CSIFT: A SIFT Descriptor with Color Invariant Characteristics

Alaa E. Abdel-Hakim and Aly A. Farag
Computer Vision and Image Processing Laboratory (CVIP)
University of Louisville, Louisville, KY 40292,USA
http://www.cvip.uofl.edu
[alaa,farag]@cvip.uofl.edu

Abstract

SIFT has been proven to be the most robust local invariant feature descriptor. SIFT is designed mainly for gray images. However, color provides valuable information in object description and matching tasks. Many objects can be misclassified if their color contents are ignored. This paper addresses this problem and proposes a novel colored local invariant feature descriptor. Instead of using the gray space to represent the input image, the proposed approach builds the SIFT descriptors in a color invariant space. The built Colored SIFT (CSIFT) is more robust than the conventional SIFT with respect to color and photometrical variations. The evaluation results support the potential of the proposed approach.

1. Introduction

Color is an important component for distinction between objects. If the color information in an object is neglected, a very important source of distinction may be lost. The objects of Figure 1 are good examples for the importance of considering color information for object distinction. In this figure, we can see clearly how the pure gray-based geometric description can cause confusion between two completely different features. Nevertheless, most of the existing approaches use gray geometric-based feature extractors. On the other hand, color-based image retrieval approaches neglect the geometrical characteristics of objects. Thus, most research studies in feature extraction for object recognition and matching problems have been focusing on either geometric or color features. Geometric features of an object are extracted in high informative regions like corners. Other kinds of approaches use the luminance and/or color signature in order to describe an object. Color histograms [25] and gray level histograms [22] are well-known luminance-based approaches. The color histograms concept has been extended to include some sort of illumination invariance by using color ratios of neighboring pixels [20] or by using illumination-invariant moments for color histogram distributions [8, 24].

For all of those approaches, the invariance with respect to imaging conditions represents the biggest challenge. Specifically, the extracted features should be invariant with respect to geometrical variations, such as translation, rotation, scaling, and affine/projective transformations. At the same time, these features should be invariant with respect to photometric variations such as illumination direction, intensity, colors, and highlights. Therefore, several research studies in the literature have been presented to develop feature descriptors that maximize the robustness with respect to these variations.

In geometrical invariant approaches, local features are preferred because of their robustness to partial appearance and their lower sensitivity to global displacements in the image [16, 23]. Nearly all geometrical invariant approaches avoid dealing with colored images; since colors add another layer of difficulty represented in the color constancy problem. Therefore, color invariance is a crucial problem which has to be solved for distinct object description and recognition. Many research studies have been presented to solve the color constancy problem [3, 6]. The normalized RGB representation [9] has been used to partially achieve the illumination invariance. Some other invariant color representations have been developed depending on statistical-based transformations [1, 21]. As a more sophisticated approach, various physical-based color invariants have been developed in [11] for invariant color representations under different imaging conditions.

Pure geometric-based approaches may have difficulties in describing *"non-geometric objects"* and they may fail in differentiating between many objects [23]. On the other hand, due to the global nature in photometric-based approaches, they suffer from partial visibility and *"extraneous features"* [23]. In spite of their relatively few number, some research studies in the literature have been presented to combine geometrical and color features. For ex-

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

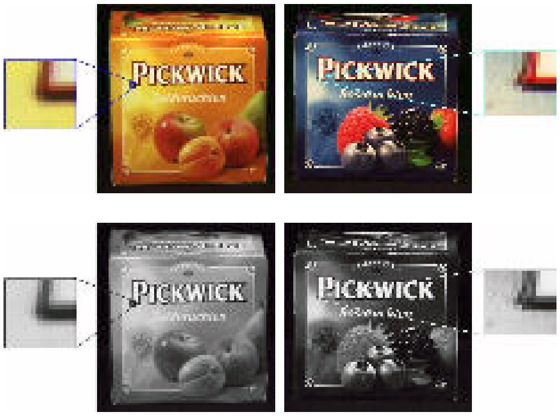


Figure 1. Neglecting the color content may affect the object distinction. Note the big similarity between the two magnified corners which occurs when discarding the color information.

ample, in [12], color and shape invariants are combined for image retrieval. However, the color invariants in that approach are very sensitive to the noise around their singularities. Also, the geometrical invariants are primitive when compared with the pure gray-based approaches.

Scale Invariant Feature Transform (SIFT) [16, 17] has been proven to be the most robust among the other local invariant feature descriptors with respect to different geometrical changes [19]. SIFT was mainly developed for gray images which limits its performance with some colored objects. However, there are some attempts in the literature which have been introduced to make use of the color information inside the SIFT descriptors. For example, in [4], the normalized RGB model has been used in combination with SIFT to achieve partial illumination invariance besides its geometrical invariance. The color invariance of this approach is still limited because of the primitive color model used. In [7], a multi-stages recognition approach has been developed in order to achieve both color and geometrical invariance. In the first stage, a color classifier is used label the different image regions. Then, the SIFT descriptors are augmented by adding the color labels. In spite of the good performance of this approach, its need for colored learning instances limits its performance in several applications.

In this paper, we present a novel Colored SIFT (CSIFT), not to just embed the color information in the descriptors, but to give the built descriptor the robustness with respect to color variations as well as the robustness of the conventional SIFT against geometrical changes. The proposed CSIFT approach is compared to the conventional SIFT approach [16, 17]. The evaluation results show that CSIFT is more stable and distinctive with respect to variations in the photometrical imaging conditions.

2. Problem Statement

The problem of object description using local invariant approaches can be looked at as the problem of transforming the object image into a set of feature vectors or descriptors. For good object description, two criteria should be satisfied in the extracted features. The first one is the stability, i.e. the extracted features should be invariant to different photometric and geometric changes. The second one is the distinctiveness, which means that the extracted features should have the minimum information to distinguish between the object which they describe and other objects. In section (3), we discuss the geometrical invariance, whereas in section (4) we focus on the color invariance. In section (5), we explain our proposed CSIFT approach for combining both geometrical and color invariants in a single descriptor. Finally, we show some evaluation results that support the potential of CSIFT.

3. Geometrical Invariance

Geometrical invariance means the invariance of the extracted features to translation, rotation, scaling, or affine transformations as well as occlusion and partial appearance. In other words, for a specific object, a feature $F(\vec{x})$ at a location $\vec{x} = (x, y)$ should satisfy the following condition:

$$F(\vec{x}) = F(T\vec{x}) \tag{1}$$

where T is a transformation which includes translation, rotation, scaling or affine transformation.

The locality of the extracted features and the way in which the descriptors are built provides the invariance with respect to these geometrical variations, as shown in section (5). The more challenging point is the invariance to scale changes.

Scale-space theory offers the main tools for selecting the most robust feature locations, or the interest points, against scale variations. Given a signal $f : \mathbb{R}^N \rightarrow \mathbb{R}$, the scale-space representation $L : \mathbb{R}^N \times \mathbb{R}_+ \rightarrow \mathbb{R}$ is defined as:

$$L(\vec{x}, t) = g(\vec{x}, t) * f(\vec{x}) \tag{2}$$

where $L(\vec{x}, 0) = f(\vec{x}) \forall \vec{x} \in \mathbb{R}^N$ and $g(\vec{x}, t)$ is the scale-space kernel. As t increases, the scale-space representation $L(\vec{x}, t)$ of the signal tends to coarser scales.

It has been proven that the Gaussian kernel is the unique kernel for generating the scale-space representation [15]. Moreover, Lindeberg [14] has shown that the normalization of the Laplacian of Gaussian, $\nabla^2 g$, with a factor $\sigma^2 = t$ is necessary to give a signal the scale-invariance property. Empirically, it has been proven that the maxima and minima of $\sigma^2 \nabla^2 g$ produces the most stable image features [18]. The normalized Laplacian of Gaussian pyramid can be approximated by a difference-of-Gaussian pyramid [17]. Hence, the locations of the maxima and minima in the difference-of-Gaussian pyramid correspond to the most stable features with respect to scale changes.

4. Color Invariance

In this paper, we use the color invariance model, which was developed by Geusebroek et.al [11] to build our CSIFT descriptors. So, in this section, we give a brief description of the invariants in this model.

In this model, the color invariants depend on the old Kubelka-Munk theory which models the reflected spectrum of colored bodies [13, 26]. The Kubelka-Munk theory models the photometric reflectance by:

$$E(\lambda, \vec{x}) = e(\lambda, \vec{x})(1 - \rho_f(\vec{x}))^2 R_\infty(\lambda, \vec{x}) + e(\lambda, \vec{x})\rho_f(\vec{x}) \quad (3)$$

where λ is the wavelength and $\vec{x}$ is a 2D vector which denotes the image position. e$(\lambda, \vec{x})$ denotes the illumination spectrum and $\rho_f(\vec{x})$ is the Fresnel reflectance at $\vec{x}$. $R_\infty(\lambda, \vec{x})$ denotes the material reflectivity. $E(\lambda, \vec{x})$ represents the reflected spectrum in the viewing direction. This model is suitable for modelling non-transparent/non-translucent materials. Some special cases can be derived from Eq. (3). For example, the Fresnel coefficient can be neglected for matte and dull surfaces. By assuming equal energy illumination, the spectral components of the source are constant over the wavelengthes and variable over the position, which is applicable for most of the practical cases. So, they can be denoted as $i(\vec{x})$. Then, Eq. (3) will be:

$$E(\lambda, \vec{x}) = i(\vec{x})[\rho_f(\vec{x}) + (1 - \rho_f(\vec{x}))^2 R_\infty(\lambda, \vec{x})] \quad (4)$$

By differentiating Eq. (4) with respect to λ, we get:

$$E_\lambda = i(\vec{x})(1 - \rho_f(\vec{x}))^2 \frac{\partial R_\infty(\lambda, \vec{x})}{\partial \lambda} \quad (5)$$

and

$$E_{\lambda\lambda} = i(\vec{x})(1 - \rho_f(\vec{x}))^2 \frac{\partial^2 R_\infty(\lambda, \vec{x})}{\partial \lambda^2} \quad (6)$$

By dividing Eq. (5) by Eq. (6), we get:

$$\begin{aligned} H = \left(\frac{E_\lambda}{E_{\lambda\lambda}}\right) &= \frac{\partial R_\infty(\lambda, \vec{x})}{\partial \lambda} / \frac{\partial^2 R_\infty(\lambda, \vec{x})}{\partial \lambda^2} \\ &= f(R_\infty(\lambda, \vec{x})) \end{aligned} \quad (7)$$

Thus, $H = \left(\frac{E_\lambda}{E_{\lambda\lambda}}\right)$ is the reflectance property which is independent of viewpoint, surface orientation, illumination direction, intensity, and Fresnel reflectance coefficient.

By considering only matte and dull surfaces for the model of Eq. (3), i.e. $\rho_f \approx 0$ and $E = i(\vec{x})R_\infty(\lambda, x)$ (which is the Lambertian model under the constraint of equal energy illumination), another object reflectance property $C_\lambda = \left(\frac{E_\lambda}{E}\right)$ is provided as an invariant to the viewpoint, surface orientation, illumination direction and illumination intensity. By adding an assumption of planar objects to the previous assumptions, $W_x = \left(\frac{E_x}{E}\right)$ is given as an invariant to the changes in the illumination intensity. For matte and dull surfaces with single illumination spectrum $N_{\lambda x} = \left(\frac{E_x E - E_\lambda E_x}{E^2}\right)$ is given as an object reflectance property that is independent of the viewpoint, surface orientation, illumination direction, illumination intensity, and illumination color. Hence, $N_{\lambda x}$ determines material transitions independent of illumination color and intensity distribution. Higher order derivatives for these invariants are used for more robust representations. For the detailed derivation of these invariants, the reader is referred to [11].

To calculate these invariants from the known RGB color space, the Gaussian color model is used as a general model for representation of spectral information and local image structure [11]. In this model, a linear transformation from the RGB space is used to obtain *spectral differential quotients*$(\hat{E}, \hat{E}_\lambda, \hat{E}_{\lambda\lambda})$. Then, *spatial differential quotients* $(\hat{E}_x, \hat{E}_{\lambda x}, \hat{E}_{\lambda\lambda x})$ are obtained by convolution with Gaussian derivative filters. A good approximation for the human vision system and for the CIE 1964 XYZ basis can be obtained by taking $\lambda_o = 520nm$ and $\sigma_\lambda = 55nm$ when calculating the first three components $(\hat{E}, \hat{E}_\lambda, \hat{E}_{\lambda\lambda})$ of the Gaussian color model [11]. Using the product of two linear transformations, one from RGB to XYZ and the other from XYZ to the Gaussian color model [11], the desired implementation of the Gaussian color model in terms of RGB can be obtained, as shown in Eq. (8). Measurement of the color invariants is obtained by substitution of E, E_λ, and $E_{\lambda\lambda}$ by $\hat{E}, \hat{E}_\lambda$, and $\hat{E}_{\lambda\lambda}$ at a given σ_x.

$$\begin{pmatrix} \hat{E} \\ \hat{E}_\lambda \\ \hat{E}_{\lambda\lambda} \end{pmatrix} = \begin{pmatrix} .06 & .63 & .27 \\ .3 & .04 & -.35 \\ .34 & -.6 & .17 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \quad (8)$$

5. CSIFT descriptors

Object recognition using local invariant features involves three main stages: interest points detection, descriptor building, and descriptor matching and pose estimation. Considering all the points in the image for object description is not feasible. Therefore, highly informative points are selected as interest points. More stable interest points means better performance. For each of these interest points, a local feature descriptor is built to distinctively describe the local region around the interest point. The final stage is matching the descriptors to decide if this point belongs to the object of interest or not. The matched points are used for further processing such as performing a global object recognition or pose estimation.

5.1. Interest points detection

Interest points should be selected so that they achieve the maximum possible repeatability under different photometric and geometric imaging conditions. As discussed in section (3), the extrema in Laplacian pyramid, which is approximated by difference-of-Gaussian for the input image in different scales, has been proven to be the most robust interest points detector to geometrical changes [5, 19]. Therefore, we detect the interest points at the extrema of a difference-of-Gaussian pyramid of the input image. We use the color invariants, which were presented in the previous section, as the working space for the input image in order to achieve the stability of the detected features to photometric changes. Similarly, as in SIFT, we expand the input image by factor of two, before building the pyramid, to preserve the highest spatial frequencies. For the Gaussian color model, we use $\sigma_x = 2$, whereas $\sigma = 1.4$ for the Gaussian filter of the pyramid levels. In order to localize the interest points, sub-pixel/sub-scale approximation is performed for the obtained extrema to achieve the maximum geometrical stability of the detected interest points [4].

In this paper, we show the results obtained by using the H invariant of Eq. (7)only. However, CSIFT is developed to be used with the other invariants as well. In the next section, we show the improvement which is obtained by using this model instead of gray level representation.

5.2. Descriptor building

After localizing the interest points, feature descriptors are built to characterize these points. These descriptors should contain the necessary distinct information for their corresponding interest points. Different schemes have been followed for descriptor building [16, 17, 19, 23]. We follow the same strategy of SIFT in building CSIFT descriptors. In other words, the local gradient-orientation histograms for the same-scale neighboring pixels of an interest point are used as the key entries of the descriptor. All orientations are assigned relative to a dominant/canonical orientation of the interest point. Thus, the built descriptor is invariant to the global object orientation. The stability to occlusion, partial appearance, and cluttered surroundings is achieved by the nature of the local description of the interest points.

Instead of using gray gradients in building the keys, we use the gradients of the color invariants which are represented in the previous section. Building CSIFT descriptors in this way makes them obtain inherently the robustness of SIFT to different geometrical transformations. At the same time, the use of color invariants in the feature descriptors, instead of using gray values, guarantees the robustness with respect to photometric changes.

5.3. Feature matching and pose estimation

The matching process is performed for the built local descriptors by finding the nearest neighbor of each feature key in a given feature descriptor database. The collection of location, scale, and canonical orientation of each match provides an estimation for a 2D transformation of the object. After rejecting outliers, the generalized Hough transform [2] is used to find a peak cluster among the estimated 2D transformations. Hence, the object pose is estimated.

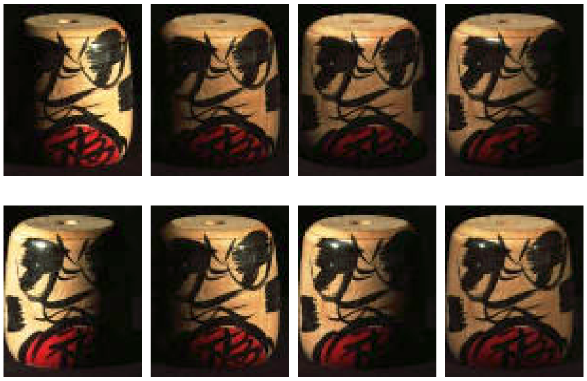


Figure 2. Sample images from ALOI [10] for a colored object under different illumination directions and intensities.

6. Experimental Results

To evaluate the proposed approach, we use the "*Amsterdam Library of Object Images (ALOI)*" [10] which is an image database of colored objects. ALOI contains a large number of objects under different imaging conditions, namely, different illumination directions, illumination intensities, illumination colors, and object viewpoints. Figure 2 shows a sample object under different illumination directions and intensities. It is noted that there are large variations in the image content with respect to the illumination changes. Therefore, we found that this database will be a good data set in order to prove the potential of our proposed CSIFT.

For evaluation purposes, we compare the performance of CSIFT with the performance of the SIFT. For fair comparison, we assign the optimum values to the SIFT parameters, as described in [17]. Since the geometrical-feature structure of SIFT and CSIFT are very close to each other, we focus on the comparison results between them with respect to photometric variations. Figure 3 shows the detected features of a sample object under different illumination directions and intensities using the H color invariant space versus those obtained using SIFT. It is clear that the number of detected features in the color invariant space is much larger than those in the gray images. It is known that as the number of the detected features increases, the performance of the recognition process is enhanced. Therefore, it is noted from the first glance at Figure 3 that CSIFT performs better with respect to the number of the detected features. The potential of CSIFT in feature detection is appreciated when

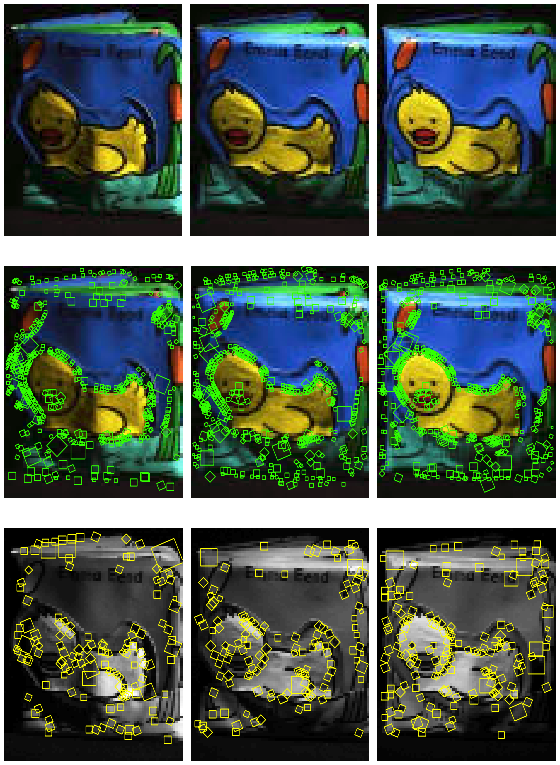


Figure 3. Detected features for a specific object under different illumination directions and intensities (Top) Original images (Middle) CSIFT detected features (Bottom) SIFT detected features. Note the stability of the detected colored features in the head and tail areas when compared to the gray detected features.

some challenging regions for SIFT are considered, e.g. the head and the tail areas of the object of Figure 3.

Although the total number of the detected features depends on thresholding constraints, e.g. the contrast threshold, CSIFT still has a large number of repeated features, which leads to a more accurate estimation of the object pose. Table 1 shows the average values of the ratio between the number of the repeated CSIFT features to the number of those obtained by SIFT after rejecting the pixels whose contrast is under a certain threshold. In general, the performance of CSIFT is at least 1.5 times better than the gray-based SIFT for low contrast rejection threshold up to 10%. For the recommended threshold of SIFT, which is 3% [17], the number of the repeated CSIFT features is, in average, 1.94 times the number of the repeated gray SIFT features.

Figure 4 shows quantitative evaluation results for CSIFT versus SIFT. In this figure, we show the repeatability and the matching results for objects imaged under different illumination conditions. Although the percentage repeatability of SIFT may be higher than CSIFT in few cases, the number of matched features of CSIFT is much larger than those of SIFT, as shown in Figures 4(b).

7. Conclusion

In this paper, we introduced CSIFT as a novel colored local invariant feature descriptor for the purpose of combining both color and geometrical information in object description. Opposite to many existing methods, the proposed approach balances between color and geometrical characteristics. We achieved the color invariance by using the color invariance model developed by Geusebroek et. al. [11], whereas the geometrical invariance is achieved by building CSIFT using a structure similar to that of the SIFT descriptors. Evaluation results proved the high performance of CSIFT when compared with the conventional SIFT descriptors.

8. Acknowledgements

The authors would like to appreciate the useful discussion with Prof. David Lowe. This research has been suppored by NSF Grant IIS-0513974.

References

[1] A. E. Abdel-Hakim and A. A. Farag. Color segmentation using an eigen color representation. In *The Eighth International Conference on Information Fusion (Fusion 2005)*, pages 230–237, Philadelphia, PA, 25-29 July 2005. 1

[2] D. Ballard. Generalized hough transform to detect arbitrary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(2):111–122, 1981. 4

[3] D. Brainard and W. Freeman. Bayesian color constancy. *the Journal of Optical Society of America*, 14(7):1393–1411, 1997. 1

[4] M. Brown and D. Lowe. Invariant features from interest point groups. In *British Machine Vision Conference*, pages 656–665, 2002. 2, 4

[5] P. J. Burt and E. H. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, Apr. 1983. 4

[6] M. D'Zmura and P. Lennie. Mechanisms of color constancy. *the Journal of Optical Society of America*, 3(10):1662–1672, 1986. 1

[7] A. A. Farag and A. E. Abdel-Hakim. Detection, categorization and recognition of road signs for autonomous navigation. In *Advanced Concepts in Intelligent Vision Systems (ACIVS'2004)*, pages 125–130, Brussel, Belgium, August-September 2004. 2

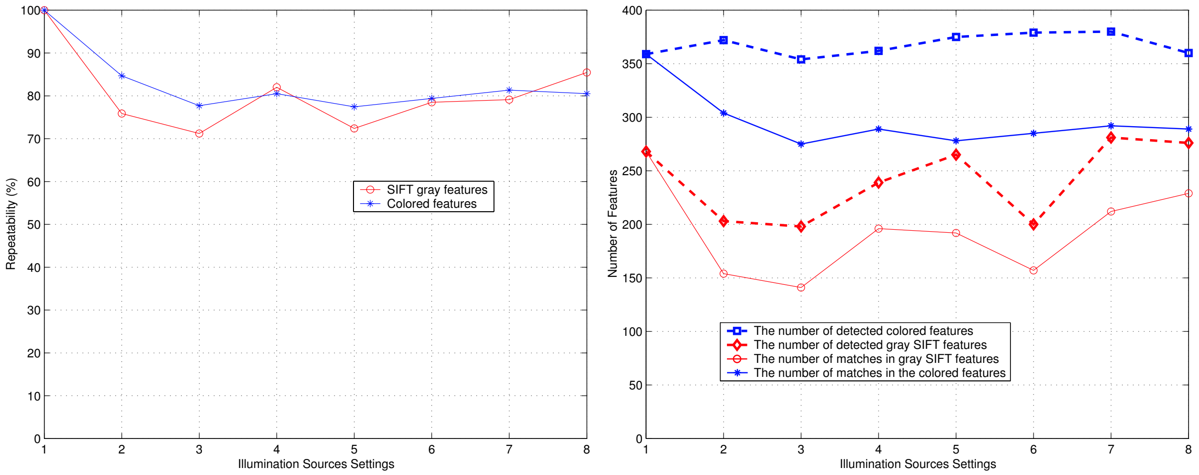
[8] G. D. Finlayson, S. S. Chatterjee, and B. V. Funt. Color angular indexing. In *Proceedings of the Second European Conference on Computer Vision*, pages 16–27, 1996. 1

[9] B. V. Funt, K. Barnard, and L. Martin. Is machine colour constancy good enough? In *Proceedings of the 5th European Conference on Computer Vision*, volume 1, pages 445–459, London, UK, 1998. Springer-Verlag. 1

0-7695-2597-0/06 $20.00 © 2006 IEEE

Table 1. The average values of the ratio between the number of the repeated CSIFT features to the number of the SIFT features for different low contrast rejection thresholds.

Contrast threshold[%]	0	1	2	**3**	4	5	6	7	8	9	10
Ratio	2.77	2.18	2.06	**1.94**	1.85	1.63	1.6	1.49	1.54	1.66	1.74



(a) Percentage repeatability of the detected features under different illumination directions and intensities

(b) Number of the detected and matched features under different illumination directions and intensities

Figure 4. Repeatability and matching results under varying illumination conditions. The results are obtained for 384x288 object images.

[10] 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, January 2005. 4

[11] 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. 1, 3, 5

[12] T. Gevers and A. W. M. Smeulders. Pictoseek: Combining color and shape invariant features for image retrieval. *IEEE Transactions on Image Processing*, 9(1):102–119, January 2000. 2

[13] P. Kubelka. New contribution to the optics of intensely light-scattering materials, part i. *the Journal of Optical Society of America*, 38(5):448–457, 1948. 3

[14] T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, 21(2):224–270, 1994. 2

[15] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, Norwell, MA, USA, 1994. 2

[16] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1150–1157, Corfu, Greece, 1999. 1, 2, 4

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004. 2, 4, 5

[18] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proceedings of the 7th European Conference on Computer Vision-Part I (ECCV '02)*, pages 128–142, Copenhagen, Denmark, 2002. Springer-Verlag. 2

[19] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 257–263, June 2003. 2, 4

[20] S. Nayar and R. Bolle. Reflectance based object recognition. *Int. J. Comput. Vision*, 17(3):219–240, 1996. 1

[21] Y. I. Ohta, T. Kanade, and T. Sakai. Color information for region segmentation. *Computer Graphics and Image Processing*, 13:222–241, 1980. 1

[22] B. Schiele and J. L. Crowley. Object recognition using multidimensional receptive field histograms. In *Proceedings of the Fourth European Conference on Computer Vision*, pages 610–619, 1996. 1

[23] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997. 1, 4

[24] D. Slater and G. Healey. The illumination-invariant recognition of 3d objects using local color invariants. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(2):206–210, 1996. 1

[25] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991. 1

[26] G. Wyszecki and W. S. Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulae*. John Wiley & sons, second edition, 1982. 3