid is used [43]. Then, for every pyramid level and color channel, the original Harris energy is calculated. Features are selected based on their persistence through the pyramid. However, a scale selection based on the local structure is not obtained by this method. The method is independent of the color space used. Faille [18] proposes a shadow, shading, illumination color and specularities invariant interest point localization which models the color information as Lambertian and specular reflection. Derivatives of the invariants are incorporated in the Harris second Moment matrix. It uses fixed scales for matching of images under varying lighting. Weijer et al. [15] extend the color Harris approach to arbitrary color spaces and suggest two approaches: A photometric quasi-invariant HSI color space providing a corner detector with better noise stability characteristics compared to existing photometric invariants and a color boosting hypothesis for defining salient colors. These two approaches

provide a robust corner estimation under varying lighting and shadowing effects for the quasi-invariant color space and a saliency measure [25]. Our contribution is to extend this approach by incorporating a scale selection strategy to detect color interest points.

In conclusion, the Harris-Laplacian is taken as the basis of our color interest point as the Hessian-Laplacian gives similar but additional locations resulting in better results due to the number (better probability of matching) and the quality of locations (better distinctiveness). MSER is less stable under varying lighting conditions and contrast, and is sensitive to parameter settings. Blob detectors depend on geometric transformations [30].

III. Color Interest Point Detection

In this section, the multi-channel Harris corner detector and a scale selection method are presented. These may be used with any color space. First, different color spaces are discussed. Then, color interest point detectors are presented and a scale selection method is proposed.

A. Color Spaces

Photometric effects such as shadowing, shading or specular effects can be modeled using an appropriate reflection model [18], [44], [45]. The quasi invariant color space proposed in [15] is derived from an orthonormal transformation from RGB. Compared to other Opponent Color Space (OCS) definitions (e.g. [46], [47]), this transformation uses a rotated chromaticity axis and a different normalization [48]. It provides specular variance and is defined by

$$OCS = \begin{pmatrix} o_1 \\ o_2 \\ o_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix}. \tag{1}$$

As this color space is often motivated by early visual processing in primates, the opponent colors blue/yellow and red/green are the end points of the o_1 and o_2 axis of the color space. As primates do not see combinations of these colors (e.g. a “blueish yellow” or a “greenish red”) it is argued that the co-occurrence of these opponent colors attracts the most attention. A polar transformation on o_1 and o_2 of the OCS leads to the HSI color space

$$HSI = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix} = \begin{pmatrix} \tan^{-1}\left(\frac{o_1}{o_2}\right) \\ \sqrt{o_1^2 + o_2^2} \\ o_3 \end{pmatrix}. \tag{2}$$

The gradient magnitude of the hue component ϕ_1 is invariant both to shading and to specular effects, as it is perpendicular both to the shadow-shading direction and to the specular direction. The main drawback is that it is unstable for large or small ϕ_2, i.e. saturation (grey-axis). To address this shortcoming, a stable photometric invariant h is obtained by the derivation in the ϕ_1 direction while scaling it by the saturation component ϕ_2 [15],

$$|h'| = \phi_2 \phi_1'. \tag{3}$$

Contrary to this approach, [16] proposed to examine the saliency of color information and its gradient magnitudes. Colors that have different occurrence probabilities $p(v)$ will also have different information content $\iota(v)$ of their descriptor $\iota(v) = -log(p(v))$. The idea behind *color boosting* is to boost rare colors for having a higher saliency. Traditionally, the gradients of color vectors with equal vector norms have equal impact on the saliency function. The goal is to find a color boosting function g so that color vectors having equal information content have equal impact on the saliency function. The saliency s at a position $\mathbf{x}$ is then given by

$$s_{\mathbf{x}} = H_\sigma(g(\mathbf{f}_x), g(\mathbf{f}_y)), \tag{4}$$

where H_σ is any saliency function at scale σ and $\mathbf{f}_x$ and $\mathbf{f}_y$ are the gradient magnitudes in the x and y direction of a color vector at location $\mathbf{x}$. In the following, the estimation of $g(\mathbf{f}_x)$ is described in detail. $g(\mathbf{f}_y)$ is obtained in a similar way.

The saliency boosting function $g : \mathbb{R}^3 \to \mathbb{R}^3$ is a transformation such that

$$p(\mathbf{f}_{1,x}) = p(\mathbf{f}_{2,x}) \leftrightarrow \|g(\mathbf{f}_{1,x})\| = \|g(\mathbf{f}_{2,x})\|, \tag{5}$$

where $\mathbf{f}_{1,x}$ and $\mathbf{f}_{2,x}$ denote gradient magnitudes in the x direction of two arbitrary color vectors. The transformation is obtained by deriving a function describing the surface of the 3 dimensional color distribution which can be approximated by an ellipsoid. The third coordinate of the color space is already aligned with the luminance which forms the longest axis of the ellipsoid. The other two axes are rotated so that they are aligned with the other two axes of the ellipsoid [15].

This can then be approximated by ellipsoids satisfying the following:

$$(\alpha k_x^1)^2 + (\beta k_x^2)^2 + (\gamma k_x^3)^2 = R^2, \tag{6}$$

where vector $\mathbf{k}$ and its elements $k_x^{[1..3]}$ is the transformation of the color derivative followed by the rotation to align the axes with those of the ellipsoid in the corresponding color space. To find the transformation

in Eq. (5), the approximated ellipsoid is scaled to a sphere so that vectors of equal saliency lead to vectors of equal length. The function g is therefore defined by

$$g(\mathbf{f}_x) = \mathfrak{M}\mathbf{k}(\mathbf{f}_x), \tag{7}$$

where the matrix $\mathfrak{M}$ consists of the elements $\mathfrak{M}_{11} = \alpha$, $\mathfrak{M}_{22} = \beta$ and $\mathfrak{M}_{33} = \gamma$, assuming $\alpha^2+\beta^2+\gamma^2 = 1$.

B. Color-based Interest Points

The second moment matrix M is a structure tensor describing the gradient distribution of a single channel image I of a local neighborhood at position $\mathbf{x}$:

$$M(\mathbf{x}, \sigma_I, \sigma_D) = \left\{ \sigma_D^2 G(\sigma_I) \otimes \begin{bmatrix} L_x^2(\sigma_D) & L_x L_y(\sigma_D) \\ L_x L_y(\sigma_D) & L_y^2(\sigma_D) \end{bmatrix} \right\} (\mathbf{x}), \tag{8}$$

where $\otimes$ denotes the convolution and $G(\sigma_I)$ the Gaussian kernel of size σ_I, $L_x(\sigma_D) = I \otimes G_x(\sigma_D)$ the convolution with the first derivative of the Gaussian kernel of scale σ_D in the x direction, and $L_y(\sigma_D) = I \otimes G_y(\sigma_D)$ in the y direction. L_x^2, L_y^2 and $L_x L_y$ are then found by multiplying these elements [16]. More generally, the second moment matrix can be computed by a transformation in the RGB space [15]. The first step is to determine the gradients of each component of the RGB color system. The gradients are then transformed into the desired color system. By multiplication and summation of the transformed gradients, all components of the second moment matrix are computed. In symbolic form, an arbitrary color space C is used with its n components $[c_1, \ldots, c_n]^T$. The elements for M are then calculated more generally as follows

$$\begin{aligned} L_x^2(\sigma_D) &= \sum_{i=1}^{n} c_{i,x}^2(\sigma_D), \\ L_x L_y(\sigma_D) &= \sum_{i=1}^{n} c_{i,x}(\sigma_D) c_{i,y}(\sigma_D), \\ L_y^2(\sigma_D) &= \sum_{i=1}^{n} c_{i,y}^2(\sigma_D), \end{aligned} \tag{9}$$

where $c_{i,x}$ and $c_{i,y}$ denote the components of the transformed color channel gradients at scale σ_D, and where the subscript x and y indicates the direction of the gradient. As shown in several experiments [36], [4], the relation $3\sigma_D = \sigma_I$ performs best. Based on the eigenvalues of M, the Harris energy $\mathfrak{C}$ is found as a corner measurement by

$$\mathfrak{C}(\mathbf{x}, \sigma_I, \sigma_D) = \det(M(\mathbf{x}, \sigma_I, \sigma_D)) - \kappa \cdot \text{trace}^2(M(\mathbf{x}, \sigma_I, \sigma_D)). \quad (10)$$

The constant κ indicates the slope of the *zero line*, i.e. the border between corner and edge.

This leads to stable locations that are robust to noise, scale changes up to $\sqrt{2}$, translation and rotation under arbitrary color spaces. For many computer vision tasks, it is crucial to provide scale invariant features. Therefore, in the next section, a principled approach for saliency based estimation of the characteristic scale in arbitrary color spaces for local features is proposed.

C. Color-based Scale Detection

Using the elements of the structure tensor for higher dimensional data in Eq. 8, it is straightforward to extend the Color Harris to the Harris-Laplacian [25] by applying Harris Corners and the Laplacian of Gaussian (LoG) on different scales. The characteristic scales are then found when both functions reach their maximum. This method is referred to as *RGB Color Harris*. Contrary to this extension, we propose a way to incorporate a global saliency measure in the process of scale selection.

The proposed scale selection is carried out on a single-channel saliency image $\widehat{I}$. It is related to the concept of Eigenimages [49] with the main difference that there is no training set of *image templates*, but the color channel information itself serves as the template for the color distribution. Let an input image I_C in a color space C consist of $I_C = \{\mathbf{f}_1, .., \mathbf{f}_m\}$ color vectors where every $\mathbf{f}_i$ consists of n color components $\mathbf{f}_i = [c_1, .., c_n]$ which are normalized and have a mean of zero. The eigenvectors of the covariance matrix $\mathbf{Q}$ of I_C are $\mathbf{e}_i$ and the corresponding eigenvalues λ_i. Since $n \ll m$, n eigenvectors can be efficiently estimated using single value decomposition [50]. The eigenvectors are in descending order with respect to their corresponding eigenvalues λ_i. $\widehat{I} = [\widehat{f}_1, .., \widehat{f}_m]$ is represented by the scalar product of the eigenvector with the highest eigenvalue $\mathbf{e}_1$ and the color vectors $\mathbf{f}_i$,

$$\widehat{f}_i = \mathbf{f}_i \mathbf{e}_1. \quad (11)$$

The LoG gives a maximum in the center of a Gaussian blob. From the LoG's scale, the size of the blob can be estimated. Therefore it is used to find the *characteristic scale* of a local structure [51]. To be more robust to noise, we take the following approach. At position $\mathbf{x}$ of scale σ, $\Lambda_{\mathbf{x},\sigma}$ is defined by

$$\Lambda_{\mathbf{x},\sigma} = \left[\left(\frac{\partial^2 \widehat{I}}{\partial x^2} + \frac{\partial^2 \widehat{I}}{\partial y^2} \right) \otimes G_{\sigma_D} \otimes \Gamma_{\sigma_D} \right] (\mathbf{x}) \quad (12)$$

where Γ_{σ_D} is the circularly symmetric raised cosine kernel, which is defined for each location (x_e, y_e) as

$$\Gamma_{\sigma_D} = \frac{1 + (cos(\frac{\pi}{\sigma_D}\sqrt{x_e^2 + y_e^2})}{3}. \tag{13}$$

A convolution with this kernel gives smoother borders than the Gaussian kernel G for scale decision and leads to detection of larger structure [52]. For computational efficiency, Λ can be approximated by the sum of the independently computed values L_x^2 and L_y^2 of $\widehat{I}$:

$$\Lambda_{\mathbf{x},\sigma_D} = \left\{ \left[\sigma_D^2 | L_x^2(\mathbf{x}, \sigma_D) + L_y^2(\mathbf{x}, \sigma_D) | \right] \otimes \Gamma_{\sigma_D} \right\} (\mathbf{x}). \tag{14}$$

It is known that corner locations shift when the scale changes – the smaller the scale change from one iteration to the next, the more precise the location is estimated. As the Harris detector is robust to scale changes up to a factor $t = \sqrt{2}$, this factor is the standard used in various applications [39], [53]. The scale space of the Harris function is obtained by calculating the Harris energy under varying σ. The number of different scales examined is of importance for the processing time. Each step must be calculated on its own (but independently and therefore possibly in parallel) and the processing time increases with the size of the kernels.

Using scale levels $l_S = 1, 2, \ldots$ with a factor t from 1.2 to $\sqrt{2}$, the Harris energy is calculated at scales $t^s\sigma$. A *potential* characteristic scale of a possible region is obtained when both the Harris Energy and the Laplacian of Gaussian are at their extremes:

$$\nabla \Lambda_{\mathbf{x},\sigma_D} = \nabla \mathfrak{C}_{H,\mathbf{x},\sigma_I,\sigma_D} = 0. \tag{15}$$

With this non-maxima suppression, the locations with their corresponding scales are found. However, there may be multiple candidates of scale sizes per location. All the candidate scales are taken into account. The size of the largest structure detected by $\Lambda_{\mathbf{x},\sigma}$ is then approximated. Having the chosen constants σ and t, the functions $\widehat{E}_{\mathbf{x}}$ give the location of highest local maximum of $\mathfrak{C}_{H,\mathbf{x},\sigma_I,\sigma_D}$ and $\widehat{\Lambda}_{\mathbf{x}}$ the largest scale of local maxima of $\Lambda_{\mathbf{x},\sigma_D}$. Therefore, $\widehat{E}_{\mathbf{x}}$ and $3t^{\arg\max(\widehat{\Lambda}_{\mathbf{x}})}\sigma_D$ define all candidates for interest points and the corresponding region size. The multi-channel Harris energy $\mathfrak{C}_H$ is the saliency property which can be used to select interest points.

D. Light Invariant Points

To extract invariant points from an arbitrary color image, the input image is transformed to the illumination invariant image $I_C = \{\mathbf{f}_1, .., \mathbf{f}_m\}$ with $\mathbf{f}_i = [\phi_1, \phi_2]$. The fully illumination variant part

of the image in HSI, ϕ_3, is discarded. For $c_{1,[x|y]}$, the stable photometric invariants $h_{[x,y]}$ derived in Eq. 3 are estimated. $c_{2,[x|y]}$ are the gradient magnitudes of ϕ_2 in the spatial direction of $[x|y]$. The structure tensor $\mathfrak{C}(\mathbf{x}, \sigma_I, \sigma_D)$ is built under increasing scales σ_D and σ_i with a constant factor $t = [1.2, \sqrt{2}]$. Scale selection is carried out on all three color components $\mathbf{x}_i = [\phi_1, \phi_2, \phi_3]$, from which the saliency image $\widehat{I}$ is built.

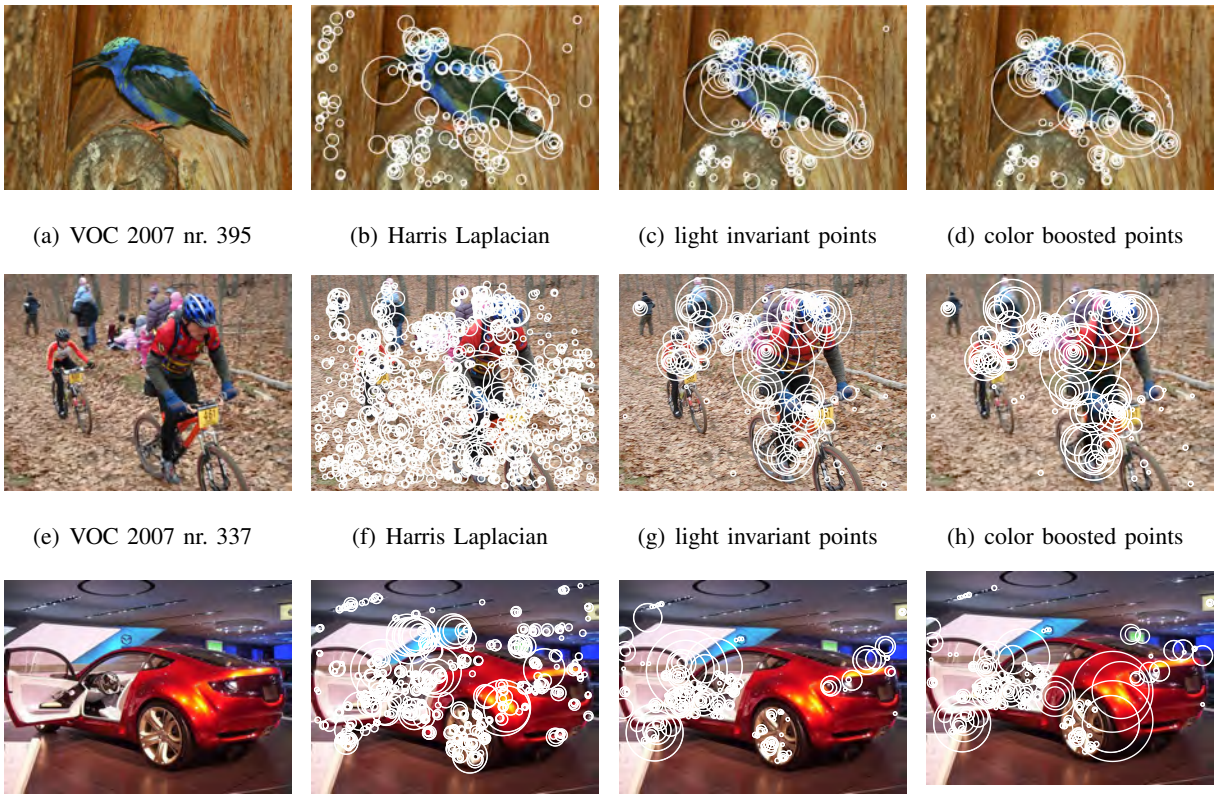
E. Color Boosted Points

Color boosted points are extracted in the OCS color space $I_C = \{\mathbf{f}_1, .., \mathbf{f}_m\}$ with $\mathbf{f}_i = [o_1, o_2, o_3]$. The saliency boosting function is estimated based on the whole set of training images. For experiments without training images (e.g. the repeatability experiments in Section IV-A), the results on the Corel dataset are used ($\alpha = 0.85$, $\beta = 0.524$, $\gamma = 0.065$) [16]. The values and a discussion about the impact of change of the parameters per dataset is given in [16]. The saliency boosting function $g(\mathbf{f_i})$ is estimated for every location, providing an image where rare colors provide higher gradient magnitudes compared to more common colors. The subsequent operations are equal to the extraction of light invariant points.

F. Discussion

To illustrate the different interest point detectors, examples are shown in Figure 2. Parameters are equal to the ones in the experiments. As the baseline, Harris Laplacian is extracted with the suggested threshold on $\mathfrak{C}$ of 500, the proposed approaches give at most 400 interest points per image. The first column gives three images from the VOC 2007 dataset. The other columns give the Harris Laplacian, light invariant points and color boosted points. Interest point locations and scales are indicated by the white circles. Generally, it can be derived that the luminance-based Harris Laplacian detects many background features that are solely detected due to shadows and shading. Intensity-based features are very sensitive to the smallest changes in the lighting conditions and thus less valuable for visual recognition and matching. The first row of Figure 2 shows that the two proposed approaches lead to very similar results. For light invariant points, shadows and specular reflections are disregarded. For color boosted points, patterns containing very common colors are disregarded. Both are able to reduce the number of non-robust features effectively.

In the presence of heavier shadowing effects in the second row of Figure 2, the color boosted points (h) are less robust and introduce small, more ambiguous features compared to the light invariant points (g). Harris Laplacian interest points (f) are located all over the shadows in the background. The third row of Figure 2) shows the ability of the color points to reduce the features caused by specular effects: the



(a) VOC 2007 nr. 395 (b) Harris Laplacian (c) light invariant points (d) color boosted points

(e) VOC 2007 nr. 337 (f) Harris Laplacian (g) light invariant points (h) color boosted points

(i) VOC 2007 nr. 134 (j) Harris Laplacian (k) light invariant points (l) color boosted points

Fig. 2. Three visual examples of the VOC 2007 dataset (row 1-3). The original images are found in the first column, luminance based Harris Laplacian in the second, light invariant points in the third, and color boosted points in the fourth. White circles indicate the location and the scale of interest points, parameters are chosen equal to the ones used in the experiments with 500 as the suggested threshold for Harris Laplacian and a maximum of 400 interest points for the proposed approaches.

color reflections on the car (i) are chosen by Harris Laplacian (j), but disregarded by the light invariant points (k), leading to a reduced number, but more stable features. The color boosted points (l) focus on salient colors, and provide therefore few features on colors of the reflection.

IV. Experiments

The stability of the proposed color interest points is tested by carrying out repeatability experiments on natural scenes in Section IV-A. For the image retrieval experiment in Section IV-B, we test whether fewer but more informative interest points will increase the retrieval precision for K-nearest neighbor (k-nn) classifiers. Finally, in the context of object recognition, the color interest points are tested on

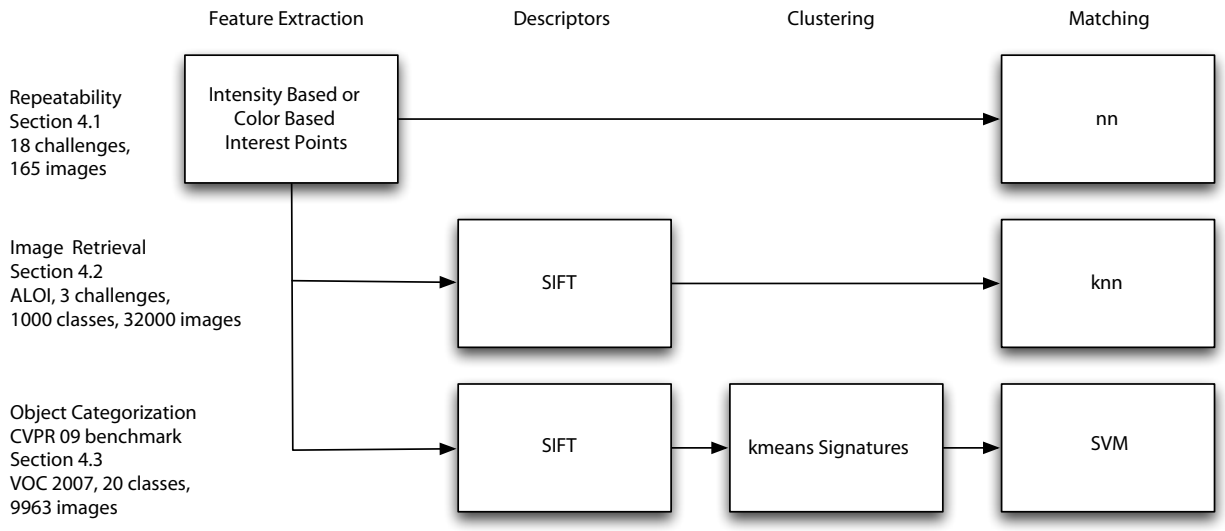


Fig. 3. The main stages of image matching and the correspondent stages of the experiments.

international object categorization benchmarks[1] [54].

In Figure 3, the steps of the experiments are shown. We evaluate the interest points in the feature extraction by using a (geometric) nearest neighbor matching (nn) for the repeatability experiments, a simple k-nn matching for image retrieval and a more robust state-of-the-art classification scheme for the object categorization benchmark.

Throughout the experiments, if we consider both of the proposed approaches, we refer to them as *color points*. For evaluating the impact of perceptual color spaces, we use the proposed scale selection in RGB and refer to it as *RGB points*. Further, we use RGB Color Harris with luminance-based Laplacian scale selection denoted by *RGB Color Harris*. It is a straightforward extension of [21]. As the state-of-the-art reference, we use the *Harris Laplacian* and its color variations. All experiments are carried out with the best performing parameters $\sigma_D = 1$, $l = 10$, $t = \sqrt{2}$ in [42] (defined in Section II). In case we choose a subset of the provided points, we rank the points by their Harris energy (according to eq. 10).

A. Repeatability

Mikolajczyk and Schmid [42] provide a way to test the quality of interest points: they measure the repeatability of interest points under different challenges[2]. We use the 18 challenges with color

[1] http://www.featurespace.org

[2] http://lear.inrialpes.fr/people/mikolajczyk/Database

Fig. 4. "Graffiti" test set used in the repeatability experiment challenging viewpoint variation.

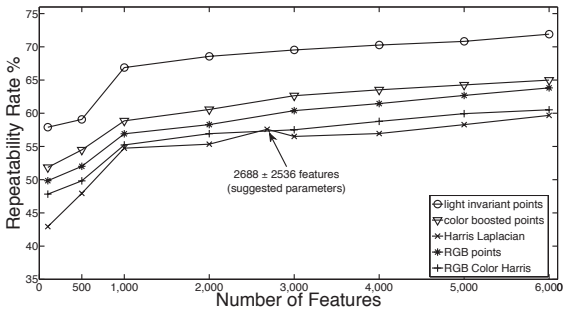


Fig. 5. The mean repeatability rate of the 18 repeatability challenges per number of points.

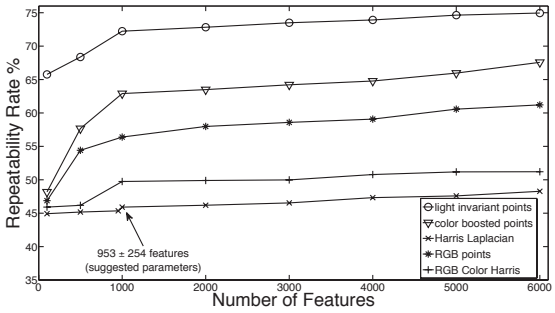


Fig. 6. The mean repeatability rate of the 5 datasets challenging lighting changes (cars, nuts, movi, fruits, toy).

information. An example set with challenging viewpoint changes is shown in Figure 4. The repeatability rate is defined as the ratio between the number of actual corresponding interest points and the total number of interest points that occur in the area common to both images. Feature detectors tend to have higher repeatability rates when they produce a richer description.

In Figure 5, the averaged results of the experiments are shown. We carry out the repeatability challenges on a varying number of features. It is observed that for all approaches, 1000 interest points per image are enough to reach more than 90% of the maximum repeatability rate. The experiment is carried out for up to 6000 interest points per image, when on average every pixel is selected by at least 10 features. We denote this as a *dense distribution* of interest points.

For the Harris Laplacian, the literature suggests a fixed threshold on the Harris energy (in the figures denoted as *suggested parameters*). This leads to a variable number of points for the images of the dataset due to their contrast. 1000 color points reach Harris Laplacian performance with the suggested parameters. This results in an average number of 2688 points in the experiment where the number of points varies from a minimum of 763 to a maximum of 9191 per image with a standard deviation of 2536 (we will denote these properties by "2688 [763,9191] $\pm$ 2536" in the following). Comparing light invariant points with Harris Laplacian, 100 light invariant points are enough to outperform the state-of-the-art.

Figure 6 shows the mean repeatability over the five datasets with varying lighting (cars, nuts, movi, fruits, toy). Increasing the number of Harris Laplacian points does not improve the repeatability against light changes significantly. In contrast, light invariant points remain more stable for all chosen numbers of points. Generally, color boosted points prove to be less repeatable than the HSI points. The reason is that the saliency function is sensitive to luminance changes.

In conclusion, the use of color information allows for extracting repeatable and scale invariant interest points. Further, the number of interest point can be reduced without loss of performance.

B. Image Retrieval

Fig. 7. Example images from the ALOI database.

This experiment evaluates the impact of different color spaces in image retrieval scenarios with varying illumination direction and intensity. The Amsterdam Library of Object Images (ALOI)[3] provides images of 1000 objects under supervised, predefined conditions recorded against a (dark) uniform background. Example images are shown in Figure 7.

The following experiment is carried out with 7000 images as the ground truth set and 1000 query images. The interest point approaches evaluated provide the locations and scales for the subsequent calculation of SIFT descriptors [40]. For matching, the similarity between two images is determined by using the 30 nearest SIFT descriptors between the two images. These 30 smallest distances are ranked in an increasing order. The final score is found by weighted sum of these distances, where the weight is the inverse rank of the descriptor. We measure the mean average precision (MAP) for the first 30 query results. Note that the only difference between the evaluated approaches is in the stage of interest point extraction.

The maximum number of N interest points implies that the N interest points with the largest Harris energies are extracted. If fewer than N interest points are detected for an image, then all of them are

[3]http://staff.science.uva.nl/~aloi/

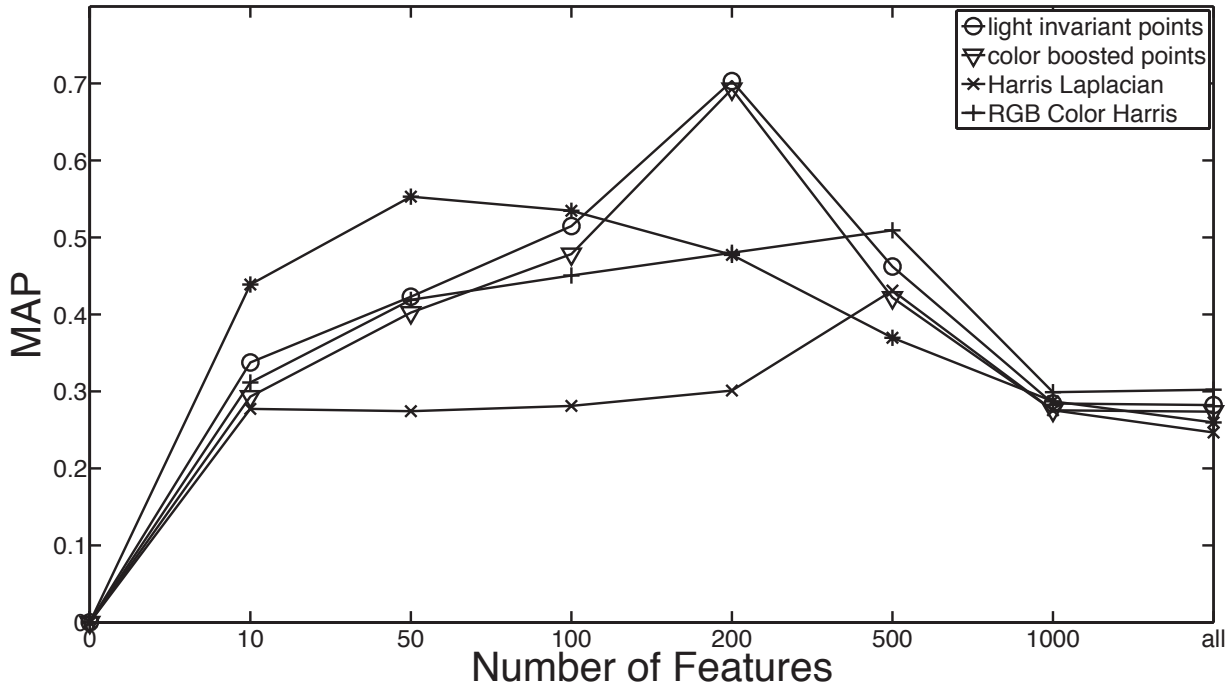


Fig. 8. Mean average precision (MAP) under varying maximum number of features for changing illumination direction on the ALOI database.

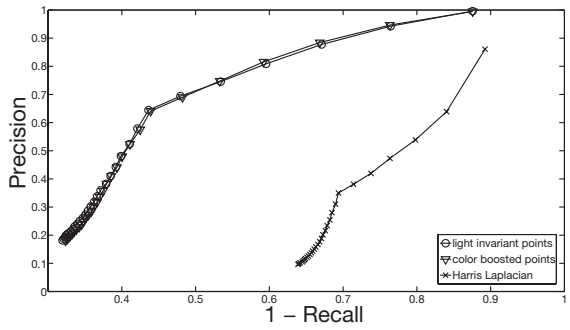


Fig. 9. Best performing color points compared to suggested parameters of Harris Laplacian for changing illumination direction on the ALOI database.

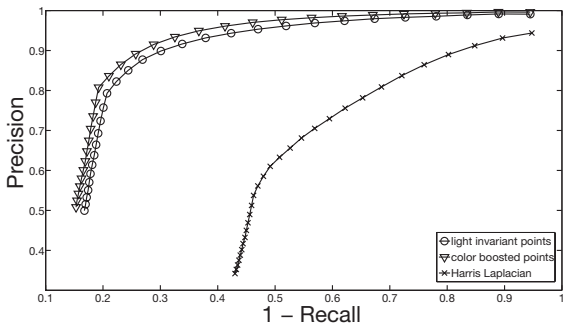


Fig. 10. Best performing color points compared to suggested parameters of Harris Laplacian for object rotation on the ALOI database.

used. First, all extractable interest points (up to 22117 maxima of the Harris energy per image) are used and then N is decreased (see Figure 8). Every approach has its best performing maximum number of interest points. Overcoming many problems of illumination changes, the color points remain more stable on the test images and thereby outperform all other approaches with a maximum number of 200 color points per image. Harris Laplacian reaches the best performance with a maximum of 500 points, which

(a) ALOI 225 (b) ALOI 245 (c) ALOI 584

Fig. 11. Sparse color points retrieve object 225 (a) perfectly with rank 8-13 going to object 245 (b). Dense points perform worse shifting object 584 (c) to rank 2-8.

approximately coincides with the suggested parameters of a fixed threshold providing 381 [12,8873] $\pm$ 393 Harris Laplacian points. The precision-recall graph showing the averages over 1000 queries for the best performing parameters of the color points and the suggested parameters of the Harris Laplacian (max. of 200 color points, 381 [12,8873] $\pm$ 393 Harris Laplacian points) is given in Figure 9.

This argument also holds for geometrical transformations of the objects: For each of the 1000 ALOI objects, 9 images are taken rotating the object 60 degrees in both directions. From 5 to 30 degree and 355 to 330 degree rotation, the steps are taken in 5 degree increments. Up to 60 degree and 300 degrees, respectively, the steps are carried out in 10 degree increments. This results in a database of 18000 images (see Figure 10).

The ALOI provides predefined changes in illumination color. The best performing parameters are evaluated on the dataset of 18000 images. The changes of color temperature are not severe enough to change locations of the detectors significantly. All approaches perform equally well with precision and recall of more than 0.99. Hence, no plot is shown.

Figure 11 shows one specific example of decrease in performance with an increasing number of interest points. Object 225 is shown in Figure 11(a): it is retrieved within the first 7 ranks being correct for 200 light invariant points. The next candidate on rank 8 is object 245 (Figure11(b)) for this set of parameters. This is not surprising as it contains similar texture. With this very few interest points, object 225 does not appear in ranks 1–7 for querying any of the other 999 objects. This means that the description is robust and distinct. Regarding all the interest points available, object 225 appears in 43 other queries in the top ranks, worsening the overall results significantly. Querying for object 225 itself, it still ranks one correct candidate at the first rank, having the following 7 from object 584 (Figure 11(c)). It seems that the only distinct features, the spikes at the border of object 225 and on the head of object 584 remain. The other features become more ambiguous, the more points we consider.

The difference in the image retrieval results is not only significant because of the images of colorful

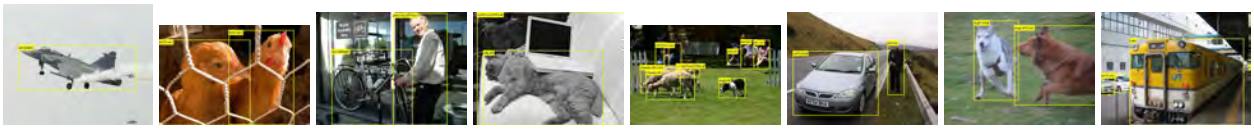


Fig. 12. Annotated sample images from the VOC 2007 dataset.

objects, our point is also emphasized by the use of a simple classification by the nearest neighbor. A more stable matching approach would gain and possibly compensate for retrieval performance, but the difference in the quality of the data would no longer be so obvious. Additionally, the approach has a runtime of $O(n^2)$, a reduction of the number of points changes the runtime of each query significantly.

In conclusion, the proposed color interest point detector has high repeatability and a reduced number of color features show no loss in performance.

C. Object Categorization

We now evaluate the color point detectors on the PASCAL VOC 2007 dataset[4] [55] in the context of object categorization [54]. This dataset consists of 9963 images, 20 classes of objects are annotated. The number of objects in one image is not fixed. The whole dataset contains 12608 objects. Twenty classes of objects (aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, diningtable, dog, horse, motorbike, person, sheep, sofa, train, and tvmonitor) are annotated with ground truth. Example images are shown in Figure 12.

The evaluated descriptors are clustered into 4000 clusters using the k-means algorithm. The image representation is obtained by a histogram of cluster occurrences. This is a 4000 bin histogram where bins correspond to clusters. Each bin contains the number of descriptors in the image that fall into the cluster corresponding to that bin. The Euclidean distance is used for quantization. For training, χ^2 distance and the generalized RBF kernel are used. Classification is performed using SVM kernel-fusion [27][5]. The performance is measured using the average precision (AP) per class or mean average precision (MAP) for all classes.

For the evaluation of detectors only, the standard SIFT is used. For descriptor evaluation only, the Harris Laplacian is used as a detector. This builds an independent evaluation of features with the best performing classification framework. 33 different approaches are evaluated.

[4]http://www.pascal-network.org/challenges/VOC/voc2007/

[5]http://www.featurespace.org/

The best performing approaches are denoted as follows.

Light invariant [400 | 800] SIFT uses the [400 | 800] most salient features and standard SIFT as a baseline description.

Dense [de1p | de2 | hap | hep] [30 | 192]c denotes the approach with either dense sampling combined with Hessian Laplacian and Harris Laplacian, dense sampling, Harris Laplacian, or Hessian Laplacian as the feature detection. The description is given either with 30 or 192 dimensional color descriptor of segmentation maps based on gradient, color and region shapes. An extensive parameter evaluation is given in [56].

HarLap [C-SIFT | oSIFT | cHoG | SURF | DAISY | 165c | 45c | 37c | 30c] uses the baseline Harris Laplacian detector and evaluates different descriptor performances: C-SIFT is a color SIFT descriptor using the intensity normalized OCS providing a 384 dimensional feature vector. It is the best performing descriptor of [57]. oSIFT denotes a modification of the baseline SIFT. The proposed rank-ordering normalizes the descriptor in order to be invariant to monotonic deformations of the baseline descriptor [58].

cHoG provides a compressed histogram of gradients that exploits gradient statistics in a canonical image patch. Tree-coding techniques are used to quantize the histograms to a length of 63 [59].

SURF is also inspired by SIFT, but uses the sum of approximated 2D Haar wavelet responses and makes efficient use of integral images [60]. The DAISY descriptor uses a circular grid and does not use weighted sums of gradient norms, but convolutions of gradients for the histogram of length 200 [61]. 165c denotes the color histogram of length 165; whereas 45c, 37c and 30c denote the color moments having a descriptor lenght of 45, 37 and 30, respectively. These color descriptors are described and evaluated in [57].

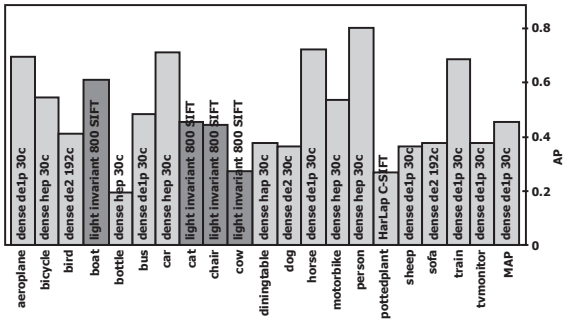


Fig. 13. Overall results for rank 1 approaches per class.

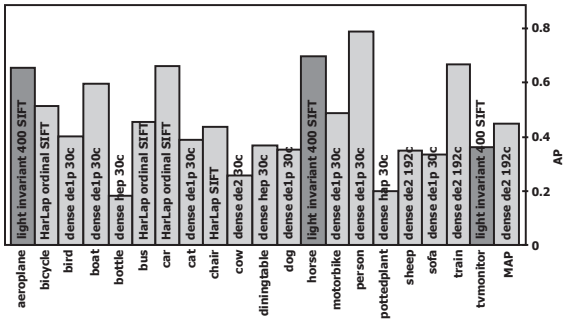


Fig. 14. Overall results for rank 2 approaches.

The results show that light invariant 800 SIFT outperforms all other approaches in 4 out of 20 classes

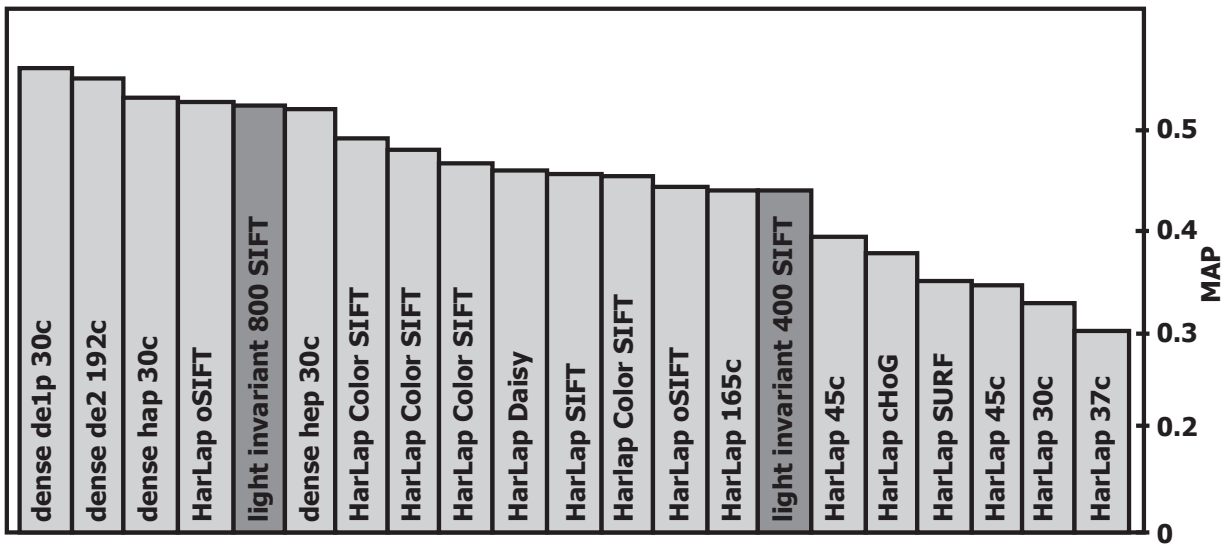


Fig. 15. Overall mean results for all classes.

(Figure. 13) and is ranked second in 3 other challenges (Figure 14). This is remarkable as the leading approaches use three detectors and dense sampling, providing a vast number of features for the subsequent steps of the categorization framework. The proposed method uses only a subset of these features and is still obtaining state-of-the-art results. The mean results over all classes are shown in Figure 15.

It is shown that, even with the baseline SIFT descriptor, light invariant points are able to outperform other approaches using more and better feature descriptions. Therefore, feature localization, scale estimation and their selection have an impact on state-of-the-art classification schemes. Using more stable and salient features, the number of features can be reduced while retaining the *meaningful* features and thus achieving state-of-the-art performance.

V. Conclusion

In this paper, a principled approach has been proposed to extract scale invariant interest points based on color and saliency. This allows the use of color-based interest points for arbitrary image matching. A PCA-based scale selection method is proposed which provides robustness to scale changes. Perceptual color spaces are incorporated and their advantages are directly passed on to the feature extraction. The use of color information allows for extracting repeatable and scale invariant interest points. Hence, more discriminative features and a sparser representation of images for image matching have been achieved. By reducing the number of features and providing a predictable number of sparse features, larger datasets can be processed in less time. Additionally, a stable number of features lead to a more predictable workload

for such tasks.

From large scale experiments, it has been shown that the proposed color interest point detector has a higher repeatability than a state-of-the-art luminance-based one. Further, a reduced number of color features increase the performance in image retrieval. Finally, for the PASCAL VOC 2007 challenge, our method gave comparable performance to the state-of-the-art in object categorization using only a subset of the features used for matching, reducing the computing time considerably.

ACKNOWLEDGMENTS

This work was partially supported by the European Commission under contract FP7-248984 GLOCAL and FP7-287704 CUBRIK.

REFERENCES

[1] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *CVPR*, 2003, pp. 264–271.

[2] C. Harris and M. Stephens, "A combined corner and edge detection," in *4th Alvey Vision Conference*, 1988, pp. 147–151.

[3] T. Kadir and M. Brady, "Saliency, scale and image description," *IJCV*, vol. 45, no. 2, pp. 83–105, 2001.

[4] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *IJCV*, vol. 60, no. 1, pp. 63–86, 2004.

[5] N. Sebe, T. Gevers, S. Dijkstra, and J. van de Weijer, "Evaluation of intensity and color corner detectors for affine invariant salient regions," in *CVPR* Workshop, 2006, p. 18.

[6] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *PAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[7] N. Sebe, T. Gevers, J. van de Weijer, and S. Dijkstra, "Corners detectors for affine invariant salient regions: Is color important?" in *CIVR*, vol. 4071, 2006, pp. 61–71.

[8] F. van de Velde, M. de Kamps, and G. T. van de Voort van de Kleij, "CLAM: Closed-loop attention model for visual search," *Neurocomputing*, no. 58-60, pp. 607–612, 2004.

[9] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *IJCV*, vol. 73, no. 2, pp. 213–238, 2007.

[10] K. Mikolajczyk, B. Leibe, and B. Schiele, "Multiple object class detection with a generative model," in *CVPR*, 2006, pp. 26–36.

[11] J. Sivic, B. Russell, A. A. Efros, A. Zisserman, and B. Freeman, "Discovering objects and their location in images," in *ICCV*, 2005, pp. 370–377.

[12] T. Tuytelaars and C. Schmid, "Vector quantizing feature space with a regular lattice," in *ICCV*, 2007, pp. 1–8.

[13] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *CVPR*, 2007, pp. 1–7.

[14] P. Turcot and D. G. Lowe, "Better matching with fewer features: The selection of useful features in large database recognition problems," in *ICCV*, Workshop, 2009.

[15] J. van de Weijer and T. Gevers, "Edge and corner detection by photometric quasi-invariants," *PAMI*, vol. 27, no. 4, pp. 625–630, 2005.

[16] J. van de Weijer, T. Gevers, and A. Bagdanov, "Boosting color saliency in image feature detection," *PAMI*, vol. 28, no. 1, pp. 150–156, 2006.

[17] A. E. Abdel-Hakim and A. A. Farag, "Csift: A sift descriptor with color invariant characteristics," in *CVPR*, 2006, pp. 1978–1983.

[18] F. Faille, "Stable interest point detection under illumination changes using colour invariants," in *BMVC*, 2005.

[19] P. Gabriel, J.-B. Hayet, J. Piater, and J. Verly, "Object tracking using color interest points," *AVSS*, pp. 159–164, 2005.

[20] V. Gouet and N. Boujemaa, "Object-based queries using color points of interest," *CVPR Workshop*, pp. 30–36, 2001.

[21] P. Montesinos, V. Gouet, and R. Deriche, "Differential invariants for color images," in *ICPR*, 1998, pp. 838–841.

[22] J. D. Rugna and H. Konik, "Color interest points detector for visual information retrieval," in *Electronic Imaging*, 2002, pp. 139–146.

[23] R. Unnikrishnan and M. Hebert, "Extracting scale and illuminant invariant regions through color," in *BMVC*, 2006.

[24] P.-E. Forssén, "Maximally stable colour regions for recognition and matching," in *CVPR*, 2007.

[25] D. A. R. Vigo, F. S. Khan, J. van de Weijer, and T. Gevers, "The impact of color in bag-of-words based on object recognition," in *ICPR*, 2010, pp. 1549–1552.

[26] J. Stöttinger, A. Hanbury, T. Gevers, and N. Sebe, "Lonely but Attractive: Sparse Color Salient Points for Object Retrieval and Categorization", in *CVPR* Workshop, 2009.

[27] F. Yan, J. Kittler, K. Mikolajczyk, and M. A. Tahir, "Non-sparse multiple kernel learning for fisher discriminant analysis," in *ICDM*, 2009, pp. 1064–1069.

[28] A. Torralba, R. Fergus, and Y. Weiss, "Small codes and large image databases for recognition," in *CVPR*, 2008, pp. 1–8.

[29] S. H. Srinivasan and N. Sawant, "Finding near-duplicate images on the web using fingerprints," in *ACM MM*, 2008, pp. 881–884.

[30] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *PAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.

[31] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

[32] R. Okada and S. Soatto, "Relevant feature selection for human pose estimation and localization in cluttered images," in *ECCV*, 2008, pp. 434–445.

[33] G. Dorko and C. Schmid, "Selection of scale-invariant parts for object class recognition," in *ICCV*, 2003, pp. 634–641.

[34] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *ICCV*, 2005, pp. 604–610.

[35] H. Moravec, *Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover*. CMU-RI-TR-80-03, 1980.

[36] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *ECCV*, 2002, pp. 128–142.

[37] A. P. Witkin, "Scale-space filtering," in *IJCAI*, 1983, pp. 1019–1022.

[38] T. Lindeberg, "Effective scale: A natural unit for measuring scale-space lifetime," in *ISRN KTH*, 1994.

[39] ——, "Feature detection with automatic scale selection," *IJCV*, vol. 30, no. 2, pp. 79–116, 1998.

[40] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[41] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions," in *BMVC*, 2002, pp. 384–393.

[42] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *IJCV*, vol. 65, no. 1/2, pp. 43–72, 2005.

[43] H. Konik, V. Lozano, and B. Laget, "Color pyramids for image processing," in *JIST*, vol. 40, no. 6, 1996, pp. 535–542.

[44] T. Gevers and A. W. M. Smeulders, "Color based object recognition," in *ICIAP*, 1997, pp. 319–326.

[45] J. Matas, D. Koubaroulis, and J. Kittler, "Colour image retrieval and object recognition using the multimodal neighbourhood signature," in *ECCV*, 2000, pp. 48–64.

[46] K. N. Plataniotis and A. N. Venetsanopoulos, *Color Image Processing and Applications*. Springer, 2000.

[47] P. Lambert and T. Carron, "Symbolic fusion of luminance-hue-chroma features for region segmentation," *PR*, vol. 32, pp. 1857–1872, 1999.

[48] A. Hanbury, "Constructing cylindrical coordinate colour spaces," *PR Letters*, vol. 29, pp. 494–500, 2008.

[49] A. Leonardis and H. Bischof, "Robust recognition using eigenimages," *CVIU*, vol. 78, no. 1, pp. 99 – 118, 2000.

[50] H. Murase and S. K. Nayar, "Visual learning and recognition of 3-d objects from appearance," *IJCV*, vol. 14, pp. 5–24, 1995.

[51] K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points," in *ICCV*, 2001, pp. 525–531.

[52] C. Kenney, M. Zuliani, and B. Manjunath, "An axiomatic approach to corner detection," in *CVPR*, 2005, pp. 191–197.

[53] D. G. Lowe, "Object recognition from local scale-invariant features," in *ICCV*, vol. 2, 1999, pp. 1150–1157.

[54] K. Mikolajczyk, M. Barnard, J. Matas, and T. Tuytelaars, "Feature detectors and descriptors: The state of the art and beyond," Feature Workshop at CVPR, Tech. Rep., 2009.

[55] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results."

[56] P. Koniusz and K. Mikolajczyk, "On a quest for image descriptors based on unsupervised segmentation maps," in *ICPR*, 2010, pp. 762 – 765.

[57] K. van de Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *PAMI*, vol. 32, pp. 1582 – 1596, 2009.

[58] M. Toews and W. Wells, "Sift-rank: Ordinal description for invariant feature correspondence," *CVPR*, pp. 172–177, 2009.

[59] V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, R. Grzeszczuk, and B. Girod, "Chog: Compressed histogram of gradients a low bit-rate feature descriptor." in *CVPR*, 2009, pp. 2504–2511.

[60] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *ECCV*, 2006, pp. 346– 359.

[61] E. Tola, V. Lepetit, P. Fua, and S. Member, "Daisy: An efficient dense descriptor applied to wide baseline stereo," *PAMI*, 2010.

Julian Stöttinger received the B.Sc. in media informatics 2004, M.Sc in computer graphics and digital image processing (Hons) 2007 and Ph.D. degree in technical sciences (Hons) 2010, all from the Vienna University of Technology. He is a post-doctoral researcher at the Knowdive group at the University of Trento, an associate center of the European Institute of Innovation and Technology. Main research interests are local features, color-based recognition, semantic image understanding, and compositional approaches for visual recognition.

Allan Hanbury received the B.Sc. degree in physics and applied mathematics in 1994, the B.Sc. (Hons) degree in physics in 1995, and the M.Sc. degree in physics in 1999, all from the University of Cape Town, South Africa. He received the Ph.D. degree in Mathematical Morphology in 2002 from Mines ParisTech, France, and the Habilitation in Practical Informatics from the Vienna University of Technology, Austria in 2008.

He is currently a Senior Researcher with the Information and Software Engineering Group, Institute of Software Technology and Interactive Systems, Vienna University of Technology, Austria. He is the Scientific Coordinator of the EU-funded KHRESMOI Integrated Project on biomedical information search and analysis and has led a number of Austrian national projects. His research interests include information retrieval, multimodal information retrieval, health information retrieval, and the evaluation of information retrieval algorithms. He is an author or coauthor of over 60 publications in refereed journals and international conferences.

Nicu Sebe is with the Faculty of Cognitive Sciences, University of Trento, Italy, where he is leading the research in the areas of multimedia information retrieval and human-computer interaction in computer vision applications. He was involved in the organization of the major conferences and workshops addressing the computer vision and human-centered aspects of multimedia information retrieval, among which as a General Co-Chair of the IEEE Automatic Face and Gesture Recognition Conference, FG 2008, ACM International Conference on Image and Video Retrieval (CIVR) 2007 and 2010, and WIAMIS 2009 and as one of the initiators and a Program Co-Chair of the Human-Centered Multimedia track of the ACM Multimedia 2007 conference. He is the general chair of ACM Multimedia 2013 and a program chair of ACM Multimedia 2011. He has served as the guest editor for several special issues in IEEE Computer, Computer Vision and Image Understanding, Image and Vision Computing, Multimedia Systems, and ACM TOMCCAP. He has been a visiting professor in Beckman Institute, University of Illinois at Urbana-Champaign and in the Electrical Engineering Department, Darmstadt University of Technology, Germany. He is the co-chair of the IEEE Computer Society Task Force on Human-centered Computing and is an associate editor of Machine Vision and Applications, Image and Vision Computing, Electronic Imaging and of Journal of Multimedia.

Theo Gevers is an associate professor of computer science at the University of Amsterdam, The Netherlands, and a full professor at the Computer Vision Center, Universitat Autnoma de Barcelona, Barcelona, Spain. At the University of Amsterdam, Theo Gevers is a teaching director of the MSc in Artificial Intelligence. He currently holds a VICI Award (for research excellence) from the Dutch Organisation for Scientific Research. He is a co-founder and chief scientific officer of ThirdSight, a spin-off of the UvA. His main research interests are in the fundamentals of image understanding, object recognition and color in computer vision. He is the chair for various conferences and is an associate editor for the IEEE Transactions on Image Processing. Further, he is a program committee member for a number of conferences, and an invited speaker at major conferences. He is a lecturer delivering postdoctoral courses given at various major conferences (CVPR, ICPR, SPIE, and CGIV). He is a member of the IEEE.