Compact Object Descriptors from Local Colour Invariant Histograms

Jan-Mark Geusebroek *
Intelligent Systems Lab Amsterdam, Informatics Institute,
University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam.
mark@science.uva.nl

Abstract

Much emphasis has recently been placed on the detection and recognition of locally (weak) affine invariant region descriptors for object recognition. In this paper, we take recognition one step further by developing features for non-planar objects. We consider the description of objects with locally smoothly varying surface. For this class of objects, colour invariant histogram matching has proven to be very encouraging. However, matching many local colour cubes is computationally demanding. We propose a compact colour descriptor, which we call Wiccest, requiring only 12 numbers to locally capture colour and texture information. The Wiccest features are shown to be fairly insensitive to photometric effects like shadow, shading, and illumination colour. Moreover, we demonstrate the features to be applicable to highly compressed images while retaining discriminative power.

1 Introduction

There has always been a drive for deriving good features. Many computer vision tasks depend heavily on local feature extraction. Object recognition is considered a typical case where local information is gathered to obtain evidence for recognition of previously seen objects. Recently, much emphasis has been placed on the detection and recognition of locally (weak) affine invariant regions [11, 12, 13, 14, 18]. The rationale here is that planar regions transform according to well known laws. Successful methods rely on fixing a local coordinate system to a salient image region, resulting in an ellipse describing local orientation and scale. After transforming the local region to its canonical form, descriptors like SIFT [11] are well able to capture the invariant region appearance. Indeed, for such a setting the detection of affine regions combined with the SIFT descriptor is shown to be better than many alternatives [12]. Given the success of these approaches, the detection and recognition of planar regions may even be considered a (close to) solved problem.

In this paper, we take object recognition one step further by developing features for non-planar patches. We consider the description of properties of locally non-planar but smoothly varying objects, see Figure 1. The problem is very different from the description of planar regions. For one, interest points are less obvious to detect, and if detected,

*This work is sponsored by the Netherlands Organisation for Scientific Research (NWO). Part of the paper is written during the author's visit to the Robotics Research Group, University of Oxford.

Figure 1: Example objects with illumination direction varied from left to right. The example objects are increasingly hard to index by current methodology. Affine region descriptors work well for the tea box, as it contains planar regions with rich internal structure. For planar regions, a difference in light direction will merely cause a variation in intensity of the surface. A slightly harder case is presented by the marmalade bottle, but may still be recognised by affine region descriptors due to its smooth convex shape and rich structure at its label. A large step further is the smooth type of concavities and convexities often occurring in natural objects, represented by the red pepper. No obvious keypoints are present, and self-shadow, shading and highlights present severe difficulties for object descriptors. The teddy is an example of a smooth shape with rough material texture. The teddy's texture contrast depends heavily on illumination direction, and is affected by local shading and self shadowing. Although these illustrations represent laboratory extremes, as no ambient illumination is present, these effects also manifest itself in normal indoor and outdoor imaging conditions.

they often do not attach to reproducible locations on the object. Furthermore, where the appearance of planar regions is marginally affected by changes in viewing conditions, this has drastic effects for locally convex or concave object, see Figure 1. Under these circumstances, descriptors should handle shadow and shading effects.

Under these circumstances, object recognition based on colour invariant features have proven to be very encouraging in the past [3, 6, 10]. Colour invariant descriptors normalise the local intensity in the image, hence do not "see" shadow and shading effects. As such, they counteract the effect of shadow and shading which cause a distorted appearance of the object as could be obtained from the intensity channel only. However, application of colour invariance in object recognition has not emerged beyond a limited number of applications, mainly being in content based image retrieval [19]. Besides the mere challenge in understanding the physics needed to grasp colour invariance, and make effective use of it, two other drawbacks are visible. For one, colour invariants used so far lack the stability and robustness to withstand compression. The argument that all images are in colour nowadays, and hence this information should be exploited, is simply countered by the fact that almost all of these images are compressed. Compression algorithms typically use most of the bits for intensity information, coding a poor estimate of the chromaticity. Although this will not too much affect indexing based on the global colour histogram, local colour statistics are severely altered.

A second issue against the use of colour is its computational complexity. Computing power and available memory increase with Moore's law, which makes channel wise processing of colour images effectively no problem at today's machines. However, for

object recognition, histograms of colour models are often evaluated, which requires a 3D colour cube to be stored in memory for each keypoint or image region. Processing and matching of these cubes require simply to much computation time, often circumvented by down sampling to a 32x32x32 cube [10] or by histogram compression [1]. The result is still an awful lot more numbers than, e.g. SIFT features, which only stores 128 values per location. These drawbacks prevent the large scale use of colour for object recognition.

In the recent past, colour invariants have been proposed based on Gaussian scale-space [9]. These features improve upon pixel-based invariants [5, 10] in that they overcome the problem of noise and compression sensitivity, as a better estimate of colour value is obtained by simply choosing a larger integration scale of the Gaussian filter. In this way, compression artifacts are averaged, yielding a more stable result for colour invariants. Van de Weijer *et al.* [21] continued this work, and dealt with the problem of stability at low intensities and low saturation. However, colour in natural images is often not simply described by a single assignment of an average colour value. Many natural objects have not a single unique colour, it is rather the blend of colours which give the objects its distinctive visual characteristics. This is the very principle which turned colour based histogram matching into a success in image retrieval [19] and object tracking [2]. We will include this principle in the very heart of our feature design.

In this paper, we build onto these past successes, and develop a new class of local colour features specifically targeted for object recognition. We derive a photometric and geometric invariant descriptor based on the local histogram of colour edges. We analyse how photometric transformations affect the (local) statistics of colour edges. Our main contribution targets the dimensionality problem for colour features. The RGB histogram (or intensity histogram alike) can have almost any shape for arbitrary images, albeit constraint by the physics of light reflection. The histogram of edges is tightly constraint by local correlations. As a consequence, edge histograms follow a simple shaped probability density. We exploit this a-priori structure in images by parameterising edge histograms according to the statistics of natural images. We derive invariant properties of these statistics, yielding only 12 numbers to adequately describe local colour edge histograms. We demonstrate our method on a database of highly curved and complex 3D objects. The features are systematically tested under changes in illumination colour, illumination direction, viewing direction, and image compression.

2 Preliminaries

2.1 Colour Image Formation Model

We start our analysis of colour features with a short rehearsal of the physics of colour image formation. Symbols and reflectance models are defined in accordance with [9]. We restrict to a model which takes a directed light source and an ambient diffuse illumination into account. Directed light is for example sunlight or spotlight, whereas ambient light is present due to the sky or the reflectance from walls and ceiling. These light sources are modeled locally, for the extent of a single feature. The photometric reflectance model consists of a diffuse body reflection component, a specular interface reflectance component, and an ambient illumination component [17],

$$E(\lambda,\vec{x}) = i(\vec{x})e(\lambda)R(\lambda,\vec{x}) + e(\lambda)\rho(\vec{x}) + a(\lambda) \tag{1}$$

where $\vec{x}$ denotes the position at the imaging plane and λ the wavelength. Further, $e(\lambda)$ denotes the direct illumination spectrum. The ambient illumination term is given by $a(\lambda)$. Note that the spectral distribution of $a(\lambda)$ maybe different from the spectrum of $e(\lambda)$, thereby including coloured cast shadows. The combined effect of intensity flux and illumination intensity variations at the object surface is given by $i(\vec{x})$, and encodes shading and non-coloured shadow effects. The Fresnel reflectance yielding specular (interface) reflection is denoted by $\rho(\vec{x})$. The material reflectivity is denoted by $R(\lambda,\vec{x})$. The reflected spectrum in the viewing direction is given by $E(\lambda,\vec{x})$. We are especially interested in the body reflectance $R(\lambda,\vec{x})$, as it is indicative for the "true" object's colour. Hence, we aim at deriving photometric invariant features, that is, features solely depending on $R(.)$.

2.2 Gaussian Based Colour Measurements

In [9], a Gaussian scale-space framework for colour features has been proposed. In short, each pixel's RGB value is transformed into an opponent colour representation. The rationale behind this transformation is that the RGB sensitivity curves of the camera are transformed to Gaussian basis functions, being the Gaussian and its first and second order derivative. Hence, the transformed values represent an opponent colour system. Advantage of the use of an opponent colour space is that colour values are decorrelated. Spatial scale is incorporated by convolving the opponent colour images with a Gaussian filter,

$$\hat{E}_{klm}(x,y,\sigma_i) = G_{kl}(x,y,\sigma_i) * E_m(x,y) \quad , \tag{2}$$

where $G_{kl}(x,y,\sigma_i)$ represents the Gaussian derivative filter of derivative order k in the x-direction and order l in the y-direction, and E_m represents the opponent colour channels E, E_λ, $E_{\lambda\lambda}$. Note that we now have a spatial Gaussian times a spectral Gaussian, yielding a combined Gaussian measurement in a spatio-spectral Hilbert space.

2.3 Local Kernel Based Histogram Estimation

Localisation and spatial extent (scale) of local histograms is obtained by weighing the contribution of pixels by a Gaussian kernel,

$$h_{x_o,y_o,\sigma_o}(i) = \sum_{x,y} G(x-x_o,y-y_o;\sigma_o)\,\delta\left[r_{\sigma_i}(x,y)-i\right] \quad , \tag{3}$$

where δ is the Kronecker delta function, $r_{\sigma_i}(x,y)$ is a discretised version of one of the Gaussian (derivative) filter responses (eq. 2), and $G(.)$ is the Gaussian kernel. The histogram $h(i)$ is constructed by taking all pixels with discretised value i, and adding there weighted contribution, weighed by kernel $G(.)$, to the histogram bin i. The parameter σ_o represent the size of the kernel, not to be mistaken for the scale σ_i of the Gaussian filters (eq. 2). Hence, we have an "inner" scale σ_i at which point measurements are taken, which are accumulated over an "outer" scale σ_o into a local histogram.

3 Compact Invariant Descriptors for Local Histograms

3.1 Quasi Colour Invariant Derivatives

Our measurement framework, laid down in Section 2.2, allows us to measure (smoothed) derivatives of the reflectance function. Hence, like in [9], we adopt a differential frame-

work to derive image features. We start by deriving the three opponent colour channels from (eq. 1), omitting function parameters for brevity,

$$\begin{aligned} E(\lambda,x) &= ieR + e\rho + a \\ E_\lambda(\lambda,x) &= i(e_\lambda R + eR_\lambda) + e_\lambda \rho + a_\lambda \\ E_{\lambda\lambda}(\lambda,x) &= i(e_{\lambda\lambda} R + 2e_\lambda R_\lambda + eR_{\lambda\lambda}) + e_{\lambda\lambda}\rho + a_{\lambda\lambda} \quad , \end{aligned} \tag{4}$$

indices denoting differentiation. Note that after any spatial derivative of these colour channels, the ambient illumination term $a(\lambda)$ will vanish, as it is locally constant with respect to the spatial coordinates. Hence,

$$\begin{aligned} E_x(\lambda,x) &= e(iR_x + i_x R + \rho_x) \\ E_{\lambda x}(\lambda,x) &= e(iR_{\lambda x} + i_x R_\lambda) + e_\lambda (iR_x + i_x R + \rho_x) \\ E_{\lambda\lambda x}(\lambda,x) &= e(iR_{\lambda\lambda x} + i_x R_{\lambda\lambda}) + 2e_\lambda (iR_{\lambda x} + i_x R_\lambda) + e_{\lambda\lambda}(iR_x + i_x R + \rho_x) \quad . \end{aligned} \tag{5}$$

These derivatives are independent of a constant additive term to each opponent colour channel. As such, the measured derivatives $\hat{E}_x$, $\hat{E}_y$, $\hat{E}_{\lambda x}$, $\hat{E}_{\lambda y}$, $\hat{E}_{\lambda\lambda x}$, $\hat{E}_{\lambda\lambda y}$, are invariant for additive changes caused by (possibly coloured) ambient illumination, (coloured) cast shadows, and camera offset values. Furthermore, the zero order effect of changing the colour of the light source is a shift in the distribution of colours in the opponent colour channels $\hat{E}_\lambda$ and $\hat{E}_{\lambda\lambda}$. By taking local differences between colour values, this constant offset again is canceled out. This is effectively incorporated in colour difference measures for well known opponent colour spaces like Luv and CIELab. Hence, taking the spatial derivative as above yields already a fair robustness against local illumination colour changes. Note that multiplicative intensity effects are still present in the derivatives. Hence, we still have to deal with global intensity changes, local intensity effects of shading and shadow, non-uniform illumination and with specularities. Before doing so, we first consider the local statistics of these quasi-invariant derivatives.

3.2 Compact Features by Histogram Parameterisation

Compactness of the local histogram representation of filter responses may be obtained by clustering techniques like principal component analysis [1] or by quantisation of the histogram through K-means clustering over a large set of filter responses [23]. Alternatively, one could parameterise the histogram by fitting a functional through the histogram values, and subsequently storing the parameters of the fitted function. This may yield a condense description, provided that the functional closely approximates the histogram. Popular functionals include fitting to a mixture of Gaussians or the normal distribution itself. However, from natural image statistics research, it is known that histograms of derivative filters can be well modeled by a simple distribution [20]. The local histogram of invariants of derivative filters can be well modeled by an integrated Weibull type distribution, also known as Generalised Laplacian,

$$p(r) = \frac{\gamma}{2\gamma^{\frac{1}{\gamma}}\beta\Gamma(1/\gamma)} \exp\left\{ -\frac{1}{\gamma}\left|\frac{r-\mu}{\beta}\right|^\gamma \right\} \quad . \tag{6}$$

In this case, r represents edge response of a derivative filter. Furthermore, $\Gamma(\alpha)$ represents the complete Gamma function, $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$. The parameter μ denotes the

origin of the distribution, the parameter β denotes the width of the distribution, and the γ parameter indicates the peakness of the distribution. Note that the integrated Weibull distribution is only under very strict circumstances close to Gaussian. This is only the case if the image depicts high frequency noise, such that the edge responses are normally distributed. For general images, the γ parameter will often be within the interval $[0.5\ldots1]$ [8]. By using the Weibull parameters, one obtains an accurate and very compact parameterisation of the derivative histogram.

Estimates for the parameters of the integral form of the Weibull distribution are obtained by the maximum likelihood method. Transforming the estimated density to the log domain and subsequently differentiation to the parameters μ, β, and γ, respectively, and setting them to zero, yields the following estimators for μ and β,

$$\hat{\mu} = \sum_i h(r_i) r_i \quad , \quad \hat{\beta} = \left[\sum_i h(r_i) \left(r_i - \hat{\mu}\right)^{\hat{\gamma}} \right]^{1/\hat{\gamma}} \quad ,$$

where $h(.)$ represents the kernel based local histogram of one of the colour invariants. The derivative of the log-likelihood to γ is given by

$$\hat{\gamma} + \log\hat{\gamma} + \Psi(1/\hat{\gamma}) - 1 + \sum_i h(r_i) \left| \frac{r_i - \hat{\mu}}{\hat{\beta}} \right|^{\hat{\gamma}} \left(1 - \hat{\gamma} \log \left| \frac{r_i - \hat{\mu}}{\hat{\beta}} \right| \right) = 0 \quad , \qquad (7)$$

where $\Psi(.)$ denotes the digamma function, the logarithmic derivative of the gamma function. The final equation is optimised by a dichotomic search scheme, implying searching for the $\hat{\beta}$ and $\hat{\gamma}$ combination (by varying $\hat{\gamma}$) for which (eq. 7) is closest to zero.

3.3 Geometric Invariance

Before continuing with the final derivation of compact invariant colour descriptors, geometrical invariance of the Weibull parameters has to be addressed. So far, Weibull parameters were estimated from the response histogram of single derivative filters. These responses do depend on the orientation of the image content. A trivial solution is to use the rotationally invariant gradient magnitude, in accordance to [15]. However, one looses discriminative power, as in that case homogeneity between edge direction over regional scale σ_o (eq. 3) is lost. Rather we aim for a coherent description of local image structure.

Consider the steerability of Gaussian derivative filters. If one takes a derivative filter in the x and y-direction, a derivative in any other direction may be achieved by the linear combination $E_\theta = E_x \cos\theta + E_y \sin\theta$ where E_θ is the resulting response of a derivative filter in the θ-direction. The E_x and E_y responses are orthogonal, each being characterised by an integral Weibull type probability density, although they may have different parameters. From probability theory, we know that a weighed sum of independent random variables result in a probability density given by the convolution of the individual densities. As a consequence, the Weibull parameters span ellipses when plotted as function of angle. The shortest and longest principal axis for β and γ, together with the orientation of the ellipse, indicate the directional structure in the underlying edges. Rotational invariance is achieved by estimating the longest and shortest principal axes of these ellipses, disregarding its orientation. Many methods exists for elliptic fitting. As a simple solution, and one we will apply, estimate the γ and β for 0°, 45°, 90°, and 135° derivative

filters (using the steering property), and use a least square fitting to obtain the shortest and longest axes for γ and for β, which characterises the local histogram invariant to rotation of the original image.

3.4 Weibull Invariant Colour Contrast Estimator

We are now in a position to derive a fully photometric and geometric invariant descriptor from the Weibull parameters of the local colour edge histogram. First consider the mean value μ of the Weibull parameters. In many cases, this will be close to zero, as there often are as much rising slopes as falling slopes in an image. Deviation from zero is due to a locally non-uniform illumination (possibly by shading). We assume the directed light source and shading component to be slowly varying over the image plane, such that the illumination is locally planar, although not uniform, over the extent σ_o of the local region from which statistics are taken. A planar illumination will cause an offset to the derivative values obtained from (eq. 5). Note that this offset, when deviating from zero, will be present in all three (derivative-) opponent colour channels. To arrive at photometric invariance, the local histogram should be zero centred by subtracting its mean value. Hence, μ is to be ignored.

We are still left with a multiplicative term due to the intensity of the directed light source, the term being reflected in the width β of the Weibull distribution. A meaningful measure representing the variation in the local derivatives is the average colour gradient magnitude. As derived above, derivatives are the lowest order measures not affected by ambient illumination and contrast manipulations. However, a larger ambient illumination component relative to the direct illumination component reduces the contrast in the image. In that case, normalisation becomes less stable as the average contrast diminishes, even when the area is well illuminated . Being pragmatic, one may as well consider the local average intensity to normalise for the intensity component, yielding only instabilities at low intensities. Hence, we consider the local (Gaussian weighted) average intensity $\bar{E}$ to normalise the width of the distribution, $\beta' = \beta/\bar{E}$, which yields robustness against low-intensity values, noise, and compression artifacts, as long as the average intensity is above (compression) noise level. The latter is easily checked.

The remaining parameter γ indicates the (local) roughness or textureness, and is a photometric invariant. In summary, by estimating of the longest and shortest principal axes of the γ and contrast normalised $\beta' = \beta/\bar{E}$ parameters for each opponent colour channel, one obtains the twelve "Wiccest" parameters describing a local region in colour and edge (texture) content.

3.5 Summary of invariances achieved

We have theoretically shown the Wiccest parameters to be either invariant or highly robust to: global intensity changes, local intensity effects of shading and (coloured) shadow, non-uniform illumination, ambient illumination, coloured illumination, Euclidean transformations, and (compression) noise. The remaining photometric effect which is affecting the Wiccest features is specular reflection. However, in many cases the specularities are either not dominating the local histogram –the specular area being much smaller than the scale σ_o of the local region–, or results in an outlier during the matching phase.

4 Experiment: Object Recognition from One Example

To assess the constancy of the proposed Wiccest feature under varying imaging conditions, the features has been applied to the ALOI collection [7]. The collection consists of 1,000 objects recorded under various imaging circumstances. Specifically, viewing angle, illumination angle, and illumination colour is systematically varied for each object. In our experiment, we investigate object recognition by indexing only one example of each object - a problem considered non-trivial given the variety in imaging conditions. As example image for each object the *l8c1* image is taken, which is a frontal view of the object, under semi-hemispherical white illumination.

To illustrate the effectiveness of the proposed features in capturing object properties, a simple algorithm for object recognition is suggested. Our setup should be considered a mere baseline indication of performance. Objects are characterised by densely sampling the image, locations being $2\sigma_o$ apart. A threshold just above noise level is set on the contrast of the intensity channel, hence disregarding locations without edge content or very low in intensity. Matching between images is straightforwardly performed by comparing the Wiccest parameters for each region of the query image against all regions in the target image, and accumulating the best scores per query region. Regions were compared by calculating the fraction between the respective parameters β and γ, $score = \sum(\min(\beta_1,\beta_2)/\max(\beta_1,\beta_2))(\min(\gamma_1/\gamma_2)/\max(\gamma_1,\gamma_2))$, where the sum accumulates the three scores of the opponent colour channels. Although the algorithm seems expensive at first sight, remember that only 12 numbers capture the local colour histogram. Hence, the proposed scheme is much more efficient than local histogram matching ([3]), and much more efficient than k-nearest-neighbour search over large collections ([11]).

We applied our algorithm to the four times down-sampled version of ALOI. In our experiment, the scale for the derivative filters was set at a typical value of $\sigma_i = 1$ pixels. The scale for the kernels was set at $\sigma_o = 7.5$ pixels. In practice, an average of 29 non-empty kernels (348 numbers) per object were indexed. For the same collection, keypoint extraction combined with the SIFT descriptor yields an average of 60 keypoints (7,924 numbers) per object. Object recognition performance was tested by evaluating correct recognition rates under varying illumination direction, varying illumination colour, and under varying object viewing direction. To evaluate the robustness against compression, we re-indexed the example images, but now after JPEG compression at a quality of 75% of the original image. We evaluated recognition rates as function of compression quality.

To indicate that our results are non-trivial and indeed progress on the state-of-the-art, we included results for the method by Lowe [11] based on Laplacian keypoints and the SIFT descriptor (program courtesy of David Lowe available from his website). Furthermore, as a second baseline, we include results for global RGB histogram matching and normalised rgb histogram matching, implemented as in [10]. Note that histogram matching is a global method, where SIFT and the proposed Wiccest features are local descriptors which can handle occlusions and clutter (although not shown in our experiments). Results for normalised rgb are similar to RGB, hence are not shown.

The results are shown in Figure 2. As expected, Lowe's method [11] based on intensity information is insensitive to illumination colour changes, whereas RGB histogram matching is extremely sensitive to an illuminant change. The proposed method is fairly colour constant, with a 27% error at the worst condition *i110*. Note that this is no sinecure for a colour based method. For variation in illumination direction, the proposed method

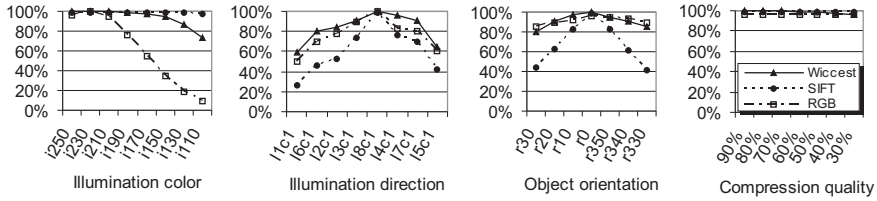

Figure 2: Object recognition performance as function of varying imaging conditions.

gradually degrades until the extreme oblique illuminations *l1c1* and *l1c5*. These conditions are exceptional, and not often encountered in practice. The method is fairly robust to a change in viewpoint, allowing the indexing of object views as far as 45 degrees apart. RGB histogram matching performs well in this case. Most important, the proposed method is shown to be very robust against compression, and keeps well up with the alternatives. Regarding computation time, matching one images against the 1,000 in the database takes less than half a second on a nowadays laptop. Just as a reference, for the same setup, SIFT takes about 7 seconds per image, 0.89 seconds for calculating the SIFT descriptor at each keypoint location, and slightly over 6 seconds for matching against the 1,000 database images. Global histogram matching takes even more time, mainly due to the huge volume of memory that has to be matched over and over again while intersecting against 1,000 histograms of 32x32x32 bins. Here, one clearly sees the advantage of compact descriptors.

5 Conclusions

We proposed a compact object descriptor, Wiccest, requiring only 12 numbers to locally capture colour and texture information. Although the assumptions underlying our method seem restricted –smoothly varying surfaces, photometrically constraint by a simple reflectance model–, we applied the method successfully a) on a large collection of objects (this paper); b) under different imaging conditions (this paper); c) under severe JPEG compression (this paper); d) in MPEG compressed video retrieval (TRECVID [4, 22] - top rank performance); e) real-time recognition of over a hundred objects by a Sony Aibo robodog [16]. Achieving these results requires highly robust and discriminative features.

Our experiments are preliminary in that they do not include significant scale change, occlusion, and background clutter. The proposed matching scheme accumulates local features. Object recognition based on local features has proven to be robust to clutter and occlusion. From our experience, the suggested approach seems fairly robust to these effects [16, 4, 22]. However, a thorough evaluation remains a point of future research. Regarding scale invariance, we included scale as a free parameter in the feature design. Hence, densely sampling Wiccest features at multiple scales implies a form of scale invariance. However, the matching of multi-scale Wiccest features needs to be solved.

The Wiccest features are shown to be highly robust to common photometric effects like shadow, shading, and illumination colour, and, most important, retain their properties and discriminative power under image compression.

References

[1] J. Berens, G. D. Finlayson, and G. Qiu. Image indexing using compressed colour histograms. *IEE Proc. Vis. Image Signal Process.*, 147:349–355, 2000.

[2] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Trans. Pattern Anal. Machine Intell.*, 25:564–577, 2003.

[3] F. Ennesser and G. Medioni. Finding Waldo, or focus of attention using local color information. *IEEE Trans. Pattern Anal. Machine Intell.*, 17:805–809, 1995.

[4] C. G. M. Snoek et al. Mediamill: Exploring news video archives based on learned semantics. In *Proc. ACM Multimedia*, pages 225–226, 2005.

[5] G. D. Finlayson. Color in perspective. *IEEE Trans. Pattern Anal. Machine Intell.*, 18(10):1034–1038, 1996.

[6] B. V. Funt and G. D. Finlayson. Color constant color indexing. *IEEE Trans. Pattern Anal. Machine Intell.*, 17(5):522–529, 1995.

[7] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders. The Amsterdam library of object images. *Int. J. Comput. Vision*, 61(1):103–112, 2005.

[8] J. M. Geusebroek and A. W. M. Smeulders. Fragmentation in the vision of scenes. In *Proc. 9th Int. Conf. Comput. Vision*, volume 1, pages 130–135. IEEE Computer Society, 2003.

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

[10] T. Gevers and A. W. M. Smeulders. Color based object recognition. *Pat. Rec.*, 32:453–464, 1999.

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

[12] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Machine Intell.*, 27:1615–1630, 2005.

[13] S. Odbrzalek and J. Matas. Object recognition using local affine frames on distinguished regions. In *Proc. BMVC*, 2002.

[14] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *Int. J. Computer Vision*, 66:231–259, 2006.

[15] C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *IEEE Trans. Pattern Anal. Machine Intell.*, 19:530–534, 1997.

[16] F. J. Seinstra and J. M. Geusebroek. Color-based object recognition by a grid-connected robot dog. In *ECCV Demonstrator*, 2006.

[17] S. A. Shafer. Using color to separate reflection components. *Color Res. Appl.*, 10(4):210–218, 1985.

[18] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, 2003.

[19] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(12):1349–1380, 2000.

[20] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S. C. Zhu. On advances in statistical modeling of natural images. *J. Math. Imaging Vision*, 18:17–33, 2003.

[21] J. van de Weijer, T. Gevers, and J. M. Geusebroek. Color edge and corner detection by photometric quasi-invariants. *IEEE Trans. Pattern Anal. Machine Intell.*, 27(4):625–630, 2005.

[22] J.C. van Gemert, J.M. Geusebroek, C.J. Veenman, C.G.M. Snoek, and Arnold W.M. Smeulders. Robust scene categorization by learning image statistics in context. In *CVPR Workshop on Semantic Learning Applications in Multimedia (SLAM)*, 2006.

[23] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Proc. ICCV*, 2005.