Per-patch Descriptor Selection Using Surface and Scene Properties

Ivo Everts, Jan C. van Gemert, and Theo Gevers

Intelligent Systems Lab Amsterdam (ISLA), University of Amsterdam,
Science Park 904, 1098 XH, Amsterdam, The Netherlands

Abstract. Local image descriptors are generally designed for describing all possible image patches. Such patches may be subject to complex variations in appearance due to incidental object, scene and recording conditions. Because of this, a single-best descriptor for accurate image representation under all conditions does not exist. Therefore, we propose to automatically select from a pool of descriptors the one that is best suitable based on object surface and scene properties. These properties are measured on the fly from a single image patch through a set of attributes. Attributes are input to a classifier which selects the best descriptor. Our experiments on a large dataset of colored object patches show that the proposed selection method outperforms the best single descriptor and a-priori combinations of the descriptor pool.

1 Introduction

Representing local image structures is important for many computer vision tasks such as (object) recognition, wide baseline matching and tracking. In these tasks, a generic image descriptor is typically chosen which should be well-suited for describing all possible image patches.

Image patch appearance is determined by combinations of material properties, such as color and texture, with accidental scene properties such as illumination conditions, viewpoint, scale, and so on. A successful image descriptor should have high discriminative power between material properties, while remaining invariant against disturbing instances of scene-accidental conditions.

Invariance, however, is inversely related with discriminative power [1, 2]. Many excellent image descriptors have been designed [3] or optimized [4] to find a good trade-off between invariance and discriminative power. Nevertheless, a single descriptor cannot be optimal in all cases. Consider for example a patch containing highlights in the top row of figure 1a. Using a highlight-invariant descriptor would increase the matching score. On the other hand, consider the bottom row of figure 1a. Using the highlight-invariant descriptor may actually remove the discriminating characteristic. A similar argument holds for material properties such as texture and color. For the example in the second row of figure 1a it makes little sense to use a color descriptor since it becomes unstable with little color present [1]. The conflicting demands on the degree of invariance and between material representations cannot be resolved by a single descriptor.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part VI, LNCS 7577, pp. 172–186, 2012.

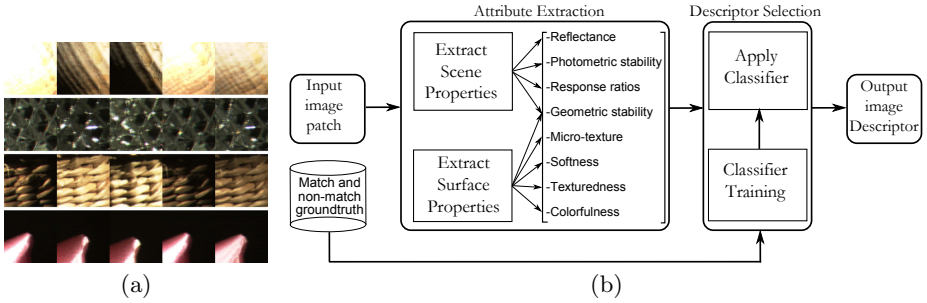


Fig. 1. (a) Example input patches to our method. The rows contain corresponding patches under a range of photometric and geometric disturbances. (b) Schematic of the descriptor selection algorithm.

In this paper, we propose a method to select the best descriptor for a single patch. For example, if we can detect the difference between a strong scene-accidental highlight and a glossy material surface in figure 1a, then we can select a suitable different descriptor in both cases. With this aim, we identify material properties [5, 6] and scene-accidental properties [1, 3] which we will use in a supervised learning scheme that can take mismatching costs into account. See figure 1b for a schematic overview of our approach. While we are not aware of any articles in which physical properties are related to descriptors on a per-patch basis, we review relevant works in the following.

2 Related Work

A local image descriptor can be optimized discriminatively out of combined variations of atomic operations such as smoothing, angular quantization, spatial pooling and feature normalization [4]. Alternatively, a projection can be learned on (sift) features to reduce descriptor size and simultaneously improve matching performance [7] or visual word assignment in retrieval [8]. These methods take a descriptor and improve its overall performance, resulting in a better single-purpose descriptor for all patches. It is, however, not possible to tune a descriptor to a single patch, as we propose in this paper.

For category-level image classification, the best image-level descriptor can be learned for each category. When such supervised information is available, various machine-learning techniques can learn the best descriptor combination by boosting [9–11], multiple kernel learning [2], topographic filter maps [12], dimension reduction [6, 13] or the Fisher criterion [14]. These methods cleverly exploit the intra- and inter-class variance between the image categories. However, when category labels are not available, as for example in feature tracking or wide-baseline matching, these methods cannot be applied. We propose a generic method that is suitable for such applications by selecting the best descriptor based on the material properties of a single patch.

Material recognition is a generalization of texture recognition, which is widely studied, see e.g. [15]. Recently, more generic material classes such as glass, metal and fabric have been proposed [6] and followed-up by [13]. These methods aim to find the named material class of a given image, such as wood, leather or stone. In this paper, however, we are not as much interested in the class per se, as this would require a large database of common material classes. Alternatively, we propose to find the best image descriptor of a patch based on its structural and surface reflectance properties.

Surface reflection can be characterized by the bidirectional reflectance distribution function (BRDF). The BRDF represents the reflection ratio for all surface locations under all possible illumination and viewing directions. Despite the complexity of the BRDF, there are methods to estimate it under constraints on object shape or illumination direction [16–18]. In this work, we are interested in unconstrained shapes and illuminations and therefore focus on simpler features.

3 Descriptor Selection Using Surface and Scene Properties

To select the best descriptor for an image patch, it is important to identify features that can estimate material properties such as colorfulness, roughness, shininess etc. Moreover, the pool of available descriptors to choose from has to be diverse enough to emphasize or ignore those properties that are important for recognition. For example, a smooth shiny patch from an apple will benefit from keeping the shininess and perhaps not focusing on edges too much. On the other hand, a cast shadow or the position of a strong highlight is scene accidental, and therefore better ignored. The material properties should be able to measure and represent such effects from an image patch whereas the image descriptors should ideally be able to distinguish between various levels of invariance. Such levels of invariance apply to the object's structure, such as edge-based vs. pixel-based, but also on photometric invariant properties such as highlights, shadows and shading.

3.1 Photometric Representations

Photometric invariance can be modeled by the dichromatic reflection model [19]. In this model, an RGB vector $\mathbf{f} = (R, G, B)^T$ is the vector summation of the body reflectance with the specular interface reflectance

$$\mathbf{f} = e(m^b \mathbf{c}^b + m^i \mathbf{c}^i), \tag{1}$$

where e is the intensity of the light source, $\mathbf{c}^b$ is the color of the body reflectance, $\mathbf{c}^i$ the color of the interface reflectance, the scalars m^b and m^i depend on the surface orientation and represent the magnitude of the body and interface reflection respectively.

Table 1. Image representations and descriptor names

	Intensity	Chromatic	Normalized Chromatic	Hue
Representation	O_3	$[O_1, O_2]$	$\left[\frac{O_1}{O_3}, \frac{O_2}{O_3}\right]$	$\frac{O_1}{O_2}$
Invariant to	-	Highlights	Shadows	Highlights & Shadows
Descriptor name	I.pix/I.grad	C.pix/C.grad	N.pix/N.grad	H.pix/H.grad

For representing image invariants we consider the transformation to the opponent color space [1, 3]. Save scaling factors, the transformation is given by

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} R-G \\ R+G-2B \\ R+G+B \end{pmatrix}. \tag{2}$$

The opponent color components are combined in four different representations. First, the chromatic components O_1 and O_2 are separated from the intensity component O_3. On itself, O_3 has no invariance properties but does generally contain most information regarding image structure. Due to the subtraction of RGB components, O_1 and O_2 are invariant with respect to shifts in illumination such as highlights. Nevertheless, O_1 and O_2 are still sensitive to illumination scalings such as shadow and shading. To this end, we also consider intensity-normalized chromatic components $\frac{O_1}{O_3}$ and $\frac{O_2}{O_3}$. These intensity-normalized invariants however, are again sensitive to illumination shifts. Therefore, the set of photometric representations is complemented with $hue = \frac{O_1}{O_2}$, which is invariant to both illumination scalings and shifts. The full set of photometric representations that we consider in this paper is given in table 1. Patch descriptors and image attributes are extracted from these representations as detailed in the following.

3.2 Image Attributes

Low-level image attributes have been used to measure a degree of *objectness* in a bounding box [20], or in a functional extension of the spatial pyramid beyond spatial information towards more generic types of pooling [21]. Here we are interested in low-level features to identify surface structure and reflection. Surface reflectance properties such as shiny, matte or gloss have been found to correlate with simple image statistics [16, 22, 23]. Surface structures such as crinkles in leather or grains in paper have been proposed to be detectable by subtracting a bilateral filtered image from the original [6]. Further, the difference in edge types in e.g. metal, glass, or paper can be linked to the variance of the gradient magnitude or orientation [13]. We will evaluate and extend these low-level surface features to link structure and reflectance surface qualities to an image descriptor. To avoid any confusion between local image descriptors which are also often called features we will use the term *attribute* for low-level image

features that measure structural and photometric surface/scene *properties*. We detail these attributes in the following.

(a) Interface reflectance. The dichromatic model in equation 1, has a term for the object reflectance $m^b\mathbf{c}^b$ and a term for the interface reflectance, $m^i\mathbf{c}^i$. The latter term, representing gloss, matte/shininess, has been found to correlate with the skew (third-moment) of the intensity histogram [22]. Other research also uses the standard deviation, 10th, and 90th percentile [23] and kurtosis [16] to represent the shape of the intensity histogram to predict interface reflectance. The amount of interface reflection is a valuable attribute for selecting between the highlight invariants ($[O_1, O_2]$, $\frac{O_1}{O_2}$) and the highlight variants ($O_3, [\frac{O_1}{O_3}, \frac{O_2}{O_3}]$). Therefore we use these intensity statistics over O_3 as patch attributes.

(b) Photometric stability. The invariant representations introduced in the previous section are insensitive to various photometric transformations. However, this comes at a price that is paid in numerical instability [1]. The $hue = \frac{O_1}{O_2}$ invariant is unstable for colors on the black-white axis (i.e. low saturation), whereas the intensity-normalized $\frac{O_1}{O_3}$ and $\frac{O_2}{O_3}$, invariants are unstable near zero intensity. The occurrence of this instability depends on the surface reflection, and varies per patch. Therefore, these features are well-suited attributes to determine if an invariant representation is suitable. To this end, we use the mean intensity $\mu(O_3)$, and mean saturation $\mu(\sqrt{O_1^2 + O_2^2})$ as photometric stability attributes. Moreover, to obtain a richer representation, we compute the same statistical values for saturation as we did for intensity in the previous paragraph.

(c) Photometric response ratio. To obtain attributes specifically tuned to each invariant representation, we relate the response in the full-color representation to the response for each invariant. Different invariant representations will respond differently to shadows, shading, gloss and highlights and consequently this difference allows descriptor diversification. To this end, we compute the average gradient ratio [24] for each invariant with respect to the full color gradient, $\frac{|\nabla O_3|}{|\nabla [R,G,B]|}$ and similarly for other representations.

(d) Geometric stability. We include a sense of the geometric stability of the patch under a viewpoint change. The basic idea is that a small geometric transformation on a stable patch should lead to a small differences in the descriptor. Large differences may indicate high sensitivity to the disturbance. This is also a structured attribute since it gives a sense of homogeneousness. The sensitivity is measured by a set of self-dissimilarities after applying a geometric transformation to the patch. Specifically, we depart from a centered sub patch (80%) cropped out of the image. The region of interest is then up- and down- scaled and translated such that a set of geometrically transformed versions of the initial patch is obtained. We take the average descriptor distances over two scales and eight directions as geometric stability attributes for each image descriptor.

(e) Micro-texture. The surface structure of a patch may be rough or smooth. Metal, for example, is typically smooth, whereas fabric is fine grained. To distinguish between rough and smooth surfaces, we follow the approach of Liu et al. [6] to detect micro-texture. Specifically, we subtract a bilateral smoothed version

from the original image patch. As attributes we use the sum of the residual, and we do this for each of the four invariant representations separately.

(f) Softness. Material may also be soft as plastic, or hard as metal. As suggested in [13], we adopt the standard deviation of the gradient orientation, and the standard deviation of the gradient magnitude to measure material softness. The authors' rationale is that soft materials have soft edges, with softly varying transitions in gradient orientation and magnitude. We compute this for each of the invariants and add the mean values to obtain a richer statistical representation.

(g) Texturedness. For a notion of material texturedness we build on the work of [25] in which it is shown that a weibull parameterization of images results in textural diversification. Specifically, the contrast distributions of natural images generally follow the 2-parameter integrated weibull distribution. We compute these two parameters, β and γ in each invariant, and also compute additional statistics by counting the number of edges above a noise threshold.

(h) Colorfulness. As a measure of colorfulness we compute a single valued hue entropy score, $-\sum(p \log_2 p)$ where p is the histogram of the hue pixels $\frac{O_1}{O_2}$.

3.3 Image Descriptors

For constructing image descriptors, we use the photometric representations as given in table 1. Besides these photometric variation, we model structure variation with multiple differential orders. Zeroth-order descriptors are histograms extracted from pixels per color channel, whereas first-order (sift) descriptors are based on the per-channel gradient orientations [3]. Note that the order of differentiation affects the invariance properties of the descriptor. We do not consider higher order representations. Spatial pooling of the descriptors is obtained by aggregating features in a 4×4 cell grid as originally proposed by Lowe. For zeroth-order descriptors we compute 8-bin histograms of pixel values. For first order descriptors the gradient orientations are quantized in 8 bins. Furthermore, feature contributions are weighted by a Gaussian window centered on the image patch. Finally, the descriptors are normalized to unit length. The invariance properties and the names of the descriptors are given in the bottom row of table 1, where *pix* denotes zeroth-order, and *grad* indicates first-order.

3.4 Descriptor Selection

We relate attributes to descriptors in a supervised learning setup. Our setup is similar to [4]. However, where they learn the best single-descriptor parameters over a training set, we leverage the patch attributes to learn the best descriptor for a single patch. We start with a ground truth set which has for each patch a corresponding transformed version of the same patch (under homography or photometry, more details below) and 100 randomly sampled non-matches. Such a set allows the computation of a matching score in average precision (AP) for each descriptor type per patch. The attributes of the patch are the input for our

supervised setup whereas our goal is to select the descriptor that gives the best average precision score.

Let $X = \{x_1, x_2, \ldots, x_n\}$ be the patch ground truth data set containing n patches, where x_i is a p-dimensional vector containing the p attribute values. The corresponding average precision scores $Y_i = \{y_1^i, y_2^i, \ldots, y_d^i\}$ for patch i are computed by ranking all retrieved patches according to each descriptor distance for all d descriptors. We aim to find a learning model $\mathbb{L}$ that maximizes the average precision for a patch, i.e., $\mathbb{L}(x_i) = \arg\max_i(Y_i)$. Note that the cost of misclassification is not uniformly distributed over the descriptor classes since each descriptor typically gives a different average precision score. The misclassification cost c_i of selecting a descriptor for a patch i is the score of the selected descriptor minus the score of the best possible (oracle) descriptor in the pool, $c_i = Y_{\mathbb{L}(x_i)} - Y_{\arg\max_i(Y_i)}$. To take non-uniform misclassification costs into account, we adopt the cost-sensitive support vector machine (SVM) approach by Zadrozny et al. [26]. This formulation incorporates the misclassification costs c_i directly in the SVM optimization problem. Note that the classes are descriptor types. For each binary g-vs-h sub-class problem (e.g. *I.grad* vs *H.pix*, see table 1) this becomes

$$\begin{aligned} &\underset{w,\ \xi,\ k}{\text{minimize}} && \frac{1}{2}w^T w + C\sum_{i=1}^{n} c_i\xi_i \\ &\text{subject to} && b_i(w^T x_i + k) \geq 1 - \xi_i, \quad \xi_i > 0, \\ &\text{where} && b_i = \begin{cases} +1 & \text{if } \ y_g^i > y_h^i; \\ -1 & \text{otherwise.} \end{cases} \end{aligned} \tag{3}$$

Thus, b_i denotes $+1$ if the average precision score of x_i of descriptor g is higher than the average precision of x_i of descriptor class h; and -1 otherwise. If two patch descriptors have the same average precision score ($y_g^i = y_h^i$) in the training phase, we assign the patch to both descriptor classes (the misclassification cost c_i will subsequently be 0). We use class voting to obtain the multiclass label from all 1-vs-1 class-pairs. In the case of equal votes we assign the sample to the descriptor with the highest a-priori score on the training set.

Note that our 1-vs-1 setup allows us to utilize the full dataset for each descriptor class-pair. If we first assign the best descriptor to each patch, a binary descriptor-pair classifier could only train on those samples where the global maximum is obtained by one of the two classifiers. Because we use the pair-wise maximum b_i for each descriptor-pair, we do not suffer from this problem.

4 Experiments

It is our hypothesis (fig. 1b) that the best descriptor is dependent on two factors: the accidental scene properties and the patch's surface material properties. To evaluate this hypothesis we create a separate dataset for each factor. The influence of the surface material is tested by extracting attributes from a clean, canonical, patch without any distortions. This allows us to judge the influence of

the surface alone. Alternatively, to evaluate the influence of the scene, we extract attributes from a photometric/geometric distorted patch. In this case, however, it cannot be helped that the surface will also have some influence. We refer to the patch used for prediction as the query patch, and the aim is to match the same patch in a set of non-matching patches.

In our experimental setup we use pairs of matching patches. For every match in the database, we sample 100 random non-matches, which is repeated 10 times. Retrieved patches are ranked based on the Euclidean distances to the query patch of the respective descriptors. From this we compute the average precision for measuring retrieval performance per descriptor. One half of the dataset is used for training, and the other half is used for testing. Note that a patch is exclusively in the train or in the test set.

4.1 Synthetic Dataset

We start with a synthetic dataset to evaluate the descriptors and attributes under controlled circumstances. The synthetic dataset is generated from the work of Barnard et al. [27] where measurements of 1995 surface reflection spectra and 287 illumination spectra are provided. The camera sensitivity function allows computation of RGB values given the surface albedo and illumination spectrum. We create 'Mondrian'-style images by inserting colored blocks of random size at random locations on a 64x64 image lattice. We keep the illuminant color fixed, rotate the colored blocks by a random degree, and introduce some skew and noise, so as to reflect a more diverse range of image content.

Pairs of matching patches are generated by applying a geometric or photometric disturbance to a synthetized image. Geometric disturbances encompass a translation or rotation. Photometric disturbances are achieved by applying a scaling (shadow and shading) or offset (highlights) to all RGB channels, see eq. 1. The location and extent of the disturbance is governed by an anisotropic Gaussian with a random location and covariance. The disturbances are progressively increased, in five steps, where we generate a dataset consisting of 1000 matching image pairs per disturbance level. This is repeated for increasing amounts of foreground blocks, which denotes a basic notion of image complexity. See figure 2 for some example patches.

Matching performance per descriptor under each of the disturbances is shown in figure 3. The figure shows that pixel-based descriptors almost always outperform gradient-based descriptors on this dataset. This is partly due to the fact

Fig. 2. Example patches of the synthetic dataset. The patches on the right side of a patch-pair are distortions of the respective patches to the left side (translation, rotation, shadow/shading, highlight, noise).

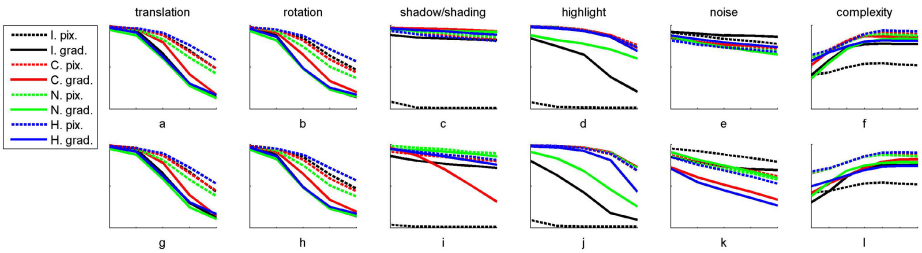


Fig. 3. Influence of geometric and photometric disturbances on descriptor matching performance. Mean average precision is plotted against the disturbance level on the x-axis. Pixel-based (pix) and gradient-based (grad) descriptors are extracted from intensity, chromatic, normalized chromatic and hue representations, denoted by I., C., N. and H (see table 1). Figures *a-f* are the result of matching experiments in which the query patch is in canonical form. In figures *g-l* the query patch is distorted. Note that the per-descriptor performances are averaged over all disturbances against known numbers of foreground blocks to obtain the 'complexity' figures in f and l.

that the set of colors in the dataset is limited and distinct. However, as can be seen in figures 3 *a-b* and *g-h*, it stands out that pixel-based descriptors are considerably less sensitive to geometric disturbances than gradient-based descriptors. Naturally, there is no difference between canonical or distorted query patches if the disturbance is purely geometric, as in figures 3*a*, 3*g* and 3*b*, 3*h*.

Pixel-based intensity-only (I.pix) descriptors fail when shadows and highlights are applied to the image (figures 3 *c-d* and *i-j*) as expected. However, under additive gaussian noise, the I.pix descriptors are superior when the patch is presented in distorted form. Gradient-based descriptors suffer more from noise in general. Furthermore, gradient-based opponent color descriptors are also sensitive to shadows, while normalized opponent color descriptors appear sensitive to highlights, which is in accordance with the respective photometric invariance classes. Gradient-based intensity descriptors also suffer from highlights because most edges comprise of color transitions which may become less prominent in the vicinity of highlights. It appears to be more difficult in general to retrieve the canonical form based on a distorted query patch than vice versa (see figures 3 *c-e* and *i-k*). This is because the disturbance increases the average similarity to all patches (you may find a highlight if you look for it). Increased image complexity (figures 3 *f* and *l*) generally results in improved matching performance for all descriptors. However, pixel-based descriptors (other than intensity) suffer significantly less from the absence of image structure.

In table 2 we show classification rates when using attributes to predict a patch's disturbance level on the synthetic data. We evaluate each disturbance individually, using an SVM trained on 90% of the data, and tested on 10% in 10 random folds. The results show that attributes can reasonably detect the disturbance levels (chance performance for the five disturbance levels is 20%).

Table 2. Per-patch classification rates for prediction of the disturbance level

Translation:	$78.5 \pm 8.3\%$	Rotation:	$78.6 \pm 14.0\%$	Noise:	$79.3 \pm 10.1\%$
Shadow:	$83.8 \pm 6.0\%$	Highlight:	$71.3 \pm 13.1\%$	Complexity:	$65.9 \pm 8.0\%$

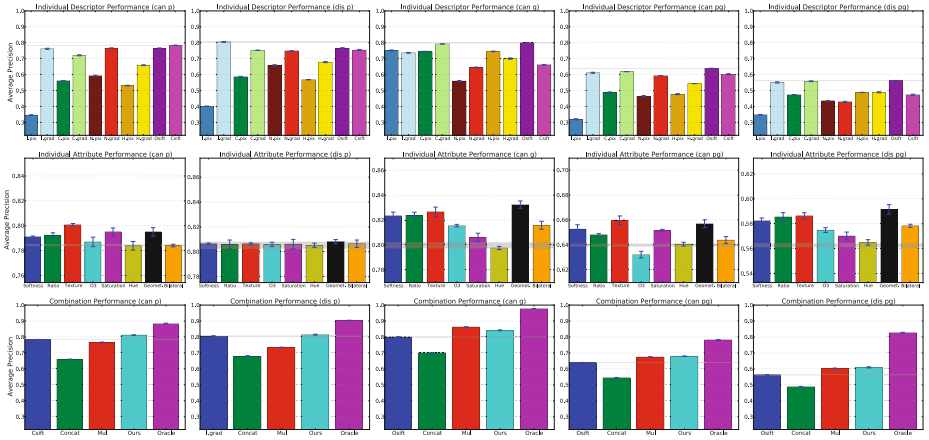
4.2 Aloi Dataset

The Aloi dataset [28] consists of images of 1000 objects under a variety of imaging conditions. These include the illumination direction, illumination color and object viewpoint. The recording setup consists of five light sources, positioned on a hemisphere aslant to the object. Three cameras are positioned next to each other underneath the light sources. Here, we consider images from the two outer cameras that are furthest apart and use eight different light source combinations. Thus, we consider two recordings for every of eight illumination conditions, leading to a total of 16 variations of object appearance.

Patches are extracted in similar spirit to [4]. First, planar homographies between the cameras are computed based on correspondences between sift descriptors extracted from interest points detected by the harris-laplace detector on images from 'canonical' illumination condition $l8$ (all light sources switched on). For this, we use a standard ransac procedure. We impose additional constraints based on camera vicinity, i.e. the transformation should be small and near-translational, regardless of object geometry. Using the obtained homography, we propagate the feature detections in the canonical image $l8$ to all other images and extract rectangular image patches proportional to the detection scale. The non-planarity of most objects causes unbiased geometric variations in the image patches. The patches are resized to 64x64 pixels and patches outside the range of 64 ± 20 pixels are discarded. The dataset consists of about 200K patches. See figure 1(a) for an example.

The descriptor pool is augmented with Osift and Csift descriptors, as these have been shown to be the single optimal choice for a wide range of matching and recognition tasks [3]. These descriptors follow from the descriptor pool (table 1) by concatenation: Osift= $[O_1, O_2, O_3]$ and Csift= $\left[\frac{O_1}{O_3}, \frac{O_2}{O_3}, O_3\right]$. Furthermore, we include combinations of all descriptors by (concat): concatenation of the full descriptor pool and (mul): multiplication of individual distances prior to ranking, both suggested by [10].

Descriptor selection is evaluated in separate settings. Each setting allows isolated analysis of attributes as representations of either object surface or scene-accidental properties. To this end, a query patch is presented in either canonical (can) or distorted (dis) form. We distinguish between distortions of a geometric (g) or of a photometric nature (p), and both (pg). A pure photometric disturbance has all patches recorded by the same camera and has therefore no geometric differences. A purely geometric disturbance considers patches for different cameras, however only under uniform illumination. Note that for this geometric distortion there is no difference between a disturbed or a canonical



(a) can p (b) dis p (c) can g (d) can pg (e) dis pg

Fig. 4. Aloi matching results individually per descriptor (row 1) and per attribute (row 2). In row 3 we show the descriptor selection performance in comparison with the single best descriptor, all descriptors combinations (mul, concat) by multiplication or concatenation, and the best possible descriptor from the descriptor pool (oracle). Several scenarios are evaluated: the query patch may be in either canonical ('can') or distorted form ('dis'), while the disturbance is either photometric ('p'), geometric ('g'), or both ('pg'). See table 3 for an overview. Pixel-based (pix) and gradient-based (grad) descriptors are extracted from intensity, chromatic, normalized chromatic and hue representations, denoted by I., C., N. and H as given in table 1. For easy reference we mark the score of the single best descriptor with a gray bar.

query patch. To these distinct evaluation settings we add the (most realistic) setting in which query patches as well as database patches may arrive in either canonical or distorted form (all). See table 3 for an overview. Matching results in terms of mean average precision are presented in figures 4 and 5.

Individual Descriptor Performance

The results show that gradient-based descriptors generally perform (much) better than pixel-based descriptors. However, if the disturbance is purely geometric (can g) pixel-based descriptors perform at their best. This is because calculating image gradients requires larger spatial support than a single pixel.

Table 3. Named Aloi experiments, results in figures 4 and 5

	Query patch		
Disturbance type	Canonical	Distorted	Both
Photometric	can p	dis p	-
Geometric	can g	can g	-
Both	can pg	dis pg	all pg

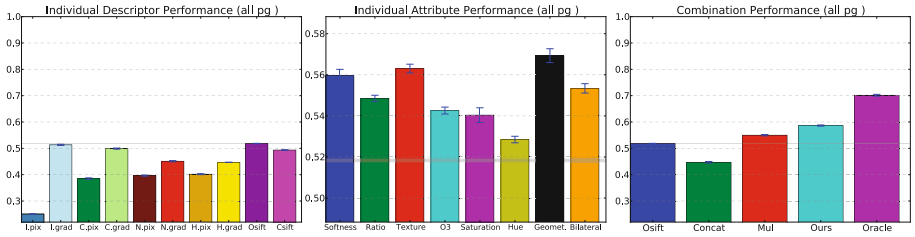


Fig. 5. Most realistic results where the query patch is either canonical or distorted, with photometric and/or geometric distortion. See the caption of figure 4 for the explanation of the symbols.

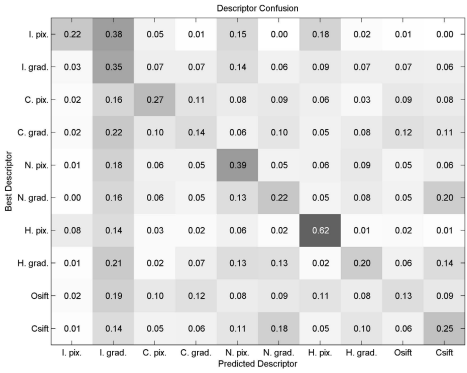


Fig. 6. Predicted and best descriptors confusion (all pg)

In accordance to other work [3], Osift is often the best performing individual descriptor. When the distortion is purely photometric, a distorted query patch (dis p) gives better performance for most descriptors than an undistorted query patch (can p). When the distortion is both photometric and geometric (can pg and dis pg) the situation is reversed. Moreover, the combined photometric and geometric distortions (can pg, dis pg, all pg) perform significantly worse then their single-distortion counterparts (can p, dis p, can g).

Individual Attribute Performance
Texture and geometric stability are the best performing attributes, whereas hue rarely helps. Overall, each attribute generally increases performance. This indirectly shows that they are helpful for descriptor selection. Interestingly, the attributes are able to predict which material will be sensitive to a photometric distortion (can p). However, they fail completely to recognize a photometric distortion (dis p) when it is present.

Descriptor Selection Performance
Our descriptor selection method is always better than the best single descriptor (typically Osift). When comparing descriptor selection to feature combination methods, we perform equal, or better. Concatenating the full descriptor pool produces a very high-dimensional descriptor yet results in poor performance. Descriptor combinations by multiplication often improve over individual

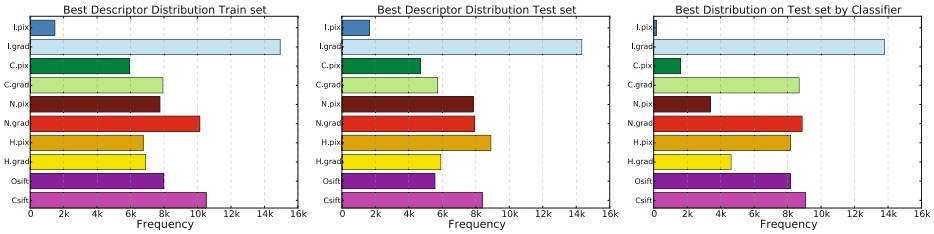


Fig. 7. Distribution of the best descriptors (all pg) on: (a) Train set (b) Test set (c) selected by our classifier

descriptors, however, it fails when there is high variance between individual descriptor performance (can p, dis p). The performance gain of descriptor selection is most prominent in the scenario in which patches appear in either canonical or distorted form under mixed disturbances (all pg) in figure 5.

Analysis of Descriptor Selection
In figure 6 we display the confusion matrix between predicted and best descriptors. The diagonal shows that we outperform random classification performance which for 10 descriptors is 0.1. There is a slight bias towards intensity sift (I.grad), because this is often the best descriptor. In figure 7(a-c) we give the distribution of the predicted and best descriptors. Pixel-based descriptors are frequently superior in both train and test-sets, but hard to select automatically by the classifier. The intensity sift (I.grad) is often the best descriptor, whereas on average Osift is slightly better, as shown in figure 5.

5 Conclusion

This paper introduces a novel descriptor selection framework. Other methods can select a single feature for a whole image, or optimize a single feature over a dataset of patches. We, in contrast, show that the most appropriate descriptor alternates per patch. Therefore, we propose to select the descriptor on a per-patch basis. The selection method operates on attributes extracted from the image, through which object surface and scene properties are measured. These attributes are indicative for the appropriate descriptor. On a large dataset of colored object patches, the proposed selection method is shown to outperform existing sophisticated image descriptors that claim to be invariant to one or more imaging conditions.

References

1. van de Weijer, J., Gevers, T., Geusebroek, J.M.: Edge and corner detection by photometric quasi-invariants. PAMI (2005)

2. Varma, M., Ray, D.: Learning the discriminative power-invariance trade-off. In: ICCV (2007)
3. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. PAMI (2010)
4. Brown, M., Hua, G., Winder, S.: Discriminative learning of local image descriptors. PAMI (2011)
5. Adelson, E.H.: On seeing stuff: the perception of materials by humans and machines. In: Society of Photo-Optical Instrumentation Engineers (SPIE) (2001)
6. Liu, C., Sharan, L., Adelson, E.H., Rosenholtz, R.: Exploring features in a bayesian framework for material recognition. In: CVPR (2010)
7. Cai, H., Mikolajczyk, K., Matas, J.: Learning linear discriminant projections for dimensionality reduction of image descriptors. TPAMI (2010)
8. Philbin, J., Isard, M., Sivic, J., Zisserman, A.: Descriptor Learning for Efficient Retrieval. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 677–691. Springer, Heidelberg (2010)
9. Dubout, C., Fleuret, F.: Tasting families of features for image classification. In: ICCV (2011)
10. Gehler, P.V., Nowozin, S.: On feature combination for multiclass object classification. In: ICCV (2009)
11. Opelt, A., Pinz, A., Fussenegger, M., Auer, P.: Generic object recognition with boosting. TPAMI 28 (2006)
12. Kavukcuoglu, K., Ranzato, M.A., Fergus, R., LeCun, Y.: Learning invariant features through topographic filter maps. In: CVPR (2009)
13. Hu, D., Bo, L., Ren, X.: Toward robust material recognition for everyday objects. In: BMVC (2011)
14. Guo, Y., Zhao, G., Pietikäinen, M., Xu, Z.: Descriptor Learning Based on Fisher Separation Criterion for Texture Classification. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part III. LNCS, vol. 6494, pp. 185–198. Springer, Heidelberg (2011)
15. Caputo, B., Hayman, E., Fritz, M., Eklundh, J.O.: Classifying materials in the real world. Image Vision Comput., 150–163 (2010)
16. Dror, R.O., Adelson, E.H., Willsky, A.S.: Recognition of surface reflectance properties from a single image under unknown real-world illumination. In: CVPR Workshop on Identifying Object Across Variations in Lighting (2001)
17. Romeiro, F., Zickler, T.: Blind Reflectometry. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 45–58. Springer, Heidelberg (2010)
18. Wang, O., Gunawardane, P., Scher, S., Davis, J.: Material classification using brdf slices. In: CVPR (2009)
19. Shafer, S.A.: Using color to separate reflection components. Color Research and Applications 10, 210–218 (1985)
20. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. PAMI 99 (2012)
21. van Gemert, J.C.: Exploiting photographic style for category-level image classification by generalizing the spatial pyramid. In: ICMR (2011)
22. Motoyoshi, I., Nishida, S., Sharan, L., Adelson, E.H.: Image statistics and the perception of surface qualities. Nature 447, 206–209 (2007)
23. Sharan, L., Li, Y., Motoyoshi, I., Nishida, S., Adelson, E.H.: Image statistics for surface reflectance perception. J. Opt. Soc. Am. A 25, 846–865 (2008)
24. Gijsenij, A., Gevers, T., van de Weijer, J.: Improving color constancy by photometric edge weighting. PAMI (2011)

25. Yanulevskaya, V., Geusebroek, J.M.: Significance of the weibull distribution and its submodels in natural image statistics. In: VISSAP (2009)
26. Zadrozny, B., Langford, J., Abe, N.: Cost-sensitive learning by cost-proportionate example weighting. In: IEEE International Conference on Data Mining (2003)
27. Barnard, K., Martin, L., Funt, B., Coath, A.: Data for colour research. Color Research and Application (2000)
28. Geusebroek, J.M., Burghouts, G.J., Smeulders, A.W.M.: The amsterdam library of object images. IJCV (2005)