Computer Vision and Image Understanding 113 (2009) 48–62

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Performance evaluation of local colour invariants

Gertjan J. Burghouts [a,*], Jan-Mark Geusebroek [b]

[a] TNO Observation Systems, Electro Optics, Oude Waalsdorperweg 63, 2597 AK, The Hague, The Netherlands
[b] Intelligent Systems Lab Amsterdam, Informatics Institute, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 30 September 2007
Accepted 8 July 2008
Available online 24 July 2008

Keywords:
Local descriptors
Colour
SIFT

ABSTRACT

In this paper, we compare local colour descriptors to grey-value descriptors. We adopt the evaluation framework of Mikolayzcyk and Schmid. We modify the framework in several ways. We decompose the evaluation framework to the level of local grey-value invariants on which common region descriptors are based. We compare the discriminative power and invariance of grey-value invariants to that of colour invariants. In addition, we evaluate the invariance of colour descriptors to photometric events such as shadow and highlights. We measure the performance over an extended range of common recording conditions including significant photometric variation. We demonstrate the intensity-normalized colour invariants and the shadow invariants to be highly distinctive, while the shadow invariants are more robust to both changes of the illumination colour, and to changes of the shading and shadows. Overall, the shadow invariants perform best: they are most robust to various imaging conditions while maintaining discriminative power. When plugged into the SIFT descriptor, they show to outperform other methods that have combined colour information and SIFT. The usefulness of `C-colour-SIFT` for realistic computer vision applications is illustrated for the classification of object categories from the VOC challenge, for which a significant improvement is reported.

1. Introduction

Many computer vision tasks depend heavily on local feature extraction and matching. Object recognition is a typical case where local information is gathered to obtain evidence for recognition of previously learned objects. Recently, much emphasis has been placed on the detection and recognition of locally (weakly) affine invariant regions [1–5]. The rationale here is that planar regions transform according to well known laws. Successful methods rely on fixing a local coordinate system to a salient image region, resulting in an ellipse describing local orientation and scale. After transforming the local region to its canonical form, image descriptors should be well able to capture the invariant region appearance. As pointed out by Mikolajczyk and Schmid [6], the detection of elliptic regions varies covariantly with the image (weak perspective) transformation, while the normalized image pattern they cover and the image descriptors derived from them are typically invariant to the geometric transformation. Recognition performance is further enhanced by designing image descriptors to be photometric invariant, such that local intensity transformations due to shading and variation in illumination have no or limited effect on the region description. State-of-the-art methods in object recognition normalize mean intensity and standard deviation of the intensity image [2,6,7]. Moreover, image measurements using a Gaussian filter and its derivatives is becoming increasingly popular as a way of detecting and characterizing image content in a geometric and photometric invariant way. Gaussian filters have interesting properties from an image processing point of view, among others, their robustness to noise [8], their rotational steerability [9], and their applicability in multi-scale settings [10]. Many of the intensity based descriptors proposed in literature are based on Gaussian (derivative) measurements [1,11–14]. Hence, as one contribution of this paper, we aim to evaluate the Gaussian derivative performance independent of the descriptor. A well engineered exponent of intensity descriptors is Lowe's SIFT descriptor [2]. Indeed, for grey-value descriptors, the detection of affine regions combined with the SIFT descriptor is demonstrated to be better than many alternatives [1]. Hence, as a second contribution of this paper, we aim to extent this descriptor to colour, and we will evaluate its performance with respect to photometric variation and discriminative power.

In this paper, we consider the extension to colour-based descriptors. colour has high discriminative power; in many cases, objects can well be recognized merely by their colour characteristics [15–20]. However, photometric invariance is less trivial to achieve, as the accidental illumination and recording conditions affect the observed colours in a complicated way. Photometric invariance has been intensively studied for colour features

* Corresponding author.
E-mail addresses: gertjan.burghouts@tno.nl (G.J. Burghouts), mark@science.uva.nl (J.-M. Geusebroek).

1077-3142/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2008.07.003

[16,17,21–24]. Geusebroek et al. [25] derived a set of colour invariant features based on the Gaussian derivative framework, facilitated by Koenderink's Gaussian colour model. The important research question is if colour-based descriptors indeed improve upon their grey-based counterparts in practise. The answer depends on the stability of the non-linear combinations of Gaussian derivatives necessary to achieve a similar level of invariance as implemented in grey-value descriptors. For instance, the values of photometric invariants are distorted when the image is JPEG compressed, as the compression distorts the pixel values and spatial layout, and more for the colour channels than for the intensity. Therefore, we aim at a comparative study of local colour descriptors, in comparison to grey-value descriptors.

To be precise on the scope of the paper, there is no need to address the issue of (affine) region detection, as many well performing methods exist [6,26–31]. Hence, we will concentrate on descriptor performance. Furthermore, to enable a fair comparison between intensity based descriptors and colour based descriptors, we demand identical geometric invariance for both intensity based features and colour based features. This requirement is conveniently fulfilled by the Gaussian measurement framework.

For the evaluation of local grey-value and colour invariants, we adopt the extensive methodology of Mikolajczyk and Schmid [1]. In this paper, the authors propose the evaluation of descriptor performance by the matching of regions from one image to another image. Correct matches are determined using the homography between the two images. From [1], we adopt the measures to evaluate discriminative power and invariance. Also, we adopt variety in recording conditions, being changes of illumination intensity, of the camera viewpoint, blurring of the image, and JPEG compression. We go beyond [1] by extending this set with images recorded under different illumination colours and illumination directions. These conditions induce a significant variation in the image recording. For an illustration of images recorded under varying illumination directions, see Fig. 1.

We extend the number of images used in the evaluation framework [1] to 26,000 images, representing 1000 objects recorded under 26 imaging conditions. Moreover, we further decompose the evaluation framework in [1] to the level of local grey-value invariants on which common region descriptors are based. We measure the performance of photometric invariants for the detection of colour transitions only. Hence, we evaluate the performance of the Gaussian grey-value and colour invariant derivatives, to indicate the merit of the invariant when plugged into a region descriptor. Finally, we establish performance criteria that are specific to colour invariants, indicating the level of invariance with respect to photometric variation, and evaluating the ability to distinguish between various photometric effects.

The paper is structured as follows: In Section 2, we shortly overview grey-value and photometric invariants and we discuss previous work on the evaluation of grey-value image invariants, which we relate to the evaluation of photometric invariants as proposed in this paper. Section 3 describes the invariant features used in our comparison. Section 4 discusses the performance measures and the datasets, and presents the experimental results. For a realistic application of the invariants, we evaluate the performance on the VOC dataset [32] in Section 5. Conclusions are drawn in Section 6.

2. Previous work

2.1. Grey-value invariants

Many techniques for the description of images have considered local features. Methods based on local intensity values in the image, see e.g. [33,34], are successfully applied to image matching. A considerable step forward was the work by Schmid and Mohr [12]. They combined Gaussian derivative measurements in a multi-scale and rotation invariant descriptor. The Gaussian derivatives were computed at Harris corner points [11], achieving general recognition under occlusion and clutter. The choice for the Gaussian filter was fundamental in there method, allowing their descriptor to capture the local differential structure of the image [35] such that scale-invariance was achieved.

To identify an appropriate and consistent scale for Gaussian-based image measurements, Lindeberg [10] determined local maxima over scale. This scheme determines the characteristic scale for the local differential image structure, and has been successfully applied to detect keypoints [2] and multiscale Harris detectors [6]. To achieve invariance to affine planar transformations, Lindeberg and Gårding [27] considered a local affine adaptation. Such an affine adaptation has recently been incorporated in Harris-affine and Hessian-affine detectors [6].

The use of the local Gaussian differential structure has received considerable interest. Gaussian derivative based descriptors have been proven to be very distinctive for matching, see e.g. [36–38]. Schiele and Crowley [13] modelled differential structure across an image by accumulating image derivatives into histograms, effectively capturing texture information. Belongie et al. [39] accumulated image derivatives in a regional grid with multiple bins to model both shape and location information, resulting in the so-called shape-context. Varma and Zisserman [40] modelled texture appearance by accumulation of the Gaussian-based MR8 filterbank. Winn et al. [41] are using a Gaussian filterbank for object recognition by a visual dictionary approach.

The most successful local image descriptor so far is Lowe's SIFT descriptor [2]. The SIFT descriptor encodes the distribution of Gaussian gradients within an image region. The SIFT descriptor is a 128-bin histogram that summarizes local oriented gradients over 8 orientations and over 16 locations. This represents the spatial intensity pattern very well, while being robust to small deformations and localization errors. Nowadays, many modifications and improvements exist, among others, PCA-SIFT [42], GLOH [1], Fast approximate SIFT [43], and SURF [44]. These region-based descriptors have achieved a high degree of invariance to overall illumination conditions for planar surfaces. Although designed to retrieve

Fig. 1. Example object recorded under semi-hemispherical illumination, and images recorded under an illuminant at decreasing altitude angles. Illuminant azimuth is to the right of the object.

identical object patches, SIFT-like features turn out to be quite successful in bag-of-feature approaches to general scene and object categorization, see e.g. [45].

2.2. Photometric invariants

Colour invariants have received extensive theoretical and experimental treatment, due to the additional discriminative power that comes with colour information in comparison to grey-value information. Additionally, colour information enables one to distinguish between true colour variation and photometric distortions, as pointed out by Gershon [46]. Indeed, for colour information to be useful, Slater and Healey [47], Finlayson [22], and Gevers and Smeulders [17], have all stressed the importance of achieving invariant colour measurements to varying lighting conditions such as a change in illumination colour, illumination direction, or camera viewpoint.

Photometric invariants can be derived from the physical laws of light reflection. Methods which normalize mean intensity and standard deviation for grey-value descriptors are assuming Lambert's law of light reflection, $I = \rho l\, n$. Here, the observed image I is the result of a multiplicative formation process, for which ρ represents the surface albedo, l the light source direction and intensity, and n the surface normal. Normalization of the (local) standard deviation removes the contribution of l in the image descriptor, whereas the mean normalization counteracts the camera sensitivity offset. However, the normalized result still depends on both the surface reflectance ρ and the geometry of the surface represented by it's normal n. Hence, shadow and shading edges are coded by image descriptors. This very effect causes nowadays image descriptors to be effective for planar patches only.

Colour images convey more information about the image formation process, and hence may improve on the features which can be discriminated. Inspired by the success of colour indexing [15], Funt and Finlayson [16,22] use the Lambertian assumption to arrive at photometric invariant indexing of images. Although the methodology to achieve photometric invariance is essentially similar to the grey-value case outlined above, they improve in discriminative power by adding the extra colour information availably from the image. Furthermore, by exploiting the extra information which comes with colour, they discount the effects of shadow and shading on their image descriptor. Gevers and Smeulders [17] elaborate on this work by deriving several sets of invariants. These sets are invariant under the more complicated photometric model proposed by Shafer [48]. In this way, they arrive at features invariant for highlights and for coloured illumination. In consecutive work [49], shadows, highlights, and true colour boundaries are separated in practise, based on a pixel-wise comparison of invariant values.

Geusebroek et al. [25] extended photometric invariance to Gaussian-based derivatives, facilitated by Koenderink's Gaussian framework. Hence, effectively combining photometric colour invariance with the highly successful Gaussian geometric invariants. The pixel-based invariants can still be represented by considering the limiting case of the spatial scale for the Gaussian filters small, such that single pixels are covered. However, tuning the filters to a larger scale allows for the more interesting class of geometric *and* photometric invariant features.

Promising recent methods aim at combining colour and shape description of the local neighbourhood. Mindru et al. [50] have considered colour moments, which are invariant to illumination colour. However, in [1], local moments-based descriptors were found to be relatively unstable. Van de Weijer and Schmid [51] augmented the SIFT descriptor with a histogram of photometric invariant values, effectively combining colour and shape information. They have shown that adding colour information to the SIFT descriptor improves its discriminative power. Likewise, Geodeme et al. [52] have used localized colour moments to reduce *a posteriori* the mismatches of SIFT descriptors. Other recent approaches have altered the SIFT descriptor itself. Abdel-Hakim and Farag [53] have based the SIFT descriptor on the hue gradient rather than the intensity gradient. Bosch and Zisserman [54] have computed SIFT from the HSV representation to provide a richer descriptor. Unfortunately, the improvement in performance for such descriptors is unclear, as no well established evaluation method is available for colour based descriptors.

2.3. Performance evaluation

For the evaluation of discriminative power of local descriptors, an extensive evaluation framework has been proposed by Mikolayzcyk et al. [1,55]. They aimed at evaluating the different stages of an nowadays object recognition framework, by decomposing the benchmark in the separate evaluation of keypoint detection and local image descriptors. Furthermore, they realized the importance of evaluating robustness against geometric and photometric distortions of the target image. They evaluate the discriminative power and invariance of descriptors over various imaging conditions. Discriminative power for any of the detector–descriptor combinations is evaluated over: illumination intensity, of the camera viewpoint, blurring of the image, and JPEG compression. Invariance is measured by the performance degradation over increasingly hard imaging conditions, e.g. increasing JPEG compression rates. Moreels and Perona [56], and Fraundorfer and Bischof [57] elaborated on this framework by considering descriptor evaluation for 3D objects.

Van de Sande et al. [58] evaluates colour SIFT descriptors for object an scene recognition. Their study provides a theoretical overview of photometric invariant properties, and an evaluation of various colour SIFT descriptor on PASCAL VOC [32] data and TRECVID data [59]. However, the authors do not link the theoretical derived invariance properties to relevant experiments. Hence, no insight is gained in the effectiveness of the various photometric invariants in discounting imaging effects. Only an overall picture of classification performance on these specific datasets is provided. We improve on this work by demonstrating both theoretically and experimentally the merit of photometric invariance in SIFT descriptors. We further break down the performance to the low-level (Gaussian) filtering and the higher-level (SIFT) feature extraction.

2.4. The contribution of this paper

With the increasing interest in distinctive and robust local features, we propose in this paper a benchmark for the evaluation of local colour invariants. The contribution of this paper is threefold:

- We establish a framework for the evaluation of colour image descriptors, including a suitable dataset and three measures of performance: discriminative power, constancy under irrelevant image distortions or imaging conditions, and the ability to distinguish true (object) variation from irrelevant (photometric) variation.
- We include colour information in SIFT descriptors, and propose three colour SIFT methods each having different characteristics with respect to photometric variation.
- We evaluate the performance of these descriptors together with the performance of the Gaussian colour invariants on which they are based. We compare with alternative colour SIFT implementations from literature.

Regarding the first contribution, we adopt the setup from [1,55] to evaluate descriptor performance over increasingly hard imaging conditions. We consider the ALOI database [60] to match regions that are computed from 26,000 images of 1000 objects in total. Ground truth is obtained by manual selection of stable Harris-affine regions inside the objects. The dataset contains both image transformations as well as photometric variation in imaging conditions, and is considered more suitable for evaluation of colour descriptors than the original database proposed by Mikolayzcyk et al. [1,55] or the one proposed by Moreels and Perona [56]. For example, the database contains six different lighting conditions and, very important for assessing colour descriptors, variation in illumination colour. Hence, allowing the assessment of colour constancy for colour image descriptors.

With respect to our second contribution, we will include the Gaussian colour invariant gradients proposed in [25] into the SIFT descriptor [2]. We will evaluate their performance with respect their grey-value counterparts, and with respect to colour SIFT descriptors from literature [53,54].

Finally, our third contribution further decomposes the evaluation framework proposed in [1,55]. Mikolayzcyk et al. evaluate discriminative power and invariance of region descriptors. SIFT-based descriptors consist of a set of Gaussian derivative image measurements and a well-designed histogram description thereof. The performance of the Gaussian filter and the non-linear combinations to obtain geometric invariance are well known and taken for granted. However, for photometric invariance, non-linear combinations may significantly alter its performance. Hence, we decompose the benchmark proposed by Mikolayzcyk et al. further in order to address this issue separately. We abstract from the descriptors here, and evaluate the underlying, local invariants only. The discriminative power and invariance will be established for local grey-value invariants, and for the Gaussian colour invariants of [25]. Furthermore, following [49], we will assess the power of an invariant to distinguish object colour variation from photometric variation.

3. Invariants

We will evaluate the performance of Gaussian-based invariant features. For completeness, and to introduce notation, we shortly rehearse grey-value differential invariants and colour invariants in this section.

3.1. Grey-value invariants

We denote a grey-value image $E(x,y)$, with a scalar value at pixel location (x,y). The filtering of a grey-value image by an (isotropic) Gaussian $G^\sigma(x,y)$ at scale σ is given by (leaving out pixel position parameters): $\hat{E}^\sigma = E * G^\sigma$, where $*$ is the convolution operator. The notational use of the hat symbol $(\hat{\cdot})$ implies dependence on the scale parameter σ, hence we leave the scale parameter out in the following and simply use $\hat{\cdot}$. More generally, we consider the filtering of an image $E(x,y)$ by a Gaussian filter G and its x- and y-derivatives,

$$\hat{E}_j = E * G_j, \tag{1}$$

where subscript $j \in \{\emptyset, x, y\}$ indicates either smoothing or spatial differentiation.

The gradient is a rotation invariant derivative measurement, given by

$$\hat{E}_w = \sqrt{\hat{E}_x^2 + \hat{E}_y^2}. \tag{2}$$

Normalizing each gradient value by the local intensity suppresses regional intensity variations [25],

$$\widehat{W}_w = \frac{\hat{E}_w}{\hat{E}}. \tag{3}$$

3.2. Colour invariants

We consider the colour-based photometric invariants from [25], which are derived from the Gaussian opponent colour model. First, we recap this colour model. Three opponent colours are obtained per pixel: $E(x,y)$, $E_\lambda(x,y)$ and $E_{\lambda\lambda}(x,y)$, representing ,respectively, the intensity, the yellow–blue channel, and the red–green channel. Here, we consider the Gaussian opponent colour model, which is computed from RGB values directly by the linear transformation [25]:

$$\begin{bmatrix} \hat{E}(x,y) \\ \hat{E}_\lambda(x,y) \\ \hat{E}_{\lambda\lambda}(x,y) \end{bmatrix} = \begin{pmatrix} 0.06 & 0.63 & 0.27 \\ 0.30 & 0.04 & -0.35 \\ 0.34 & -0.60 & 0.17 \end{pmatrix} \begin{bmatrix} R(x,y) \\ G(x,y) \\ B(x,y) \end{bmatrix}, \tag{4}$$

where $\hat{E}$, $\hat{E}_\lambda$ and $\hat{E}_{\lambda\lambda}$ denote the intensity, blue–yellow and green–red channel. The transformation effectuates the decorrelation of RGB values.

Gaussian (derivative) filtering and construction of the gradient for each opponent colour channel is similar to the grey-value case. The colour-based counterpart of Eq. (2) becomes

$$\hat{E}_{\lambda^i w} = \sqrt{\hat{E}_{\lambda^i x}^2 + \hat{E}_{\lambda^i y}^2}. \tag{5}$$

Likewise, the colour invariants $\widehat{W}_{\lambda w}$ and $\widehat{W}_{\lambda\lambda w}$ are a generalization of the grey-value invariant $\widehat{W}_w$ from Eq. (3),

$$\widehat{W}_{\lambda^i w} = \frac{\hat{E}_{\lambda^i w}}{\hat{E}}. \tag{6}$$

Note that for $i = 0$ in λ^i, the results for Eqs. (5) and (6) indeed is exactly the grey-value invariants $\widehat{E}_w$ and $\widehat{W}_w$ (Eqs. 5 and 6) by the very construction of the opponent colour space: the first channel ($i = 0$) is the intensity channel. The photometric invariants $\widehat{W}_w$, $\widehat{W}_{\lambda w}$ and $\widehat{W}_{\lambda\lambda w}$ are invariant to regional variations of the intensity.

Likewise, other photometric invariants can be constructed. The invariants $\widehat{W}_{\lambda^i x}$ compute first the gradient and normalize it by the local intensity later. Alternatively, the intensity normalized colour values $\frac{\hat{E}_\lambda(x,y)}{\hat{E}(x,y)}$ and $\frac{\hat{E}_{\lambda\lambda}(x,y)}{\hat{E}(x,y)}$ can be differentiated with respect to x or y, which, using the chain rule for differentiation, yields

$$\widehat{C}_{\lambda j} = \frac{\hat{E}_{\lambda j}\hat{E} - \hat{E}_\lambda \hat{E}_j}{\hat{E}^2}, \tag{7}$$

$$\widehat{C}_{\lambda\lambda j} = \frac{\hat{E}_{\lambda\lambda j}\hat{E} - \hat{E}_{\lambda\lambda} \hat{E}_j}{\hat{E}^2}, \tag{8}$$

where subscript $j \in \{x,y\}$ indicates spatial differentiation. Under Lambertian reflection, the normalization of colour values by the local intensity results in colour values independent of the intensity distribution. Hence, $\widehat{C}_{\lambda j}$ and $\widehat{C}_{\lambda\lambda j}$ and their derivatives are invariant to shadow and shading. The shadow and shading invariant gradients are obtained from $\widehat{C}_{\lambda w} = \sqrt{\widehat{C}_{\lambda x}^2 + \widehat{C}_{\lambda y}^2)}$ and $\widehat{C}_{\lambda\lambda w} = \sqrt{\widehat{C}_{\lambda\lambda x}^2 + \widehat{C}_{\lambda\lambda y}^2}$.

A next step is to include the Fresnel reflectance, hence additionally modelling highlights. In this case, the local colour ratio, $\frac{\hat{E}_\lambda(x,y)}{\hat{E}_{\lambda\lambda}(x,y)}$, is invariant to the intensity distribution and the Fresnel coefficient (see [25] for details). Invariance to the Fresnel coefficient implies invariance to highlights in the image. Again applying the chain rule to obtain spatial derivatives yields

(a) 100 example objects

(b) Reference image and testing conditions

Fig. 2. Randomly selected objects from the ALOI collection are depicted in (a). Imaging conditions are shown in (b), respectively: the reference image, blurring (σ = 2.8 pixels, image size 192 × 144), JPEG compression (50%), illumination direction change (to 30° altitude, from the right), viewpoint change (30°), illumination colour change (3075 $K \rightarrow$ 2175 K).

$$\widehat{H}_j = \frac{\widehat{E}_{\lambda\lambda}\widehat{E}_{\lambda j} - \widehat{E}_{\lambda}\widehat{E}_{\lambda\lambda j}}{\widehat{E}_{\lambda}^2 + \widehat{E}_{\lambda\lambda}^2}, \tag{9}$$

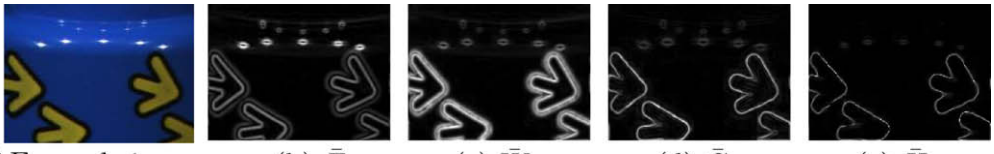
where subscript $j \in \{x,y\}$ indicates spatial differentiation. This yields the gradient $\widehat{H}_w = \sqrt{\widehat{H}_x^2 + \widehat{H}_y^2}$, which is invariant to shadow, shading and highlights.

To illustrate the gradient measurements by the photometric invariants, we combine the invariants in each of the sets $\{\widehat{W}_w, \widehat{W}_{\lambda w}, \widehat{W}_{\lambda\lambda w}\}$ and $\{\widehat{C}_{\lambda w}, \widehat{C}_{\lambda w}\}$ to obtain a single value per pixel ($\widehat{H}_w$ already yields a single value per pixel). The combined edge strength is measured by root of the squared sum. For W we compute $\overline{W}_w = \sqrt{\widehat{W}_w^2 + \widehat{W}_{\lambda w}^2 + \widehat{W}_{\lambda\lambda w}^2}$, whereas for C we have $\overline{C}_w = \sqrt{\widehat{C}_{\lambda w}^2 + \widehat{C}_{\lambda\lambda w}^2}$. Furthermore, we define $\overline{E}_w$ as the non-normalized combined edge strength over all colour channels, that is, similar to $\overline{W}_w$ but without the local intensity normalization. The total edge strengths $\overline{E}_w$, $\overline{W}_w$, $\overline{C}_w$, $\overline{H}_w \equiv \widehat{H}_w$, each illustrating one set of photometric invariants, are depicted in Fig. 3. Note that the shading is removed by $\overline{C}_w$ (d), and that the non-saturated highlights are removed by $\overline{H}_w$ (e).

4. Performance evaluation

We compare the local grey-value and colour invariants based on three evaluation criteria:

- Discriminative power. We establish the power of each invariant to discriminate between image regions. Discriminative power is measured by the quality of region matching, similar to [1]. The successful matching strategy as proposed by Lowe [2], is based on the rationale that for the recognition of an object, it suffices to correctly match only a few regions of that object. In our exper-



(a) Example image (b) $\bar{E}_w$ (c) $\bar{W}_w$ (d) $\bar{C}_w$ (e) $\bar{H}_w$

Fig. 3. Photometric invariant gradients. $\overline{E}_w$ is not photometric invariant, $\overline{W}_w$ is invariant to illumination intensity, $\overline{C}_w$ is invariant to shadow and shading, $\overline{H}_w$ is invariant to shadow, shading and highlights.

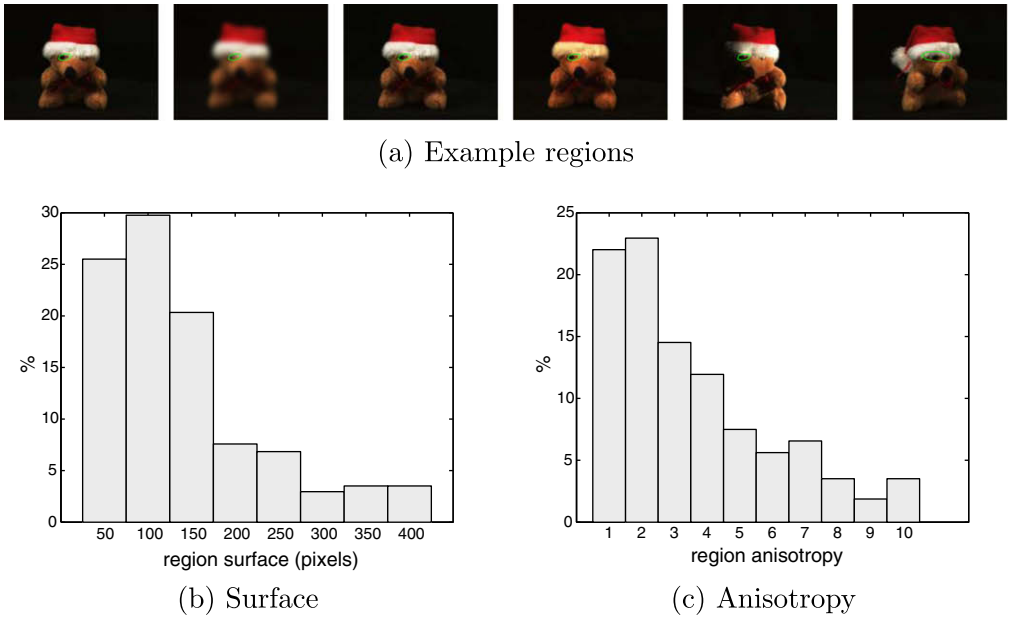


Fig. 4. (a) Image regions for, respectively: the reference image, blurring, JPEG compression, illumination colour change, illumination direction change, and viewpoint change. For all imaging conditions except the change of viewpoint, the camera is fixed, so the regions are set identical. For the camera viewpoint change, we have manually selected the most stable region. Histogram of (b) the size of the region surfaces, and (c) of the anisotropy (where anisotropy = 1 indicates isotropy).

imental framework, we push this to the extreme, and consider the matching of one region of an object against a database of 1000 regions: one noisy realization of the same object matched against 999 of other objects. Under noisy conditions we consider image deformations caused by blurring, JPEG compression and out-of-plane object rotation (viewpoint change), and photometric variation induced by changes in illumination direction and illumination colour. Precision and recall characteristics reflect the discriminative power of the invariant under evaluation.

- Invariance or robustness. As above, but now we establish the degradation of the number of correct matches as function of an imaging condition or image transformation which increasingly deteriorates, similar to [55]. As with discriminative power, the conditions we test are blurring, JPEG compression, illumination direction, viewpoint change, illumination colour. The degradation in the recall reflects the constancy of the invariant under examination.
- Information content. We establish the power of each invariant to discriminate between true colour transitions while remaining constant under non-object related transitions induced by shadow, shading, and highlights. Hence, we assess simultaneously for each invariant its power to discriminate between colour transitions, and its invariance to photometric distortions. Note that this is different from the two experiments above, as here we evaluate the property to discriminate between the variant and invariant aspects in the photometric condition, in isolation of a possible effect on recognition performance.

4.1. Experimental setup

We consider for 1000 objects from the ALOI database [25], the following imaging conditions: JPEG compression, blurring, and changes of the viewpoint, illumination direction and illumination colour. Fig. 2 illustrates the imaging conditions for some of the objects.

For each object image, we determine its regions. To be consistent with literature, we determine Harris-affine regions [6]. As pointed out in [1], to establish the correct matching of regions, one should either fix the camera viewpoint, or one should consider the homography limiting oneself to more or less flat scenes. For 3D objects, the assertion of a flat scene fails. To overcome this problem, we consider images that have been recorded with fixed camera viewpoint. However, the condition of viewpoint change has to be settled. Therefore, for each object, we manually selected the single region inside the object which is most consistent between the original and the image recorded under a viewpoint change. We copied the region from the original to all remaining imaging conditions, see Fig. 4 for an example. Note that, as we are dealing with regions inside objects only, the black background does not affect the experiments. Furthermore, trying to find one region from the 1000 selected regions could be seen as searching the one region in an image of 1000 cluttered objects, for which all selected regions are visible. Together with the variation in image transformations and imaging conditions, a total of 26,000 regions are available. The regions vary significantly in size and anisotropy, see Fig. 4a and b, respectively. The ground truth of regions is publicly available on the website of the ALOI database [61].

Next, we compute the invariants from each region. To be consistent with literature, we normalize the regions as in [6]. We consider two experiments:

- Single location computation. In the first experiment, we compute the invariant gradients from one location. We do so by computing them at a fixed scale (i.e. one third of the region size). For each region, we determine the location in which the image gradient $\bar{E}_w$ is maximum. For all copied regions (see for region extraction the description above), this location is identical. From this location, we compute all invariants.
- SIFT-based computation. In the second experiment, we compute the SIFT descriptor from the normalized region identical to Mikolayzcyk's computation [1], but with the grey-value gradient inside the SIFT descriptor replaced by one of the invariant colour gradients. [1]

[1] software available at: http://www.science.uva.nl/~mark

Table 1
Grey-value and colour invariants

Invariant	Gradients	Property	Eq.	Colour-SIFT name
E-grey	$\{E_w\}$	Not photometric invariant	(2)	–
E-colour	$\{E_w,E_{\lambda w},E_{\lambda\lambda w}\}$	Not photometric invariant	(5)	–
W-grey	$\{W_w\}$	Invariant to local intensity level	(3)	(W-colour-) SIFT
W-colour	$\{W_w,W_{\lambda w},W_{\lambda\lambda w}\}$	Invariant to local intensity level	(6)	W-colour- SIFT
C-colour	$\{W_w,C_{\lambda w},C_{\lambda\lambda w}\}$	Invariant to local intensity level, plus invariant to shadow and shading	(7)	C-colour- SIFT
H-colour	$\{W_w,H_w\}$	Invariant to local intensity level, plus invariant to shadow and shading, and highlights	(9)	H-colour- SIFT

Grey-value and colour invariants used in the experiments.

For the performance evaluation, we consider the following sets of invariant gradients, see Table 1. The appendix -SIFT implicates SIFT-based computation, otherwise single location Gaussian invariants are considered. Original SIFT is also included in the experiments and is equivalent to W-grey-SIFT. To ensure results improve in discriminative power with respect to intensity based descriptors—one of our goals in adding colour information—we include the intensity gradient W_w in the H and C colour based descriptors. Although this seems contradictory at first sight, the orthogonalization of intensity and intensity-normalized colour information proofs effective in matching.

For fair comparison to the original SIFT descriptor, we reduce the dimensionality of all colour SIFT descriptors to 128 numbers using PCA reduction (the covariances have been determined over 200 example regions computed from the reference images). Furthermore, we will evaluate the hue-based SIFT descriptor of Abdel-Hakim and Farag [53], termed hue-colour-SIFT, and the HSV-based SIFT descriptor of Bosch and Zisserman [54], termed hsv-colour-SIFT.

4.2. Discriminative power

The objective of this experiment is to establish the distinctiveness of the invariants. To that end, we match image regions computed from a distorted image to regions computed from the reference images as in [1]. The discriminative power is measured by determining the recall of the regions that are to be matched, and the precision of the matches:

$$recall = \frac{\#correct\ matches}{\#correspondences}, \tag{10}$$

$$precision = \frac{\#correct\ matches}{\#correct\ matches + \#false\ matches}. \tag{11}$$

Here, recall indicates the number of correctly matched regions relative to the ground truth of corresponding regions in the dataset. Precision indicates the relative amount of correct matches in all the returned matches. The definition of recall is specific to the problem of matching based on a ground truth of one-to-one correspondences, hence it deviates from the definition as used in information retrieval. The aim in our experiment is to match correctly all regions (recall of one) with ideally no mismatches (precision of one).

We consider the nearest-neighbour matching as employed in [1]. Distances between values of photometric invariants are computed from the Mahalanobis distance (the covariances have been determined over 200 examples computed from reference images). Over various thresholds, the number of correct and false matches are evaluated to obtain a recall vs. precision curve. A good descriptor would produce a small decay in this curve, reflecting the maintenance of a high precision while matching more image regions.

We randomly draw a test set of regions and use 1000-fold cross validation to measure performance over our dataset. However, discriminative power varies between the features. To end up with graphs which allow a comparison between various levels of colour invariance, we vary the number of regions to match per experiment. The number of regions to which a single region is compared is set to 20 for the invariants computed from one location. We consider a successful distinction between 20 image points to be the minimal requirement of a point-based descriptor. For the SIFT-based computation of invariants, we increase this number, as the region-based description is more distinctive. The number of regions to which one region is compared is between 100 or 500, depending on the hardness of the imaging condition. We consider a successful distinction between 100 regions to be the minimal requirement of a region-based descriptor. We consider a successful distinction between 500 regions to be sufficient for realistic computer vision tasks, this is in line with validation in [1,55].

4.2.1. Experimental results: discriminative power

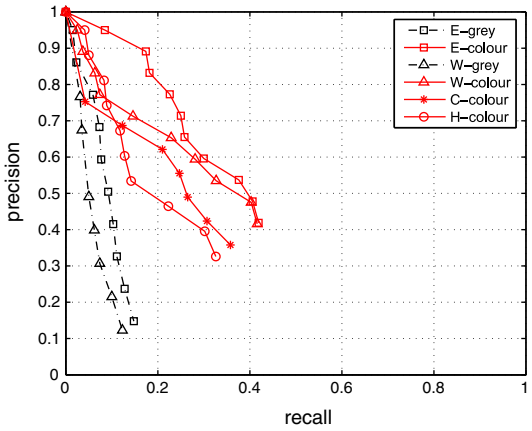
The results of the region matching for invariant gradients are shown in Fig. 5. The organization of all figures is as follows, see also the legends. All photometric invariants are plotted using solid lines. All colour-based invariants are plotted using red lines, opposed to grey-value invariants which are plotted in black lines.

Overall, the performance of H-colour is disappointing and apparently lacks discriminative power. Two effects play a role. First, this descriptor misses one colour channel of information, and better discriminative power could be achieved when adding a saturation channel. However, in that case one would, at best, expect a performance similar to W-colour. We will see a comparison later on when establishing performance for the colour SIFT descriptors. A second effect is the instabilities caused by the normalization in the denominator of Eq. (9). The expression becomes unstable for colours which are unsaturated, hence being greyish. Blurring by the Gaussian filter enhances this effect, as colour at boundaries—which we are evaluating in this setup—are mixed. Hence, H-colour seems unsuitable for region descriptors based on Gaussian derivatives.

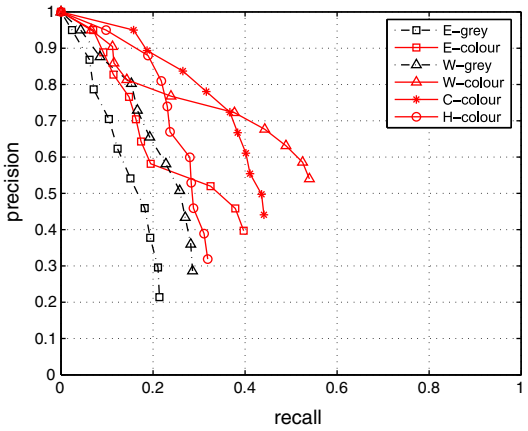
Furthermore, grey-value derivatives E-grey and W-grey are outperformed by colour based descriptors, except when illumination colour is changed (Fig. 5e). In that case, normalized intensity W-grey performs reasonable, but is still outperformed by many colour based invariants.

In detail, the effect of blurring, shown in Fig. 5a, causes the image values to be smoothed. Hence, details are lost, but no photometric variation is introduced. The colour gradient with no photometric invariant properties, E-colour, performs best. Besides the decay in performance due to additional blur, the graph clearly illustrates the gain in discriminative power when using colour information.

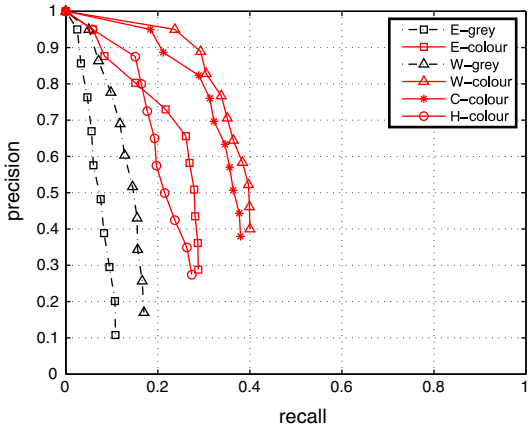
The compression of images by JPEG, shown in Fig. 5b, causes the colour values to be distorted more than the intensity channel. Still, colour information is distinctive, as the colour gradient that is invariant to the intensity level, W-colour, performs best. At the beginning of the recall–precision curves, one clearly sees the advantage of orthogonalizing intensity and colour information, as W-colour, C-colour, and H-colour perform significantly better than E-colour, for which all channels are correlated with intensity. In the latter case, all values of the SIFT descriptor will be



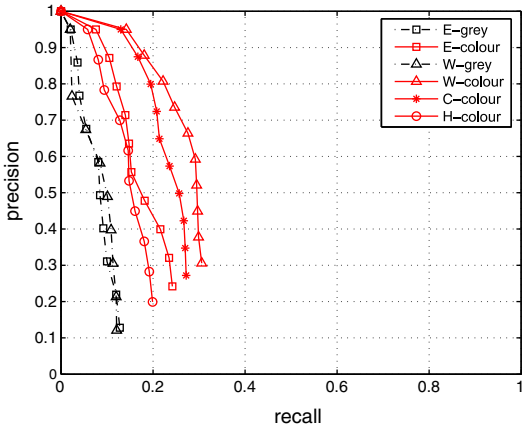
(a) Blurring ($\sigma = 1$ pixel), 1 vs 20



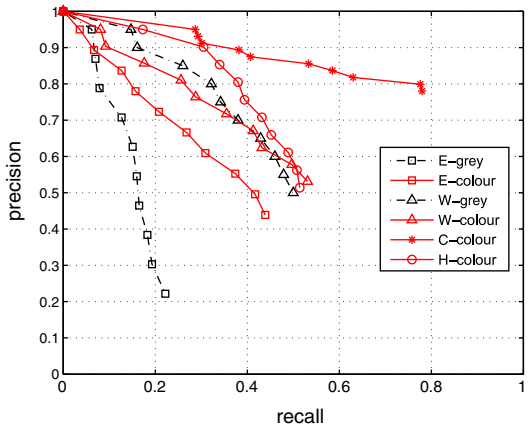
(b) JPEG compression (50%), 1 vs 20



(c) Illumination direction (30°), 1 vs 20



(d) Viewpoint change (30°), 1 vs 20



(e) Illumination colour (2100K), 1 vs 20

Fig. 5. Discriminative power of photometric invariant gradients.

severely corrupted by the JPEG compression. For the invariant colour descriptors, the intensity channel will be relatively mildly corrupted by the compression, whereas the colour channels still add extra discriminative power. Compression effects become more influential at the tail of the recall-precision curves, where one sees H-colour to drop off quit early due to instability of the descriptor, followed by `C-colour`. Although `W-colour` had a slower start, it ends up doing quite well due to the more stable calculation of the non-linear derivative combination.

For changes of the illumination direction, Fig. 5c, the main imaging effects are darker and lighter image patches, and shadow and shading changes. However, for the small scale at which we measure the Gaussian derivative descriptors, we expect intensity changes to dominate over shadow and shading edges. Shadow and shading (geometry) edges are expected to become more important when assessing SIFT based descriptors, which capture information over a much larger region. Hence, both colour gradients that are invariant to intensity changes, `W-colour` and

C-colour, are performing well. Clearly, the colour invariant descriptors outperform grey-value descriptors and non-invariant colour descriptors.

The results of a change in viewpoint, Fig. 5d, clearly demonstrate the advantage of adding colour information. The patches, manually indicated to be stable, merely contain a change in information content due to an projective transformation and small errors in the affine region detection. Furthermore, the light field will be distributed somewhat different over the image, causing W-colour and C-colour to perform superior over grey-value descriptors, non-invariant colour descriptors, and the H-colour descriptor.

For varying illumination colour, Fig. 5e, obviously the colour values become distorted. The colour gradient invariant to shadow, C-colour, shows to be very robust here. Although C-colour is based on colour, its gradients are computed in such a way that can be shown to be reasonably colour constant [25]. Furthermore, one would expect the grey-value descriptors not to be affected by illumination colour changes. However, a change in overall intensity is also present, making direct use of E-grey infeasible. The intensity normalized invariant W-grey performs reasonable, but lacks the discriminative power which comes with the use of colour.

4.2.2. Experimental results: discriminative power for colour-SIFT descriptors

Fig. 6 shows the discriminative power of the invariants when they are plugged into the SIFT descriptor. The figure has an identical organization as Fig. 5. The only exception in the experimental setup is that the number of regions, to which a single region is matched, is increased. This number varies over the imaging conditions, and is either 100 or 500, to obtain suitable resolution in the performance graphs. Furthermore, note that two extra methods from literature have been added, being the hue-colour-SIFT descriptor [53], and the hsv-colour-SIFT descriptor [54].

Overall, the relative performance of SIFT-based computation of invariants corresponds largely to relative performance of invariants from single points. Colour-based SIFT invariant to shadow and shading effects, C-colour-SIFT, performs best.

Generally, the SIFT-based computation improves significantly the discriminative power compared to single-point computation. Almost all colour and grey-value descriptors perform well under blurring (Fig. 6a), JPEG compression (Fig. 6b), and illumination colour changes (Fig. 6e). Note that the C-colour-SIFT descriptor performs equally well as the intensity based SIFT descriptor in the last case, implying a high degree of colour constancy for this descriptor.

Discriminative power drops when considering illumination direction or viewpoint changes, see Fig. 6a and b. These cases are much harder to distinguish using a SIFT descriptor. In these cases, the grey-value based SIFT is outperformed by the colour-based SIFT descriptors. In particular, the colour-based SIFT invariant to shadow and shading effects, C-colour-SIFT, is very discriminative in these cases. This can be explained by the large spatial area over which the SIFT descriptor captures image structure. Hence, shadow and shading (object geometry) effects are more likely to be captured by the SIFT descriptor, but the effects being cancelled by the *C* invariant.

The shadow and highlight invariant H-colour-SIFT is generally not very distinctive compared to W-colour-SIFT and C-colour-SIFT. Lack of discriminative power affects the performance for hue-colour-SIFT, H-colour-SIFT, and SIFT under blurring. Furthermore, the hue-based descriptors hue-colour-SIFT and H-colour-SIFT are affected by JPEG compression, and by illumination colour changes. The distinctiveness of hue-colour-SIFT is generally much less than of H-colour-SIFT. Hence, using the hue alone is not a distinctive region property. The distinctiveness of hsv-colour-SIFT is generally somewhat higher than of H-colour-SIFT. Thus, the saturation *s* in the *hsv* colour space is a distinctive property. But, the distinctiveness of hsv-colour-SIFT is generally less than of W-colour-SIFT and C-colour-SIFT, due to instability as argued before.

4.3. Invariance

The objective of this experiment is to establish the constancy of the invariants against varying imaging conditions. Likewise [55], we measure the degradation of recall (Eq. (11)) over increasingly hard imaging conditions. The experimental setup is identical to the previous experiment. The aim in this experiment is to minimize the degradation over more distorted images.

4.3.1. Experimental results: invariance

The results of the region matching over increasingly hard imaging conditions is shown in Fig. 7. The organization of the figure is identical to Figs. 5 and 6. The present graphs are orthogonal to Figs. 5 and 6, in that now the amount of degradation is varied, at a fixed recall which corresponds to the end-point of the curves in Figs. 5 and 6. Any decline in performance indicates lack of constancy with respect to the tested condition. Ideally, the decline would be zero (horizontal line), indicating perfect invariance to the set of imaging conditions.

For image blurring, Fig. 7a, no significant imaging effects are observed. Hence, all descriptors have equal performance with respect to constancy, although initial discriminative power varies from a recall of 0.2 for grey-value derivatives to more than 0.7 for colour based derivatives. For JPEG compression, Fig. 7b, the grey-value invariants I-grey and W-grey are slightly more constant than the colour invariants, as the image intensity is less affected by JPEG compression than the image chromaticity. For changes in the illumination direction, Fig. 7c, due to the small scale of the derivative descriptors, the main imaging effect is the change of region intensity. Hence, W-grey, W-colour, C-colour and H-colour are very stable. For a viewpoint change, Fig. 7d, only marginal imaging effects are observed. Hence, all measures perform equally well with respect to constancy. For varying illumination colour (e), besides the intensity based measures E-grey and W-grey, C-colour is very invariant. This measure has theoretically been shown to be reasonably colour constant [25].

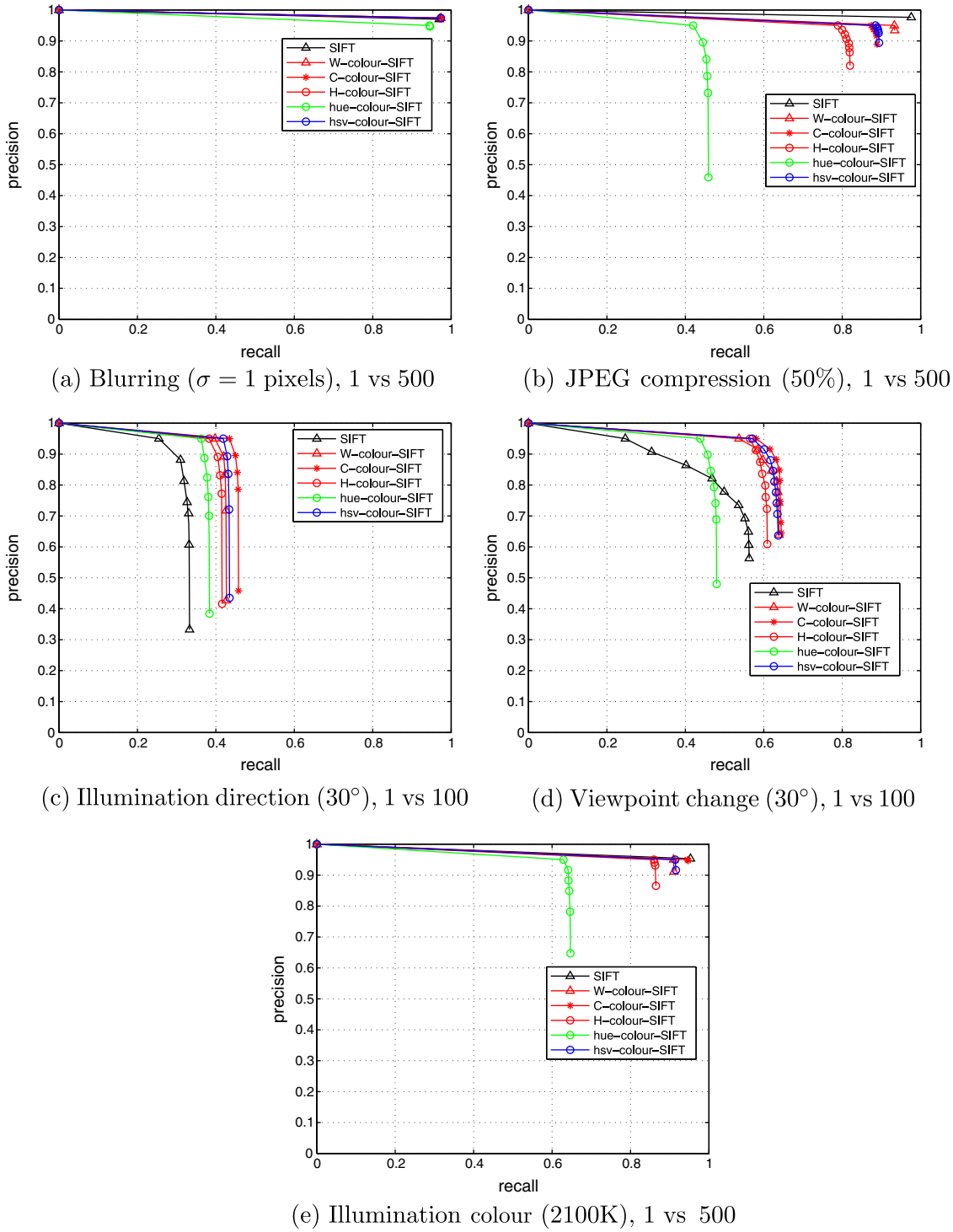
4.3.2. Experimental results: invariance for colour-SIFT descriptors

We repeat the invariance experiment but now the invariants are plugged into the SIFT descriptor. The results are shown in Fig. 8.

Overall, most descriptors are performing well for blurring (Fig. 8a), JPEG compression (Fig. 8b), and illumination colour change (Fig. 8e). Exceptions again are the hue based descriptors H-colour-SIFT and hue-colour-SIFT, which lack discriminative power, and are more affected by these conditions. A change in illumination direction or viewpoint is much harder for the SIFT descriptor to deal with, even with colour invariance build in. Overall, the C-colour-SIFT seems the best choice, for which shadow and shading edges are discounted. This descriptor has invariance comparable to the intensity based SIFT descriptor, but gains considerably in discriminative power.

4.4. Information content

The objective of this final experiment is to establish the information content of the photometric invariants. Information content refers to the ability of an invariant to distinguish between colour transitions and photometric events such as shadow, shading and highlights. Ideally, the invariant's values covaries with colour tran-



(a) Blurring ($\sigma = 1$ pixels), 1 vs 500

(b) JPEG compression (50%), 1 vs 500

(c) Illumination direction (30°), 1 vs 100

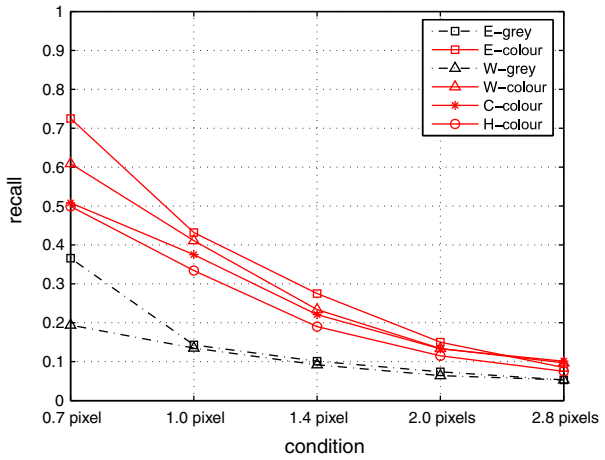
(d) Viewpoint change (30°), 1 vs 100

(e) Illumination colour (2100K), 1 vs 500

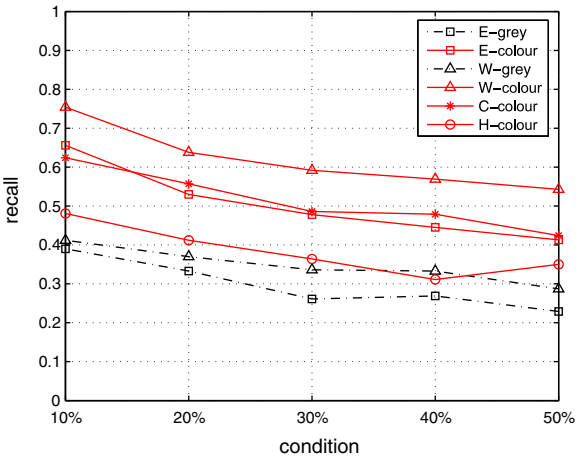
Fig. 6. Discriminative power of photometric invariant gradients, when plugged into the SIFT descriptor.

sitions and it's value is constant to photometric events to which it is designed to be invariant. We illustrate the information content of `W-colour` and `C-colour`, see Fig. 9. For the first object, new image edges are introduced by changing the illumination direction in Fig. 9b and c. Hence, the matching is better with the shadow and shading invariant descriptor `C-colour-SIFT`. Fig. 9e and f show an example where no shadow/shading invariance performs better. Here, no new edges are introduced by the change in illumination direction, and only the local intensity is affected due to relatively large-scale shading effects.

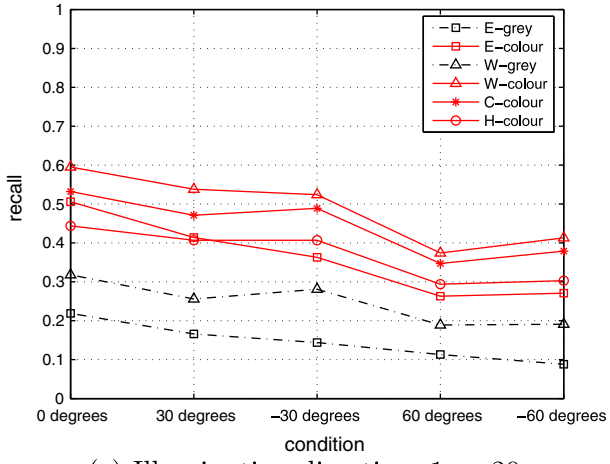
To establish the information content, we measure the discriminative power and invariance over individual image regions. Each image region is labelled whether it contains a colour transition, or a shadow, shading or highlight transition. In this way, the information content evaluates the invariant's discriminative power and invariance over various photometric events. To that end, we construct a large annotated dataset. This dataset contains tens of images with in the order of hundreds of labelled image points located at the various photometric events. The images are selected from the CURET dataset [62]. The selected texture images contain



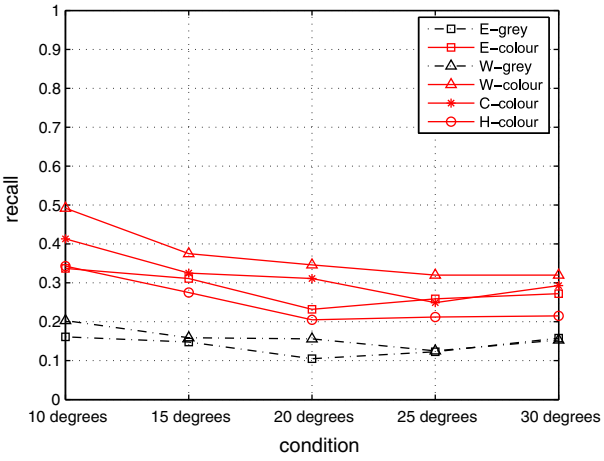
(a) Blurring, 1 vs 20



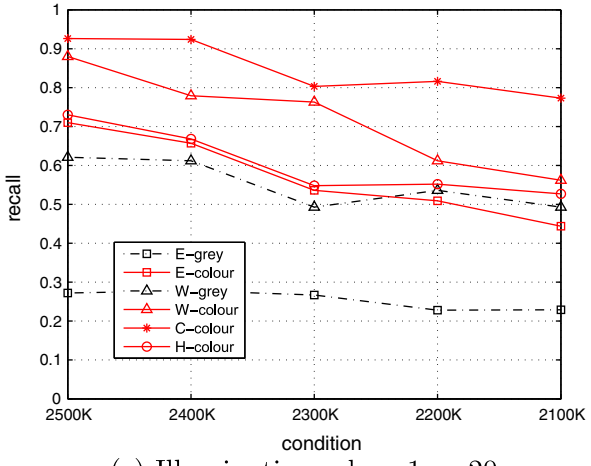
(b) JPEG compression, 1 vs 20



(c) Illumination direction, 1 vs 20



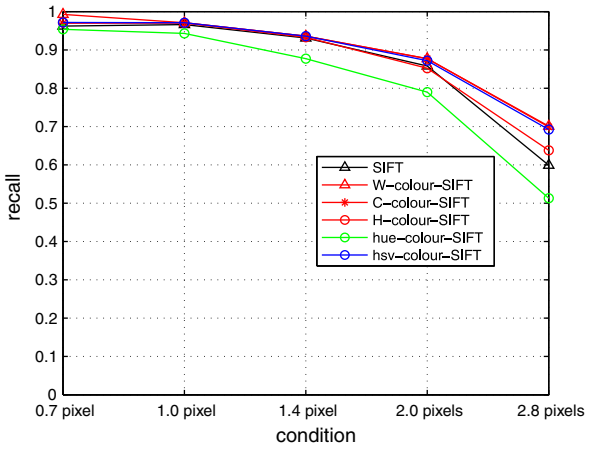
(d) Viewpoint change, 1 vs 20



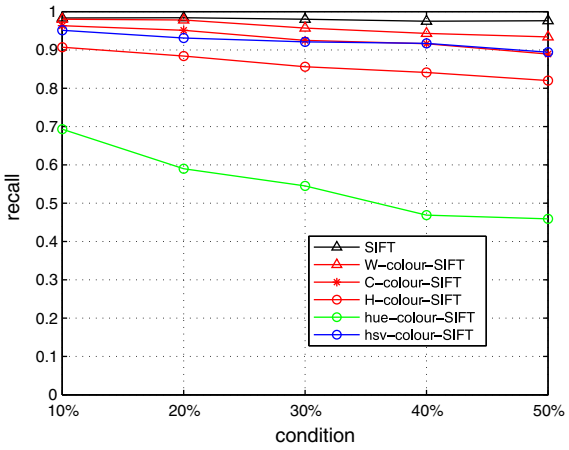
(e) Illumination colour, 1 vs 20

Fig. 7. Invariance of photometric invariant gradients over increasingly hard imaging conditions.

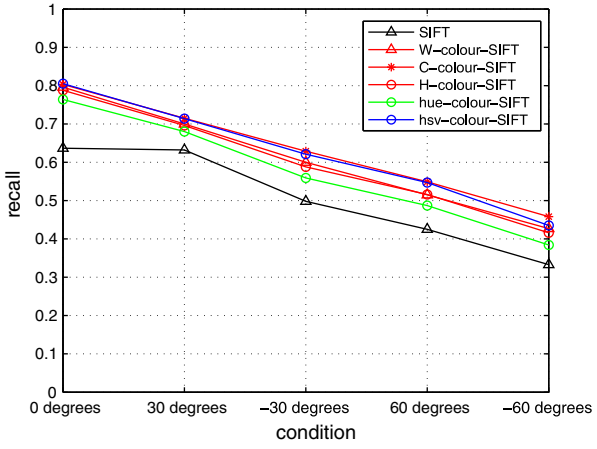
many edges, where we annotated for each image whether the texture was generated mainly by either shadow/shading (sponge, cracker b, lambswool, quarry tile, wood b, and rabbit fur) or highlight effects (aluminium foil, rug a, and styrofoam). From these images, regions have been detected by applying a Harris corner detector [11]. Fig 10a and b illustrate, for two fragments of texture images, shadow/shading and highlight edges, respectively. In addition, we have collected image points located at colour transitions. To that end, images have been taken from PANTONE colour patches [63], see Fig. 10c for an illustration. From the PANTONE patch combinations, we have selected the 100 combinations that have the largest hue difference, hence selecting patches which reflect true changes in object colour rather than intensity or saturation differences.



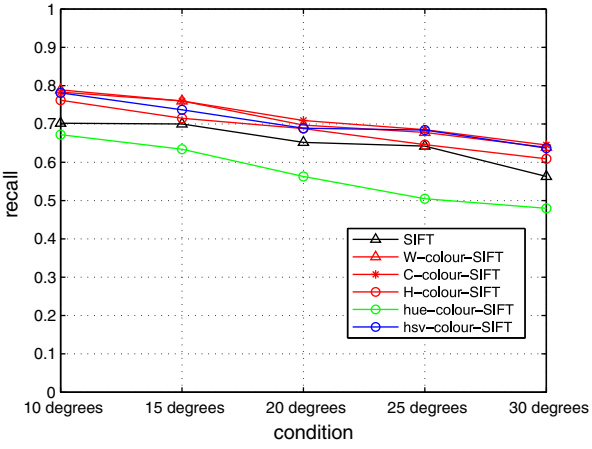
(a) Blurring, 1 vs 500



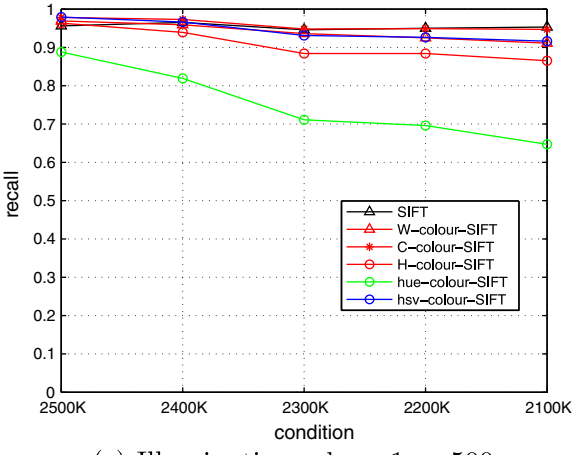
(b) JPEG compression, 1 vs 500



(c) Illumination direction, 1 vs 100



(d) Viewpoint change, 1 vs 100



(e) Illumination colour, 1 vs 500

Fig. 8. Invariance of photometric invariant gradients over increasingly hard imaging conditions, when plugged into the SIFT descriptor.

We measure an invariant's power to distinguish between colour transitions and disturbing photometric events by the Fisher criterion. From many colour transitions, we compute a first cloud of points; from transitions of a particular disturbing photometric event, we compute a second point cloud. The Fisher criterion expresses the separation between the two clouds of points, termed $\{x_1\}$ and $\{x_2\}$, respectively:

$$\textit{information} = \frac{|\mu(\{x_1\}) - \mu(\{x_2\})|^2}{\sigma^2(\{x_1\}) + \sigma^2(\{x_2\})}. \tag{12}$$

4.4.1. Experimental results: information content

The values of photometric invariants to various photometric events are shown in Fig. 11. The plots show values relative to the

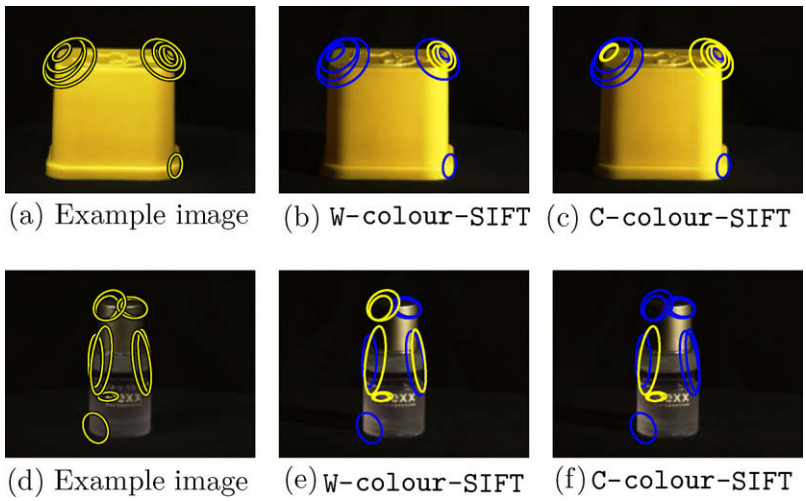


Fig. 9. Illustration of matching for two objects. One is better matched with `C-colour-SIFT`, the other with `W-colour-SIFT`, respectively. Correct matches are shown in yellow, false matches are shown in blue.

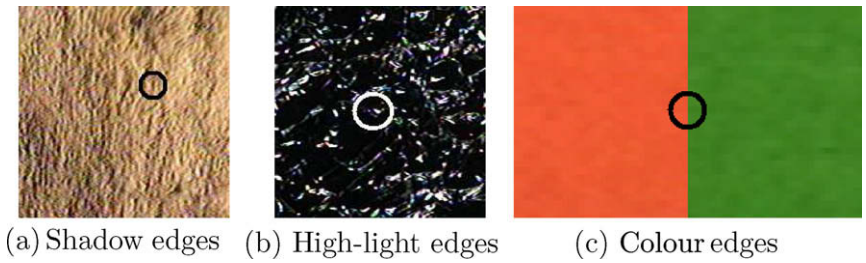


Fig. 10. Examples of the photometric events dataset. Detected points are given a label whether the point is located on a (a) shadow/shading edge, (b) highlight edge, or (c) colour edge.

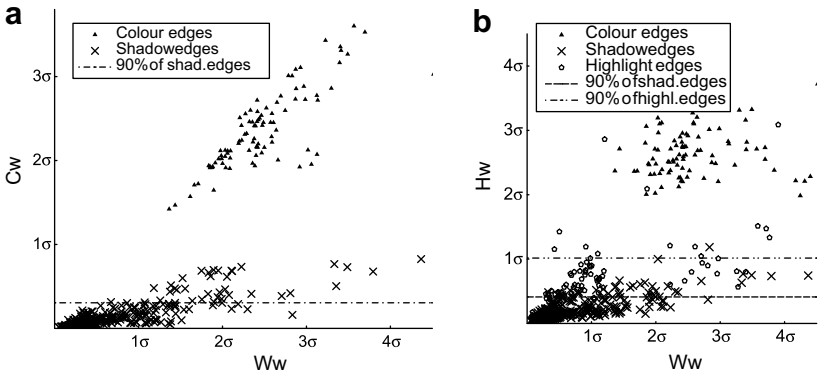


Fig. 11. Scatter plots of invariant values to photometric events. The figures depict (a) $\overline{C}_w$ vs. $\overline{W}_w$ and (b) $\overline{H}_w$ vs. $\overline{W}_w$. All invariants are sensitive to colour edges. $\overline{C}_w$ and $\overline{H}_w$ are invariant to shadow and shading, where $\overline{H}_w$ is additionally invariant to highlights. The horizontal lines describe a 90% interval of the invariant values. This gives an indication of the invariant's ability to distinguish between values to colour edges and to disturbing photometric events.

total colour edge strength $\overline{W}_w$. We do so, to express simultaneously the power of $\overline{W}_w$ and of the shadow and shading invariants $\overline{C}_w$ and $\overline{H}_w$ to distinguish between photometric events and true colour edges. As expected, the values of the invariants $\overline{C}_w$ and $\overline{H}_w$ are close to zero for shadow/shading edges (note that values of the reference invariant $\overline{W}_w$ are indeed significant to shadow/shading edges). For shadow/shading disturbances, we obtain $\mathit{information}(\overline{C}_w) = 2.6$, and $\mathit{information}(\overline{H}_w) = 4.9$. Thus, the invariant $\overline{H}_w$ separates shadow/shading from object transitions much better than $\overline{C}_w$. Furthermore, the value of $\overline{H}_w$ is also low for highlights, see Fig. 11b. However, as expected, not all of the values are close to zero due to pixel saturations at highlights. As a result, the invariance and the information content of $\overline{H}_w$ are somewhat lower for highlight disturbances than for shadow/shading disturbances, $\mathit{information}(\overline{H}_w) = 2.9$.

Overall, the photometric invariant `H-colour` is more constant to shadow and shading than `C-colour`. Both perform well when separating colour transitions from shadow and shading transitions. The separation of colour transitions and highlights by `H-colour` is harder due to saturated highlights. As a consequence, most of the highlights are separated well, but some highlights are misclassified as colour transitions.

5. Evaluation on the PASCAL-VOC 2006 dataset

In this final experiment, we evaluate the performance of the colour-SIFT descriptors on the VOC dataset [32] containing 10 categories of natural and man-made objects in realistic settings. As an experimental framework, we consider the bag-of-feature approach, see e.g. [45]. We outline the approach shortly. Images are encoded by vector quantizing the appearance space by mapping descriptor vectors obtained from the image onto a codebook. The codebook contains descriptor vectors that are representative of the dataset. A common scheme is to construct the codebook by storing the cluster centres obtained from *k*-means clustering [5,64]. We create codebook representations according to the method of Perronnin et al. [65]. They have proposed a distinctive histogram representation that is tuned to the categories to be classified. The codebook is constructed by clustering 50,000 descriptor vectors into 256 cluster centres.

It is important to notice that we deviate from [65] only in that we do not obtain cluster centres from Gaussian-mixture modelling, but from *k*-means. We do so for reason of speed, and also to prevent reduction of the dimensionality of the descriptors to 50 as done in [65]. As a consequence of the different clustering, a lower performance is achieved with our implementation than reported in [32]. Even though the performance may be less, our main point here is a relative performance of the grey and colour-based SIFT descriptors.

The VOC dataset consists of a training, validation and testing set. We prefer the *k*-nearest neighbour classifier as it performs best (tested among the linear SVM, nearest mean, Fisher and logistic regression classifiers). Optimal *k* is determined from performance on the validation set. The performance for the `SIFT` and `C-colour-SIFT` descriptors is determined from the test set. The objective is to compare qualitatively the performance of the `SIFT` and `C-colour-SIFT` descriptors within a successful bag-of-feature approach.

The performance of the `SIFT` and `C-colour-SIFT` descriptors for codebook-based classification is depicted in Fig. 12. As a classification performance measure, we consider the area under the curve (auc). For the cat, car and horse categories, the classification accuracy of `SIFT` and `C-colour-SIFT` is similar, while for one category (cows) the performance of `C-colour-SIFT` is somewhat less than of `SIFT` (3%). For the other categories, `C-colour-SIFT` classifies the images significantly better than does `SIFT`, up to approximately 10% improvement for the bike, bus and sheep categories.

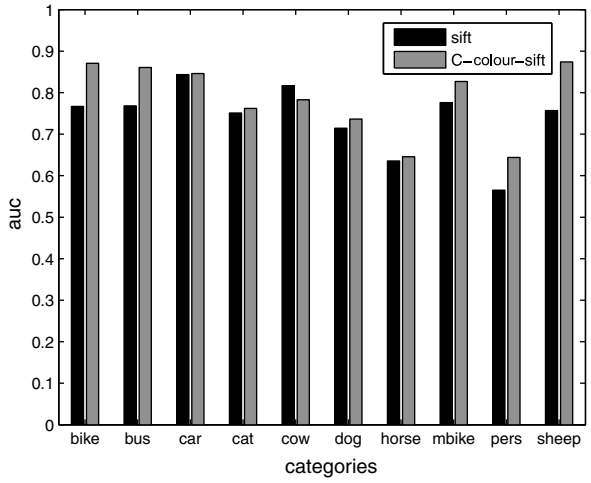


Fig. 12. VOC classification results obtained with gray (`SIFT`) and colour-SIFT (`C-colour-SIFT`) descriptors.

The colour SIFT methods, like original SIFT, are based on edge detection, in this case, chromatic edge detection. Hence, the colour itself does not play a significant role, only colour contrast is what is emphasized. The improved results are due to a better discrimination between coloured patches, and to the increased invariance to shading and shadow effects. The results on this dataset are in agreement with the recent alternative study by Van de Sande et al. [58]. We conclude that for this realistic categorization task, the `C-colour-SIFT` descriptor is the preferred choice over the traditional `SIFT` descriptor.

6. Conclusions

In this paper, we have presented an experimental evaluation of local colour invariants in the presence of realistic geometric transformations and photometric changes. The goal was to compare local invariants computed on regions from 3D objects. The evaluation was designed to assess performance of local invariants, which can be directly plugged into many of the descriptors that are available from literature. The setup is to evaluate of each invariant its distinctiveness, invariance, and information content. The evaluation protocol, together with test data and ground-truth, is available from the internet, allowing evaluation and comparison of future colour descriptors.

We have considered the grey-value based gradient `I-grey`. The grey-value photometric invariant `W-grey` is derived from `I-grey` by locally normalizing it by the image intensity. We have considered their extensions to colour, yielding `I-colour` and `W-colour`. Further, we have taken into account more advanced photometric invariants, being the shadow and shading invariant `C-colour`, and the shadow, shading and highlight invariant `H-colour`.

Our experimental evaluation showed the most distinctive colour invariant to be `C-colour`, which is designed to be constant to changes in illumination conditions, and to the geometry of the object. That is, shadow and shading effects are ignored. Furthermore, the invariant is reasonably colour constant. Our experiments showed the descriptor to outperform alternatives with respect to discriminative power, while being more constant to illumination direction, viewpoint, and illumination colour changes. Hence, the `C-colour` based invariant is applicable in many computer vision tasks.

We have plugged the local invariants into the SIFT descriptor. Our experiments showed the `C-colour-SIFT` based descriptor to outperform the traditional intensity based SIFT, due to it's significant increase in discriminative power, while being equally constant to the tested conditions as traditional SIFT. Furthermore, `C-colour-SIFT` outperforms hue-based SIFT [53] and HSV-based SIFT [54] proposed in literature. The usefulness of `C-colour-SIFT` for realistic computer vision applications is illustrated for the classification of object categories from the VOC challenge [32], for which a significant improvement is reported.

References

[1] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (10) (2005) 1615–1630.
[2] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.
[3] S. Odbrzalek, J. Matas, Object recognition using local affine frames on distinguished regions, in: Proceedings of the Brittish Machine Vision Conference, 2002.
[4] F. Rothganger, S. Lazebnik, C. Schmid, J. Ponce, 3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints, International Journal of Computer Vision 66 (2006) 231–259.

[5] J. Sivic, A. Zisserman, Video google: a text retrieval approach to object matching in videos, in: Proceedings of the International Conference on Computer Vision, 2003.
[6] K. Mikolajczyk, C. Schmid, Scale and affine invariant interest point detectors, International Journal of Computer Vision 60 (1) (2004) 63–86.
[7] S. Lazebnik, C. Schmid, J. Ponce, A sparse texture representation using local affine regions, IEEE Transactions on Pattern Analysis & Machine Intelligence 27 (8) (2005) 1265–1278.
[8] L.M.J. Florack, B. Haar Romeny, M. Viergever, J.J. Koenderink, The gaussian scale-space paradigm and the multiscale local jet, International Journal of Computer Vision 18 (1996) 61–75.
[9] W.T. Freeman, E.H. Adelson, The design and use of steerable filters, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (9) (1991) 891–906.
[10] T. Lindeberg, Feature detection with automatic scale selection, International Journal of Computer Vision 30 (2) (1998) 117–154.
[11] C. Harris, M. Stephans, A combined corner and edge detector, in: Proceedings of the 4th Alvey Vision Conference, Manchester, 1988, pp. 189–192.
[12] C. Schmid, R. Mohr, Local grayvalue invariants for image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (5) (1997) 530–535.
[13] B. Schiele, J.L. Crowley, Recognition without correspondence using multidimensional receptive field histograms, International Journal of Computer Vision 36 (1) (2000) 31–50.
[14] A. Ferencz, E. Learned-Miller, J. Malik, Building a classification cascade for visual identification from one example, in: Proceedings of the International Conference Computer Vision, IEEE Computer Society, 2003, pp. 286–293.
[15] M. Swain, D. Ballard, Color indexing, International Journal of Computer Vision 7 (1) (1991) 11–32.
[16] B.V. Funt, G.D. Finlayson, Color constant color indexing, IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (5) (1995) 522–529.
[17] T. Gevers, A.W.M. Smeulders, Color based object recognition, Pattern Recognition 32 (1999) 453–464.
[18] P. Montesinos, V. Gouet, R. Deriche, D. Pelé, Matching color uncalibrated images using differential invariants, Image and Vision Computing 18 (9) (2000) 659–671.
[19] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1349–1380.
[20] C. Carson, S. Belongie, H. Greenspan, J. Malik, Blobworld: image segmentation using expectation-maximization and its application to image querying, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (8) (2002) 1026–1038.
[21] G.D. Finlayson, M.S. Drew, B. Funt, Color constancy: generalized diagonal transforms suffice, Journal of the Optical Society of America A 11 (11) (1994) 3011–3019.
[22] G.D. Finlayson, Color in perspective, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (10) (1996) 1034–1038.
[23] G.D. Finlayson, S.D. Hordley, P.M. Hubel, Color by correlation: a simple, unifying framework for color constancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (11) (2001) 1209–1221.
[24] J. Weijer, T. Gevers, J.-M. Geusebroek, Color edge and corner detection by photometric quasi-invariants, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (4) (2005) 325–630.
[25] J.M. Geusebroek, R. Boomgaard, A.W.M. Smeulders, H. Geerts, Color invariance, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (12) (2001) 1338–1350.
[26] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, in: Proceeding of the British Machine Vision Conference, 2002, pp. 384–393.
[27] T. Lindeberg, J. Garding, Shape-adapted smoothing in estimation of 3D shape cues from affine deformations of local 2D brightness structure, Image and Vision Computing 15 (6) (1997) 415–434.
[28] T. Kadir, M. Brady, Scale, saliency and image description, International Journal on Computer Vision 2 (45) (2001) 83–105.
[29] T. Tuytelaars, L. Van Gool, Matching widely separated views based on affine invariant regions, International Journal of Computer Vision 59 (1) (2004) 61–85.
[30] G. Heidemann, Focus-of-attention from local color symmetries, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (7) (2004) 817–830.
[31] J. Weijer, T. Gevers, Boosting color saliency in image feature detection, IEEE Transactions on Pattern Analysis and Machine Intelligence.
[32] A. Zisserman, M. Everingham, C. Williams, L. Van Gool, The pascal visual object classes challenge 2006 (voc2006), 2006.
[33] A. Johnson, M. Hebert, Object recognition by matching oriented points, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1997, pp. 684–689.
[34] R. Zabih, J. Woodfill, Non-parametric local transforms for computing visual correspondance, in: Proceedings of the European Conference on Computer Vision, 1994, pp. 151–158.
[35] J.J. Koenderink, The structure of images, Biological Cybernetics 50 (1984) 363–370.
[36] R. Basri, D.W. Jacobs, Recognition using region correspondences, International Journal of Computer Vision 25 (2) (1997) 145–166.
[37] A. Baumberg, Reliable feature matching ocross widely separated views, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2000, pp. 774–781.
[38] F. Schaffalitzky, A. Zisserman, Multi-view matching for unordered image sets, in: Proceedings of the European Conference on Computer Vision, 2002, pp. 414–131.
[39] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 4 (2002) 509–522.
[40] M. Varma, A. Zisserman, A statistical approach to texture classification from single images, International Journal of Computer Vision 62 (1-2) (2005) 61–81.
[41] J. Winn, A. Criminisi, T. Minka, Object categorization by learned universal visual dictionary, in: Proceedings of the International Conference Computer Vision, IEEE Computer Society, 2005, pp. 1800–1807.
[42] Y. Ke, R. Sukthankar, Pca-sift: a more distinctive representation for local image descriptors, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2005.
[43] M. Grabner, H. Grabner, H. Bischof, Fast approximated sift, in: Proceedings of the Asian Conference Computer Vision, 1 (2006) 918–927.
[44] H. Bay, T. Tuytelaars, L. Van Gool, Surf: speeded up robust features, in: Proceedings of the European Conference on Computer Vision, 2006.
[45] F. Jurie, B. Triggs, Creating efficient codebooks for visual recognition, in: Proceedings of International Conference on Computer Vision, 2005.
[46] R. Gershon, D. Jepson, J.K. Tsotsos, Ambient illumination and the determination of material changes, Journal of the Optical Society of America A 3 (1986) 1700–1707.
[47] D. Slater, G. Healey, The illumination-invariant recognition of 3D objects using local color invariants, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (2) (1996) 206–210.
[48] S.A. Shafer, Using color to separate reflection components, Color Research and Application 10 (4) (1985) 210–218.
[49] T. Gevers, H.M.G. Stokman, Classification of color edges in video into shadow-geometry, highlight, or material transitions, IEEE Transactions on Multimedia 5 (2) (2003) 237–243.
[50] F. Mindru, T. Tuytelaars, L. Van Gool, T. Moons, Moment invariants for recognition under changing viewpoint and illumination, Computer Vision and Image Understanding 94 (2004) 3–27.
[51] J. Weijer, C. Schmid, Coloring local feature extraction, in: European Conference on Computer Vision, Springer, 2006, pp. 334–348.
[52] T. Geodeme, T. Tuytelaars, G. Vanacker, M. Nuttin, L. Van Gool, Omnidirectional sparse visual path following with occlusion-robust feature tracking, in: Proceedings of the International Conference on Computer Vision, 2005.
[53] A.E. Abdel-Hakim, A.A. Farag, Csift: a sift descriptor with color invariant characteristics, in: Proceedings of the Computer Vision and Pattern Recognition, 2006, pp. 1978–1983.
[54] A. Bosch, A. Zisserman, X. Munoz, Scene classification via pLSA, in: Proceedings of the European Conference on Computer Vision, 2006.
[55] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. Van Gool, A comparison of affine region detectors, International Journal of Computer Vision 65 (1/2) (2005) 43–72.
[56] P. Moreels, P. Perona, Evaluation of features detectors and descriptors based on 3D objects, in: Proceedings of the International Conference on Computer Vision, vol. 1, 2005, pp. 800–807.
[57] F. Fraundorfer, H. Bischof, A novel performance evaluation method of local detectors on non-planar scenes, in: Proc. Workshop Empirical Evaluation Methods in Computer Vision (with CVPR), 2005.
[58] K.E.A. van de Sande, T. Gevers, C.G.M. Snoek, Evaluation of color descriptors for object and scene recognition, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2008.
[59] C.G.M. Snoek, M. Worring, J.C. van Gemert, J.M. Geusebroek, A.W.M. Smeulders, The challenge problem for automated detection of 101 semantic concepts in multimedia, in: Proceedings of the ACM International Conference on Multimedia, Santa Barbara, USA, 2006, pp. 421–430.
[60] J.M. Geusebroek, G.J. Burghouts, A.W.M. Smeulders, The Amsterdam library of object images, International Journal of Computer Vision 61 (1) (2005) 103–112.
[61] J.M. Geusebroek, G.J. Burghouts, A.W.M. Smeulders, Amsterdam Library of Object Images (ALOI). Available from: http://www.science.uva.nl/~aloi.
[62] K.J. Dana, B. Ginneken, S.K. Nayar, J.J. Koenderink, Reflectance and texture of real world surfaces, ACM Transactions on Graphics 18 (1) (1999) 1–34.
[63] PANTONE, ed. 1992–1993, Group Basf, Paris, France, Pantone is a trademark of Patone inc.
[64] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: Proceedings of the European Conference on Computer Vision, 2004.
[65] F. Perronnin, C. Dance, G. Csurka, M. Bressan, Adapted vocabularies for generic visual categorization, in: Proceedings of the European Conference on Computer Vision, Springer Verlag, 2006.