Evaluating Color Descriptors for Object and Scene Recognition

Koen E. A. van de Sande, *Student Member, IEEE,* Theo Gevers, *Member, IEEE,*
and Cees G. M. Snoek, *Member, IEEE*

***Abstract*—Image category recognition is important to access visual information on the level of objects and scene types. So far, intensity-based descriptors have been widely used for feature extraction at salient points. To increase illumination invariance and discriminative power, color descriptors have been proposed. Because many different descriptors exist, a structured overview is required of color invariant descriptors in the context of image category recognition.**

Therefore, this paper studies the invariance properties and the distinctiveness of color descriptors[1] in a structured way. The analytical invariance properties of color descriptors are explored, using a taxonomy based on invariance properties with respect to photometric transformations, and tested experimentally using a dataset with known illumination conditions. In addition, the distinctiveness of color descriptors is assessed experimentally using two benchmarks, one from the image domain and one from the video domain.

From the theoretical and experimental results, it can be derived that invariance to light intensity changes and light color changes affects category recognition. The results reveal further that, for light intensity shifts, the usefulness of invariance is category-specific. Overall, when choosing a single descriptor and no prior knowledge about the dataset and object and scene categories is available, the OpponentSIFT is recommended. Furthermore, a combined set of color descriptors outperforms intensity-based SIFT and improves category recognition by 8% on the PASCAL VOC 2007 and by 7% on the Mediamill Challenge.

***Index Terms*—Image/video retrieval, evaluation/methodology, color, invariants, pattern recognition.**

I. Introduction

Image category recognition is important to access visual information on the level of objects (buildings, cars, *etc*.) and scene types (outdoor, vegetation, *etc*.). In general, systems for category recognition on images [1], [2], [3], [4], [5] and video [6], [7], [8] use machine learning based on image descriptions to distinguish object and scene categories. However, there can be large variations in viewing and lighting conditions for real-world scenes, complicating the description of images and consequently the image category recognition task. This is illustrated in figure 1. A change in *viewpoint* will yield shape variations such as the orientation and scale of the object. Salient point detection methods and corresponding region descriptors can robustly detect regions which are translation-, rotation- and scale-invariant, addressing these viewpoint changes [9], [10], [11]. In addition, changes in the *illumination* of a scene can greatly affect the performance of object and scene type recognition if the descriptors used are not robust to these changes. To increase photometric invariance and discriminative power, color descriptors have been proposed which are robust against certain photometric changes [12], [13], [14], [15], [16]. As there are many different methods to obtain color descriptors, however, it is unclear what similarities these methods have and how they are different. To arrange color invariant descriptors in the context of image category recognition, a taxonomy is required based on principles of photometric changes.

Therefore, this paper studies the *invariance* properties and the *distinctiveness* of color descriptors in a structured way. First, a taxonomy of invariant properties is presented. The taxonomy is derived by considering the diagonal model of illumination change [17], [18], [19]. Using this model, a systematic approach is adopted to provide a set of invariance properties which achieve different amounts of *invariance*, such as invariance to light intensity changes, light intensity shifts, light color changes and light color changes and shifts. Color descriptors are tested experimentally with respect to this set of invariance properties through an object recognition dataset with known illumination changes [20]. Then, the *distinctiveness* of color descriptors is analyzed experimentally using two benchmarks from the image domain [21] and the video domain [22]. The benchmarks are very different in nature: the image benchmark consists of photographs and the video benchmark consists of keyframes from broadcast news videos. However, they share a common characteristic: both contain the illumination conditions as encountered in the real world. Based on extensive experiments on this large set of real-world image data, the usefulness of the different invariant properties is derived. As a result, new color descriptors can be designed according to the obtained invariance criteria. Finally, recommendations are given on which color descriptors to use under which circumstances and datasets.

This paper is organized as follows. In section II, the reflectance model is presented. Further, its relation to the diagonal model of illumination change is discussed. In section III, a taxonomy of color descriptors and their invariance properties is given. The experimental setup is presented in section IV. In section V, a discussion of the results is given. Finally, in section VI, conclusions are drawn.

Manuscript received August 28, 2008; revised March 8, 2009; revised June 11, 2009; accepted July 11, 2009.

[1]Software to compute the color descriptors from this paper is available from http://www.colordescriptors.com

Digital Object Indentifier 10.1109/TPAMI.2009.154

Fig. 1. Illustration of variations in viewing and illumination conditions for real-world scenes containing potted plants. The potted plants vary in imaging scale and are imaged under outdoor lighting, indoor lighting and a combination of the two, respectively. Images are from an image benchmark [21].

II. Reflectance Model

An image $\mathbf{f}$ can be modelled under the assumption of Lambertian reflectance as follows:

$$\mathbf{f}(\mathbf{x}) = \int_{\omega} e(\lambda)\rho_k(\lambda)s(\mathbf{x},\lambda)d\lambda, \quad (1)$$

where $e(\lambda)$ is the color of the light source, $s(\mathbf{x},\lambda)$ is the surface reflectance and $\rho_k(\lambda)$ is the camera sensitivity function ($k \in \{R,G,B\}$). Further, ω and $\mathbf{x}$ are the visible spectrum and the spatial coordinates respectively.

Shafer [23] proposes to add a diffuse term to the model of eq. (1). In fact, the term includes a wider range of possible causes than only diffuse light, such as interreflections, infrared sensitivity of the camera sensor, scattering in the medium or lens. The diffuse light is considered to have a lower intensity and to originate from all directions in equal amounts:

$$\mathbf{f}(\mathbf{x}) = \int_{\omega} e(\lambda)\rho_k(\lambda)s(\mathbf{x},\lambda)d\lambda + \int_{\omega} A(\lambda)\rho_k(\lambda)d\lambda, \quad (2)$$

where $A(\lambda)$ is the term that models the diffuse light.

By computing the derivative of image $\mathbf{f}$, it can be easily derived that the effect of $a(\lambda)$ is cancelled out, since it is independent of the surface reflectance term. Then, the reflection model of the spatial derivative of $\mathbf{f}$ at location $\mathbf{x}$ on scale σ is given by:

$$\mathbf{f}_{\mathbf{x},\sigma}(\mathbf{x}) = \int_{\omega} e(\lambda)\rho_k(\lambda)s_{\mathbf{x},\sigma}(\mathbf{x},\lambda)d\lambda. \quad (3)$$

Hence, derivatives will yield invariance to diffuse light. The reflection model of eq. (1) corresponds to the diagonal model of illumination change under the assumption of narrow band filters. This is detailed in the next section.

A. Diagonal Model

Changes in the illumination can be modeled by a diagonal mapping or *von Kries Model* [18] as follows:

$$\mathbf{f}^c = \mathcal{D}^{u,c}\mathbf{f}^u, \quad (4)$$

where $\mathbf{f}^u$ is the image taken under an unknown light source, $\mathbf{f}^c$ is the same image transformed, so it appears as if it was taken under the reference light (called canonical illuminant), and $\mathcal{D}^{u,c}$ is a diagonal matrix which maps colors that are taken under an unknown light source u to their corresponding colors under the canonical illuminant c:

$$\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix}. \quad (5)$$

To include the 'diffuse' light term, Finlayson *et al.* [24] extended the diagonal model with an offset $(o_1, o_2, o_3)^T$, resulting in the diagonal-offset model:

$$\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix} + \begin{pmatrix} o_1 \\ o_2 \\ o_3 \end{pmatrix}. \quad (6)$$

The diagonal model with offset term corresponds to eq. (2) assuming narrow-band filters measured at wavelengths λ_R, λ_G and λ_B at position $\mathbf{x}$ with surface reflectance $s(\mathbf{x},\lambda_C)$ as follows:

$$\begin{pmatrix} e^c(\lambda_R) \\ e^c(\lambda_G) \\ e^c(\lambda_B) \end{pmatrix} = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \begin{pmatrix} e^u(\lambda_R) \\ e^u(\lambda_G) \\ e^u(\lambda_B) \end{pmatrix} + \begin{pmatrix} A(\lambda_R) \\ A(\lambda_G) \\ A(\lambda_B) \end{pmatrix}. \quad (7)$$

As the surface reflectance $s(\mathbf{x},\lambda_C)$ is equal for both the canonical and the unknown illuminant, equation (7) is a simplification of $e^c(\lambda_R)s(\mathbf{x},\lambda_R) = ae^u(\lambda_R)s(\mathbf{x},\lambda_R) + A(\lambda_R)$, $e^c(\lambda_G)s(\mathbf{x},\lambda_G) = be^u(\lambda_G)s(\mathbf{x},\lambda_G) + A(\lambda_G)$ and $e^c(\lambda_B)s(\mathbf{x},\lambda_B) = ce^u(\lambda_B)s(\mathbf{x},\lambda_B) + A(\lambda_B)$.

For broad-band cameras, spectral sharpening can be applied to obtain narrow-band filters [17]. Note that similar to eq. (3), when image derivatives are taken (first or higher order image statistics), the offset in the diagonal-offset model will cancel out.

B. Photometric Analysis

Based on the diagonal model and the diagonal-offset model, five types of common changes in the image values $\mathbf{f}(\mathbf{x})$ are categorized in this section.

Firstly, for eq. (5), when the image values change by a constant factor in all channels (*i.e.* $a = b = c$), this is equal to a *light intensity change*:

$$\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & a \end{pmatrix} \begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix}. \quad (8)$$

In addition to differences in the intensity of the light source, light intensity changes also include (no-colored) shadows and shading. Hence, when a descriptor is invariant to light intensity changes, it is *scale-invariant* with respect to (light) intensity.

Secondly, an equal shift in image intensity values in all channels, *i.e. light intensity shift*, where ($o_1 = o_2 = o_3$) and ($a = b = c = 1$) will yield:

$$\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix} + \begin{pmatrix} o_1 \\ o_1 \\ o_1 \end{pmatrix}. \quad (9)$$

Light intensity shifts are due to diffuse lighting including scattering of a white light source, object highlights (specular component of the surface) under a white light source, inter-reflections and infrared sensitivity of the camera sensor. When a descriptor is invariant to a light intensity shift, it is *shift-invariant* with respect to light intensity.

Thirdly, the above classes of changes can be combined to model both intensity changes and shifts:

$$\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & a \end{pmatrix} \begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix} + \begin{pmatrix} o_1 \\ o_1 \\ o_1 \end{pmatrix}; \qquad (10)$$

i.e. an image descriptor robust to these changes is scale-invariant and shift-invariant with respect to light intensity.

Fourthly, in the full diagonal model (*i.e*. allowing $a \neq b \neq c$), the image channels scale independently (eq. (5)). This allows for *light color changes* in the image. Hence, this class of changes can model a change in the illuminant color and light scattering, amongst others.

Finally, the full diagonal-offset model (eq. (6)) models arbitrary offsets ($o_1 \neq o_2 \neq o_3$), besides the light color changes ($a \neq b \neq c$) offered by the full diagonal model. This type of change is called *light color change and shift*.

In conclusion, five types of common changes have been identified based on the diagonal-offset model of illumination change, i.e. variations to light intensity changes, light intensity shifts, light intensity changes and shifts, light color changes and light color changes and shifts.

III. Color Descriptors and Invariant Properties

In this section, color descriptors are presented and their invariance properties are summarized. First, color descriptors based on histograms are discussed. Then, color moments and color moment invariants are presented. Finally, color descriptors based on SIFT are discussed. These three types of descriptors were chosen due to their distinct nature and wide-spread use. Color histograms do not contain local spatial information and are inherently pixel-based. Color moments do contain local photometrical and spatial information derived from pixel values. SIFT descriptors contain local spatial information and are derivative-based.

See table I for an overview of the descriptors and their invariance properties. We define *invariance* of a descriptor to condition A as follows: under a condition A, the descriptor is independent of changes in condition A. The independence is derived analytically under the assumption that no color clipping occurs. Color clipping occurs when the color of a pixel falls outside the valid range and is subsequently clipped to the minimum or maximum of the range. For example, for a very large scaling of the intensity in eq. (8), color clipping occurs if the scaled values exceed 255, the maximum value typically used for image storage.

A. Histograms

***RGB* histogram** The RGB histogram is a combination of three 1-D histograms based on the R, G and B channels of the RGB color space. This histogram possesses no invariance properties.

Opponent histogram The opponent histogram is a combination of three 1-D histograms based on the channels of the opponent color space:

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix}. \qquad (11)$$

The intensity information is represented by channel O_3 and the color information by O_1 and O_2. Due to the subtraction in O_1 and O_2, the offsets will cancel out if they are equal for all channels (e.g. a white light source). This is verified by substituting the unknown illuminant from eq. (9) with offset o_1:

$$\begin{aligned} \begin{pmatrix} O_1 \\ O_2 \end{pmatrix} &= \begin{pmatrix} \frac{R^c-G^c}{\sqrt{2}} \\ \frac{R^c+G^c-2B^c}{\sqrt{6}} \end{pmatrix} \\ &= \begin{pmatrix} \frac{(R^u+o_1)-(G^u+o_1)}{\sqrt{2}} \\ \frac{(R^u+o_1)+(G^u+o_1)-2(B^u+o_1)}{\sqrt{6}} \end{pmatrix} \\ &= \begin{pmatrix} \frac{R^u-G^u}{\sqrt{2}} \\ \frac{R^u+G^u-2B^u}{\sqrt{6}} \end{pmatrix}. \end{aligned} \qquad (12)$$

Therefore, these O_1 and O_2 are shift-invariant with respect to light intensity. The intensity channel O_3 has no invariance properties.

Hue histogram In the HSV color space, it is known that the hue becomes unstable near the grey axis. To this end, Van de Weijer *et al*. [14] apply an error propagation analysis to the hue transformation. The analysis shows that the certainty of the hue is inversely proportional to the saturation. Therefore, the hue histogram is made more robust by weighing each sample of the hue by its saturation. The H color model is scale-invariant and shift-invariant with respect to light intensity [14].

***rg*histogram** In the normalized RGB color model, the chromaticity components r and g describe the color information in the image (b is redundant as $r + g + b = 1$):

$$\begin{pmatrix} r \\ g \\ b \end{pmatrix} = \begin{pmatrix} \frac{R}{R+G+B} \\ \frac{G}{R+G+B} \\ \frac{B}{R+G+B} \end{pmatrix}. \qquad (13)$$

Because of the normalization, r and g are scale-invariant and thereby invariant to light intensity changes, shadows and shading [25] from eq. (8):

$$\begin{aligned} \begin{pmatrix} r \\ g \end{pmatrix} &= \begin{pmatrix} \frac{R^c}{R^c+G^c+B^c} \\ \frac{G^c}{R^c+G^c+B^c} \end{pmatrix} = \begin{pmatrix} \frac{aR^u}{aR^u+aG^u+aB^u} \\ \frac{aG^u}{aR^u+aG^u+aB^u} \end{pmatrix} \\ &= \begin{pmatrix} \frac{aR^u}{a(R^u+G^u+B^u)} \\ \frac{aG^u}{a(R^u+G^u+B^u)} \end{pmatrix} = \begin{pmatrix} \frac{R^u}{R^u+G^u+B^u} \\ \frac{G^u}{R^u+G^u+B^u} \end{pmatrix}. \end{aligned} \qquad (14)$$

Transformed color distribution An RGB histogram is not invariant to changes in lighting conditions. However, by normalizing the pixel value distributions, scale-invariance and shift-invariance is achieved with respect to light intensity. Because each channel is normalized independently, the descriptor

TABLE I
INVARIANCE OF DESCRIPTORS (SECTION III) AGAINST TYPES OF CHANGES IN THE DIAGONAL-OFFSET MODEL AND ITS SPECIALIZATIONS (SECTION II-B). INVARIANCE IS INDICATED WITH '+', LACK OF INVARIANCE IS INDICATED WITH '-'. THE INVARIANCE OF A DESCRIPTOR TO CONDITION A IS DEFINED AS FOLLOWS: UNDER A CONDITION A, THE DESCRIPTOR IS INDEPENDENT OF CHANGES IN CONDITION A. THE INDEPENDENCE IS DERIVED ANALYTICALLY UNDER THE ASSUMPTION THAT NO COLOR CLIPPING OCCURS.

	Light intensity change $\begin{pmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & a \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$	**Light intensity shift** $\begin{pmatrix} R \\ G \\ B \end{pmatrix} + \begin{pmatrix} o_1 \\ o_1 \\ o_1 \end{pmatrix}$	**Light intensity change and shift** $\begin{pmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & a \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} + \begin{pmatrix} o_1 \\ o_1 \\ o_1 \end{pmatrix}$	**Light color change** $\begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$	**Light color change and shift** $\begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} + \begin{pmatrix} o_1 \\ o_2 \\ o_3 \end{pmatrix}$
RGB Histogram	-	-	-	-	-
O_1, O_2	-	+	-	-	-
O_3, Intensity	-	-	-	-	-
Hue	+	+	+	-	-
Saturation	-	-	-	-	-
r, g	+	-	-	-	-
Transformed color	+	+	+	+	+
Color moments	-	+	-	-	-
Moment invariants	+	+	+	+	+
SIFT (∇I)	+	+	+	-	-
HSV-SIFT	-	-	-	-	-
HueSIFT	+	+	+	-	-
OpponentSIFT	+	+	+	-	-
C-SIFT	+	-	-	-	-
rgSIFT	+	-	-	-	-
Transf. color SIFT	+	+	+	+	+
RGB-SIFT	+	+	+	+	+

is also normalized against changes in light color and arbitrary offsets:

$$\begin{pmatrix} R' \\ G' \\ B' \end{pmatrix} = \begin{pmatrix} \frac{R-\mu_R}{\sigma_R} \\ \frac{G-\mu_G}{\sigma_G} \\ \frac{B-\mu_B}{\sigma_B} \end{pmatrix}, \tag{15}$$

with μ_C the mean and σ_C the standard deviation of the distribution in channel C computed over the area under consideration (e.g. a patch or image). This yields for every channel a distribution where $\mu = 0$ and $\sigma = 1$.

B. Color Moments and Moment Invariants

A color image corresponds to a function I defining RGB triplets for image positions (x, y): $I : (x, y) \mapsto (R(x, y), G(x, y), B(x, y))$. By regarding RGB triplets as data points coming from a distribution, it is possible to define moments. Mindru *et al.* [26] have defined *generalized color moments* M_{pq}^{abc}:

$$M_{pq}^{abc} = \int \int x^p y^q [I_R(x, y)]^a [I_G(x, y)]^b [I_B(x, y)]^c dx dy. \tag{16}$$

M_{pq}^{abc} is referred to as a generalized color moment of *order* $p + q$ and *degree* $a + b + c$. Note that moments of order 0 do not contain any spatial information, while moments of degree 0 do not contain any photometric information. Thus, moment descriptions of order 0 are rotationally invariant, while higher orders are not. A large number of moments can be created with small values for the order and degree. However, for larger values the moments are less stable. Typically, generalized color moments up to the first order and the second degree are used.

By using the proper combination of moments, it is possible to normalize against photometric changes. These combinations are called *color moment invariants*. Invariants involving only a single color channel (*e.g.* out of a, b and c two are 0) are called 1-band invariants. Similarly there are 2-band invariants involving only two out of three color bands. 3-band invariants involve all color channels, but these can always be created by using 2-band invariants for different combinations of channels.

Color moments The color moment descriptor uses all generalized color moments up to the second degree and the first order. This lead to nine possible combinations for the degree: M_{pq}^{000}, M_{pq}^{100}, M_{pq}^{010}, M_{pq}^{001}, M_{pq}^{200}, M_{pq}^{110}, M_{pq}^{020}, M_{pq}^{011}, M_{pq}^{002} and M_{pq}^{101}†. Combined with three possible combinations for the order: M_{00}^{abc}, M_{10}^{abc} and M_{01}^{abc}, the color moment descriptor has 27 dimensions. These color moments only have shift-invariance. This is achieved by subtracting the average in all input channels before computing the moments.

Color moment invariants Color moment invariants can be constructed from generalized color moments. All 3-band invariants are computed from Mindru *et al.* [26]. To be comparable, the $\tilde{C}_{02}$ invariants are considered. This gives a total of 24 color moment invariants, which are invariant to all the properties listed in table I.

C. Color SIFT Descriptors

SIFT The SIFT descriptor proposed by Lowe [9] describes the local shape of a region using edge orientation histograms. The gradient of an image is shift-invariant: taking the derivative cancels out offsets (section II-B). Under light intensity changes, *i.e.* a scaling of the intensity channel, the gradient direction and the relative gradient magnitude remain the same. Because the SIFT descriptor is normalized, the gradient magnitude changes have no effect on the final descriptor. The SIFT descriptor is not invariant to light color changes, because the intensity channel is a combination of the R, G and B channels. To compute SIFT descriptors, the version described by Lowe [9] is used.

HSV-SIFT Bosch *et al.* [16] compute SIFT descriptors over all three channels of the HSV color model. This gives 3x128 dimensions per descriptor, 128 per channel. As stated earlier,

†Because it is constant, the moment M_{pq}^{000} is excluded.

the H color model is scale-invariant and shift-invariant with respect to light intensity. However, due to the combination of the HSV channels, the complete descriptor has no invariance properties. Further, the instability of the hue for low saturation is not addressed here.

HueSIFT Van de Weijer *et al*. [14] introduce a concatenation of the hue histogram (see section III-A) with the SIFT descriptor. When compared to HSV-SIFT, the usage of the weighed hue histogram addresses the instability of the hue near the grey axis. Because the bins of the hue histogram are independent, the periodicity of the hue channel for HueSIFT is addressed. Similar to the hue histogram, the HueSIFT descriptor is scale-invariant and shift-invariant.

OpponentSIFT OpponentSIFT describes all the channels in the opponent color space (eq. (11)) using SIFT descriptors. The information in the O_3 channel is equal to the intensity information, while the other channels describe the color information in the image. These other channels do contain some intensity information, but due to the normalization of the SIFT descriptor they are invariant to changes in light intensity.

C-SIFT In the opponent color space, the O_1 and O_2 channels still contain some intensity information. To add invariance to intensity changes, [13] proposes the C-invariant which eliminates the remaining intensity information from these channels. The use of color invariants as input for SIFT was first suggested by Abdel-Hakim and Farag [12]. The C-SIFT descriptor [15] uses the C invariant, which can be intuitively seen as the normalized opponent color space $\frac{O_1}{O_3}$ and $\frac{O_2}{O_3}$. Because of the division by intensity, the scaling in the diagonal model will cancel out, making C-SIFT scale-invariant with respect to light intensity. Due to the definition of the color space, the offset does not cancel out when taking the derivative: it is not shift-invariant.

***rg*SIFT** For the rgSIFT descriptor, descriptors are added for the r and g chromaticity components of the normalized RGB color model from eq. (13), which is already scale-invariant.

Transformed color SIFT For the transformed color SIFT, the same normalization is applied to the RGB channels as for the transformed color histogram (eq. (15)). For every normalized channel, the SIFT descriptor is computed. The descriptor is scale-invariant, shift-invariant and invariant to light color changes and shift.

RGB-SIFT For the RGB-SIFT descriptor, SIFT descriptors are computed for every RGB channel independently. An interesting property of this descriptor, is that its descriptor values are equal to the transformed color SIFT descriptor. This is explained by looking at the transformed color space (eq. (15)): this transformation is already implicitly performed when SIFT is applied to each RGB channel independently. Because the SIFT descriptor operates on derivatives only, the subtraction of the means in the transformed color model is redudant, as this offset is already cancelled out by taking derivatives. Similarly, the division by the standard deviation is already implicitly performed by the normalization of the vector length of SIFT descriptors. Therefore, as the RGB-SIFT and transformed color SIFT descriptors are equal, we will use the RGB-SIFT name throughout this paper.

D. Conclusion

In this section, three different groups of color descriptors were discussed: histograms in different color spaces, color moments and moment invariants and color extensions of SIFT. For each color descriptor, the invariance with respect to illumination changes in the diagonal-offset model were analyzed. The results are summarized in table I.

IV. Experimental Setup

In this section, the experimental setup to evaluate the different color descriptors is outlined. The *invariance* properties of the color descriptors, which were derived analytically in the previous section, are verified experimentally as well using a dataset with known illumination conditions. The *distinctiveness* of the color descriptors is assessed experimentally through their discriminative power on the dataset with known imaging conditions, an image benchmark and a video benchmark.

First, implementation details of the descriptors in an object and scene recognition setting are discussed. Then, the datasets used for evaluation are described. After discussing these benchmarks and their datasets, evaluation criteria are given.

A. Feature Extraction Pipelines

To emperically test the different color descriptors, the descriptors are computed at scale-invariant points [5], [9]. See figure 2 for an overview of the processing pipeline. In the pipeline shown, scale-invariant points are obtained with the Harris-Laplace point detector on the intensity channel. Other region detectors [10], such as the dense sampling detector, Maximally Stable Extremal Regions [27] and Maximally Stable Color Regions [28], can be plugged in. For the experiments, the Harris-Laplace point detector is used because it has shown good performance for category recognition [5]. This detector uses the Harris corner detector to find potential scale-invariant points. It then selects a subset of these points for which the Laplacian-of-Gaussians reaches a maximum over scale. The color descriptors from section III are computed over the area around the points. The size of this area depends on the maximum scale of the Laplacian-of-Gaussians [10].

To obtain fixed-length feature vectors per image, the bag-of-words model is used [29]. The bag-of-words model is also known as 'textons' [30], 'object parts' [31] and 'codebooks' [32], [33]. The bag-of-words model performs vector quantization of the color descriptors in an image against a visual codebook. A descriptor is assigned to the codebook element which is closest in Euclidian space. To be independent of the total number of descriptors in an image, the feature vector is normalized to sum to 1.

The visual codebook is constructed by applying k-means clustering to $200,000$ randomly sampled descriptors from the set of images available for training. In this paper, visual codebooks with $4,000$ elements are used.

Color descriptor software implementing this processing pipeline is available from our website[2]. It performs point sampling, color descriptor computation and vector quantization.

[2]http://www.colordescriptors.com

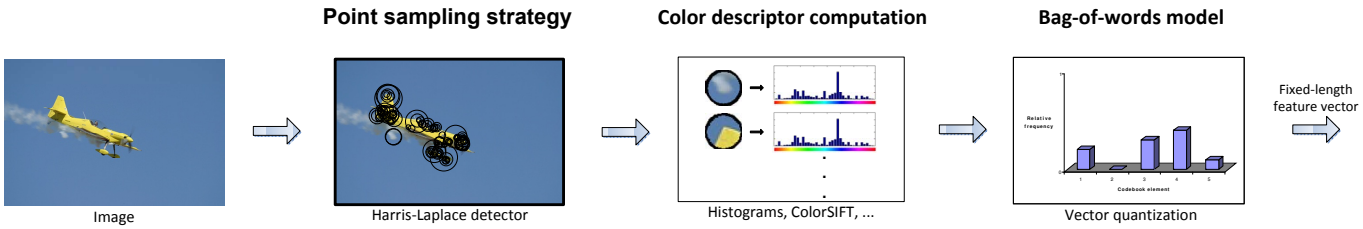

Fig. 2. The stages of the primary feature extraction pipeline used in this paper. First, the Harris-Laplace salient point detector is applied to the image. Then, for every point a color descriptor is computed over the area around the point. All the color descriptors of an image are subsequently vector quantized against a codebook of prototypical color descriptors. This results in a fixed-length feature vector representing the image.

After these steps, an image is represented by a fixed-length feature vector.

B. Classification

For datasets where only a single training example is available per object or scene category, a nearest neighbor classifier is used with χ^2 distances between feature vectors F and F':

$$dist_{\chi^2}(\vec{F},\vec{F'}) = \frac{1}{2}\sum_{i=1}^{n}\frac{(\vec{F_i}-\vec{F'_i})^2}{\vec{F_i}+\vec{F'_i}}, \tag{17}$$

with n the size of the feature vectors. For notational convenience, $\frac{0}{0}$ is assumed to be equal to 0 iff $\vec{F_i} = \vec{F'_i} = 0$.

For datasets with multiple training examples, the support vector machines classifier is used. The decision function of a support vector machines classifier for a test sample with feature vector $\vec{F'}$ has the form:

$$g(\vec{F'}) = \sum_{\vec{F}\in trainset} \alpha_{\vec{F}} y_{\vec{F}} k(\vec{F},\vec{F'}) - \beta, \tag{18}$$

where $y_{\vec{F}}$ is the class label of $\vec{F}$ (-1 or $+1$), $\alpha_{\vec{F}}$ is the learned weight of train sample $\vec{F}$, β is a learned threshold and $k(\vec{F},\vec{F'})$ is the value of a kernel function based on the χ^2 distance, which has shown good results in object recognition [5]:

$$k(\vec{F},\vec{F'}) = e^{-\frac{1}{D}dist_{\chi^2}(\vec{F},\vec{F'})}, \tag{19}$$

where D is a scalar which normalizes the distances. We set D to the average χ^2 distance between all elements of the train set.

The LibSVM implementation [34] is used to train the classifier. As parameters for the training phase, the weight of the positive class is set to $\frac{\#pos+\#neg}{\#pos}$ and the weight of the negative class is set to $\frac{\#pos+\#neg}{\#neg}$, with $\#pos$ the number of positive instances in the train set and $\#neg$ the number of negative instances. The cost parameter is optimized using 3-fold cross-validation with a parameter range of 2^{-4} through 2^4.

To use multiple features, instead of relying on a single feature, the kernel function is extended in a weighted fashion for m features:

$$k(\{\vec{F}_{(1)},...,\vec{F}_{(m)}\},\{\vec{F'}_{(1)},...,\vec{F'}_{(m)}\}) = e^{-\frac{1}{\sum_{j=1}^{m} w_j}\left(\sum_{j=1}^{m}\frac{w_j}{D_j}dist(\vec{F}_{(j)},\vec{F'}_{(j)})\right)}, \tag{20}$$

with w_j the weight of the j^{th} feature, D_j the normalization factor for the j^{th} feature and $\vec{F}_{(j)}$ the j^{th} feature vector.

An example of the use of multiple features is the spatial pyramid [3]; it is illustrated in figure 3. When using the spatial pyramid, additional features are extracted for specific parts of the image. For example, in a 2x2 subdivision of the image, feature vectors are extracted for each image quarter with a weight of $\frac{1}{4}$ for each quarter. Similarly, a 1x3 subdivision consisting of three horizontal bars, which introduces three new features (each with a weight of $\frac{1}{3}$). In this setting, the feature vector for the entire image has a weight of 1.

C. Experiment 1: Illumination Changes

The Amsterdam Library of Object Images (ALOI) dataset [20] contains more than 48,000 images of 1,000 objects, under various illumination conditions. Light intensity scaling (eq. (8)) and light intensity shifts (eq. (9)) are not present in the dataset, therefore we have artificially added these two condition changes to the dataset. The effect of simultaneous light intensity changes and shifts (eq. (10)) is a combination of the previous two properties. Since these two properties are already evaluated individually, we refrain from evaluating this combined property. The light color change images from ALOI directly correspond to our light color changes (eq. (5)). The light color is varied by changing the illumination color temperature, resulting in objects illuminated under a reddish to white light. For completeness, the other conditions present in the ALOI dataset are also included: objects lighted by a different number of white lights at increasingly oblique angles (between one and three white lights around the object, introducing selfshadowing for up to half of the object), object rotation images and images with different levels of JPEG compression.

Because only a single training example is available per object category, the nearest neighbour classifier is used for the ALOI dataset.

D. Experiment 2: Image Benchmark

The PASCAL Visual Object Classes Challenge [21] provides a yearly benchmark for comparison of object classification systems. The PASCAL VOC Challenge 2007 dataset contains nearly 10,000 images of 20 different object categories, e.g. bird, bottle, car, dining table, motorbike and people. The dataset is divided into a predefined train set (5011 images) and test set (4952 images).

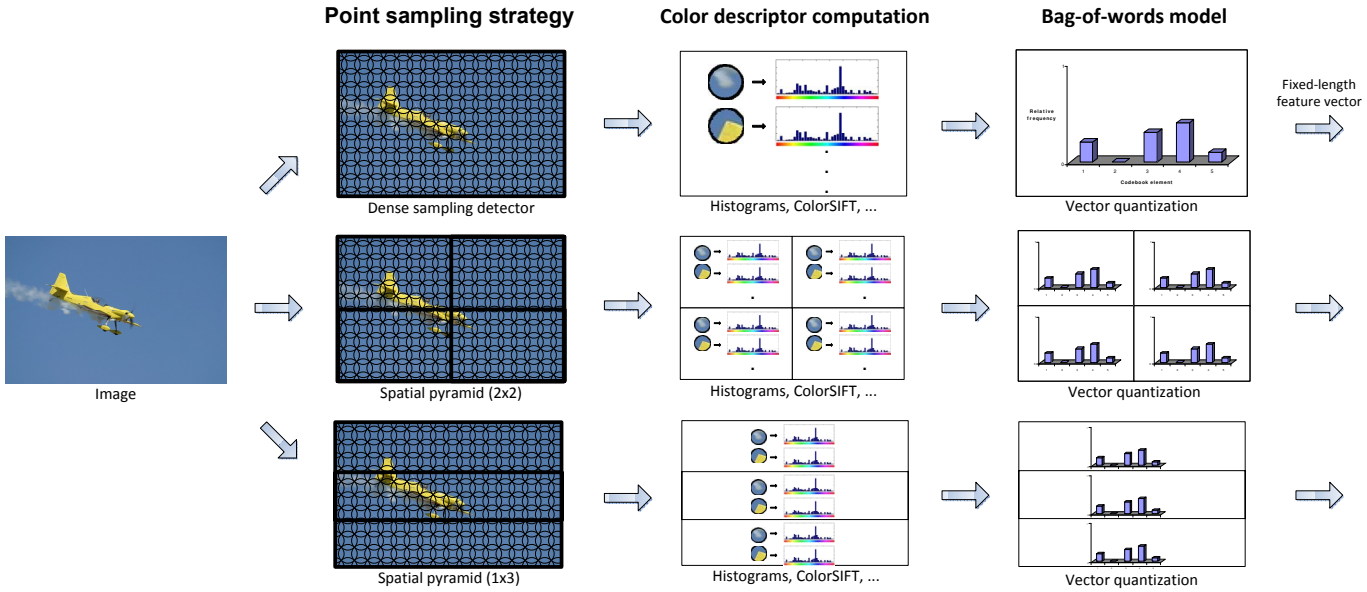

Fig. 3. Examples of additional feature extraction pipelines used in this paper, besides the primary pipeline shown in figure 2. The pipelines shown are examples of using a different point sampling strategy or a spatial pyramid [3]. The spatial pyramid constructs feature vectors for specific parts of the image. For every pipeline, first, a point sampling method is applied to the image. Then, for every point a color descriptor is computed over the area around the point. All the color descriptors of an image are subsequently vector quantized against a codebook of prototypical color descriptors. This results in a fixed-length feature vector representing the image.

E. Experiment 3: Video Benchmark

The Mediamill Challenge by Snoek *et al.* [22] provides an annotated video dataset, based on the training set of the NIST TRECVID 2005 benchmark [7]. Over this dataset, repeatable experiments have been defined. The experiments decompose automatic category recognition into a number of components, for which they provide a standard implementation. This provides an environment to analyze which components affect the performance most.

The dataset of 86 hours is divided into a Challenge training set (70% of the data or $30,993$ shots) and a Challenge test set (30% of the data or $12,914$ shots). For every shot, the Challenge provides a single representative keyframe image. So, the complete dataset consists of $43,907$ images, one for every video shot. The dataset consists of television news from November 2004 broadcasted on six different TV channels in three different languages: English, Chinese and Arabic. On this dataset, the 39 LSCOM-Lite categories [35] are employed. These include object categories like aircraft, animal, car and faces, and scence categories such as desert, mountain, sky, urban and vegetation.

F. Evaluation Criteria

Experiments on the ALOI dataset perform object recognition using one example: given a query image of an object under unknown illumination conditions, the top-ranked result should be equal to the original image of the object for successful recognition. The percentage of objects where the top-ranked result is indeed the correct object is used as the performance on the ALOI dataset.

For our benchmark results, the average precision is taken as the performance metric for determining the accuracy of ranked category recognition results. The average precision is a single-valued measure that is proportional to the area under a precision-recall curve. This value is the average of the precision over all images/keyframes judged to be relevant. Hence, it combines both precision and recall into a single performance value. For the PASCAL VOC Challenge 2007, the official standard is the 11-point interpolated average precision, and for TRECVID, the official standard is the non-interpolated average precision. The interpolated average precision is an approximation of the non-interpolated average precision. As the difference between the two is generally very small, we will follow the official standard for each dataset and refer to them as average precision scores. When performing experiments over multiple object and scene categories, the average precisions of the individual categories are aggregated. This aggregation, mean average precision, is calculated by taking the mean of the average precisions. As average precision depends on the number of correct object and scene categories present in the test set, the mean average precision depends on the dataset used.

To obtain an indication of significance, the bootstrap method [36], [37] is used to estimate confidence intervals for mean average precision. In bootstrap, multiple test sets T_B are created by selecting images at random from the original test set T, with replacement, until $|T| = |T_B|$. This has the effect that some images are replicated in T_B, whereas other images may be absent. This process is repeated 1000 times to generate 1000 test sets, each obtained by sampling from the original test set T. The statistical accuracy of the mean average precision score can then be evaluated by looking at the standard deviation of the mean average precision scores over the different bootstrap test sets.

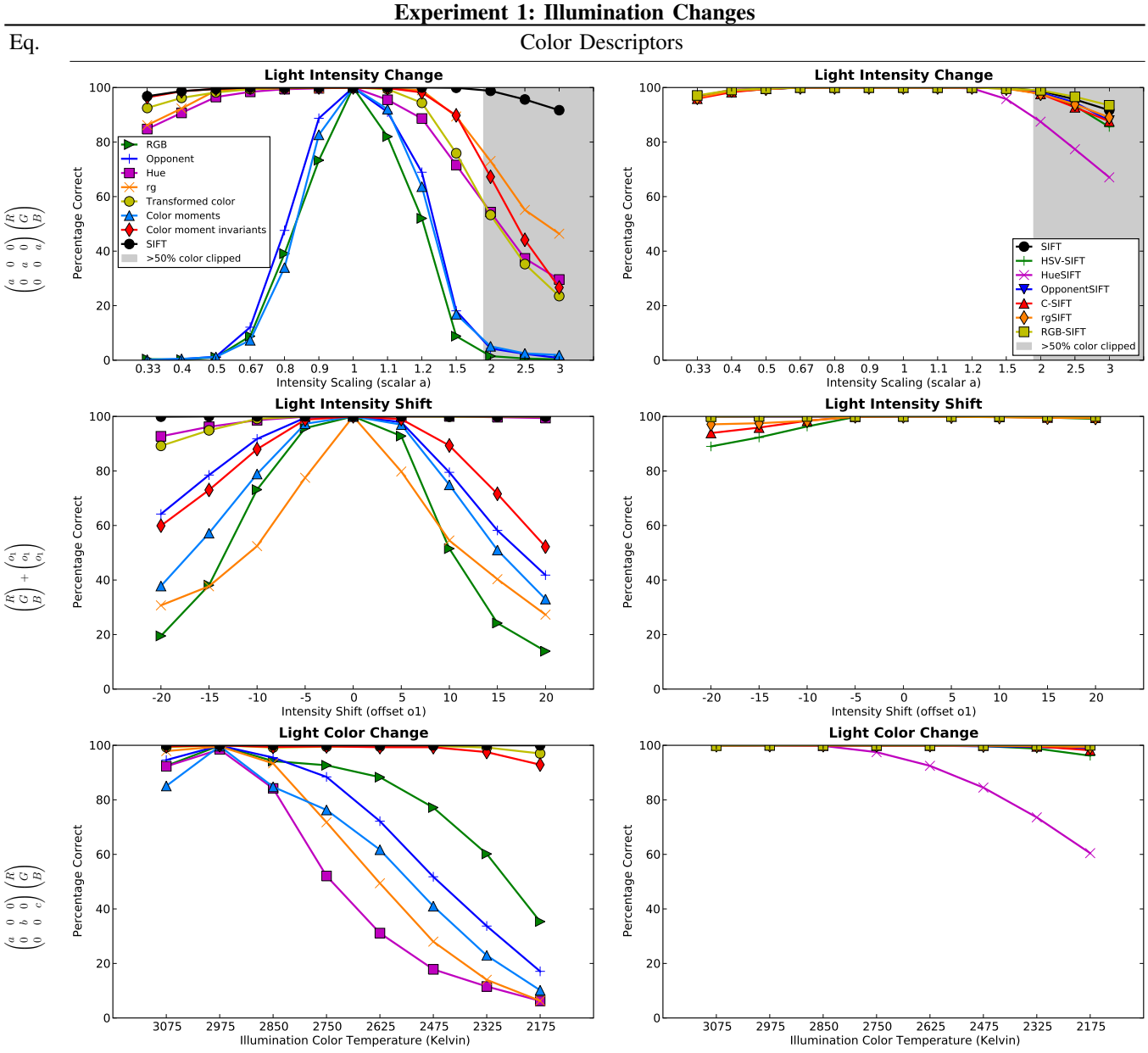

Fig. 4. Evaluation of the invariance properties of color descriptors under different illumination conditions, averaged over 1000 objects from the ALOI dataset [20]. Performance is measured using the percentage of correctly identified objects. For clarity of presentation, the results have been split into two parts. To allow for easier comparison, SIFT is shown in both the graphs on the left and the graphs on the right. The rows correspond to the invariant properties from section II, as listed in the graph titles and the equations shown. For light intensity shifts, the axis unit corresponds to image values in the range $[0, 255]$. For the light color changes, the light color is varied by changing the illumination color temperature, resulting in objects illuminated under a white to reddish light. Conditions where, on average, more than 50% of the object area is affected by color clipping (due to image values falling outside the range $[0, 255]$) are marked with a grey background.

V. Results

A. Experiment 1: Illumination Changes

From the results in figure 4, the theoretical invariance properties of color descriptors are validated. By observing the results with respect to light intensity changes, the color descriptors without invariance to this property, such as the RGB histogram, the opponent color histogram and color moments, do not perform well. There is a clear distinction in performance between these descriptors and the invariant descriptors, such as the hue histogram, color moment invariants and SIFT. Overall, within this group of invariant descriptors, the SIFT and color SIFT descriptors perform much better than histogram-based descriptors; they have higher discriminative power. HueSIFT, which is a combination of the hue histogram and the SIFT descriptor, falls between these descriptor classes in terms of performance. The HSV-SIFT descriptor, which is not invariant to light intensity changes, is the lowest-scoring SIFT descriptor after HueSIFT. For very large scaling factors, the performance of all descriptors drops. This is due to color clipping: scaled image values outside the range $[0, 255]$ are clipped to 255. In figure 4, a grey background indicates under which conditions, on average, more than half of all object

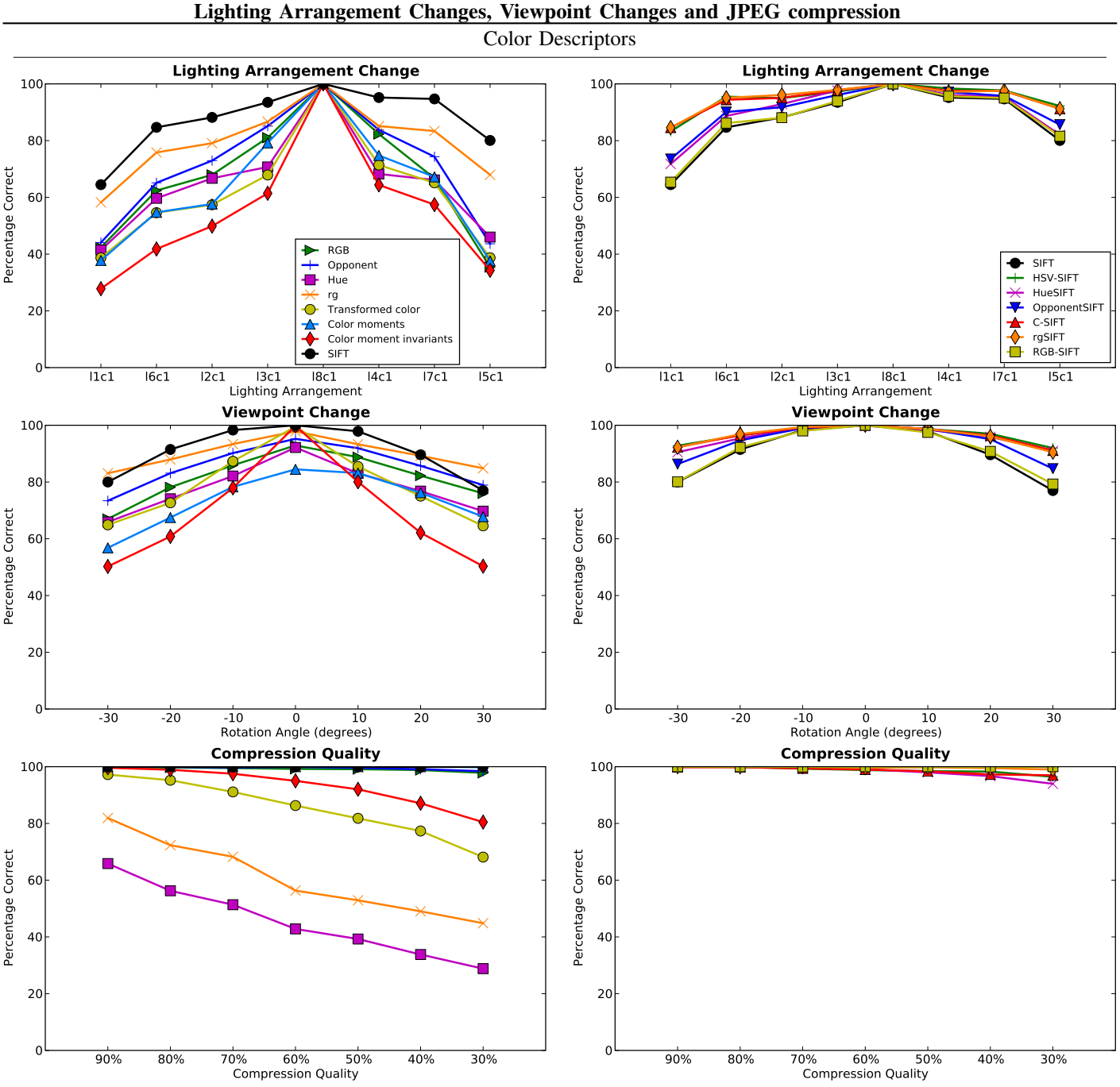

Fig. 5. For completeness, this figure contains the results for color descriptors under different lighting arrangements at increasingly oblique angles (between one and three of the lights around the object are on, introducing selfshadowing for up to half of the object), different viewpoint angles and different degrees of JPEG compression, averaged over 1000 objects from the ALOI dataset [20]. Performance is measured using the percentage of correctly identified objects. For clarity of presentation, the results have been split into two parts. To allow for easier comparison, SIFT is shown in both the graphs on the left and the graphs on the right.

pixels have been clipped.

For light intensity shifts, it is shown that the color descriptors which lack invariance, the RGB histogram, the opponent color histogram and the rghistogram, indeed have reduced performance. Additionally, color moments and color moment invariants are affected when the shift amount increases, these descriptors can only handle small light intensity shifts. The three color SIFT descriptors which lack shift-invariance, HSV-SIFT, C-SIFT and rgSIFT, show reduced performance for large shifts when compared to other SIFT variants, confirming their lack of invariance.

For light color changes, it is observed that histograms do not perform well. This is consistent with their lack of invariance. The exceptions are the transformed color histogram and the color moment invariants, which do possess invariance to light color changes and indeed perform much better. For the SIFT-based descriptors, only HSV-SIFT and HueSIFT degrade in performance as the light color changes. This is due to their lack of invariance. Of interest is that some of the descriptors which are not invariant to light color changes, *e.g.* OpponentSIFT, C-SIFT and rgSIFT, are (in practice) largely robust to the light color changes present in the ALOI dataset.

Besides the evaluation of the invariant properties, there are also different conditions which can be evaluated using ALOI. For the lighting arrangement changes, shown in figure 5, between one and three white lights around the object are turned on. This leads to shadows, shading and white highlights, *e.g.* to both light intensity scaling and shifts (eq. (10)), but also to partial visibility due to lack of light on certain parts of the object. In this setting, both the invariant properties and the discriminative power of color descriptors play an important role. The intensity scale-invariant C-SIFT and rgSIFT perform well, ahead of the OpponentSIFT descriptor, which is also shift-invariant. For the RGB-SIFT descriptor, which is invariant to light color changes in addition to begin scale-invariant and shift-invariant, the increased invariance comes at the price of reduced discriminative power: it is behind C-SIFT, rg-SIFT and OpponentSIFT under this condition. For this condition, light intensity shifts and light color changes do not occur and therefore OpponentSIFT and RGB-SIFT are too invariant. A similar pattern is observed from the results in figure 5 for viewpoint changes due to object rotation. The scale-invariant C-SIFT and rgSIFT perform best, and the light intensity shift invariance offered by OpponentSIFT and RGB-SIFT is not needed, nor is the light color invariance of RGB-SIFT.

From the results shown in figure 5 for JPEG compression quality, it can be seen that the hue histogram, the rghistogram, the transformed color histogram and the color moment invariants are not robust to even moderate amounts of compression: compression artifacts cause large deviations in these descriptors.

In conclusion, changes in lighting conditions affect color descriptors. However, for object recognition, not just the invariance of a color descriptor to lighting conditions is important, but also the distinctiveness of the descriptor. An invariant descriptor is only useful for visual categorization when it has sufficient discriminative power as well. Finally, certain color descriptors are sensitive to compression artifacts, reducing their usefulness. Although the best choice of color descriptor depends on the condition, the descriptors with the best overall performance are C-SIFT, rgSIFT, OpponentSIFT and RGB-SIFT.

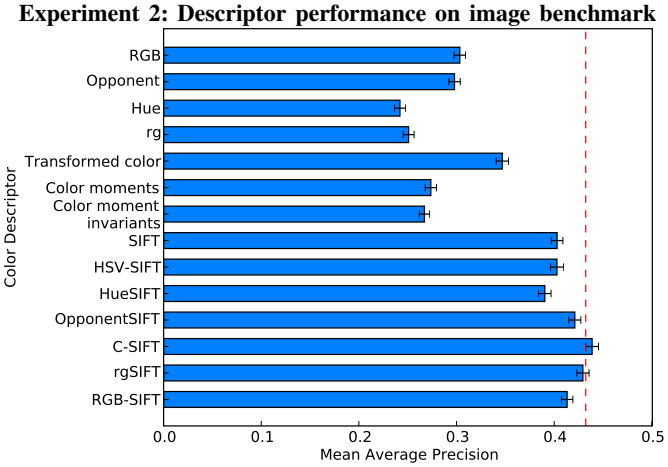

Fig. 6. Evaluation of color descriptors on an image benchmark, the PASCAL VOC Challenge 2007 [21], averaged over the 20 object categories. Error bars indicate the standard deviation in mean average precision, obtained using bootstrap. The dashed lines indicate the lower bound of the C-SIFT confidence interval.

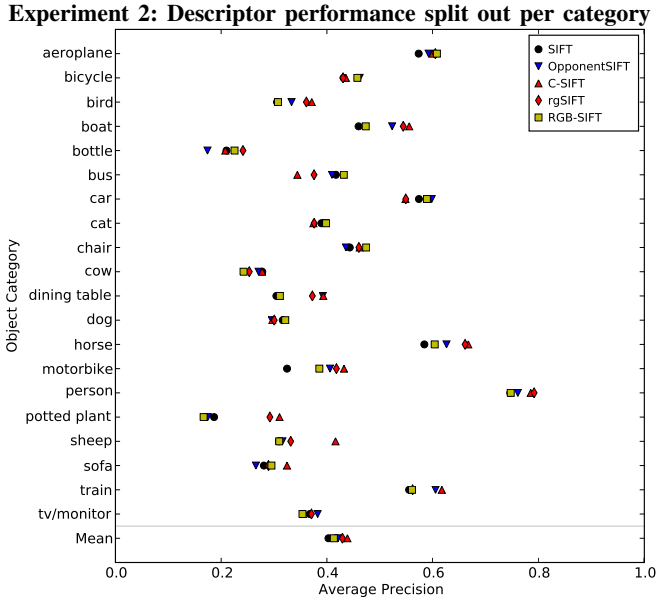

Fig. 7. Evaluation of color descriptors on an image benchmark, the PASCAL VOC Challenge 2007, split out per object category. SIFT and the best four color SIFT variants from figure 6 are shown.

B. Experiment 2: Image Benchmark

From the results shown in figure 6, it is observed that for object category recognition the SIFT variants perform significantly better than color moments, moment invariants and color histograms. The moments and histograms are not very distinctive when compared to SIFT-based descriptors: they contain too little relevant information to be competitive with SIFT.

For SIFT and the four best color SIFT descriptors from

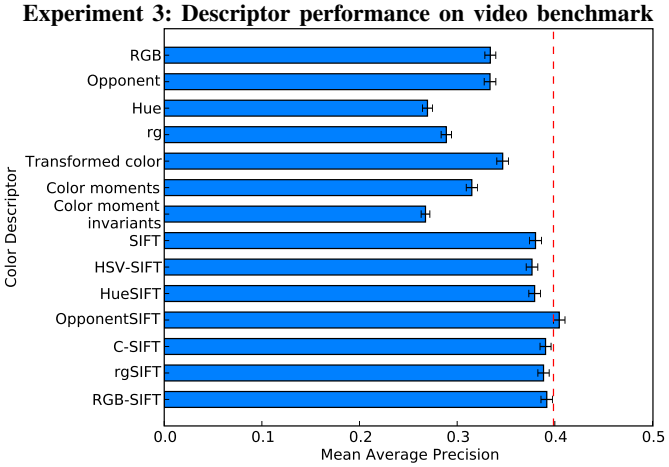

Fig. 8. Evaluation of color descriptors on a video benchmark, the Mediamill Challenge [22], averaged over 39 object and scene categories. Error bars indicate the standard deviation in mean average precision, obtained using bootstrap. The dashed line indicates the lower bound of the OpponentSIFT confidence interval.

figure 6 (OpponentSIFT, C-SIFT, rgSIFT and RGB-SIFT), the results per object category are shown in figure 7. For bird, boat, horse, motorbike, person, potted plant and sheep, it can be observed that the descriptors which perform best have scale-invariance for light intensity (C-SIFT and rgSIFT). Of these two scale-invariant descriptors, C-SIFT has the highest overall performance. The performance of the OpponentSIFT descriptor, which is also shift-invariant compared to C-SIFT, indicates that only scale-invariance, i.e. invariance to light intensity changes, is important for these object categories. RGB-SIFT includes additional invariance against light intensity shifts and light color changes and shifts when compared to C-SIFT. However, this additional invariance makes the descriptor less discriminative for these object categories, because a reduction in performance is observed. This is illustrated by the examples shown in figure 1 for potted plant, which are ranked significantly higher for C-SIFT and rgSIFT compared to OpponentSIFT and RGB-SIFT.

In conclusion, C-SIFT is significantly better than all other descriptors except rgSIFT (see figure 6) on the image benchmark. The corresponding invariant property of both of these descriptors is given by eq. (8). However, the difference between the rgSIFT descriptor and OpponentSIFT, which corresponds to eq. (10), is not significant. Therefore, the best choice for this dataset is C-SIFT.

C. Experiment 3: Video Benchmark

From the visual categorization results shown in figure 8, the same overall pattern as for the image benchmark is observed: SIFT and color SIFT variants perform significantly better than the other descriptors. The shift-invariant OpponentSIFT has left C-SIFT behind and is now the only descriptor which is significantly better than all other descriptors. An analysis on the individual object and scene categories shows that the OpponentSIFT descriptor performs best for building, meeting, mountain, office, outdoor, sky, studio, walking/running and weather news. All these concepts occur under a wide range of light intensities and different amounts of diffuse lighting. Therefore, its invariance to light intensity changes and shifts makes OpponentSIFT a good feature for these categories, and explains why it is better than C-SIFT and rgSIFT for the video benchmark. RGB-SIFT, with additional invariance to light color changes and shifts, does not differ significantly from C-SIFT and rgSIFT. For some categories, there is a small performance gain, for others there is a small loss. This contrasts with the results on the image benchmark, where a performance reduction was observed.

In conclusion, OpponentSIFT is significantly better than all other descriptors on the video benchmark (see figure 8). The corresponding invariant property is given by eq. (10).

D. Comparison with state-of-the-art

So far, the performance of single descriptors has been analyzed. It is worthwhile to investigate combinations of several descriptors, since they are not completely redundant. State-of-the-art results on the PASCAL VOC Challenge 2007 also employ combinations of several methods. Table II gives an overview of combinations on this dataset. For example, the best entry in the PASCAL VOC Challenge 2007, by Marszałek *et al.* [38], has achieved a mean average precision of 0.594 using SIFT and HueSIFT descriptors, the spatial pyramid [3], additional point sampling strategies besides Harris-Laplace such as Laplacian point sampling and dense sampling, and a feature selection scheme. When the feature selection scheme is excluded and simple flat fusion is used, Marszałek reports a mean average precision of 0.575.

To illustrate the potential of the color descriptors from table I, a simple fusion experiment has been performed with SIFT and the best four color SIFT variants (section IV-B details how the combination is constructed). To be comparable, a setting similar to Marszałek is used: both Harris-Laplace point sampling and dense sampling are employed and the same spatial pyramid is used (see figure 2 for an overview of the feature extraction pipelines used). In this setting, the best single color descriptor achieve a mean average precision 0.566. The combination gives a mean average precision of 0.605. This convincing gain of 7% suggests that the color descriptors are not entirely redundant. Compared to the intensity-based SIFT descriptor, the gain is 8%. Further gains should be possible, if the descriptors with the right amount of invariance are fused, preferably using an automatic selection strategy.

As shown in table III, similar gains are observed on the Mediamill Challenge: mean average precision increases by 7% when combinations of color descriptors are used, instead of intensity-based SIFT only. Relative to the best single color descriptor, an increase of 3% is observed. Furthermore, when the descriptors of this paper are compared to the baseline provided by the Mediamill Challenge, there is a relative improvement of 104%.

For reference, combinations of color descriptors from this paper were submitted to the PASCAL VOC 2008 benchmark [40] and the TRECVID 2008 evaluation campaign [7]. In both cases, top performance was achieved. The color descriptors as presented in this paper were the foundation of

TABLE II
IN THIS TABLE, COMBINATIONS OF DESCRIPTORS ON THE IMAGE BENCHMARK ARE COMPARED TO MARSZAŁEK *et al.* [38], WHO OBTAINS STATE-OF-THE-ART RESULTS ON THIS DATASET. ADDING COLOR DESCRIPTORS IMPROVES OVER INTENSITY-BASED SIFT ALONE BY 8%.

Combinations on image benchmark

Author	Point sampling	Descriptor	Spatial pyramid	Mean average precision
This paper	Harris-Laplace, dense sampling	SIFT	1x1+2x2+1x3	0.558
This paper	Harris-Laplace, dense sampling	C-SIFT	1x1+2x2+1x3	0.566
Marszałek *et al.* [38]	Harris-Laplace, dense sampling, Laplacian	SIFT, HueSIFT, other	1x1+2x2+1x3	0.575
Marszałek *et al.* [38]	Harris-Laplace, dense sampling, Laplacian	SIFT, HueSIFT, other; with feature selection	1x1+2x2+1x3	0.594
This paper	Harris-Laplace, dense sampling	SIFT, OpponentSIFT, rgSIFT, C-SIFT, RGB-SIFT	1x1+2x2+1x3	0.605

TABLE III
IN THIS TABLE, COMBINATIONS OF DESCRIPTORS ON THE VIDEO BENCHMARK ARE COMPARED TO THE BASELINE SET BY THE MEDIAMILL CHALLENGE [22] FOR THE 39 LSCOM-LITE CATEGORIES [35]. ADDING COLOR DESCRIPTORS IMPROVES OVER INTENSITY-BASED SIFT ALONE BY 7%.

Combinations on video benchmark

Author	Point sampling	Descriptor	Spatial pyramid	Mean average precision
Snoek *et al.* [22]	Grid	Weibull [39]	1x1	0.250
This paper	Harris-Laplace, dense sampling	SIFT	1x1+2x2+1x3	0.476
This paper	Harris-Laplace, dense sampling	OpponentSIFT	1x1+2x2+1x3	0.494
This paper	Harris-Laplace, dense sampling	SIFT, OpponentSIFT, rgSIFT, C-SIFT, RGB-SIFT	1x1+2x2+1x3	0.510

TABLE IV
IN THIS TABLE, RESULTS OF DESCRIPTOR COMBINATIONS FROM THIS PAPER AS SUBMITTED TO THE CLASSIFICATION TASK OF THE PASCAL VOC CHALLENGE 2008 [40] ARE SHOWN.

PASCAL VOC 2008 evaluation: best overall performance

Author	Point sampling	Descriptor	Spatial pyramid	Mean average precision
This paper and Tahir *et al.* [41]	Harris-Laplace, dense sampling	SIFT, OpponentSIFT, rgSIFT, C-SIFT, RGB-SIFT	1x1+2x2+1x3	0.549

TABLE V
IN THIS TABLE, RESULTS OF DESCRIPTOR COMBINATIONS FROM THIS PAPER AS SUBMITTED TO THE NIST TRECVID 2008 VIDEO BENCHMARK [7] ARE SHOWN.

NIST TRECVID 2008 evaluation: best overall performance

Author	Point sampling	Descriptor	Spatial pyramid	Inferred mean average precision
This paper and Snoek *et al.* [43]	Harris-Laplace, dense sampling	SIFT, OpponentSIFT, rgSIFT, C-SIFT, RGB-SIFT	1x1+2x2+1x3	0.194

these submissions. For additional details, see table IV [41], [42] and table V [43].

E. Discussion

Using the ALOI dataset, the theoretical invariance properties of color descriptors were verified experimentally. However, possessing invariance properties alone is not sufficient to address category recognition: the descriptor should also be distinctive and robust to compression artifacts. Several histogram-based descriptors and color moment invariants were found to be sensitive to even moderate amounts of compression, thereby reducing their usefulness. On the other hand, the results show that the SIFT descriptor and most color extensions of the SIFT descriptor are robust to compression artifacts. Also, these SIFT-based descriptors outperform histogram-based and moment-based descriptors on both image and video category recognition. Therefore, the rest of this discussion will focus on the properties of these descriptors in particular.

The results on two category recognition benchmarks show that SIFT-based descriptors which perform well are all invariant to light intensity changes. For light intensity shifts, the usefulness of invariance depends on the object or scene category. For those categories in real-world datasets where large variations in lighting conditions occur frequently, invariance to light intensity shifts is useful. Examples for the image benchmark are shown in figure 9: normally, sofas are found indoor. However, the dataset contains samples where the sofa is photographed outside on the street. As the ranking positions show, the OpponentSIFT descriptor, which is invariant to both light intensity changes and shifts, places these samples higher in the ranking. However, the converse also occurs, as the example of the potted plants shows. The descriptors

Positions in Rankings for Image Benchmark

Color Descriptor	Sofa 	Sofa 	Bus 	Bus 
OpponentSIFT	**769**	**1053**	**21**	190
C-SIFT	1782	2813	103	591
rgSIFT	3075	1445	161	486
RGB-SIFT	1917	3522	**6**	**11**

Color Descriptor	Potted Plant 	Potted Plant 	Potted Plant 
OpponentSIFT	194	709	1583
C-SIFT	**8**	**19**	**43**
rgSIFT	**10**	**18**	**63**
RGB-SIFT	264	2627	706

Fig. 9. From the PASCAL VOC Challenge 2007 [21], several positive examples for the object categories sofa, bus and potted plant are shown, together with their position in the ranked list of category recognition results for four different color descriptors. If, for one or more color descriptors, the ranked position is notably better than for the other color descriptors, it has been bold-faced. The ranking has 4952 elements.

Positions in Rankings for Video Benchmark

Color Descriptor	Building 	Building 	Vegetation 	Vegetation 
OpponentSIFT	**26**	**34**	1035	304
C-SIFT	677	2719	**53**	**1**
rgSIFT	1113	1512	**111**	**46**
RGB-SIFT	**102**	**35**	954	921

Fig. 10. From the Mediamill Challenge [22], several positive examples for the categories building and vegetation are shown, together with their position in the ranked list of category recognition results for four different color descriptors. If, for one or more color descriptors, the ranked position is notably better than for the other color descriptors, it has been bold-faced. The ranking has 12914 elements.

which are only scale-invariant place the samples higher in the ranking, and the shift-invariant OpponentSIFT and RGB-SIFT descriptors lag behind. For the video benchmark, figure 10 shows similar examples of both phenomena for buildings and vegetation.

From the results, it can be noticed that invariance to light color changes and shifts is domain-specific. For the image dataset, a significant reduction in performance was observed, whereas for the video dataset, there was no performance difference. However, there are specific samples where invariance to light color changes provides a benefit. An example is shown in figure 9 for busses: the bus illuminated by a setting sun benefits from light color invariance, as does the bus illuminated by red light tubes. Invariance to light intensity changes and shifts is not sufficient for the latter sample. However, the overall performance is not improved by light color invariance, presumably because light color changes are quite rare in both benchmarks due to the white balancing performed during data recording.

TABLE VI
THE RECOMMENDED CHOICE OF DESCRIPTORS FOR DIFFERENT DATASETS: THE PASCAL VOC 2007, MEDIAMILL CHALLENGE AND DATASETS WHERE NO PRIOR KNOWLEDGE ABOUT THE LIGHTING CONDITIONS OR THE OBJECT AND SCENE CATEGORIES IS AVAILABLE. WITHOUT SUCH PRIOR KNOWLEDGE, OPPONENTSIFT IS THE BEST CHOICE.

Recommended Color Descriptors Per Dataset

PASCAL VOC 2007	Mediamill Challenge	Unknown Data
1. C-SIFT	1. OpponentSIFT	1. OpponentSIFT
2. OpponentSIFT	2. RGB-SIFT	2. C-SIFT
3. RGB-SIFT	3. C-SIFT	3. RGB-SIFT
4. SIFT	4. SIFT	4. SIFT

Overall, when choosing a single descriptor and no prior knowledge about the dataset and object and scene categories is available, the best choice is OpponentSIFT. The correspondig invariance property is scale- and shift-invariance, given by eq. (10). Second best is C-SIFT for which the corresponding invariance property is scale-invariance, given by eq. (8). Table VI summarizes the recommendations for the datasets from this paper and datasets where no prior knowledge is available.

To obtain state-of-the-art performance on real-world datasets with large variations in lighting conditions, multiple color descriptors should be chosen, each one with a different amount of invariance. As shown earlier, even a simple combination of color descriptors improves over the individual descriptors, suggesting that they are not completely redundant. This is illustrated by the keyframes shown in figure 10: depending on the visual category, the OpponentSIFT and C-SIFT descriptors both show their strong points. Results on the two categorization benchmarks have shown that the choice of a single descriptor for all categories is suboptimal (see figure 7). While the addition of color improves category recognition by 8–10% over intensity-based SIFT only, further gains should be possible if the descriptor with the appropriate amount of invariance is selected per category, using either a feature selection strategy or domain knowledge.

VI. Conclusion

In this paper, the invariance properties of color descriptors are studied using a taxonomy of invariance with respect to photometric transformations, see table I for an overview. These invariance properties were validated using a dataset with known photometric changes. In addition, the distinctiveness of color descriptors is assessed experimentally using two benchmarks from the image domain and the video domain. On these benchmarks, the addition of color descriptors over SIFT improves category recognition by 8% and 7%, respectively.

From the theoretical and experimental results, it can be derived that invariance to light intensity changes and light color changes affects object and scene category recognition. The results reveal further that, for light intensity shifts, the usefulness of invariance is category-specific. Therefore, a color descriptor with an appropriate level of invariance should be selected for automated recognition of individual object and scene categories. Overall, when choosing a single descriptor and no prior knowledge about the dataset and object and scene categories is available, the OpponentSIFT is recommended. Finally, a proper combination of color descriptors improves over the individual descriptors.

Acknowledgments

This work was supported by the EC-FP6 VIDI-Video project.

References

[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 1–60, 2008.

[2] R. Fergus, F.-F. Li, P. Perona, and A. Zisserman, "Learning object categories from Google's image search," in *IEEE International Conference on Computer Vision*, Beijing, China, 2005, pp. 1816–1823.

[3] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories." in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, New York, USA, 2006, pp. 2169–2178.

[4] J. Vogel and B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *International Journal of Computer Vision*, vol. 72, no. 2, pp. 133–157, 2007.

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

[6] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, and J. Luo, "Large-Scale Multimodal Semantic Concept Detection for Consumer Video," in *ACM International Workshop on Multimedia Information Retrieval*, Augsburg, Germany, 2007, pp. 255–264.

[7] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *ACM International Workshop on Multimedia Information Retrieval*, Santa Barbara, USA, 2006, pp. 321–330.

[8] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *ACM International Conference on Image and Video Retrieval*, Amsterdam, The Netherlands, 2007, pp. 494–501.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints." *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1-2, pp. 43–72, 2005.

[11] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: A survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2008.

[12] A. E. Abdel-Hakim and A. A. Farag, "CSIFT: A SIFT descriptor with color invariant characteristics," in *IEEE Conference on Computer Vision and Pattern Recognition*, New York, USA, 2006, pp. 1978–1983.

[13] J. M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts, "Color invariance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 12, pp. 1338–1350, 2001.

[14] J. van de Weijer, T. Gevers, and A. Bagdanov, "Boosting color saliency in image feature detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 150–156, 2006.

[15] G. J. Burghouts and J. M. Geusebroek, "Performance evaluation of local color invariants," *Computer Vision and Image Understanding*, vol. 113, pp. 48–62, 2009.

[16] A. Bosch, A. Zisserman, and X. Muoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 04, pp. 712–727, 2008.

[17] G. D. Finlayson, M. S. Drew, and B. V. Funt, "Spectral sharpening: sensor transformations for improved color constancy," *Journal of the Optical Society of America A*, vol. 11, no. 5, p. 1553, 1994.

[18] J. von Kries, "Influence of adaptation on the effects produced by luminous stimuli," *In MacAdam, D.L. (Ed.), Sources of Color Vision. MIT Press, Cambridge, MS.*, 1970.

[19] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluation of color descriptors for object and scene recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, USA, June 2008.

[20] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, "The Amsterdam library of object images," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 103–112, 2005.

[21] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results." [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2007/

[22] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *ACM International Conference on Multimedia*, Santa Barbara, USA, 2006, pp. 421–430.

[23] M. Shafer, "Using color to seperate reflection components," *Color Research and Applications*, vol. 10, no. 4, pp. 210–218, 1985.

[24] G. D. Finlayson, S. D. Hordley, and R. Xu, "Convex programming colour constancy with a diagonal-offset model," in *IEEE International Conference on Image Processing*, 2005, pp. 948–951.

[25] T. Gevers, J. van de Weijer, and H. Stokman, *Color image processing: methods and applications: color feature detection: an overview*. CRC press, 2006, ch. 9, pp. 203–226.
[26] F. Mindru, T. Tuytelaars, L. Van Gool, and T. Moons, "Moment invariants for recognition under changing viewpoint and illumination," *Computer Vision and Image Understanding*, vol. 94, no. 1-3, pp. 3–27, 2004.
[27] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761 – 767, 2004.
[28] P.-E. Forssén, "Maximally stable colour regions for recognition and matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, USA, June 2007.
[29] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision*, Nice, France, 2003, pp. 1470–1477.
[30] T. K. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *International Journal of Computer Vision*, vol. 43, no. 1, pp. 29–44, 2001.
[31] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2003, pp. 264–271.
[32] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition." in *IEEE International Conference on Computer Vision*, Beijing, China, 2005, pp. 604–610.
[33] B. Leibe and B. Schiele, "Interleaved object categorization and segmentation," in *British Machine Vision Conference*, Norwich, UK, 2003, pp. 759–768.
[34] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[35] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE Multimedia*, vol. 13, no. 3, pp. 86–91, 2006.
[36] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, August 2006.
[37] B. Efron, "Bootstrap methods: Another look at the jackknife," *Annals of Statistics*, vol. 7, pp. 1–26, 1979.
[38] M. Marszałek, C. Schmid, H. Harzallah, and J. van de Weijer, "Learning object representations for visual object class recognition," 2007, Visual Recognition Challenge workshop, in conjunction with IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil. [Online]. Available: http://lear.inrialpes.fr/pubs/2007/MSHV07
[39] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, C. G. M. Snoek, and A. W. M. Smeulders, "Robust scene categorization by learning image statistics in context," in *CVPR Workshop on Semantic Learning Applications in Multimedia (SLAM)*, 2006.
[40] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results." [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2008/
[41] M. A. Tahir, K. E. A. van de Sande, J. R. R. Uijlings, and *et al.* , "University of Amsterdam and University of Surrey at PASCAL VOC 2008," 2008, PASCAL Visual Object Classes Challenge Workshop, in conjunction with IEEE European Conference on Computer Vision, Marseille, France. [Online]. Available: http://staff.science.uva.nl/~ksande/pub/vandesande-pascalvoc2008.pdf
[42] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.
[43] C. G. M. Snoek, K. E. A. van de Sande, O. de Rooij, B. Huurnink, J. C. van Gemert, J. R. R. Uijlings, and *et al.* , "The MediaMill TRECVID 2008 semantic video search engine," in *Proceedings of the 6th TRECVID Workshop*, Gaithersburg, USA, November 2008.

Koen van de Sande Koen E.A. van de Sande received a BSc in Computer Science (2004), a BSc in Artificial Intelligence (2004) and a MSc in Computer Science (2007) from the University of Amsterdam, The Netherlands. Currently, he is pursuing the PhD degree at the University of Amsterdam. His research interests include computer vision, visual categorization, (color) image processing, statistical pattern recognition and large-scale benchmark evaluations. He is a co-organizer of the annual VideOlympics. He is a student member of the IEEE.

Theo Gevers Theo Gevers is an Associate Professor of Computer Science at the University of Amsterdam, The Netherlands. At the University of Amsterdam he is a teaching director of the MSc of Artificial Intelligence. He currently holds a VICI-award (for excellent researchers) from the Dutch Organisation for Scientific Research. His main research interests are in the fundamentals of content-based image retrieval, colour image processing and computer vision specifically in the theoretical foundation of geometric and photometric invariants. He is co-chair of the Internet Imaging Conference (SPIE 2005, 2006), co-organizer of the First International Workshop on Image Databases and Multi Media Search (1996), the International Conference on Visual Information Systems (1999, 2005), the Conference on Multimedia & Expo (ICME, 2005), and the European Conference on Colour in Graphics, Imaging, and Vision (CGIV, 2012). He is guest editor of the special issue on content-based image retrieval for the International Journal of Computer Vision (IJCV 2004) and the special issue on Colour for Image Indexing and Retrieval for the journal of Computer Vision and Image Understanding (CVIU 2004). He has published over 100 papers on colour image processing, image retrieval and computer vision. He is program committee member of a number of conferences, and an invited speaker at major conferences. He is a lecturer of post-doctoral courses given at various major conferences (CVPR, ICPR, SPIE, CGIV). He is member of the IEEE.

Cees Snoek Cees G.M. Snoek received the MSc degree in business information systems (2000) and the PhD degree in computer science (2005) both from the University of Amsterdam, where he is currently a senior researcher at the Intelligent Systems Lab. He was a visiting scientist at Carnegie Mellon University, in 2003. His research interests focus on visual categorization, statistical pattern recognition, social media retrieval, and large-scale benchmark evaluations, especially when applied in combination for video search. He has published over 70 refereed book chapters, journal and conference papers in these fields, and serves on the program committee of several conferences. Dr. Snoek is the lead researcher of the award-winning MediaMill Semantic Video Search Engine, which is a consistent top performer in the yearly NIST TRECVID evaluations. He is co-initiator and co-organizer of the annual VideOlympics and a lecturer of post-doctoral courses given at international conferences and summer schools. He is a member of the IEEE. Dr. Snoek received a young talent (VENI) grant from the Netherlands Organization for Scientific Research in 2008.