Invariant color descriptors for efficient object recognition

van de Sande, K.E.A.

Publication date
2011

Chapter 3

Illumination-Invariant Descriptors for Discriminative Visual Object Categorization*

3.1 Introduction

Illumination-invariant descriptors are important for a number of applications, such as object category recognition and scene classification. The goal of object category recognition is to determine the presence or absence of objects of a certain class within an image. The popular bag-of-words model provides high detection accuracy [36]. The different stages of the bag-of-words model are well-studied: first, there is the point sampling strategy [80, 108]. Recent benchmark results [35, 36] and [65] show that densely sampling in a regular grid is the preferred approach for object recognition. Region descriptors [82] are then computed over the area around these points. The SIFT [75] and SURF [6] descriptors are the most well-known region descriptors. Because changes in the *illumination* of a scene can greatly affect the performance of object recognition, the descriptors need to be robust to these changes. Finally, the point descriptors are vector-quantized against a codebook of prototypical descriptors [40, 65, 71, 72, 95, 109].

Object classification is largely improved by descriptor design [24, 82, 107, 112]. In [112], several (color) SIFT descriptors are evaluated for object recognition under varying imaging conditions: OpponentSIFT [10], C-SIFT [1], HSV-SIFT [9], HueSIFT [118], RGB-SIFT [112] and SIFT [75]. Of these descriptors, three are at least invariant to light intensity changes and shifts. However, the number of these descriptors is limited because only a limited set (usually 1 or 3) is used of predefined color models such as HSV, $O_1O_2O_3$ and RGB. For object category recognition, in general, a single set of descriptors is employed for all object categories [36]. A predefined set of color channels may negatively influence the discriminative power of visual object categorization.

Therefore, in this chapter, the aim is to generate a general class of discriminative, illumination-

*Submitted to *International Journal of Computer Vision* [113]

invariant descriptors for object category recognition which go beyond existing color models. We design this new class of illumination-invariant descriptors based on SIFT, with the aim to remain invariant to light intensity changes and shifts under different normalizations. We propose to select discriminative descriptors based on learning, by selecting the best descriptor through cross-validation and by multiple kernel learning (MKL) [47, 103, 122] to learn weights for the different descriptors. The combination of newly developed descriptors and selection strategy is used to improve visual object categorization accuracy.

The contribution of this chapter is two-fold. First, we develop a new and general class of color descriptors which (a) is illumination-invariant (b) is not limited to existing color spaces (c) is intuitive and simple to compute and (d) allows emphasis on certain color models. Second, we adapt color descriptors to the object category by (a) selecting one descriptor per object category by learning or (b) learning the weights for a combination of color descriptors in a MKL framework.

We evaluate the color descriptors on the challenging PASCAL VOC object classification datasets, where the new class of color descriptors outperforms current descriptors in terms of classification performance. Another application is object localisation, where accuracy is increased by selecting a more discriminative color descriptor. Finally, we suggest using an extended opponent color descriptor instead of current color SIFT descriptors.

The chapter is organized as follows. In Section 3.2, the class of illumination-invariant SIFT descriptors is derived and instantiated. Section 3.3 outlines the methods and experimental setup. In Section 3.4, results of experiments with descriptor selection and optimization are detailed. Finally, in Section 3.5, conclusions are drawn.

3.2 Illumination-Invariant Descriptors

3.2.1 Introduction

In this section, we have propose a new class of color descriptors. Instead of using a limited and predefined set of color models for the descriptors, we identify a class of descriptors which are illumination-invariant to common photometric changes within the diagonal model of illumination change.

3.2.2 Diagonal Model

Photometric changes can be modeled by a diagonal mapping or *von Kries Model* [126] as follows:

$$\mathbf{f}^c = \mathcal{D}^{u,c}\mathbf{f}^u, \tag{3.1}$$

where $\mathbf{f}^u$ is the image taken under an unknown light source, $\mathbf{f}^c$ is the transformed image of the same scene, 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:

$$\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix}. \tag{3.2}$$

To include a 'diffuse' light term, Finlayson *et al*. [43] extended the diagonal model with an offset $(o_1, o_2, o_3)^T$. The 'diffuse' light term includes a wide range of possible causes besides diffuse light, such as interreflections, infrared sensitivity of the camera sensor, scattering in the medium or lens. The extended model results in the diagonal-offset model:

$$\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix} + \begin{pmatrix} o_1 \\ o_2 \\ o_3 \end{pmatrix}. \tag{3.3}$$

The diagonal model with offset term corresponds to the Lambertian reflectance model extended with a 'diffuse' lighting term [92], when assuming narrow-band filters. For broad-band cameras, spectral sharpening can be applied to obtain narrow-band filters [42].

In [107], the OSID descriptor is proposed, which is invariant to any monotonically increasing brightness change, while the diagonal model only handles linear changes. In practice, such nonlinear brightness changes are quite rare, except perhaps for gamma corrections. It has been shown empirically in [112], using a dataset with 1000 objects under known illumination conditions [48], that descriptors which are analytically illumination-invariant in the diagonal model are indeed robust to real-world lighting changes. Therefore, the diagonal model with offset is used in this chapter.

3.2.3 A Novel Class of Illumination-Invariant SIFT Descriptors

The Scale-Invariant Feature Transform (SIFT) descriptor by Lowe [75] measures the area around a sampling point and describes this area using an edge orientation histogram. The descriptor is computed from spatial and gradient information only: the spatial position of a pixel determines in which of the 4x4 spatial quadrants it falls, the gradient orientation determines in which of the 8 orientation bins a pixel falls. The weight of a pixel within the bin depends on the gradient magnitude at that pixel, plus a Gaussian weighting determined by the distance of the pixel to the descriptor center. The weight of a pixel can be distributed over multiple bins when it is close to the boundary of a quadrant or orientation bin; however, the total weight of the pixel will not change by this operation.

SIFT descriptors have been evaluated for object recognition [10, 112]: OpponentSIFT, C-SIFT, RGB-SIFT and standard intensity SIFT. These descriptors have different levels of invariance. In general, it is concluded that invariance to light intensity changes and light intensity shifts is most important for object recognition [112]. It is shown that intensity SIFT, OpponentSIFT and RGB-SIFT are invariant to these changes and shifts. Illumination changes are represented

within the diagonal model (3.3) as follows:

$$\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & a \end{pmatrix} \begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix} + \begin{pmatrix} o_1 \\ o_1 \\ o_1 \end{pmatrix}; \tag{3.4}$$

where the image values change by the same factor in all channels ($a = b = c$ and $o_1 = o_2 = o_3$ compared to (3.3)), *i.e.* intensity changes. In fact, SIFT, OpponentSIFT and RGB-SIFT descriptors are computed from (color) channels which are linear combinations of R, G and B. For example, Intensity$= 0.2126R + 0.7152G + 0.0722B$, $O_1 = \frac{R-G}{\sqrt{2}}$ and $O_2 = \frac{R+G-2B}{\sqrt{6}}$. The complete set of existing descriptors uses only the standard color spaces and therefore is limited in size. These standard color spaces use both linear and nonlinear color transformations. Because the above set of existing descriptors is limited to a few non-related descriptors, the question is whether there exists a larger class of illumination-invariant descriptors which can be generated in a systematic way.

To this end, we propose a new class of illumination-invariant descriptors based on SIFT: any linear combination of the R, G and B channels will result in an illumination-invariant SIFT descriptor. Further, the descriptors remain invariant under different normalizations.

To be precise, let the linear combination from R, G and B be defined as channel C:

$$C = dR + eG + fB, \tag{3.5}$$

where d, e and f are scalar values. To compute the SIFT descriptor on channel C, the gradient magnitude and orientation are required, *i.e.* first order derivatives. The first order derivative of C in direction x is:

$$C_x = \frac{\delta}{\delta x}(dR + eG + fB). \tag{3.6}$$

If we introduce light intensity changes and shifts from Eq. (3.4) by substitution, the following is obtained:

$$\begin{aligned} C^c_x &= \frac{\delta}{\delta x}(dR^c + eG^c + fB^c) \\ &= \frac{\delta}{\delta x}(d(aR^u + o_1) + e(aG^u + o_1) + f(aB^u + o_1)) \\ &= \frac{\delta}{\delta x}(a(dR^u + eG^u + fB^u) + o_1(d + e + f))) \\ &= a\frac{\delta}{\delta x}(dR^u + eG^u + fB^u) = aC^u_x. \end{aligned} \tag{3.7}$$

Obviously, the offset is cancelled out by taking the derivative and the derivative is scaled by a factor a. The gradient angle determines in which orientation bin of the SIFT descriptor a pixel falls. The gradient angle is $\arctan \frac{C_y}{C_x}$. Combined with Eq. (3.7), the scaling factor a is cancelled out. Therefore, the gradient orientation does not change for light intensity changes and shifts. Each pixel will fall in the same SIFT descriptor bin for both the canonical and the unknown illuminants.

The weight of a pixel within the SIFT descriptor depends on two factors: the distance to the center of the descriptor, which is used for the Gaussian weighting function, and the gradient magnitude of the pixel. When the same image region is described, but under different illuminants, the same spatial sampling grid is used for both and therefore the same Gaussian weighting is obtained. Then, the question is how the scaling of the first order derivatives by a (*i.e.* Eq. (3.7)) influences the gradient magnitude $||\nabla C|| = \sqrt{C_x^2 + C_y^2}$ as follows:

$$\begin{aligned} ||\nabla C^c|| &= \sqrt{(C_x^c)^2 + (C_y^c)^2} = \sqrt{(aC_x^u)^2 + (aC_y^u)^2} \\ &= \sqrt{a^2(C_x^u)^2 + a^2(C_y^u)^2} = a\sqrt{(C_x^u)^2 + (C_y^u)^2} \\ &= a||\nabla C^u||. \end{aligned} \tag{3.8}$$

The gradient magnitude is scaled by a under light intensity changes and shifts. Because this scaling influences all pixels, the total weight in each SIFT descriptor bin will be multiplied by a as well. Due to the L^2-normalization performed in the final stage of the SIFT descriptor, this scaling has no effect on the final descriptor:

$$\begin{aligned} bin_i' &= \frac{a \cdot bin_i}{||bin||_2} = \frac{a \cdot bin_i}{\sqrt{\sum_j^m (a \cdot bin_j)^2}} = \frac{a \cdot bin_i}{\sqrt{\sum_j^m (a^2 bin_j^2)}} \\ &= \frac{a \cdot bin_i}{\sqrt{a^2 \sum_j^m (bin_j^2)}} = \frac{a \cdot bin_i}{a\sqrt{\sum_j^m (bin_j^2)}} = \frac{bin_i}{\sqrt{\sum_j^m (bin_j^2)}}, \end{aligned} \tag{3.9}$$

where m is the number of bins in the SIFT descriptor, and bin_i the i^{th} bin of the SIFT descriptor before normalization. This proves that the SIFT descriptor will be the same under the canonical and the unknown illuminants for channel C. Because C is a linear combination of R, G and B, it follows that SIFT descriptors computed from linear combinations of RGB are invariant to light intensity changes and shifts.

Moreover, this invariance is preserved for any L^p norm, since:

$$\begin{aligned} ||bin||_p &= \Big(\sum_j^m (a \cdot bin_j)^p \Big)^{\frac{1}{p}} = \Big(a^p \sum_j^m (bin_j)^p \Big)^{\frac{1}{p}} \\ &= a \Big(\sum_j^m (bin_j)^p \Big)^{\frac{1}{p}}, \end{aligned} \tag{3.10}$$

so a will cancel out again. The derivation of photometric invariance for SIFT in this section also holds for other gradient-based descriptors, such as SURF [6] and to Histogram-of-Oriented Gradients [24] when ϵ in their normalization is equal to 0.

In conclusion, it is proven that the SIFT descriptor computed for a linear combination of R, G and B is invariant to light intensity changes and shifts. Furthermore, this invariance holds for any L^p normalization of the SIFT descriptor. This gives us a new class of illumination-invariant color descriptors, which includes the standard SIFT, OpponentSIFT and RGB-SIFT descriptors. The new descriptors are based on simple to compute linear combinations of the RGB channels.

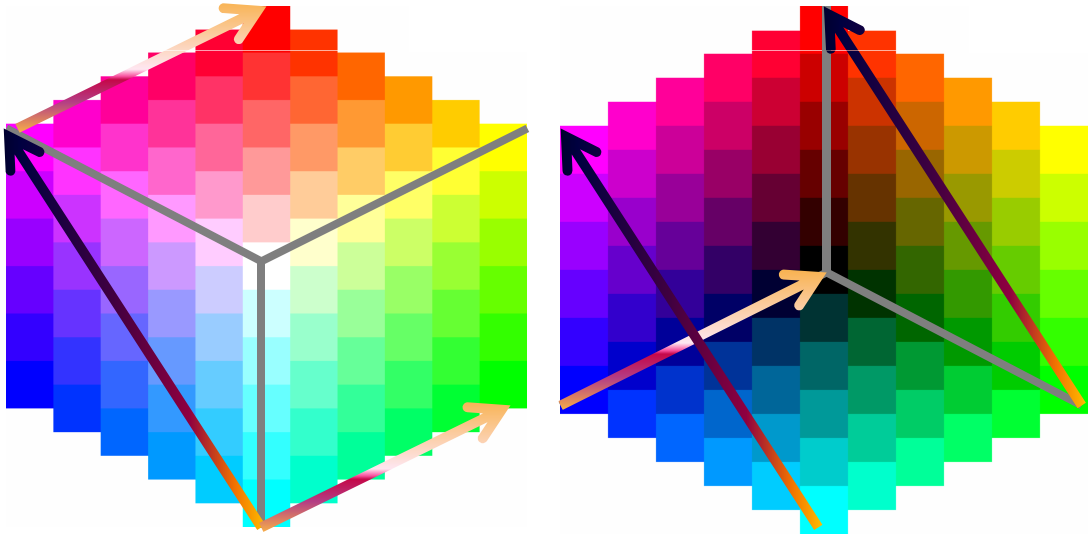

Figure 3.1: Stylized drawing of the RGB cube: view of the front planes of the cube (on the left) and the back planes of the cube (on the right). If we sample linear combinations of R, G and B with coefficients d, e, f, then we can interpret the coefficients as a vector in the space of the RGB cube. The dark arrows are such vectors and they all point in the same direction. Because the SIFT descriptor operates on gradient information only, these arrows will lead to the same descriptor: only the direction matters, not the starting point. Also, the length of the vector has no effect on the final descriptor. Therefore, only unique directions in the RGB cube constitute a unique instantiation of an illumination-invariant descriptor.

3.2.4 Instantiating Illumination-Invariant Descriptors

Uniform Sampling of the RGB Cube

The number of linear combinations of R, G and B is infinite. Because of descriptor normalization, scaled versions of the same channel are trivial ($(d, e, f) = k(d, e, f)$) with coefficients d, e, f as weights for R, G and B. If we interpret the coefficients as a vector in the space of the RGB cube (see Figure 3.1), then the length of the vector is not important. To compute a SIFT descriptor on a color channel, only the gradient information is used, not the absolute image values. This implies that parallel lines in the RGB cube will lead to the same gradient information: only the direction of the coefficient vector is important, not the absolute starting point. This is illustrated in Figure 3.1. Therefore, only unique directions in the RGB cube constitute a unique instantiation of an illumination-invariant descriptor.

Therefore, in this chapter, to approximately uniformly sample the possible directions in the RGB cube [21], linear combinations with integer coefficients from -2 to 2 are taken:

$$\mathcal{C} = \{dR + eG + fB | d, e, f \in [-2, -1, 0, 1, 2]\}. \tag{3.11}$$

Excluding scaled versions of the same channel ($(d, e, f) = k(d, e, f)$) and excluding the trivial channel $(0, 0, 0)$, this set contains 49 linear combinations, which includes the channels from

OpponentSIFT, O_1 (1, -1, 0) and O_2 (1, 1, -2), as well as the channels from RGB-SIFT. For completeness, we also consider the intensity channel $I = 0.2126R + 0.7152G + 0.0722B$, which is used in intensity SIFT, in our experiments.

Different Normalizations

As was shown in Section 3.2.3, illumination invariance holds for any L^p normalization of the SIFT descriptor. To allow for multiple candidate normalizations, we consider the following set of norms $\mathcal{L}$:

$$\mathcal{L} = \{L^2, L^4, L^8, L^{16}\}. \tag{3.12}$$

The influence is that, for large p, SIFT bins with small values will be discounted, and bins with large values are emphasized due to the normalization. Therefore, large color transitions are favored over small transitions. The dominant color transition in a patch will receive the extra weight for large p. The consequence is that emphasis is placed on certain color edges, whose color gradient corresponds best to the direction (d, e, f) chosen in the RGB cube.

3.2.5 Discussion

Instead of using a limited and predefined set of color models, in this section, we have proposed a new class of color descriptors consisting of linear combinations of R, G and B with any L^p normalization. We have derived that any descriptor from this class is invariant to light intensity changes and shifts, and, being a linear combination, is simple to compute. For the new descriptors it holds that linear combinations $\mathcal{C}$ are directions in the RGB cube and different normalizations $\mathcal{L}$ emphasize certain color edges corresponding to this direction. The aim is now to select the right color descriptors from $\mathcal{C}$ and $\mathcal{L}$ per object category.

3.3 Methods and Experimental Setup

In the previous section, we developed new illumination-invariant color descriptors. In section 3.3.1, we detail how to use these descriptors to represent an image using a feature vector, *i.e.* feature extraction. In section 3.3.2, we discuss how to train an object category model based on features. The experiments investigate how to select the best color descriptor per object category: experiment 1 (section 3.3.3) selects a single descriptor per category based on learning and experiment 2 (section 3.3.4) uses multiple kernel learning to obtain a weighted combination of descriptors per category, instead of selecting a single descriptor. Finally, we discuss the dataset and evaluation criteria used in the experiments.

3.3.1 Feature Extraction

For image representation, the bag-of-words model is used [65, 71, 95]. The feature extraction stages are illustrated in figure 3.2. First, points are densely sampled uniformly throughout the image with a spacing of 6 pixels. Results in object recognition benchmarks [36] show that dense

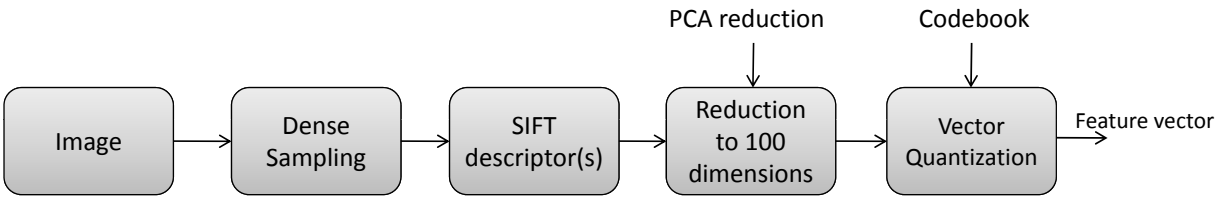

Figure 3.2: Feature extraction with the bag-of-words model. First, points are densely sampled in the image. Over the area around each sampled point, SIFT descriptors are computed. On which (color) channels these SIFT descriptors are computed is a parameter of the pipeline. To allow fair comparisons of 1-channel and two-channel descriptors, all descriptors are reduced to 100 dimensions using a PCA reduction estimated from the training set. After that, the reduced descriptors are vector quantized against a codebook with prototypical (reduced) descriptors. The occurrence frequency of each prototypical codebook element forms the fixed-length feature vector for the image.

sampling is one of the best sampling methods. After sampling, descriptors are then computed over the area around each sampled point. Depending on the number of channels described, the descriptor will have a varying length (128 for one channel, 256 for two channels, *etc*.). To allow for a fair comparison between these descriptors, their total length is reduced to 100 dimensions using PCA. Without this step, a two channel descriptor would have more degrees of freedom than a single channel descriptor. The PCA-reduced descriptors are then vector-quantized against a codebook of prototypical descriptors: a descriptor is assigned to the codebook element closest in Euclidian space. The number of descriptors assigned to each codebook element then forms the representation of the image.[†]

The PCA reduction to 100 dimensions is computed from 250,000 randomly sampled descriptors from the training set. The codebook of size 4,096 is built using k-means clustering on 250,000 randomly sampled descriptors from the training set. Both sets of 250,000 descriptors are independently sampled, *i.e*. they are different. Naturally, each different type of descriptor has its own codebook and PCA reduction.

To add coarse spatial information to the bag-of-words pipeline, the spatial pyramid by Lazebnik [69] is included. In the spatial pyramid, specific parts of the image are also represented as if they were complete images, *e.g*. with subdivisions of 2x2, the image quarters each have their own representation. All these representations are then concatenated to obtain a single feature vector. In this chapter, image subdivisions of 2x2 and 1x3 (horizontal bars) are used.

3.3.2 Category Model Training

To learn object models, Zhang *et al*. [130] show that the χ^2 kernel is a good choice for the Support Vector Machines classifier. Therefore, the Support Vector Machines classifier is used

[†]The software used to compute the descriptor set from section 3.2.4 will be made available online.

with the χ^2 kernel function to train object models:

$$k(\vec{F}, \vec{F'}) = e^{-dist_{\chi^2}(\vec{F},\vec{F'})}, \tag{3.13}$$

with $dist_{\chi^2}$ the χ^2 distance, and with $\vec{F}$ and $\vec{F'}$ the feature vectors of two images.

We use the Shogun toolbox [103] to train the Support Vector Machines classifiers. Besides training a classifier from a single feature, this toolbox also supports multiple kernel learning (MKL) to train a classifier from multiple features. During the training of a MKL classifier, a set of weights for a linear combination of the input features is learned.

3.3.3 Experiment 1: Candidate Descriptor Selection

In the first experiment, we compare the performance of selecting illumination-invariant descriptors per object category. Candidate color channels, inside the illumination-invariant descriptor, are the 49 elements of $\mathcal{C}$ (see section 3.2.4) and the intensity channel; all of which are invariant to light intensity changes and shifts (section 3.2.3). To select the descriptor to use for a specific object category, the one with the highest cross-validation score on the train set after 3-fold cross-validation with 10 repetitions is chosen. The final object category model is then trained with that descriptor using the full training set.

Besides the selection of a color channel, the experiment studies the selection of a normalization. The experiment is carried out once with the default L^2 norm for each channel, and once with multiple candidate norms $\mathcal{L}$ (see section 3.2.4) under consideration. Still, only a single combination of color channel and normalization can be chosen for each object category.

3.3.4 Experiment 2: Multiple Kernel Learning

In the second experiment, we allow the use of multiple candidate color channels and normalizations per object category through multiple kernel learning. By providing many candidate channels and normalizations as input to the MKL process, it will automatically determine the weight of the different descriptors for a linear combination with good classification accuracy. Compared to experiment 1, the trained classifier is no longer restricted to a single feature.

3.3.5 Experiment 3: Optimized Multi-Channel Descriptors

In the third experiment, we seek to construct a new color space which captures most of the weight assigned to different features in the multiple kernel learning of experiment 2. The new color space will consist of a set of channels $\mathcal{S}$ with $\mathcal{S} \subset \mathcal{C}$. Viewing the candidate color channels from set $\mathcal{C}$ as directions in a vector space, the approximation error of a channel C depends on the angle between C and the closest channel $S \in \mathcal{S}$ and the total weight w_C assigned to the channel over all concepts:

$$E_{approx} = \sum_{C \in \mathcal{C}} w_c^2 \arg\min_{S \in \mathcal{S}} \angle(C, S) \tag{3.14}$$

We consider the weights from a MKL system trained with L^1 norm and the 49 one-channel descriptors from $\mathcal{C}$. In the experiment, we consider the color spaces S with 1 through 8 channels which minimize the above approximation error. To penalize models with additional free parameters, we compute the bias corrected version of Akaike's information criterion [2, 11]:

$$AIC = n \log \frac{E_{approx}}{n} + 2K + \frac{2K(K+1)}{n-K-1}, \tag{3.15}$$

with $n = 49$ the number of data points, *i.e.* input channels in the dataset, K the number of parameters that were fitted. Each color channel defines a direction, and therefore there are two degrees of freedom per channel (two angles are sufficient to describe a channel). Besides this measure, we computed the Bayesian information criterion (BIC) using:

$$BIC = n \log \frac{E_{approx}}{n} + K \log n. \tag{3.16}$$

The preferred model is the one with the lowest AIC or BIC value.

3.3.6 Datasets

The PASCAL Visual Object Classes (VOC) Challenge [36] 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. airplane, bottle, car, dog, motorbike and person. The dataset is split into a fixed training set and test set with 5011 and 4952 images, respectively. The PASCAL VOC Challenge 2009 has 3473 images for training and 3581 images for validation. It is recommended to carry out comparitive experiments on the validation set.

Images in the dataset are stored as JPEGs, where the storage format is $sRGB$. In $sRGB$, there is a nonlinear transformation between the intensity of a pixel and the actual number stored. This transformation roughly corresponds to a power law with $\gamma = 2.2$. Because the same transformation is applied to all pixels in all the images, analytically, all derivations from Section 3.2 can still be applied, substituting R by R^γ, G by G^γ, *etc*.Empirically, the effect of $sRGB$ compared to the linear RGB space is limited: in the standard VOC experiment, inclusion of a gamma correction of 2.2 to address the nonlinear transformation has very little impact: SIFT performance changes by 0.0036 and OpponentSIFT performance by 0.0007. Therefore, we leave out this additional transformation in our experiments.

3.3.7 Evaluation criteria

For our results, the average precision is used. 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 judged to be relevant. Hence, it combines both precision and recall into a single performance value. When performing experiments over multiple object categories, the average precisions of the individual categories are aggregated into the Mean Average Precision (MAP).

Experiment 1: Candidate Descriptor Selection

Candidate Descriptor(s)	MAP with L^2 norm	MAP with norm $\in \mathcal{L}$
Intensity SIFT	0.4902	0.4959
OpponentSIFT (baseline)	0.4975	0.5109
Intensity SIFT or OpponentSIFT per category	0.4975	0.5077
One channel $C \in \mathcal{C}$ SIFT per category	0.4881	0.4861
Two channel I and $C \in \mathcal{C}$ SIFT per category	**0.5085**	**0.5130**

Table 3.1: Performance results for experiment 1 on the PASCAL VOC 2007 test set, where a single descriptor is selected per object category for different sets of candidate descriptors and normalizations.

3.4 Results

3.4.1 Experiment 1: Candidate Descriptor Selection

PASCAL VOC 2007 test set: The goal of this experiment is to investigate the object recognition performance of different sets of illumination-invariant descriptors. See table 3.1 for all results. The OpponentSIFT baseline with L^2 normalization has a MAP of 0.4975. When using the set of 49 candidate channels $\mathcal{C}$ to select a single channel per object category, it can be derived that the performance is reduced. The channels selected using cross-validation do not perform better than either intensity SIFT or OpponentSIFT. When performing descriptor selection with just intensity SIFT and OpponentSIFT, we notice that there is noise in the selection process: performance is not improved, while performance should have improved had the best descriptor per object category been selected.

Besides the noisy descriptor selection process, another cause for lack of improvement in the candidate channel $\mathcal{C}$ experiment is the lack of intensity information in the candidates. When introducing color channels in the descriptor, it is important to include the intensity channel as well [112]. Indeed, when we extend the experiment to two channels, one of which is always the intensity channel I, results do improve over the OpponentSIFT baseline. When we allow different descriptor normalizations $\mathcal{L}$ instead of a fixed L^2 norm, the MAP improves further.

Overall, selecting the color channel and descriptor normalization per object category improves by 3% over the baseline (from 0.4975 to 0.5130). We note that selecting a single best feature per object category based on cross-validation scores is sensitive to noise and leads to suboptimal results.

3.4.2 Experiment 2: Multiple Kernel Learning

PASCAL VOC 2007 test set

In experiment 1, we saw that selecting a single best feature per object category is sensitive to noise. To reduce this noise, in experiment 2, we allow the use of multiple features per object category through multiple kernel learning. This also eliminates the descriptor selection process,

Experiment 2: Multiple Kernel Learning

Descriptors	MAP with L^2 norm	MAP with norms $\in \mathcal{L}$
Intensity SIFT	0.4902	0.5169
OpponentSIFT (baseline)	0.4975	0.5203
Intensity SIFT and OpponentSIFT	0.5187	0.5357
One channel $C \in \mathcal{C}$ SIFT	0.5351	0.5405
Two channel I and $C \in \mathcal{C}$ SIFT	**0.5463**	**0.5507**

Table 3.2: Performance results for experiment 2 on the PASCAL VOC 2007 test set, where multiple kernel learning is used on different sets of descriptors and normalizations.

as multiple kernel learning will automatically learn weights for the different input features. See table 3.2 for all multiple kernel learning results.

In table 3.2, we observe that using both SIFT and OpponentSIFT now clearly improves over the OpponentSIFT baseline. The gains from using additional norms are also increased. When using the set of 49 candidate channels $\mathcal{C}$ to select a single channel per object category, it can be derived that the performance now improves, even without explicitly including intensity information. Still, when we include the intensity channel again, *e.g.* the two channel setting, results improve over the single channel per object category like they did in experiment 1. Compared to the baseline, the improvement is 10% (from 0.4975 to 0.5463) when using a fixed L^2. With multiple normalizations $\mathcal{L}$, the improvement is 11% (from 0.4975 to 0.5507).

In a multiple kernel learning experiment, it is also interesting to compare to a combination of intensity SIFT and OpponentSIFT, instead of to the OpponentSIFT baseline. In this comparison, the improvement is 6% (from 0.5187 to 0.5507).

In figure 3.3, the results are split out per object category for the baseline and the two-channel results. The largest gains are observed for potted plants, birds, bicycles, motorbikes, busses and trains. In figure 3.4, for birds and trains, several images are shown. For birds, especially the third image, with a small bright red bird, would not have been recognized without additional color descriptors and higher norms to emphasize strong color edges. As a side-effect, an airplane seen through brightly colored tree branches is also placed higher in the ranking. For trains, the commonality between images which have been placed higher in the ranking is that the trains themselves are relatively small or truncated, but there is a clear presence of railroad tracks. Through MKL, weight can be assigned to a color descriptor which triggers on the color edges of railroad tracks.

Continuing with figure 3.3, for all object categories except bottle, there is a consistent small improvement by using multiple norms. For bottle, it is shown that the performance using a fixed L^2 norm is higher than when using multiple norms, though it still exceeds the baseline. In this case, the classifier has fitted too much to the training examples by using the additional norms.

In conclusion, applying MKL to our candidate descriptors improves performance by 11% compared to an OpponentSIFT baseline, *i.e.* to the best single descriptor from related work [112]. Compared to a MKL combination of intensity SIFT and OpponentSIFT, the improvement is 6%.

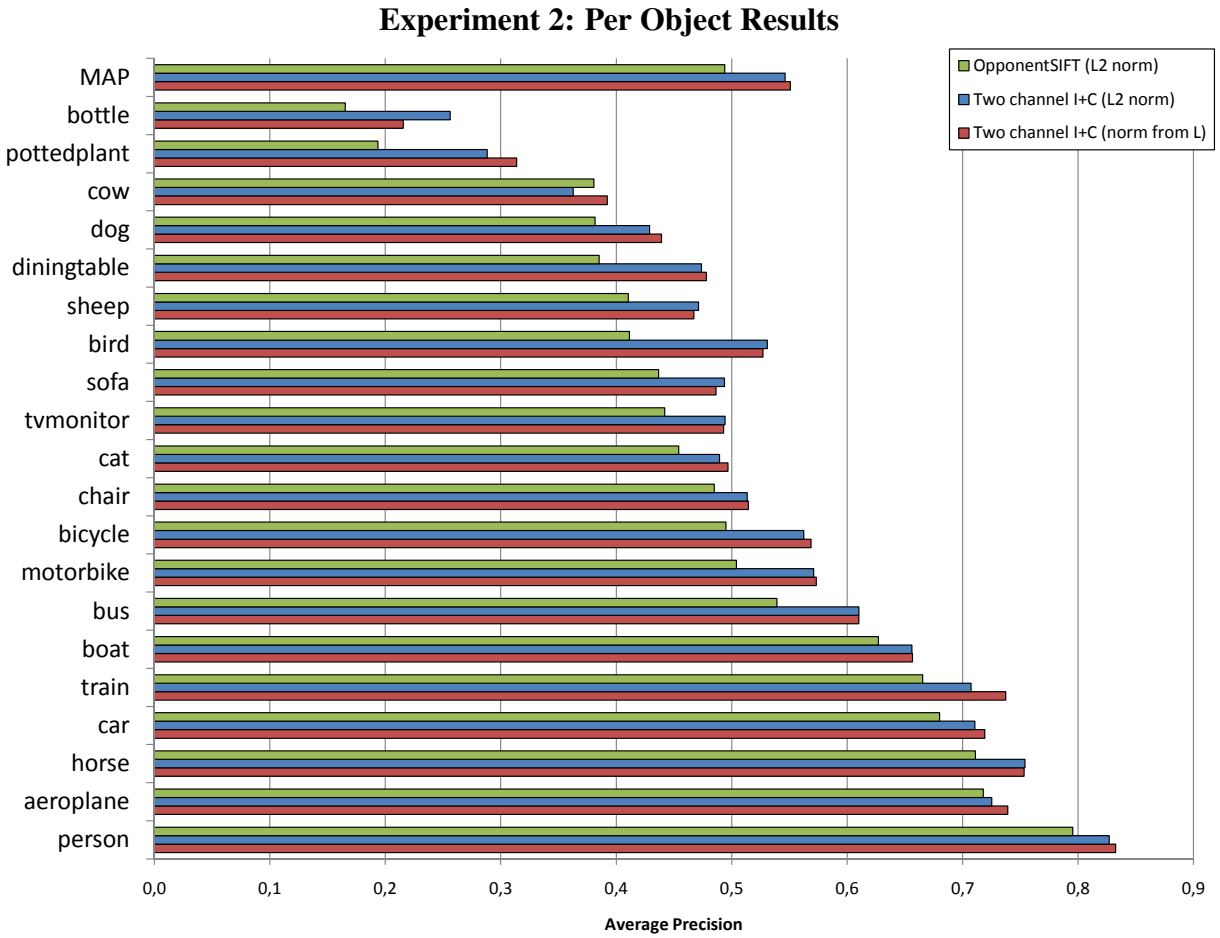

Figure 3.3: Detailed per-object results from table 3.2 for the Opponent SIFT baseline and the two-channel scores. Results are sorted by the average precision for the baseline.

PASCAL VOC 2009 validation set

In table 3.3, we show the results of experiment 2 on the PASCAL VOC 2009 validation set. Compared to the PASCAL VOC 2007 test set results from table 3.2, we observe the pattern as in that table: applying MKL to our candidate descriptors improves performance by 12% compared to an OpponentSIFT baseline. Compared to a MKL combination of intensity SIFT and OpponentSIFT, the improvement is 7%.

3.4.3 Experiment 3: Optimized Multi-Channel Descriptors

PASCAL VOC 2009 validation set

In experiment 2, we observed that MKL is a good way to learn the weights for multiple features. However, because most features receive a non-zero weight for at least one concept, a drawback of the MKL approach is that almost all features still need to be computed. In this experiment,

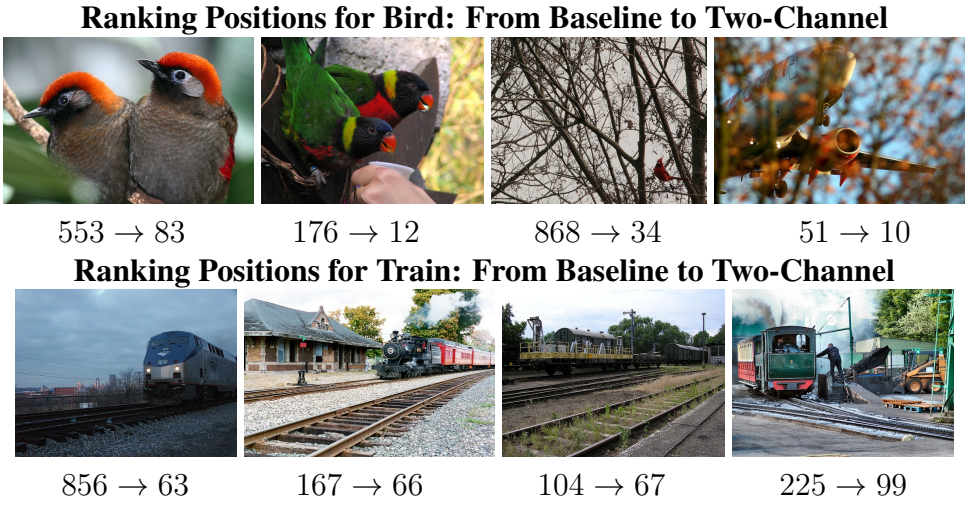

Figure 3.4: Qualitative results for birds and trains from figure 3.3: first number is the position in the ranking for the OpponentSIFT baseline, second number for the two channel I and $C \in \mathcal{C}$ descriptors with norms $\in \mathcal{L}$. All images except the top-right one really contain birds/trains.

Experiment 2: Multiple Kernel Learning

Descriptors	MAP with L^2 norm	MAP with norms $\in \mathcal{L}$
Intensity SIFT	0.4535	0.4729
OpponentSIFT (baseline)	0.4566	0.4770
Intensity SIFT and OpponentSIFT	0.4725	0.4874
One channel $C \in \mathcal{C}$ SIFT	0.4911	0.4948
Two channel I and $C \in \mathcal{C}$ SIFT	**0.4963**	**0.5031**

Table 3.3: Performance results for experiment 2 on the PASCAL VOC 2009 validation set, where multiple kernel learning is used on different sets of descriptors and normalizations.

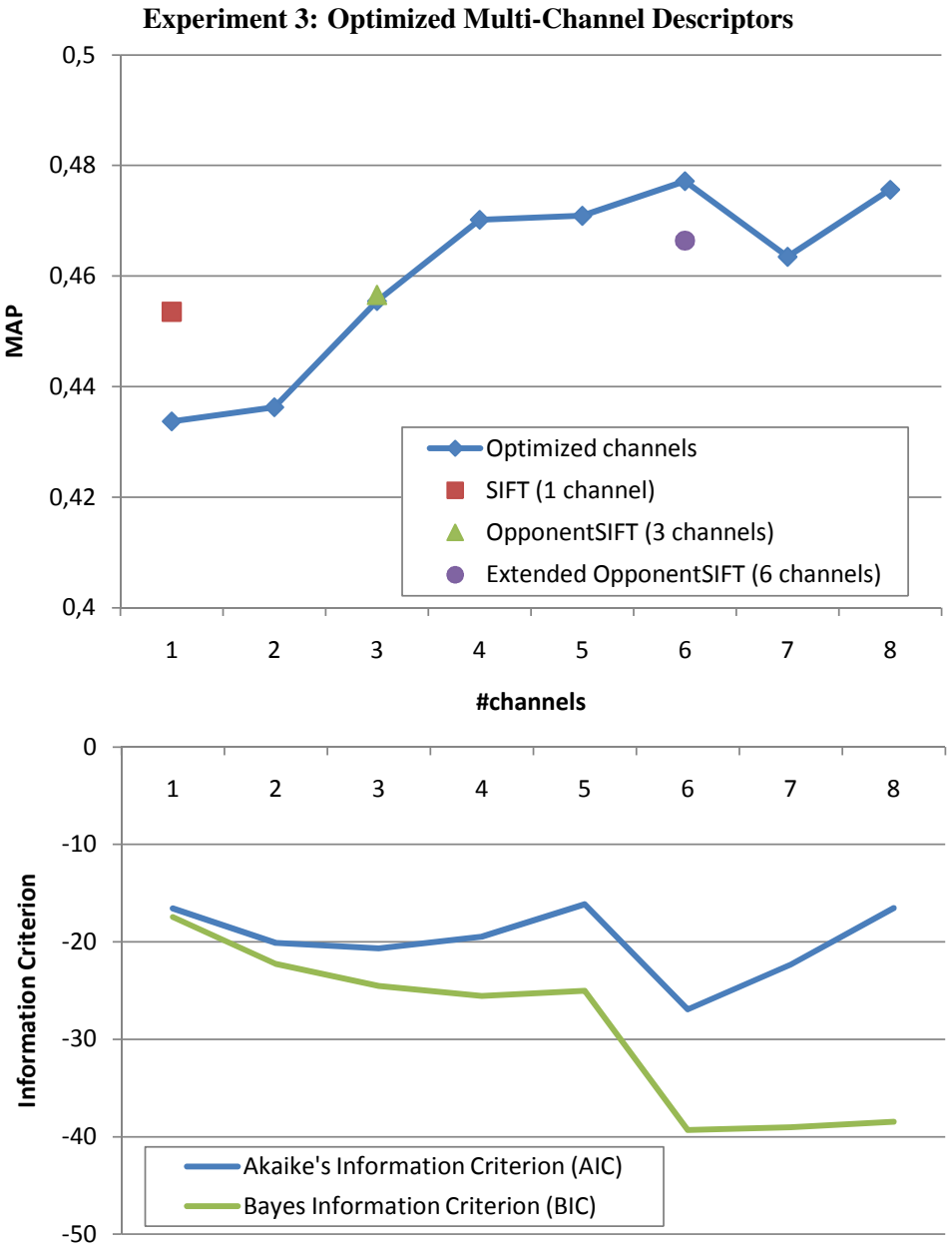

Figure 3.5: Performance results (top plot) of descriptors with 1 through 8 color channels, selected based on the MKL weights from experiment 2. For reference, results for the best single-channel descriptor (intensity SIFT) and the best 3-channel descriptor (OpponentSIFT) [112] are also shown. The bottom plot shows two information criteria (BIC and AIC), which penalize overly complex models. The best choice is 6 channels, which corresponds to the maximum MAP (see top plot). By analyzing the color channels selected, we observe that the optimized channels are similar to an extension of the opponent color space, with finer sampling in the chromaticity plane. Therefore, we propose a 6-channel Extended OpponentSIFT descriptor which exploits this observation.

we construct new color spaces based on the MKL weights assigned to features on the PASCAL VOC 2009 training set. We construct color spaces which capture the most weight with 1 to 8 channels according to the procedure from Section 3.3.5. The advantage of these optimized color spaces is that they only need to have a single feature extracted. Also, the feature extraction time is largely independent from the number of color channels, because of the PCA reduction to 100 dimensions early in the pipeline (see Figure 3.2).

Results for the optimized color spaces are shown in Figure 3.5. We observe that selecting a single channel performs worse than intensity SIFT. With 3 channels, the optimized descriptor performs equal to OpponentSIFT, which also has 3 channels. Going beyond 3 channels, the optimized descriptors outperform OpponentSIFT (MAP=0.4566). The best result is obtained with a 6-channel color space: a MAP of 0.4772, which is a relative improvement of 4.5%. To penalize overly complex spaces, *i.e.* many channels, we consider the two information criteria (BIC and AIC) from Section 3.3.5 with respect to how well they fit the MKL weights. The best scores are obtained for 6 channels, which corresponds to the highest MAP.

To analyze the color channels which have been selected for the 6-channel space, we analyzed their directions in the opponent color space (data not shown). Here, we observe that there is one channel which is close to the intensity channel, and the five other channels are close to or in the chromaticity plane. The lack of directions close to the intensity channel suggests that, to improve classification accuracy, it is more sensible to have additional channels to discriminate between different colors (chromaticity) than small variations of the intensity channel. Psychophysical studies have shown that multiple chromatic mechanisms (channels) in the isoluminant plane, presumably operating at higher cortical levels, are required to describe experimental data on visual search [29], image segmentation [57] and discriminability of chromatic distributions [53]. To verify the importance of multiple channels in the isoluminant plane, we constructed an extended opponent color space, which consists of the intensity channel I and 5 channels evenly spaced in the plane perpendicular to the intensity direction. The Extendend OpponentSIFT descriptor achieves a MAP of 0.4664, which is 2.1% better than OpponentSIFT without any optimization on the training data. When computational resources are constrained and a single descriptor must be chosen, we suggest using this descriptor instead of current color SIFT descriptors.

3.4.4 Application: Object Localisation

Several approaches [68,123] to object localisation use the bag-of-words model to classify whether objects are present within a rectangular area of the image. Following the experimental setup of Vedaldi *et al*. [123], we construct a simple localisation system based on their classification strategy with the histogram intersection kernel for SVM. The codebook sizes and spatial pyramid are the same as in previous experiments. We compare the baseline used in the other experiments, intensity SIFT plus OpponentSIFT, to a combination of intensity SIFT plus the extended 6-channel OpponentSIFT descriptor. The first achieves a MAP of 0.2183, while the use of the extended OpponentSIFT descriptor results in a MAP of 0.2267. Overall, this is an improvement

of 4% with roughly the same amount of computation[‡].

3.5 Conclusion

In this chapter, we have derived that any linear combination of the R, G and B channels will result in an illumination-invariant SIFT descriptor. Also, the descriptor will remain invariant to light intensity changes and shifts even if the L^2 normalization is changed to a generic L^p normalization. Selecting the most discriminative illumination-invariant descriptor per object category improves over the OpponentSIFT baseline by 3%. By using multiple kernel learning instead of selecting a single descriptor, this improvement increases to 11%. Comparing multiple kernel learning results to the strong combination of intensity SIFT and OpponentSIFT, the improvement is 6%. By analyzing the weights assigned to different color channels by multiple kernel learning, we devised a single optimized color descriptor which performs 4.5% better than OpponentSIFT (the recommended descriptor from [112]) on the PASCAL VOC 2009 object classification task and 4% better on the object localisation task. Furthermore, we have found that the color channels in the new descriptor can be viewed as an extension of the opponent color space with additional sampling in the color plane. Exploiting this observation, we propose the Extended OpponentSIFT descriptor which is 2.1% better than OpponentSIFT but does not require any optimization on the training data.

[‡]Both combinations consist of two features. Due to the PCA reduction to 100 dimensions, the number of color channels in each feature does not impact on computation times significantly.