Evaluation of Color Descriptors for Object and Scene Recognition

Koen E. A. van de Sande and Theo Gevers and Cees G. M. Snoek
ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
{ksande,gevers,cgmsnoek}@science.uva.nl

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. To increase illumination invariance and discriminative power, color descriptors have been proposed only recently. As many descriptors exist, a structured overview of color invariant descriptors in the context of image category recognition is required.

Therefore, this paper studies the invariance properties and the distinctiveness of color descriptors in a structured way. The invariance properties of color descriptors are shown analytically using a taxonomy based on invariance properties with respect to photometric transformations. The distinctiveness of color descriptors is assessed experimentally using two benchmarks from the image domain and 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 changes, the usefulness of invariance is category-specific.

1. 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 in both image retrieval [10, 22, 24] and video retrieval [3, 19] use machine learning based on image descriptions to distinguish object and scene categories. However, there can be large variations in lighting and viewing conditions for real-world scenes, complicating the description of images. A change in viewpoint will yield shape variations such as the orientation and scale of the object in the image plane. Salient point detection methods and corresponding region descriptors can robustly detect regions which are translation-, rotation- and scale-invariant, addressing the problem of viewpoint changes [11, 13]. In addition, changes in the illumination of a scene can greatly affect the performance of object recognition if the descriptors used are not robust to these changes. To increase illumination invariance and discriminative power, color descriptors have been proposed [2, 8, 21]. However, as there are many different methods, a taxonomy is required based on principles of illumination changes to arrange color invariant descriptors in the context of image category recognition.

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 [6, 23]. 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. Then, the *distinctiveness* of color descriptors is analyzed experimentally using two benchmarks from the image domain and the video domain. The benchmarks are very different in nature: the image benchmark consists of photographs and the video benchmark consists of news broadcast videos. Based on extensive experiments on a large set of real-world images and videos, the usefulness of the different invariant properties can be derived.

This paper is organized as follows. In section 2, the reflectance model is presented. Further, its relation to the diagonal model of illumination change is discussed. In section 3, a taxonomy is given of color descriptors and their invariance properties. The experimental setup is presented in section 4. In section 5, a discussion of the results is given. Finally, in section 6, conclusions are drawn.

2. Reflectance Model

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

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

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

where $e(\lambda)$ is the color of the light source, $s(\mathbf{x}, \lambda)$ is de 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 [17] proposes to add a 'diffuse' light term to the model of eq. (1). The diffuse light is considered to have a lower intensity and originates from all directions in equal amounts:

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

where $a(\lambda)$ is the term that models the diffuse light. For example, using this equation, objects under daylight should be better modelled, since daylight consists of both a point source (the sun) and diffuse light coming from the sky.

By computing the derivative of image $\mathbf{f}$, it can be easily derived that the effect of the diffuse light source $a(\lambda)$ as in eq. (2) 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, simply taking derivatives will yield invariance to diffuse light. This reflection model corresponds to the diagonal model under the assumption of narrow band filters. This is detailed in the next section.

2.1. Diagonal Model

Changes in the illumination can be modeled by a diagonal mapping or *von Kries Model* [23]. The diagonal mapping is given 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:

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

However, under some conditions, the diagonal model is too simplistic. For example, a shift in the color values due to increased 'diffuse' light cannot be modelled. To overcome this, Finlayson *et al*. [7] extended the diagonal model with an offset, resulting in the diagonal-offset model:

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

Deviations from the diagonal model are reflected in the offset term $(o_1, o_2, o_3)^T$. The diagonal model with offset term corresponds to eq. (2) assuming narrow-band filters. For broad-band cameras, spectral sharpening can be applied to obtain this [6]. 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.

2.2. Photometric Analysis

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

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*:

$$\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & a \end{pmatrix} \begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix}. \quad (7)$$

Light intensity changes include shadows and lighting geometry changes such as shading. Hence, when a descriptor is invariant to light intensity changes, it is *scale-invariant* with respect to (light) intensity.

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:

$$\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix} + \begin{pmatrix} o_1 \\ o_1 \\ o_1 \end{pmatrix}. \quad (8)$$

Light intensity shifts correspond to object highlights under a white light source and scattering of a white source. When a descriptor is invariant to a light intensity shift, it is *shift-invariant* with respect to light intensity.

The above classes of changes can be combined to model both intensity changes and shifts:

$$\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & a \end{pmatrix} \begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix} + \begin{pmatrix} o_1 \\ o_1 \\ o_1 \end{pmatrix}; \quad (9)$$

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

In the full diagonal model (*i.e*. allowing $a \neq b \neq c$), the image channels scale independently; this is equal to 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.

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*.

	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	+/-	+	+/-	+/-	+/-
W-SIFT	+	+	+	+/-	+/-
rgSIFT	+	+	+	+/-	+/-
Transf. color SIFT	+	+	+	+	+

Table 1. Invariance of descriptors (section 3) against types of changes in the diagonal-offset model and its specializations (section 2.2). Invariance is indicated with '+', lack of invariance is indicated with '-'. A '+/-' indicates that the intensity SIFT part of the descriptor is invariant, but the color part is not.

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 color changes and light color changes and shifts.

3. 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. See table 1 for an overview of the descriptors and their invariance properties.

3.1. 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, see table 1.

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

$$\begin{pmatrix} O1 \\ O2 \\ O3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix}. \tag{10}$$

The intensity is represented in channel $O3$ and the color information is in channels $O1$ and $O2$. Due to the subtraction in $O1$ and $O2$, the offsets will cancel out if they are equal for all channels (e.g. a white light source). Therefore, these color models are shift-invariant with respect to light intensity. The intensity channel $O3$ has no invariance properties. The histogram intervals for the opponent color space have ranges different from the RGB model.

Hue histogram In the HSV color space, it is known that the hue becomes unstable around the grey axis. To this end, Van de Weijer *et al*. [21] apply an error analysis to the hue. 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 and the S color models are scale-invariant and shift-invariant with respect to light intensity.

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

$$\begin{pmatrix} r \\ g \\ b \end{pmatrix} = \begin{pmatrix} \frac{R}{R+G+B} \\ \frac{G}{R+G+B} \\ \frac{B}{R+G+B} \end{pmatrix}. \tag{11}$$

Because of the normalization, r and g are scale-invariant and thereby invariant to light intensity changes, shadows and shading [9].

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

$$\begin{pmatrix} R' \\ G' \\ B' \end{pmatrix} = \begin{pmatrix} \frac{R-\mu_R}{\sigma_R} \\ \frac{G-\mu_G}{\sigma_G} \\ \frac{B-\mu_B}{\sigma_B} \end{pmatrix}, \tag{12}$$

with μ_C the mean and σ_C the standard deviation of the distribution in channel C. This yields for every channel a distribution where $\mu = 0$ and $\sigma = 1$.

3.2. 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*. [14] have defined *generalized color moments* M_{pq}^{abc}:

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

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 $M_{pq}^{101\dagger}$. 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*. [14]. 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 1.

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

3.3. Color SIFT Descriptors

SIFT The SIFT descriptor proposed by Lowe [11] 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 2.2). 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. Light color changes have no effect on the descriptor because the input image is converted to grayscale, after which the intensity scale-invariance argument applies. To compute SIFT descriptors, the version described by Lowe [11] is used.

HSV-SIFT Bosch *et al*. [2] compute SIFT descriptors over all three channels of the HSV color model, instead of over the intensity channel only. This gives 3x128 dimensions per descriptor, 128 per channel. Drawback of this approach is that the instability of the hue for low saturation is ignored.

The properties of the H and the S channels also apply to this descriptor: it is scale-invariant and shift-invariant. However, the H and the S SIFT descriptors are not invariant to light color changes; only the intensity SIFT descriptor (V channel) is invariant to this. Therefore, the descriptor is only partially invariant to light color changes.

HueSIFT Van de Weijer *et al*. [21] introduce a concatenation of the hue histogram (see section 3.1) with the SIFT descriptor. When compared to HSV-SIFT, the usage of the weighed hue histogram addresses the instability of the hue around the grey axis. Because the bins of the hue histogram are independent, there are no problems with the periodicity of the hue channel for HueSIFT. Similar to the hue histogram, the HueSIFT descriptor is scale-invariant and shift-invariant. However, only the SIFT component of this descriptor is invariant to illumination color changes or shifts; the hue histogram is not.

OpponentSIFT OpponentSIFT describes all the channels in the opponent color space (eq. (10)) 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. However, these other channels do contain some intensity information: hence they are not invariant to changes in light intensity.

W-SIFT In the opponent color space (eq. (10)), the O_1 and O_2 channels still contain some intensity information. To add invariance to intensity changes, [8] proposes the W invariant which eliminates the intensity information from these channels. The W-SIFT descriptor uses the W invariant, which can be defined for the opponent color space as $\frac{O_1}{O_3}$ and $\frac{O_2}{O_3}$. Because of the division by intensity, the scaling in the diagonal model will cancel out, making W-SIFT scale-invariant with respect to light intensity. As for the other colorSIFT descriptors, the color component of the de-

scriptor is not invariant to light color changes.

***rg*SIFT** For the rgSIFT descriptor, descriptors are added for the r and g chromaticity components of the normalized RGB color model from eq. (11), which is already scale-invariant. Because the SIFT descriptor uses derivatives of the input channels, the rgSIFT descriptor becomes shift-invariant as well. However, the color part of the descriptor is not invariant to changes in illumination color.

Transformed color SIFT For the transformed color SIFT, the same normalization is applied to the RGB channels as for the transformed color histogram (eq. (12)). For every normalized channel, the SIFT descriptor is computed. The descriptor is scale-invariant, shift-invariant and invariant to light color changes and shift.

4. Experimental Setup

In this section, the experimental setup to evaluate the different color descriptors is outlined. First, implementation details of the descriptors in an object and scene recognition setting are discussed. Then, the two benchmarks used for evaluation are described: an image benchmark and a video benchmark. After discussing these benchmarks and their datasets, evaluation criteria are given.

4.1. Implementation

To emperically test color descriptors, these descriptors are used inside local features based on scale-invariant points [11, 24]. Scale-invariant points are obtained with the Harris-Laplace point detector [13]. 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 3 are computed over the area around the points. To achieve comparable descriptors for different scales, all regions are proportionally resampled to a uniform patch size of 60 by 60 pixels.

The method by Zhang *et al*. [24] is used to learn object appearance models from region descriptors: the descriptors present in an image are reduced to 40 clusters using the k-means algorithm. Then, the Earth Movers Distance [16] between the cluster sets of different images is used in the kernel function of the Support Vector Machine algorithm.

For the video benchmark, the χ^2 distance is used in the kernel function of the Support Vector Machine algorithm, because it requires significantly less computational time, while offering performance similar to the Earth Movers Distance [24]. Because the χ^2 distance requires fixed-length feature vectors, the visual codebook model is used as described in [18], amongst others. The visual codebook with 4000 elements is constructed using k-means clustering.

Figure 1. Object categories of the PASCAL Visual Object Challenge 2007 [5], used in the image benchmark of experiment 1.

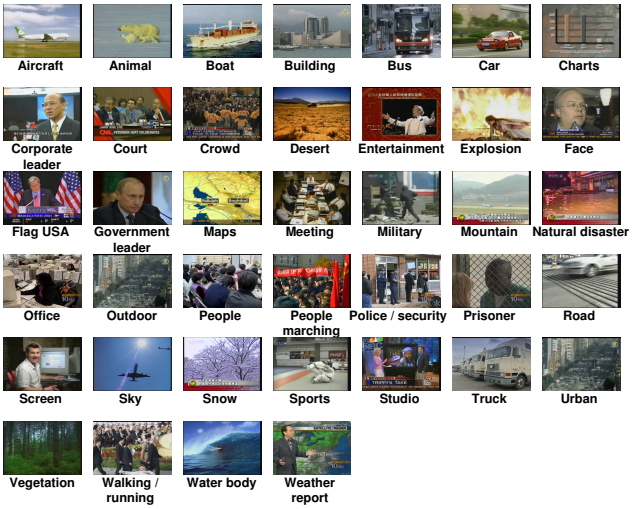

Figure 2. Object and scene categories of the LSCOM-Lite concept ontology [15], used in the video benchmark of experiment 2.

4.2. Experiment 1: Image Benchmark

The PASCAL Visual Object Classes Challenge [5] 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 (see figure 1), 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).

4.3. Experiment 2: Video Benchmark

The Mediamill Challenge by Snoek *et al*. [20] provides an annotated video dataset, based on the training set of the NIST TRECVID 2005 benchmark [19]. 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 30993 shots) and a Challenge test set (30% of the data or 12914 shots). For every shot, the Challenge provides a single representative keyframe im-

age. So, the complete dataset consists of 43907 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 [15] are employed, listed in figure 2.

4.4. Evaluation Criteria

The average precision is taken as the performance metric for determining the accuracy of ranked category recognition results, following the standard set in the PASCAL VOC Challenge 2007 and TRECVID. 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 shots judged relevant. Let $\rho^k = \{l_1, l_2, ..., l_k\}$ be the ranked list of items from test set A. At any given rank k, let $|R \cap \rho^k|$ be the number of relevant shots in the top k of ρ, where R is the set of relevant shots and $|X|$ is the size of set X. Average precision, AP, is then defined as:

$$AP(\rho) = \frac{1}{|R|} \sum_{k=1}^{|A|} \frac{|R \cap \rho^k|}{k} \psi(l_k) \qquad (13)$$

with indicator function $\psi(l_k) = 1$ if $l_k \in R$ and 0 otherwise. $|A|$ is the size of the answer set, *e.g.* the number of items present in the ranking. When performing experiments over multiple object classes, the average precisions of the individual classes can be aggregated. This aggregation is called mean average precision (MAP). MAP is calculated by taking the mean of the average precisions. Note that MAP depends on the dataset used: scores of different datasets are not easily comparable.

To obtain an indication of significance, the bootstrap method [1, 4] is used to estimate confidence intervals for MAP. In bootstrap, multiple test sets A_B are created by selecting images at random from the original test set A, with replacement, until $|A| = |A_B|$. This has the effect that some images are replicated in A_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 A. The statistical accuracy of the MAP score can then be evaluated by looking at the standard deviation of the MAP scores over the different bootstrap test sets.

5. Results

5.1. Experiment 1: Image Benchmark

From the results shown in figure 3, it is observed that 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.

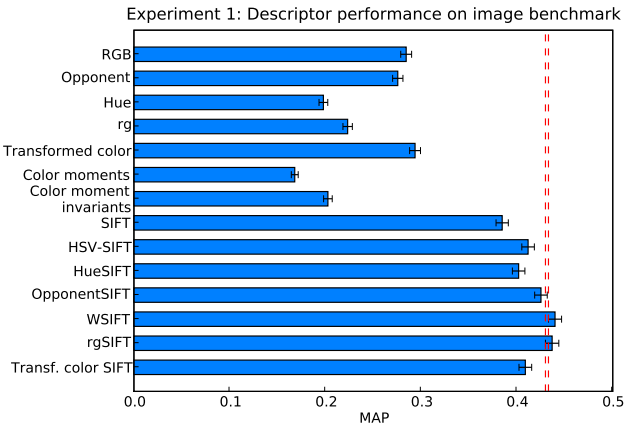

Figure 3. Evaluation of color descriptors on an image benchmark, the PASCAL VOC Challenge 2007 [5], averaged over the 20 object categories from figure 1. Error bars indicate the standard deviation in MAP, obtained using bootstrap. The dashed lines indicate the lower bound of the rgSIFT and WSIFT confidence intervals.

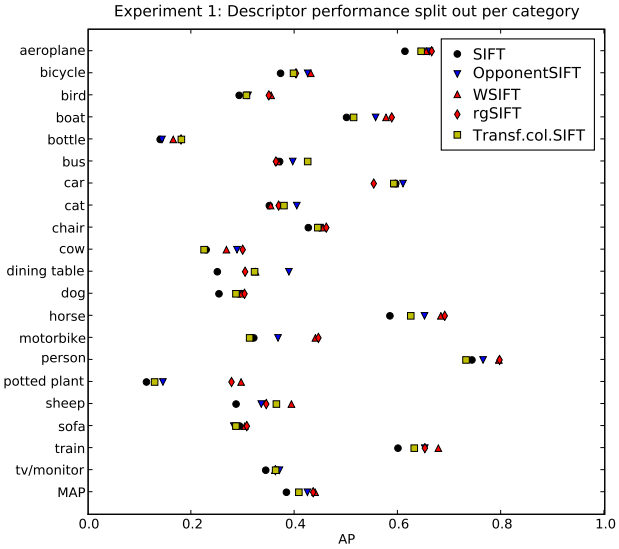

Figure 4. Evaluation of color descriptors on an image benchmark, the PASCAL VOC Challenge 2007, split out per object category. SIFT and the best four color SIFT variants from figure 3 are shown.

For SIFT and the four best color SIFT descriptors from figure 3 (OpponentSIFT, WSIFT, rgSIFT and transformed color SIFT), the results per object category are shown in figure 4. For bird, horse, motorbike, person and potted plant, it can be observed that the descriptors which perform best have scale-invariance and shift-invariance for light intensity (WSIFT and rgSIFT). The performance of the OpponentSIFT descriptor, which lacks scale-invariance compared to WSIFT, yields that scale-invariance, i.e. invariance to light intensity changes is important for these object cate-

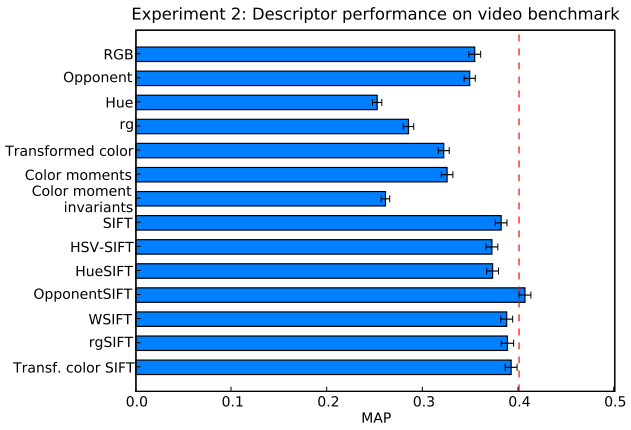

Figure 5. Evaluation of color descriptors on a video benchmark, the Mediamill Challenge [20], averaged over 39 object and scene categories from figure 2. Error bars indicate the standard deviation in MAP, obtained using bootstrap. The dashed line indicates the lower bound of the OpponentSIFT confidence interval.

gories. Transformed color SIFT includes additional invariance against light color changes and shifts when compared to WSIFT and rgSIFT. However, this additional invariance makes the descriptor less discriminative for these object categories, because a significant reduction in performance is observed.

In conclusion, WSIFT and rgSIFT are significantly better than all other descriptors except OpponentSIFT (see figure 3) on the image benchmark. The corresponding invariant property is given by eq. (9). Second-best is OpponentSIFT, which corresponds to eq. (8).

5.2. Experiment 2: Video Benchmark

From the results shown in figure 5, the same overall pattern as for the image benchmark is observed: SIFT and color SIFT variants perform significantly better than the other descriptors. However, the scale-invariant WSIFT and rgSIFT are no longer the best descriptors. Instead, the scale-variant OpponentSIFT is now significantly better than all other descriptors. An analysis on the individual object and scene categories shows that the OpponentSIFT descriptor performs best for building, outdoor, sky, studio, walking/running and weather news. All these concepts either occur indoor or outdoor, but not both. Therefore, the intensity information present in the OpponentSIFT is very distinctive for these categories. This explains why OpponentSIFT is slightly better than WSIFT and rgSIFT for the video benchmark, instead of the other way around like in the image benchmark. Transformed color SIFT, with additional invariance to light color changes and shifts, does not differ significantly from WSIFT 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 5). The corresponding invariant property is given by eq. (8).

5.3. Discussion

The results show that, for light intensity changes, the usefulness of invariance depends on the object or scene category. Because the lighting information itself is discriminative, categories which only appear under certain lighting conditions do not benefit from invariance. However, for most categories, invariance to light intensity changes is useful, because in real-world datasets there are often large variations in lighting conditions.

Because almost all color descriptors are shift-invariant, the effect of light intensity variations on the performance cannot be observed easily. The color descriptors which are sensitive to light intensity shifts are the three color histograms. Given that SIFT and its color variants show best performance, it can be derived that shift-invariance has no adverse effects on performance.

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.

In conclusion, when no prior knowledge about the dataset and object and scene categories is available, the best choice for a color descriptor is OpponentSIFT. The correspondig invariance property is shift-invariance, given by eq. (8). When such knowledge is available, WSIFT and rgSIFT are potentially better choices. The correspondig invariance property is scale-invariance, given by eq. (9).

5.4. Combinations

So far, the performance of single descriptors has been analyzed. It is worthwhile to investigate combinations of several descriptors, since they are potentially complementary. State-of-the-art results on the PASCAL VOC Challenge 2007 also employ combinations of several methods. For example, the best entry in the PASCAL VOC Challenge 2007, by Marszałek *et al*. [12], has achieved an MAP of 0.594 using SIFT and HueSIFT descriptors, the spatial pyramid [10], additional point sampling strategies besides Harris-Laplace such as Laplacian point sampling and dense sampling, and an advanced fusion scheme. When the advanced fusion scheme is excluded and simple flat fusion is used, Marszałek reports an MAP of 0.575.

To illustrate the potential of the color descriptors from table 1, a simple flat fusion experiment has been performed with SIFT and the best four color SIFT variants. To be comparable, a setting similar to Marszałek is used: both

Harris-Laplace point sampling and dense sampling are employed, using the spatial pyramid up to level 1 and the χ^2 SVM kernel. In this setting, the color descriptors achieve an MAP≈0.50. The combination gives an MAP of 0.562 ($\sigma = 0.007$). This convincing gain suggests that the color descriptors are complementary. Otherwise, overall performance would not have improved significantly. Further gains should be possible, if the descriptors with the right amount of invariance are fused, preferably using an automatic selection strategy.

6. Conclusion

In this paper, the invariance properties of color descriptors are studied using a taxonomy of invariance with respect to photomeric transformations, see table 1 for an overview. The distinctiveness of color descriptors is assessed experimentally using two benchmarks from the image domain and the video domain.

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 show further that, for light intensity changes, the usefulness of invariance is category-specific.

Acknowledgments

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

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, August 2006.

[2] A. Bosch, A. Zisserman, and X. Muoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Trans. Pattern Analysis and Machine Intell.*, 30(04):712–727, 2008.

[3] 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 *Proc. ACM Multimedia Information Retrieval*, pages 255–264, Augsburg, Germany, 2007.

[4] B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

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

[6] G. D. Finlayson, M. S. Drew, and B. V. Funt. Spectral sharpening: sensor transformations for improved color constancy. *J. Optical Society of America A*, 11(5):1553, 1994.

[7] G. D. Finlayson, S. D. Hordley, and R. Xu. Convex programming colour constancy with a diagonal-offset model. In *IEEE Int. Conf. on Image Processing*, pages 948–951, 2005.

[8] J. M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts. Color invariance. *IEEE Trans. Pattern Analysis and Machine Intell.*, 23(12):1338–1350, 2001.

[9] T. Gevers, J. van de Weijer, and H. Stokman. *Color image processing: methods and applications: color feature detection*, chapter 9, pages 203–226. CRC press, 2006.

[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conf. Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178, New York, USA, 2006.

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int'l. J. of Computer Vision*, 60(2):91–110, 2004.

[12] 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 ICCV, Rio de Janeiro, Brazil.

[13] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Int'l. J. of Computer Vision*, 65(1-2):43–72, 2005.

[14] F. Mindru, T. Tuytelaars, L. Van Gool, and T. Moons. Moment invariants for recognition under changing viewpoint and illumination. *Computer Vision and Image Understanding*, 94(1-3):3–27, 2004.

[15] M. R. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, and A. Hauptmann. A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005. Technical report, IBM, 2005.

[16] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *IEEE Int. Conf. on Computer Vision*, pages 59–66, 1998.

[17] M. Shafer. Using color to seperate reflection components. *Color Research and Applications*, 10(4):210–218, 1985.

[18] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE Int. Conf. on Computer Vision*, pages 1470–1477, Nice, France, 2003.

[19] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *Proc. ACM Multimedia Information Retrieval*, pages 321–330, Santa Barbara, USA, 2006.

[20] 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 *Proc. ACM Multimedia*, Santa Barbara, USA, 2006.

[21] J. van de Weijer, T. Gevers, and A. Bagdanov. Boosting color saliency in image feature detection. *IEEE Trans. Pattern Analysis and Machine Intell.*, 28(1):150–156, 2006.

[22] 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 Appl. in Multimedia*, 2006.

[23] 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*, 1970.

[24] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *Int'l. J. of Computer Vision*, 73(2):213–238, 2007.