Combining Color and Shape Information for Illumination-Viewpoint Invariant Object Recognition

Aristeidis Diplaros, *Student Member, IEEE*, Theo Gevers, *Member, IEEE*, and Ioannis Patras, *Member, IEEE*

***Abstract*—In this paper, we propose a new scheme that merges color- and shape-invariant information for object recognition. To obtain robustness against photometric changes, color-invariant derivatives are computed first. Color invariance is an important aspect of any object recognition scheme, as color changes considerably with the variation in illumination, object pose, and camera viewpoint. These color invariant derivatives are then used to obtain similarity invariant shape descriptors. Shape invariance is equally important as, under a change in camera viewpoint and object pose, the shape of a rigid object undergoes a perspective projection on the image plane. Then, the color and shape invariants are combined in a multidimensional color-shape context which is subsequently used as an index. As the indexing scheme makes use of a color-shape invariant context, it provides a high-discriminative information cue robust against varying imaging conditions. The matching function of the color-shape context allows for fast recognition, even in the presence of object occlusion and cluttering. From the experimental results, it is shown that the method recognizes rigid objects with high accuracy in 3-D complex scenes and is robust against changing illumination, camera viewpoint, object pose, and noise.**

***Index Terms*—Color-shape context, composite information, geometric invariants, image retrieval, object recognition, photometric invariants.**

I. Introduction

In a general context, object recognition involves the task of identifying a correspondence between a three-dimensional (3-D) object and some part of a two-dimensional (2-D) image taken from an arbitrary viewpoint in a cluttered real-world scene.

Many practical object recognition systems are appearance or model based. The recognition consists of matching the stored models (model based) or images (appearance based), encapsulated in the representation scheme, against the target image to determine which model (image) corresponds to which portion of the target image. Several systems have been developed to deal with the problem of model-based object recognition by solving the correspondence problem by tree search. However, the computational complexity is exponential for nontrivial images. Therefore, in this paper, we focus on appearance-based object recognition.

Most of the work on appearance-based object recognition based on shape information is by matching sets of shape image features (e.g., edges, corners, and lines) between a model and a target image. In fact, the projective invariance of cross ratios and its generalization to cross ratios of areas of triangles and volumes of tetrahedra has been used for viewpoint-invariant object recognition and significant progress has been achieved [1]. Other shape invariants are computed based on moments, Fourier transform coefficients, edge curvature, and arc length [2], [3]. Unfortunately, shape features are rarely adequate for discriminatory recognition of 3-D objects from arbitrary viewpoints. Shape-based object recognition is often inadequate especially in case of large data sets [4]. Another approach to appearance-based object recognition is to use color (reflectance) information. It is well known that color provides powerful information for object recognition, even in the total absence of shape information. A very common recognition scheme is to represent and match images on the basis of color (invariant) histograms [5], [6]. The color-based matching approach is widely in use in various areas such as object recognition, content-based image retrieval and video analysis.

Recently, another approach is taken which combines color and shape information into a unified framework. Interesting work is done by [7], [8] using moment invariants to merge geometric and photometric-invariant information of planar surfaces for object recognition. Further, in [4], color and shape invariants are combined for object recognition based on geometric algebraic invariants computed from color co-occurrences. Although the method is efficient and robust, the discriminative power decreases by the amount of invariance. Color, shape, and texture are combined in [9] for visual object recognition. However, the scheme is heavily dependent on the illumination conditions.

Therefore, in this paper, we study computational models and techniques to merge color and shape *invariant* information to recognize objects in 3-D scenes. Color and shape invariants are important aspects of any object recognition scheme as color and shape vary considerably with change in illumination, object pose, and camera viewpoint. After computing the color and shape invariants, a vector-based framework is proposed to index images on the basis of color and shape information. This color-shape context scheme is designed according to the following principles: *1) Generality*: The class of objects from which the images are taken from is the class of multicolored planar objects

Manuscript received April 5, 2004; revised November 29, 2004. This work was supported by the the Dutch Ministry of Economic Affairs under IOP project number IBV99001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jianying Hu.

A. Diplaros and T. Gevers are with the Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands (e-mail: diplaros@science.uva.nl; gevers@science.uva.nl).

I. Patras is with the Department of Computer Science, The University of York, York YO10 5DD U.K. (e-mail: i.patras@cs.york.ac.uk).

Digital Object Identifier 10.1109/TIP.2005.860320

in 3-D real-world scenes. *2) Invariance*: The scheme should be able to deal with images obtained from arbitrary unknown viewpoints discounting deformations of the shape (viewpoint, object pose) and color (shadowing, shading, and illumination). *3) Stability*: The scheme should be robust against substantial sensing and measurement noise.

This paper is organized as follows. First, we propose a scheme to compute illumination-invariant derivatives in a robust way. Then, shape invariance is discussed in Section III. In Section IV, color- and shape-invariant information is combined. Matching is discussed in Section IV-B. Finally, experiments are given in Section V.

II. Photometric Invariants for Color Images

The color of an object varies with changes in the color of the light source [e.g., spectral power distribution (SPD)] and object-illumination geometry (e.g., angle of incidence and reflectance). Hence, in outdoor images, the color of the illuminant (i.e., daylight) varies with the time of day, cloud cover, and other atmospheric conditions. Consequently, the color of an object may change drastically due to varying imaging conditions.

A. Illumination-Invariant Derivatives

Derived from neighboring image points, various color ratios have been proposed which are independent of the illumination [5], [10]. Unfortunately, these color ratios assume planar object surfaces, and, therefore, they might be negatively affected by the shape of the object.

Therefore, we focus on the following color ratio [11]:

$$M\left(C^i_{\vec{x}_1}, C^i_{\vec{x}_2}, C^j_{\vec{x}_1}, C^j_{\vec{x}_2}\right) = \frac{C^i_{\vec{x}_1} C^j_{\vec{x}_2}}{C^i_{\vec{x}_2} C^j_{\vec{x}_1}}, C^i \neq C^j \tag{1}$$

expressing the color ratio between two neighboring image locations $\vec{x}_1$ and $\vec{x}_2$, for C^i, $C^j \in \{C^1, C^2, \ldots, C^N\}$, giving the measured sensor response obtained by a narrow-band filter with central wavelengths i and j.

For a standard RGB color camera, we have

$$m_1\left(R_{\vec{x}_1}, R_{\vec{x}_2}, G_{\vec{x}_1}, G_{\vec{x}_2}\right) = \frac{R_{\vec{x}_1} G_{\vec{x}_2}}{R_{\vec{x}_2} G_{\vec{x}_1}} \tag{2}$$

$$m_2\left(R_{\vec{x}_1}, R_{\vec{x}_2}, B_{\vec{x}_1}, B_{\vec{x}_2}\right) = \frac{R_{\vec{x}_1} B_{\vec{x}_2}}{R_{\vec{x}_2} B_{\vec{x}_1}} \tag{3}$$

$$m_3\left(G_{\vec{x}_1}, G_{\vec{x}_2}, B_{\vec{x}_1}, B_{\vec{x}_2}\right) = \frac{G_{\vec{x}_1} B_{\vec{x}_2}}{G_{\vec{x}_2} B_{\vec{x}_1}}. \tag{4}$$

The color ratio is independent of the illumination, a change in viewpoint, and object geometry [11].

Taking the natural logarithm of both sides of (2) results in

$$\begin{aligned} &\ln m_1\left(R_{\vec{x}_1}, R_{\vec{x}_2}, G_{\vec{x}_1}, G_{\vec{x}_2}\right) \\ &= \ln\left(\frac{R_{\vec{x}_1} G_{\vec{x}_2}}{R_{\vec{x}_2} G_{\vec{x}_1}}\right) \\ &= \ln R_{\vec{x}_1} + \ln G_{\vec{x}_2} - \ln R_{\vec{x}_2} - \ln G_{\vec{x}_1} \\ &= \ln\left(\frac{R_{\vec{x}_1}}{G_{\vec{x}_1}}\right) - \ln\left(\frac{R_{\vec{x}_2}}{G_{\vec{x}_2}}\right). \end{aligned} \tag{5}$$

Equal derivations are obtained for m_2 and m_3.

Hence, (5) shows that the logarithm of color ratio can be seen as differences at two neighboring locations $\vec{x}_1$ and $\vec{x}_2$ in the image domain of $\ln(R/G)$. Further, if $\vec{x}_1 \to \vec{x}_2$, then the color ratios are identical to the directional derivative of the $\ln(R/B)$ image

$$\begin{aligned} &\ln m_1\left(R_{\vec{x}_1}, R_{\vec{x}_2}, G_{\vec{x}_1}, G_{\vec{x}_2}\right) \\ &= \ln\left(\frac{R}{G}\right)\Bigg|_{\vec{x}_1} - \ln\left(\frac{R}{G}\right)\Bigg|_{\vec{x}_2} \\ &= |\vec{x}_2 - \vec{x}_1| \nabla_{(\vec{x}_2 - \vec{x}_1)} ln\left(\frac{R}{G}\right). \end{aligned} \tag{6}$$

Differentiation is obtained by computing the difference in a particular direction between neighboring pixels of $\ln(R/G)$. The resulting derivation is independent of the illumination color, and also a change in viewpoint, the object geometry, and illumination intensity.

To obtain the gradient magnitude and direction, the Canny's edge detector is taken (derivative of the Gaussian with $\sigma = 1.0$) on image $\ln(R/G)$ with nonmaximum suppression in a standard way. The results obtained so far for $\ln m_1$ also holds for $\ln m_2$ and $\ln m_3$. So, the derivatives of $\ln(R/G)$, $\ln(R/B)$ and $\ln(G/B)$ images share the same invariant properties with the $\ln m_1$, $\ln m_2$, and $\ln m_3$ color ratios, respectively. The gradient magnitude $|\nabla F|$ of the illumination-invariant derivatives is

$$|\nabla F| = \sqrt{\sum_{i=1}^{3}\left[\left(\left|\frac{\partial c_i}{\partial x}\right|\right)^2 + \left(\left|\frac{\partial c_i}{\partial y}\right|\right)^2\right]} \tag{7}$$

where c_i is the notation for the $\ln(R/G)$, $\ln(R/B)$, and $\ln(G/B)$ images. For the ease of exposition, we concentrate on $\ln(R/G)$ in the following discussion.

B. Noise Robustness of Illumination-Invariant Derivatives

The above-defined illumination-invariant derivatives may become unstable when intensity is low. In fact, these derivatives are undefined at the black point $(R = G = B = 0)$ and they become very unstable at this singularity, where a small perturbation in the RGB values (e.g., due to noise) will cause a large jump in the transformed values. As a consequence, false color constant derivatives are introduced due to sensor noise. These false gradients can be eliminated by determining a threshold value corresponding to the minimum acceptable gradient norm. We aim at providing a method to determine automatically this threshold by computing the uncertainty for the color constant gradients through noise propagation, as follows.

Additive Gaussian noise is widely used to model thermal noise and is the limiting behavior of photon counting noise and film grain noise. Therefore, in this paper, we assume that sensor noise is normally distributed.

Then, for an indirect measurement, the true value of a measurand u is related to its N arguments, denoted by u_j, as follows:

$$u = q(u_1, u_2, \cdots, u_N). \tag{8}$$

Assume that the estimate $\hat{u}$ of the measurand u can be obtained by substitution of $\hat{u}_j$ for u_j. Then, when $\hat{u}_1, \cdots, \hat{u}_N$ are measured with corresponding standard deviations $\sigma_{\hat{u}_1}, \cdots, \sigma_{\hat{u}_N}$, we obtain [12]

$$\hat{u} = q(\hat{u}_1, \cdots, \hat{u}_N). \tag{9}$$

Then, it follows that if the quantities $\hat{u}_1, \cdots, \hat{u}_N$ are draws from independent random variables, with relatively small deviations, the predicted uncertainty in q is given by [12]

$$\sigma_q^2 = \sum_{j=1}^{N} \left(\frac{\partial q}{\partial \hat{u}_j} \sigma_{\hat{u}_j} \right)^2 \tag{10}$$

the so-called squares-root sum method. Although (10) is deduced for random errors, it is used as an universal formula for various kinds of errors.

Focusing on the x direction, the edge detection operation for the $h(x,y) = \ln(R_{xy}/G_{xy})$ image with a Gaussian derivative is computed as

$$\begin{aligned} f(x,y) &= (h * g_1)(x,y) \\ &= \sum_m \sum_n \ln\left(\frac{R_{mn}}{G_{mn}}\right) g_1(x-m, y-n) \end{aligned} \tag{11}$$

where $g_1(x,y) = \partial g_0(x,y)/\partial x$ and $g_0(x,y)$ is a 2-D Gaussian. Since we have assumed that sensor noise has the same standard deviations (σ_R, σ_G, and σ_B) over the whole image, the substitution of (11) in (10) gives the uncertainty for the illumination-invariant derivative

$$\sigma_f^2(x,y) = \sum_m \sum_n \left[\left(\frac{\partial f(x,y)}{\partial R_{mn}} \sigma_R \right)^2 + \left(\frac{\partial f(x,y)}{\partial G_{mn}} \sigma_G \right)^2 \right]. \tag{12}$$

Keeping in mind that (11) is just a sum of $(\ln(R_{mn}/G_{mn}) g_1(x-m, y-n))$ terms, then, for an individual image location (m,n), the partial derivative of $f(x,y)$ with respect to the particular measurand in R channel is

$$\frac{\partial f(x,y)}{\partial R_{mn}} = \left(\frac{1}{R_{mn}}\right) g_1(x-m, y-n) \tag{13}$$

and, similarly, for $\partial f(x,y)/\partial G_{mn}$. Then, (12) becomes

$$\begin{aligned} \sigma_f^2(x,y) &= \sum_m \sum_n \left[\left(\left(\frac{1}{R_{mn}}\right) g_1(x-m, y-n) \sigma_R \right)^2 \right. \\ &\qquad \left. + \left(\left(\frac{1}{G_{mn}}\right) g_1(x-m, y-n) \sigma_G \right)^2 \right] \\ &= \sum_m \sum_n \left[\left(\frac{\sigma_R^2}{R_{mn}^2} + \frac{\sigma_G^2}{G_{mn}^2} \right) g_1^2(x-m, y-n) \right] \\ &= \left(\left(\frac{\sigma_R^2}{R^2} + \frac{\sigma_G^2}{G^2} \right) * g_1^2 \right)(x,y). \end{aligned} \tag{14}$$

Similar results are obtained for the y direction and for the $\ln(R/B)$ and $\ln(G/B)$ images. Assuming normally distributed random quantities, the standard way to calculate the standard deviations σ_R, σ_G, and σ_B is to compute the mean and variance estimates derived from a homogeneously colored surface patches in an image under controlled imaging conditions. From the analytical study of (14), it can be derived that our edge detection becomes unstable around the black point $R = G = B = 0$.

To further propagate the uncertainties, the uncertainty in the gradient norm is determined by (10). Using (7) and (10), we obtain

$$\sigma_{|\nabla F|}^2 = \frac{\sum_{i=1}^{3} \left[\left(\left| \frac{\partial c_i}{\partial x} \right| \right)^2 \sigma_{\left| \frac{\partial c_i}{\partial x} \right|}^2 + \left(\left| \frac{\partial c_i}{\partial y} \right| \right)^2 \sigma_{\left| \frac{\partial c_i}{\partial y} \right|}^2 \right]}{|\nabla F|^2} \tag{15}$$

where $\sigma_{|\partial c_i/\partial x|}$ are calculated by replacing the σ_R, σ_G, R, G, and g_1 terms in (14) with the appropriate ones [e.g., for the derivative of $\ln(G/B)$ image in y direction should be σ_G, σ_B, G, B and $g_1(x,y) = \partial g_0(x,y)/\partial y$, respectively]. In this way, the effect of measurement uncertainty due to noise is propagated throughout the edge detection operation and finally to the gradient norm.

For a Gaussian distribution, 99% of the values fall within a 3σ margin. If the value of the gradient norm in a image location is detected to exceed $3\sigma_{|\nabla F|}$, we assume that there is only 1% chance that this gradient value corresponds to no color transition.

III. Geometric Invariant Transformation

In this section, shape-invariant descriptors are discussed expressing shape properties of an object independent of a specific coordinate transformation.

A. Affine Deformations and Inverse Transformation

The geometric transformations, considered in this paper, are up to affine transformations

$$\vec{p'} = \mathbf{A}\vec{p} + \mathbf{B} \tag{16}$$

where a point $\vec{p} = (x, y)$ in one image is transformed into the corresponding point $\vec{p'} = (x', y')$ in a second image, by the transformation matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad B = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}. \tag{17}$$

When objects are relative far away from the camera, this transformation approximates the projective transformation of 3-D planar objects.

For a binary image $f(x,y)$, the geometric moment of order $(p+q)$ is defined as

$$M_{pq} = \sum_{x=1}^{J} \sum_{y=1}^{K} x^p y^q f(x,y) \tag{18}$$

where the ratio's $\overline{x} = M_{10}/M_{00}$, $\overline{y} = M_{01}/M_{00}$ define its centroid $\vec{c}$. Transforming the image with respect to $\vec{c} = (\overline{x}, \overline{y})$ yields invariance to translation. The principal axis is obtained by rotating the axis of the central moments until M_{11} is zero. Then, the angle θ between the original and the principal axis is defined as follows [13]

$$\tan 2\theta = \frac{2M_{11}}{M_{20} - M_{02}}. \tag{19}$$

This angle may be computed with respect to the minor or the major principal axis. To determine a unique orientation, the additional condition $M_{20} > M_{02}$ and $M_{30} = 0$ is required. Setting the rotation matrix to

$$A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{20}$$

will provide rotation invariance.

In conclusion, the shape information will be normalized with respect to translation and rotation using a collection of low-order moments. Scale invariance and robustness to affine transformation are discussed in the following section.

IV. Object Indexing and Matching

In this section, we propose a new color-shape representation scheme by combining shape- and color-invariant information in a principled way to obtain a unified indexing framework.

A. Color-Shape Indexing Scheme

As in our previous work [14], the color-shape indexing scheme is as follows. For color image $\mathcal{I} : \mathbb{R}^2 \rightarrow \mathbb{R}^3$, illumination-invariant edges are computed to yield a binary image $\mathcal{E} : \mathbb{R}^2 \rightarrow \{0, 1\}$, as described in Section II-A. Then, for each edge point $\vec{e} = (x, y)$ in image $\mathcal{E}$ (i.e., where $\mathcal{E}(x, y) = 1$), we calculate the color-invariant ratios m_1, m_2, m_3 from image $\mathcal{I}$. The two image locations $\vec{x_1}$ and $\vec{x_2}$ in $\mathcal{I}$ necessary to compute the color ratios are chosen as

$$\vec{x_1} = \vec{e} + \vec{u} \quad \text{and} \quad \vec{x_2} = \vec{e} - \vec{u}$$

where $\vec{u}$ is parallel to the gradient direction of $\mathcal{I}$. Because $\vec{x_1}$ and $\vec{x_2}$ must be neighboring, in order for the color ratios to be invariant, $|\vec{u}|$ should be small (i.e., three pixels in our experiments).

We choose an image location $\vec{c} = (x_c, y_c)$ that we call a central point. The central point $\vec{c}$ defines a local coordinate system. We define $\vec{p_n} \in \mathbb{R}^5$ as

$$\vec{p}_n = (x - x_c, y - y_c, m_1, m_2, m_3) | \mathcal{E}(x, y) = 1. \tag{21}$$

Each vector $\vec{p}_n$ is decomposed as follows:

$$\vec{p}_n = \vec{p}_{S_n} + \vec{p}_{C_n} \tag{22}$$

where

$$\begin{aligned} \vec{p}_{S_n} &= (x - x_c, y - y_c, 0, 0, 0) \\ \vec{p}_{C_n} &= (0, 0, m_1, m_2, m_3). \end{aligned} \tag{23}$$

Note that the set $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$ of n points in a five-dimensional space represents both the shape and color information in the image. The first two dimensions hold the location of the edges and the other three dimensions corresponds to the colors around the edges. Each of these n points corresponds to a vector originating from the central point $\vec{c}$. To provide noise robustness, the distribution of these vectors is considered. Then, the color-shape representation scheme $h^c_{\mathcal{P}}$ of set $\mathcal{P}$ is defined as the estimate of the distribution of all p_n vectors, as defined in (22). In other words, the histogram of the decomposed and quantized versions of p_n vectors, with respect to the distance and orientation from a central point c, is now called the color-shape context of the image I at position c. The vector $\vec{p}_S$ is represented by its polar coordinates while vector $\vec{p}_C$ is represented by its spherical coordinates.

To construct the color-shape context of image I at the position c, the following histogram is computed:

$$h^c_{\mathcal{P}}(k) = \frac{\#\{q : q \in bin(k)\}}{n} \tag{24}$$

where $q \in \mathcal{P}$ and $bin(k)$ is a bin corresponding to a partition of the feature space. This bin is defined as a Cartesian product of the spatial and color bins. For partitioning the spatial space, we use a log-polar with equally spaced radial bins scheme, as in [15]. An equally-spaced bin scheme is used to partition the 3-D color-invariant space.

In summary, the color-shape indexing scheme has intrinsic robustness against illumination changes and small amounts of displacements due to noise. As we will see later using the moments framework, as described Section III, we can also achieve shape invariant for translation, rotation and scale coordinate transformations.

B. Object Matching Without Occlusion and Cluttering

To recognize objects without any object occlusion and cluttering (i.e., one object per image), we set the position c as the center of gravity of the spatial information. Note that the spatial information is normalized with respect to translation that the center of gravity is at the origin (0,0). Using the moments framework. To achieve scale invariance we normalize all $|\vec{p}_{S_n}|$ with respect to the mean distance between all point pairs in $\mathcal{E}$. Rotation invariance is obtained by rotating all $|\vec{p}_{S_n}|$ by θ which is computed from (19). Consider an image a with corresponding color-shape context h_a. Because $h_a(k) \in [0, 1]$ and $\sum_k h_a(k) = 1$, the cost function to compute the distance with another color-shape context h_b is given by

$$C_{ab} = \frac{1}{2} \sum_{k=1}^{K} \frac{(h_a(k) - h_b(k))^2}{(h_a(k) + h_b(k))}. \tag{25}$$

The complexity of this operation is only dependent on the number K which is constant for all images in the data set and is usually small.

C. Object Matching With Occlusion and Cluttering

For object recognition in complex scenes, which may contain cluttering and occlusion, a modified matching strategy is adopted. This is due to the fact that the occlusion and cluttering may result in a shape context that contains information about the background and/or other objects in the scene. For these shape contexts, it is apparent that the influence of cells that contain such irrelevant information should be reduced in the matching scheme. Furthermore, the edge points from other objects will effect the estimation of the moments that are used to determine

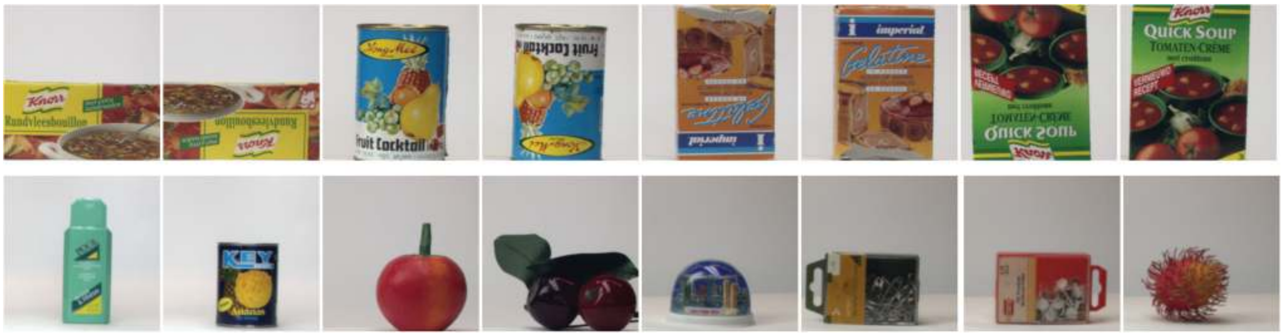


Fig. 1. Various images which are included in the Amsterdam image dataset of 500 images. The images are representative for the images in the dataset. Objects were recorded in isolation (one per image). (Color version available online at http://ieeexplore.ieee.org.)

the center and the scale of the color-shape context. To this end, we compute multiple color-shape representations per image and match them with a modified distance function. The multiple color-shape contexts are generated at specific image locations, which are the same for all images. The scale and orientation of these color-shape contexts are fixed for all images in the dataset. The modification of the cost function aims at reducing the influence of large costs introduced by occlusions. More specifically, we introduce an occlusion field that indicates in which spatial cell occlusion occurs and modify the cost function in order to reduce the cost of matching when the occlusion fields are spatially coherent.

More specifically, the occlusion field of spatial cell k is defined as

$$O_k = \begin{cases} 1: & \frac{d_k}{q_k} > T \\ 0: & \frac{d_k}{q_k} \leq T \end{cases} \tag{26}$$

where d_k denotes the accumulated cost of matching in the spatial cell k and is defined as

$$d_k = \frac{1}{2} \sum_{\{n|f(n)=k\}} \frac{(h_a(n) - h_b(n))^2}{(h_a(n) + h_b(n))} \tag{27}$$

where $f(n)$ denotes a mapping from the index of the color-shape context to the appropriate spatial cell index k. q_k denotes the percentage of points of the two images in the spatial cell k and is defined as

$$q_k = \sum_{\{n|f(n)=k\}} (h_a(n) + h_b(n)). \tag{28}$$

The cost function is now as follows:

$$C'_{ab} = C_{ab} - \sum_k O_k \frac{1}{|N_k|} \sum_{l \in N_k} (d_k - T q_k O_l) \tag{29}$$

where N_k is a 4-neighborhood of spatial cell k. The way that C'_{ab} is defined is that it assigns a cost at the occluded spatial cell k that varies between $q_k T$ and d_k depending on the spatial coherency of the occlusion fields in the neighborhood of k. In the extreme case, if none of the $l \in N_k$ is occluded the local cost at the spatial cell k is d_k while if all $l \in N_k$ are occluded it becomes $q_k T$.

V. Experiments

In this section, we consider the performance of the proposed method. Therefore, in Section V-A, the datasets, used in our experiments, are discussed. In Section V-B, matching quality measures are presented. In the remaining sections, we test our method with respect to the following criteria: *1) Generality*: Object recognition in the presence of object occlusion and cluttering. *2) Invariance*: Robustness against varying imaging conditions such as camera viewpoint, object pose, and illumination. *3) Stability*: Robustness against substantial sensing and measurement noise.

A. Datasets

For comparison reasons, we have selected the following datasets: Amsterdam and Columbia—COIL-100, which are publicly available and often used in the context of object recognition [4], [16], and the Corel image collection mostly used in the context of image retrieval.

Amsterdam Dataset: In Fig. 1, various images from the image database are shown. These images are recorded by the SONY XC-003P CCD color camera and the Matrox Magic Color frame grabber. Two light sources of average day-light color are used to illuminate the objects in the scene. The database consists of $N_1 = 500$ target images taken from colored objects, tools, toys, food cans, art artifacts, etc. Objects were recorded in isolation (one per image). The size of the images are 256 × 256 with 8 bits per color. The images show a considerable amount of shadows, shading, and highlights. A second, independent set (the query set) of $N_2 = 70$ query or test recordings was made of randomly chosen objects already in the database. These objects were recorded again one per image with a new, arbitrary position and orientation with respect to the camera, some recorded upside down, some rotated, some at different distances.

COIL-100: To test the algorithm with respect to variability in appearance, the COIL-100 has been selected. The COIL-100 has been gathered at the Columbia University [17]. This dataset is well suited for multiview object recognition, i.e., there are various images taken from the same object (e.g., back, front, and from a side; see Fig. 2).

COREL: A subset of the Corel image collection has been selected. The subset consists of 25 categories and has a total of 2600 images. These categories cover a wide rage of subjects,

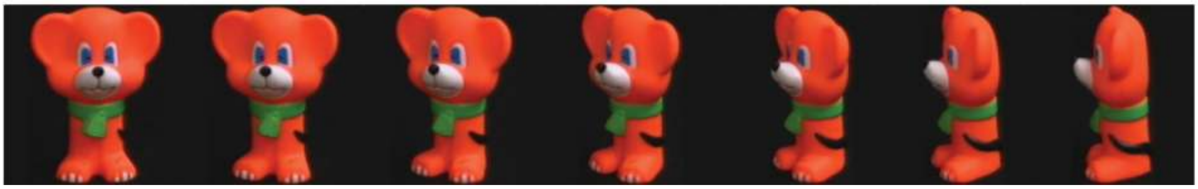


Fig. 2. Different images recorded from the same object under varying viewpoint. The COIL-100 database consisting of 7200 images of 100 different objects with 72 different views each. (Color version available online at http://ieeexplore.ieee.org.)

which includes among others buildings, ships, mountains, cars, flowers, and sunsets. This dataset can be considered as a representative collection of images found on the Internet, since, it consists of photos taken by both amateurs and professionals under different imaging conditions. From this dataset, we randomly selected 10% of the images from every category as a query set. This is a total of 260 queries. In order to establish a ground truth for our tests, we assume all pictures that belong to a category to be relevant, an assumption which is not always true.

B. Error Measures

For comparison reasons, two different quality measures are used in our experiments.

1) For a measure of object match quality, let rank r^{Q_i} denote the position of the correct match for test image Q_i, $i = 1, \ldots, N_2$, in the ordered list of N_1 match values. The rank r^{Q_i} ranges from $r = 1$ from a perfect match to $r = N_1$ for the worst possible match.

   Then, for one experiment with N_2 test images, the average ranking percentile is defined as

$$\overline{r} = \left(\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{N_1 - r^{Q_i}}{N_1 - 1} \right) 100\%. \tag{30}$$

2) To evaluate image match quality, we have:

- *precision*: the percentage of similar images retrieved with respect to the total number of retrieved images;
- *recall*: the percentage of similar images retrieved with respect to the total number of similar shapes in the dataset.

   Specifically, let $\widehat{A}$ be the set of similar images in the dataset, and let A denote the set of the retrieved images from the system. Precision p and recall r are then given as

$$p = \left(\frac{|A \cap \widehat{A}|}{|A|} \right) 100\%, \quad r = \left(\frac{|A \cap \widehat{A}|}{|\widehat{A}|} \right) 100\%. \tag{31}$$

   Performance will be illustrated in terms of precision-recall graphs. In this case, the horizontal axis corresponds to the measured recall while the vertical axis corresponds to precision. Each query retrieves the best 100 matches and each point in a precision-recall graph is the average over all queries. Precision and recall values are computed from each answer set after each matching (from 1 to 100), and, therefore, each plot contains exactly 100 points. A method outperforms another one if it achieves better precision and recall. Methods achieving higher precision and recall for large answer sets are considered better than others (based on the assumption that typical users retrieve 10 to 20 images on average).

C. Viewpoint

To test the effect of a change in viewpoint, the COIL-100 dataset has been used. This dataset consists of 7200 images taken from 100 different objects which have been put perpendicularly in front of the camera. Then, a total of 72 recordings were generated by varying the angle between the camera (every 5°) with respect to the object (see Fig. 2). The experiment is conducted as follows. First, images have taken on input from views ranging from 0° to 70° (15 different views). Then, the purpose was to recognize the corresponding object from the 0° views of all 100 objects. This was done for all 100 objects and for all 15 views of each object. To test the robustness of the method for varying viewpoint differentiated by color, shape and composite information, three different color-shape context schemes have been constructed. First, both color and shape-invariant information have been included in the color-shape context scheme denoted by $\mathcal{H}_{CS}$. Second, only color is considered which is denoted by $\mathcal{H}_C$. Third, only shape information is taken into account expressed by $\mathcal{H}_S$.

For comparison reasons, the well-known matching scheme is computed based on histogram intersection [6] denoted by $\mathcal{HI}_{\mathrm{RGB}}$. $\mathcal{HI}_{\mathrm{inv-rgb}}$ represents histogram intersection based on normalized rgb color (i.e., invariant to intensity changes).

The performance of the recognition scheme is given in Fig. 3. When the performance of different invariant image indices is compared, matching based on color invariants produces the highest discriminative power. In fact, superior performance is shown where 97% of the images are still recognizable up to 70° of a change in viewpoint. The matching based on both color and shape invariants produces also excellent results where 95% of the images are still recognizable up to 70° of a change in viewpoint. Shape-based invariant recognition yields poor discriminative power with 72% at 70° of viewpoint change.

In conclusion, recognition based on only color, and composite information, produce the highest discriminative power. The small performance gain in using only color is that the objects are very colorful suppressing the additional effect of using shape information. Obviously, the color-shape context outperforms the histogram intersection method even when the images have been transformed to the normalized rgb color space (i.e., invariant to intensity changes in the illumination). Finally, color-shape based recognition is almost as robust to a change in viewpoint as the color based recognition. Even when the object-side is nearly vanishing, object identification is still acceptable.

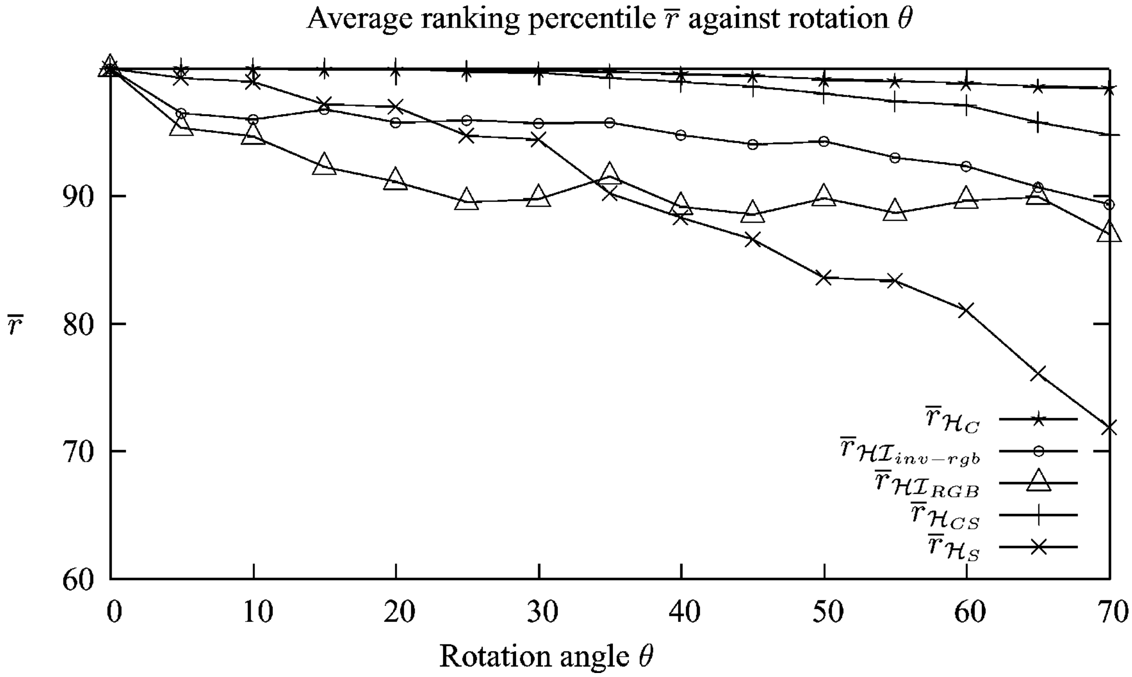


Fig. 3. Discriminative power of the color-shape matching process under varying viewpoint differentiated by color information, shape information, and the integration of both. The average ranking percentile of color-shape, color, and shape contexts are denoted by $\overline{r}_{\mathcal{H}_{CS}}$, $\overline{r}_{\mathcal{H}_C}$, and $\overline{r}_{\mathcal{H}_S}$, respectively. Also, the average ranking percentile of the Histogram Intersection with and without conversion to normalized rgb color space is denoted with $\mathcal{HI}_{\mathrm{inv-rgb}}$ and $\mathcal{HI}_{\mathrm{RGB}}$, respectively.

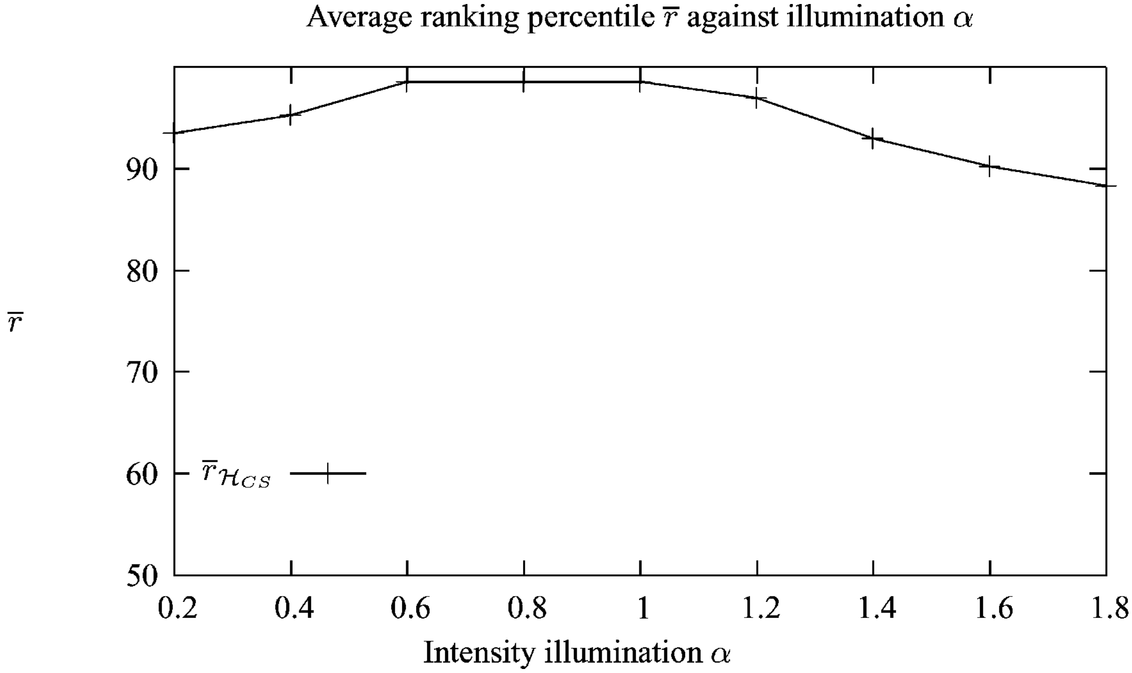


Fig. 4. Discriminative power of the color-shape matching process plotted against the varying illumination intensity.

D. Illumination

The effect of a change in the illumination intensity is equal to the multiplication of each RGB color by a uniform scalar factor α. To measure the sensitivity of the color-shape context, RGB images of the Amsterdam test set are multiplied by a constant factor varying over $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8\}$. The discriminative power of the matching process plotted against the varying illumination intensity is shown in Fig. 4.

Color-shape context is, to a large degree, robust to illumination intensity changes.

E. Noise

To measure the sensitivity of our method with respect to varying signal-to-noise ratio (SNR), ten objects are randomly chosen from the Amsterdam image dataset. Each object is recorded again under a global change in illumination intensity (i.e., dimming the light source) generating images with

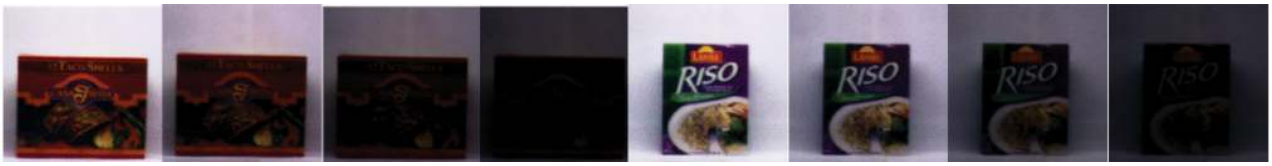


Fig. 5. Two objects under varying illumination intensity generating each four images with $\mathrm{SNR} \in \{24, 12, 6, 3\}$. (Color version available online at http://ieeexplore.ieee.org.)

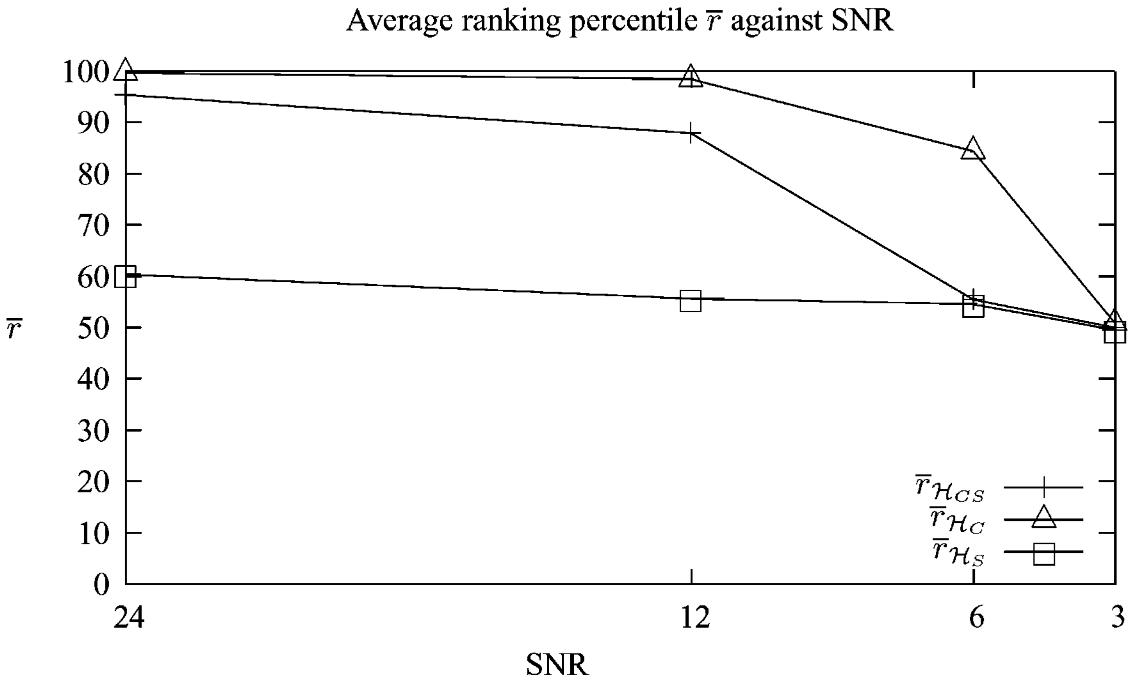


Fig. 6. Discriminative power of the color-shape matching process plotted against the varying SNR.

$\mathrm{SNR} \in \{24, 12, 6, 3\}$ (see Fig. 5). These low-intensity images can be seen as images of snap shot quality, a good representation of views from everyday life as it appears in home video, the news, and consumer digital photography in general. The discriminative power of the color-shape matching process plotted against the varying SNR is shown in Fig. 6.

For $3 < \mathrm{SNR} < 12$, the results show a rapid decrease in the performance. For these SNRs, the color-shape based recognition scheme still outperforms the shape based recognition. For SNRs $\leqslant 3$, the performance of all methods converge. This is because the images are getting too dark to recognize anything at all (color and shape).

In conclusion, the method is robust to low SNRs. Even when the object is nearly visible, object identification is still sufficient.

F. Occlusion and Cluttering

To test our method for images taken from 3-D objects in complex scenes, we used a subset of the Amsterdam dataset. An image depicting a cluttered scene with four occluded objects was used as query. The dataset consisted of 100 randomly selected images. Only two of objects where considered known, that is, were present in the query and also present in the dataset. These objects are shown in Fig. 7. The average ranking percentile for this experiment was 99%. Despite substantial object cluttering and occlusion, the method is capable to identify the two objects.

G. Image Retrieval

In this section, we address the use of the color-shape context feature in the context of image retrieval. Although the scope of this paper is object recognition the image retrieval experiment provides an excellent opportunity to demonstrate the expressive description of the visual information by the color-shape feature. For this experiment, the Corel dataset is used. The total number of images is 2600 from which 260 are randomly selected to generate the query set.

For this experiment, we adopted a slightly different framework from the one proposed for object recognition and described in Section III. For each image, we calculate five partially overlapping color-shape context features in fixed positions (i.e., the center and close to the corners of the image), with fixed orientations and fixed scales, same for all images in the dataset. In effect, we are using a semi-global framework, where each one of the five shape context feature describes a local neighborhood of the image. The matching function using (29) becomes

$$C_{\mathrm{CBIR}} = \sum_{i=1}^{5} \arg \min_{j \in [1,5]} C'_{a_i b_j}. \tag{32}$$

The matching based on color-shape is denoted by $\mathcal{H}_{CS}$. Again, for comparison, matching based on histogram intersection derived from the RGB values is denoted by $\mathcal{HI}_{\mathrm{RGB}}$ and

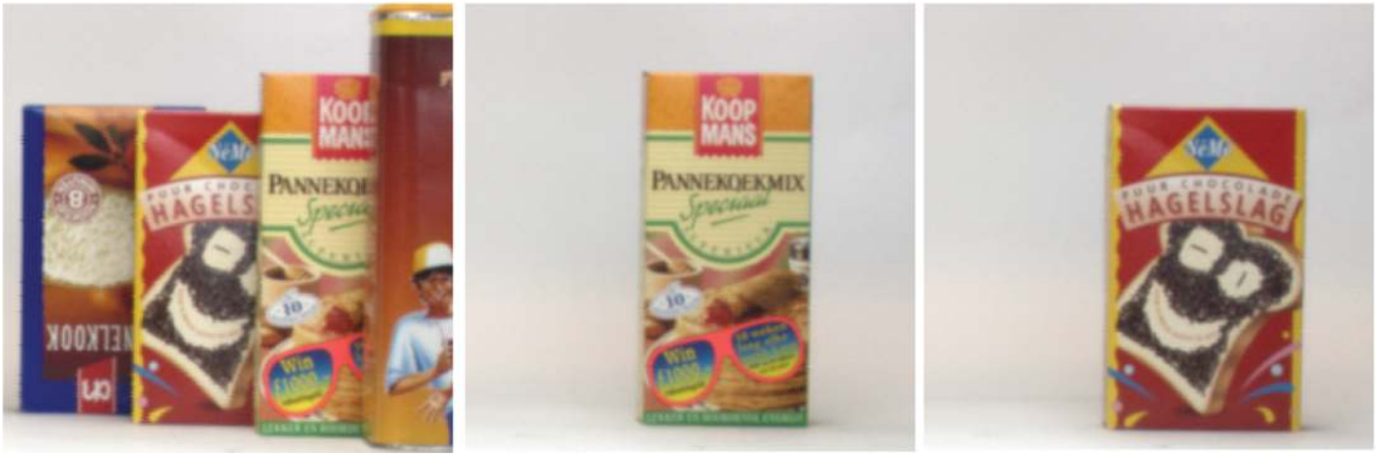


Fig. 7. Left: Image containing cluttering and occlusion. Right: The objects that were recognized from a dataset of 100 objects. (Color version available online at http://ieeexplore.ieee.org.)

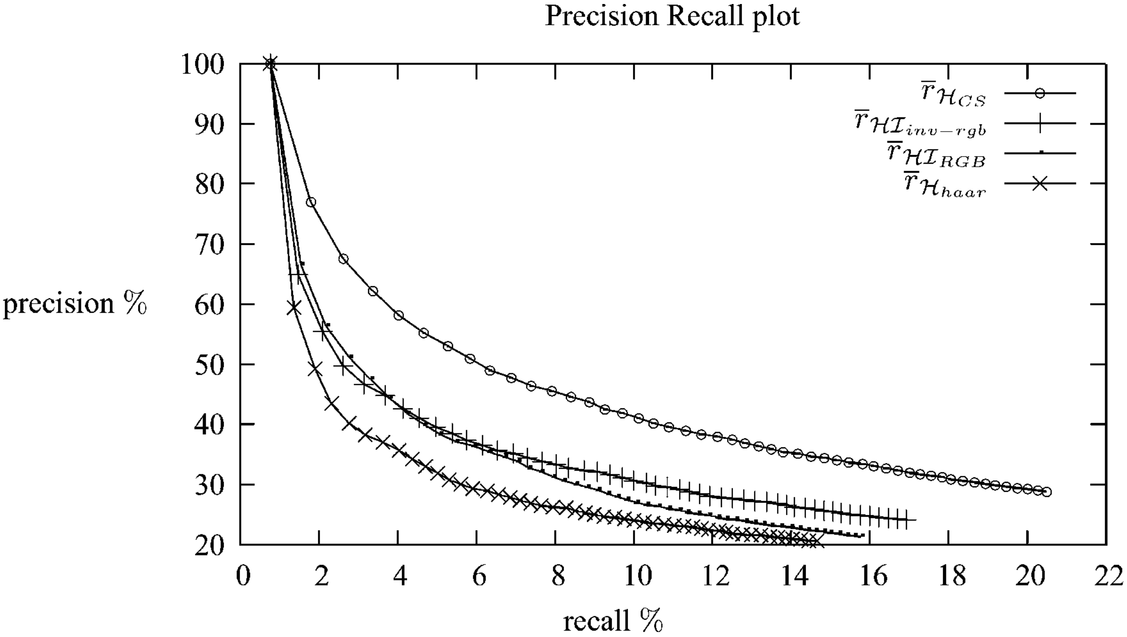


Fig. 8. Precision recall plot of our retrieval experiment on the Corel dataset. The precision recall curve based on color-shape context is denoted by $\overline{r}_{\mathcal{H}_{CS}}$. The precision recall curve based on salient points is denoted by $\overline{r}_{\mathcal{H}_{\mathrm{haar}}}$. The precision recall curve based on histogram intersection with and without conversion to normalized rgb color space is denoted by $\mathcal{HI}_{\mathrm{inv-rgb}}$ and $\mathcal{HI}_{\mathrm{RGB}}$, respectively.

based on normalized rgb is given by $\mathcal{HI}_{\mathrm{inv-rgb}}$. In additional comparisons, matching is performed based on salient points [18], [19] denoted by $\mathcal{H}_{\mathrm{haar}}$. The salient points detector is based in the wavelet transform using the Haar wavelet function. 200 salient points are computed in every image from which local color moments and texture features are computed.

The performance of the recognition scheme is given in Fig. 8. The method based on color and shape outperforms the other methods achieving approximately 15% better precision and recall than any other method. The performance of histogram intersection based on normalized rgb is slightly better than the one derived from RGB. The content-based image retrieval method based on salient points performs the worst.

Fig. 9 shows two examples of image retrieval using our method. In both examples the top left image is the query. The top 12 retrieved images from the 2600 images dataset are shown in Fig. 9(a). Fig. 9(b) shows the top 12 retrieved images in the whole Corel dataset which consists of 47 395 images.

VI. Conclusion

In this paper, computational models and techniques have been proposed to merge color and shape *invariant* information in the context of object recognition and image retrieval. A vector-based framework has been presented to index images on the basis of illumination (color) invariants and viewpoint (shape) invariants. The matching function of the color-shape context allows for fast recognition even in the presence of object occlusion and cluttering. From the experimental results, it is shown that the method recognizes rigid objects with high accuracy in 3-D complex scene robust to changing illumination, camera viewpoint, object pose, and noise.

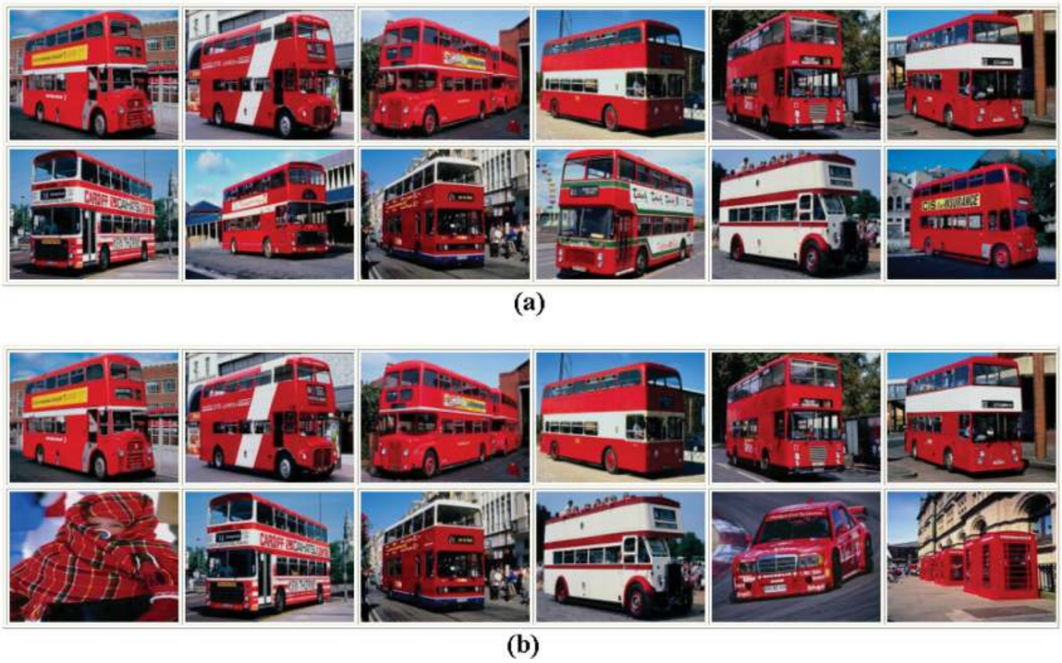


Fig. 9. Retrieval example from the Corel dataset, the top left image is the query. (a) Top 12 retrieved images from the 2600 images dataset. (b) Top 12 retrieved images from whole Corel dataset (47 395 images). (Color version available online at http://ieeexplore.ieee.org.)

ACKNOWLEDGMENT

The authors would like to thank N. Sebe for providing the results of his content-based image retrieval method based on salient points. The authors would also like to thank N. Vlassis and J. van de Weijer for their valuable comments.

REFERENCES

[1] C. A. Rothwell, A. Zisserman, D. A. Forsyth, and J. L. Mundy, "Planar object recognition using projective shape representation," *Int. J. Comput. Vis.*, vol. 16, no. 1, pp. 57–99, 1995.

[2] T. H. Reiss, *Recognizing Planar Objects Using Invariant Image Features*. New York: Springer-Verlag, 1993.

[3] I. Weiss, "Geometric invariants and object recognition," *Int. J. Comput. Vis.*, vol. 10, no. 3, pp. 207–231, 1993.

[4] T. Gevers and A. W. M. Smeulders, "Image indexing using composite color and shape invariant features," in *Proc. 6th Int. Conf. Computer Vision*, Bombay, India, 1998, pp. 576–581.

[5] B. V. Funt and G. D. Finlayson, "Color constant color indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 5, pp. 522–529, May 1995.

[6] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, 1991.

[7] L. van Gool, T. Moons, and D. Ungureanu, "Affine/photometric invariants for planar intensity patterns," in *Proc. 4th Eur. Conf. Computer Vision*, vol. 1, 1996, pp. 642–651.

[8] F. Mindru, T. Moons, and L. van Gool, "Recognizing color patterns irrespective of viewpoint and illumination," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1999, pp. 368–373.

[9] B. W. Mel, "Seemore: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition," *Neur. Comput.*, vol. 9, no. 4, pp. 777–804, 1997.

[10] S. K. Nayar and R. M. Bolle, "Reflectance based object recognition," *Int. J. Comput. Vis.*, vol. 17, no. 3, pp. 219–240, 1996.

[11] T. Gevers and A. W. M. Smeulders, "Color-based object recognition," *Pattern Recognit.*, vol. 32, no. 3, pp. 453–464, 1999.

[12] J. R. Taylor, *An Introduction to Error Analysis, the Study of Uncertainties in Physical Measurements*. Sausalito, CA: Univ. Science Books, 1982.

[13] A. P. Reeves, R. P. Prokop, S. E. Andrews, and F. P. Kuhl, "Three-dimensional shape analysis using moments and fourier descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 6, pp. 937–943, Jun. 1988.

[14] A. Diplaros, T. Gevers, and I. Patras, "Color-shape context for object recognition," presented at the IEEE Workshop on Color and Photometric Methods in Computer Vision, in conjuction with the 9th Int. Conf. Computer Vision, Nice, France, 2003.

[15] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[16] N. Sebe, M. S. Lew, and D. P. Huijsmans, "Toward improved ranking metrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1132–1143, Oct. 2000.

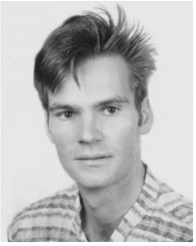
[17] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia Object Image Library: Coil-100," Columbia Univ., New York, Tech. Rep. CUCS-006-96, Feb. 1996.

[18] N. Sebe and M. S. Lew, "Comparing salient point detectors," *Pattern Recognit. Lett.*, vol. 24, no. 1–3, pp. 89–96, 2003.

[19] Q. Tian, N. Sebe, M. S. Lew, E. Loupias, and T. S. Huang, "Image retrieval using wavelet-based salient points," *J. Electron. Imag.*, vol. 10, no. 4, pp. 835–849, 2001.

Aristeidis Diplaros (S'02) received the diploma in electronic and computer engineering from the Technical University of Crete, Crete, Greece, in 2001. He is currently pursuing the Ph.D. degree at the Faculty of Science, University of Amsterdam, Amsterdam, The Netherlands.

His research interests include computer vision and image retrieval.



Theo Gevers (M'01) is an Associate Professor of computer science at the University of Amsterdam, Amsterdam, The Netherlands.

His main research interests are in the fundamentals of image database system design, image retrieval by content, theoretical foundations of geometric and photometric invariants, and color image processing.

Ioannis Patras (S'97–M'02) received the B.Sc. and M.Sc. degrees in computer science from the Computer Science Department, University of Crete, Heraklion, Greece, in 1994 and in 1997, respectively, and the Ph.D. degree from the Department of Electrical Engineering, Delft University of Technology, Delft (TU Delft), The Netherlands, in 2001.

From 2001 to 2003, he was a Postdoctorate Researcher in the area of multimedia analysis at the University of Amsterdam, Amsterdam, The Netherlands. From 2003 to 2005, he was a Postdoctorate Researcher in the area of vision-based human machine interaction at TU Delft. Since 2005, he has been a Lecturer in computer vision at the Department of Computer Science, University of York, York, U.K. His research interests lie mainly in the areas of computer vision and pattern recognition and their applications in multimedia data management, multimodal human computer interaction, and visual communications.