Color-Shape Context for Object Recognition

Aristeidis Diplaros, Theo Gevers
Faculty of Science,
University of Amsterdam,
The Netherlands
{*diplaros, gevers*}*@science.uva.nl*

Ioannis Patras
Department of Information Technology and Systems,
Delft University of Technology,
The Netherlands
I.Patras @its.tudelft.nl

Abstract

In this paper, we study computational models and techniques to merge color and shape invariant information to recognize objects. We propose a feature, which we call color shape context, and it is a histogram that combines the spatial (shape) and color information of the image in one compact representation. This histogram codes the locality of color transitions in an image. Illumination invariant derivatives are first computed and provide the edges of the image, which is the shape information of our feature. These edges are used to obtain similarity (rigid) invariant shape descriptors. The color transitions that take place on the edges are coded in an illumination invariant way and are used as the color information. The color and shape information are combined in one multidimensional vector. Our experiments show that the feature is invariant to the similarity transformations of shape such as translation, rotation and scaling and also to noise and illumination changes of color. We conducted our experiments in three databases whose size ranges from 500 to 7200 images. We report considerably better results than only color-based or only shape-based methods. We also found experimentally that the feature exhibits robustness to viewpoint changes for the COIL-100 dataset.

Keywords:

Photometric and geometric invariants, color-shape context, object recognition.

1 Introduction

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

Many practical object recognition systems are appearance- or model-based. To succeed they address two major interrelated problems: object representation and object matching. The representation should be good enough to allow for reliable and efficient matching. The recognition consists of matching the stored models (model-based) or images (appearance-based), encapsulated in a 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 query 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 [20]. Other shape invariants are computed based on moments, Fourier transform coefficients, edge curvature and arc length [19], [23]. Unfortunately, shape features are rarely adequate for discriminatory object recognition of 3-D objects from arbitrary viewpoints. The shape-based approach is often insufficient especially in case of large data sets [10]. Another way 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 common recognition scheme is to represent and match images on the basis of color invariant histograms [17], [8]. The color-based matching approach is widely in use in various areas such as object recognition, content-based image retrieval and video analysis.

Little research has been done on how to combine color

and shape information for object recognition. Important work is done by [3], [13] using moment invariants that combine geometric and photometric changes for planar objects. Further, in [10] 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 [12] for visual object recognition. However, the scheme is heavily dependent on severe illumination changes.

Therefore, in this paper, we study computational models and techniques to merge color and shape *invariant* information to recognize objects in 3D- scenes. Shape deformations occur by a change in viewpoint, object position etc. Deformations for the color channels occur due to shading, illumination, shadows etc. To this end, a vector-based framework is used to index images on the basis of color, shape and composite information. The scheme makes use of a color-shape context providing a high-discriminative cue in vector form to be used as an index. In this paper we do not consider robustness against cluttering and occlusion.

The recognition 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 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.

The 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 3. In Section 4, color and shape invariant information is combined. Matching is discussed in Section 4.2. Finally, experiments are given in Section 5.

2 Photometric Invariants for Color Images

The color of an object varies with changes in illuminant color (Spectral Power Distribution) and geometry (i.e. 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.

2.1 Illumination Invariant Derivatives

Various illumination-independent color ratios have been proposed [8], [14]. These color ratios are derived from neighboring points. A drawback, however, is that these color ratios might be negatively affected by the geometry and pose of the object.

Therefore, we focus on the following color ratio [9]:

$$M(C^1_{\vec{x}_1}, C^1_{\vec{x}_2}, C^2_{\vec{x}_1}, C^2_{\vec{x}_2}) = \frac{C^1_{\vec{x}_1} C^2_{\vec{x}_2}}{C^1_{\vec{x}_2} C^2_{\vec{x}_1}}, C^1 \neq C^2, \quad (1)$$

expressing the color ratio between two neighboring image locations, for $C^1, C^2 \in \{C^1, C^2, ..., C^N\}$ giving the measured sensor pulse response at different wavelengths, where $\vec{x}_1$ and $\vec{x}_2$ denote the image locations of the two neighboring pixels.

For a standard RGB color camera, we have:

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

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

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

The color ratio is independent of the illumination, a change in viewpoint, and object geometry [9].

For the ease of exposition, we concentrate on m_1 based on the RG-color bands in the following discussion. Without loss of generality, all results derived for m_1 will also hold for m_2 and m_3.

Taking the natural logarithm of both sides of Eq. 2 results for m_1 in:

$$\ln m_1(R^{\vec{x}_1}, R^{\vec{x}_2}, G^{\vec{x}_1}, G^{\vec{x}_2}) = \ln(\frac{R^{\vec{x}_1} G^{\vec{x}_2}}{R^{\vec{x}_2} G^{\vec{x}_1}}) =$$

$$\ln R^{\vec{x}_1} + \ln G^{\vec{x}_2} - \ln R^{\vec{x}_2} - \ln G^{\vec{x}_1} = \ln(\frac{R^{\vec{x}_1}}{G^{\vec{x}_1}}) - \ln(\frac{R^{\vec{x}_2}}{G^{\vec{x}_2}}) \quad (5)$$

Hence, the color ratios can be seen as differences at two neighboring locations $\vec{x}_1$ and $\vec{x}_2$ in the image domain of the logarithm of R/G:

$$\nabla_{m_1}(\vec{x}_1, \vec{x}_2) = (\ln(\frac{R}{G}))^{\vec{x}_1} - (\ln(\frac{R}{G}))^{\vec{x}_2} \quad (6)$$

By taking these differences in a particular direction between neighboring pixels, the finite-difference differentiation is obtained of the logarithm of image R/G which is independent of the illumination color, and also a change in viewpoint, the object geometry, and illumination intensity. We have taken the gradient magnitude by applying Canny's edge detector (derivative of the Gaussian with $\sigma = 1.0$) on image $\ln(R/G)$ with non-maximum suppression in a standard way to obtain gradient magnitudes at local edge maxima denoted by $\mathcal{G}_{m_1}(\vec{x})$, where the Gaussian smoothing suppresses the sensitivity of the color ratios to noise.

The results obtained so far for m_1 hold also for m_2 and m_3, yielding a 3-tuple $(\mathcal{G}_{m_1}(\vec{x}), \mathcal{G}_{m_2}(\vec{x}), \mathcal{G}_{m_3}(\vec{x}))$ denoting gradient magnitude at local edge maxima in images $\ln(R/G)$, $\ln(R/B)$ and $\ln(G/B)$ respectively. For pixels on a uniformly colored region (i.e. with fixed surface albedo), in theory, all three components will be zero whereas at least one the three components will be non-zero for pixels on locations where two regions of distinct surface albedo meet.

Higher order derivatives are taken to obtain salient points such as corners and T-junctions. These higher order derivatives are computed again by Gaussian derivatives applied on the illumination invariant color model. For the ease of exposition we concentrate on $\ln m_1$ in the following discussion.

To compute higher order derivatives, we apply the partial derivatives of the Gaussian up to order 2 for image $\ln \frac{R}{G}$. Let $\{\nabla_{m_{1_x}}, \nabla_{m_{1_y}}, \nabla_{m_{1_{xx}}}, \nabla_{m_{1_{yy}}}, \nabla_{m_{1_{xy}}}\}_\sigma$ denote the set of the first five partial Gaussian derivatives. From these partial derivatives, we have computed the Laplacian and the Hessian. Note that the Laplacian operator finds its roots in the modeling of certain psychophysical processes in mammalian vision and hence is suited to compute 'visual salient' points.

2.2 Noise Robustness of Illumination Invariant Derivatives

The above defined illumination 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 modulus. 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

$$u = q(u_1, u_2, \cdots, u_N) \tag{7}$$

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 [18]

$$\hat{u} = q(\hat{u}_1, \cdots, \hat{u}_N). \tag{8}$$

Then, it follows that if the uncertainties in $\hat{u}_1, \cdots, \hat{u}_N$ are independent, random and relatively small, the predicted uncertainty in q is given by [18]

$$\sigma_q = \sqrt{\sum_{j=1}^{N} (\frac{\partial q}{\partial \hat{u}_i} \sigma_{\hat{u}_i})^2} \tag{9}$$

the so-called squares-root sum method. Although (9) is deduced for random errors, it is used as an universal formula for various kinds of errors.

Focusing on the first derivative, the substitution of (6) in (9) gives the uncertainty for the illumination invariant coordinates

$$\sigma_{\nabla_{m_1}}(\vec{x}_1, \vec{x}_2) = \sqrt{\frac{\sigma^2_{R_{\vec{x}_1}}}{R^2_{\vec{x}_1}} + \frac{\sigma^2_{G_{\vec{x}_1}}}{G^2_{\vec{x}_1}} + \frac{\sigma^2_{R_{\vec{x}_2}}}{R^2_{\vec{x}_2}} + \frac{\sigma^2_{G_{\vec{x}_2}}}{G^2_{\vec{x}_2}}} \tag{10}$$

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 Eq.10, it can be derived that color ratio becomes unstable around the black point $R = G = B = 0$.

Further, to propagate the uncertainties from these color components through the Gaussian gradient modulus, the uncertainty in the gradient modulus is determined by convolving the confidence map with the Gaussian coefficients. As a consequence, we obtain:

$$\sigma_{\nabla F} \leq \frac{\sum_i \left[(\partial c_i/\partial x) \cdot \sigma_{\partial c_i/\partial x} + (\partial c_i/\partial y) \cdot \sigma_{\partial c_i/\partial y}\right]}{\sqrt{\sum_i \left[(\partial c_i/\partial x) + (\partial c_i/\partial y)\right]}}, \tag{11}$$

where i is the dimensionality of the color space and c_i is the notation for particular color channels. In this way, the effect of measurement uncertainty due to noise is propagated through the color constant ratio gradient.

For a Gaussian distribution 99% of the values fall within a 3σ margin. If a gradient modulus is detected which exceeds $3\sigma_{\nabla F}$, we assume that there is 1% chance that this gradient modulus corresponds to no color transition.

3 Geometric Invariant Transformation

In this section, shape transformations are discussed to measure shape properties of a set of coordinates (i.e. edges, corners and T-junctions) of an image object independent of a coordinate transformation. To this end, we consider the edges, computed from the illumination invariant derivatives proposed in previous section, as the spatial information which is normalized by the aspect ratio given in following Section.

3.1 Affine Deformations and Inverse Transformation

The geometric deformations considered in this paper are up to affine transformations:

$$\vec{x'} = \mathbf{A}\vec{x} + \mathbf{B} \tag{12}$$

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

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, B = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{13}$$

This transformation is consider to approximate the projective transformation of 3-D planar objects , therefore is valid when the object is relative far away from the camera.

It is known that the spatial moment is defined as:

$$M_u(m, n) = \sum_{j=1}^{J} \sum_{k=1}^{K} x_k^m y_j^n f(j, k) \tag{14}$$

Further, the ratio's $\overline{x}_k = \frac{M(1,0)}{M(0,0)}$, $\overline{y}_k = \frac{M(0,1)}{M(0,0)}$ define the centroid. Transforming the image with respect to $b = [\overline{x}_k \ \ \overline{y}_k]^T$ yields invariance to translation. The principle 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 principle axis, is defined as follows [15]:

$$\tan 2\theta = \frac{2M_{11}}{M_{20} - M_{02}} \tag{15}$$

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

$$A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{16}$$

will provide rotation invariance.

A change in scale by a factor of δ in the x-direction and γ in the y-direction is given by

$$M_u^{'}(m, n) = \delta^{m+1}\gamma^{n+1} M_u(m, n)$$

where

$$M_u^{'}(0, 0) = M_u(0, 0) = 1 \text{ and } M_u^{'}(2, 0) = M_u^{'}(0, 2)$$

then we obtain $\delta = 1/\gamma$ where

$$\delta = (M_u(0, 2)/M_u(2, 0))^{0.25} \tag{17}$$

Setting the transformation matrix to

$$A = \begin{bmatrix} \delta & 0 \\ 0 & 1/\delta \end{bmatrix} \tag{18}$$

we obtain aspect ratio normalization.

In conclusion, we normalize our spatial information with respect to translation, rotation and aspect-ratio using a collection of low-order moments. The scale invariance and the robustness to affine transformation are discussed in the following section.

4 Indexing and Matching

In this section, we propose an alternative indexing scheme to combine shape and color information. The scheme is called the color-shape context which is related to the shape context proposed by [1] for shape matching. The difference is that the color-shape context combines both color and spatial information into one unifying indexing framework. Moreover, the scheme is robust to affine deformations of shape and color.

Let the image database consist of a set $\{I_d\}_{d=1}^{N_b}$ of color images. Color-shape contexts are created for each image I_d to represent the distribution of quantized invariant values in a multidimensional invariant space. Color-shape contexts are formed on the basis of color, shape and combination of both.

4.1 Color-Shape Representation Scheme

The color-shape framework is as follows. For a color image $\mathcal{I} : \mathbb{R}^2 \rightarrow \mathbb{R}^3$, illumination invariant edges are computed to obtain a binary image $\mathcal{E} : \mathbb{R}^2 \rightarrow \{0, 1\}$. Then at the edge points we define:

$$\vec{p_n} = (x, y, m_1, m_2, m_3) |\ \ \mathcal{E}(x, y) = 1 \tag{19}$$

where x, y are pixel coordinates in image $\mathcal{E}$ and m_1, m_2, m_3 are the color invariant ratios calculated in image $\mathcal{I}$. Further, $\vec{p_n} \in \mathbb{R}^5$.

We can decompose each of these $\vec{p}_n$ vectors as follows:

$$\vec{p}_n = \vec{p}_{S_n} + \vec{p}_{C_n} \tag{20}$$

where:

$$\vec{p}_{S_n} = (x, y, 0, 0, 0)$$
$$\vec{p}_{C_n} = (0, 0, m_1, m_2, m_3)$$

Note that the set $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$ of n points in a 5-dimensional space representing both the shape and color information in the image. Each of these n points is now considered to represent a vector originating from a central

point. To provide noise robustness, we consider the distribution of these vectors over relative positions. Then, the color-shape context $h^c_{\mathcal{P}}$ of set $\mathcal{P}$ is defined as the coarse histogram of the distribution of all p_n vectors, as defined in Eq.20, over the relative position c. In other words, the histogram of the decomposed and quantized versions of the p_n vectors, with respect to the distance and orientation from a central point c, is called color-shape context of the image. The vector $\vec{p}_S$ is represented by its polar coordinates while the vector $\vec{p}_C$ is represented with its spherical coordinates.

To construct the color-shape context of each image I using a relative position c, we first translate the vectors $\vec{p}_{S_n}$ with respect to point c and then the following histogram is computed:

$$h^c_{\mathcal{P}}(k) = \frac{\#\{q : q \in bin(k)\}}{n} \tag{21}$$

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 compute the log-polar with equally spaced radial bins. An equally-spaced bins scheme is used to partition the 3-D color invariant space.

Note that the color-shape context feature is rotation, scale, translation and aspect ratio change invariant. Further, the color-shape context has intrinsic robustness against small amounts of displacement noise, since it is based on the distribution of illumination invariant derivatives instead of their exact positions. Total affine invariance, which would include skew invariance, is not claimed, but the experimental results show that the feature is robust to affine transformations.

4.2 Matching in simple scenes

For object recognition in simple scenes (i.e. one object per image) we set the relative position c as the center of gravity of the spatial information. Note that the shape information is normalized and that the center of gravity is at the origin $(0, 0)$. To achieve scale invariance we normalize all $|\vec{p}_{S_n}|$ with respect to the mean distance between all point pairs in $\mathcal{E}$. Let's 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:

$$C_{ab} = \frac{1}{2}\sum_{k=1}^{K} \frac{(h_a(k) - h_b(k))^2}{(h_a(k) + h_b(k))} \tag{22}$$

The complexity of this operation is only dependent on the number of K which is constant for all images in the database and is usually small. Note that the color-shape context feature is rotation, scale, translation and aspect ratio change invariant. Furthermore, the color-shape context exhibits intrinsic robustness against small amounts of displacement noise, since it is based on the distribution of illumination invariant derivatives instead of their exact positions.

4.3 Matching in complex scenes

For object recognition in complex scenes, which may contain cluttering and occlusion, a modified matching strategy is adopted. The reason is that the framework based on moments, used to determine the center and the scale of the color-shape context, can no longer be used since the edge points may belong to more than one object. To this end, we build multiple representations of color-shape context per image and match them with a modified distance function. The multiple color-shape contexts are build at different centers, scales and orientations. The modification of the cost function aims at reducing the influence of large costs introduced at 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:

$$O_k = \begin{cases} 1 & : \quad d_k/q_k > T \\ 0 & : \quad d_k/q_k \leq T \end{cases}$$

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

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

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:

$$q_k = \sum_{\{n \mid f(n)=k\}} (h_a(n) + h_b(n))$$

The cost function is now as follows:

$$C'_{ab} = C_{ab} - \sum_k O_k \frac{1}{|N_k|} \sum_{l \in N_k} (d_k - T q_k O_l) \tag{23}$$

where N_k is the 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 cases, 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 is $q_k T$.

5 Experiments

In this section, we consider the performance of the proposed object recognition scheme. Therefore, in section 5.1, the dataset and matching quality are discussed. Then, in the remaining sections, we test our object recognition scheme on two different datasets with respect to viewing and illumination conditions.

Further, for comparison, the illumination invariant derivatives $m_1 m_2 m_3$ are used as a direct index resulting in a 3-dimensional histogram. In this context, pixels from the same boundary will generate the same gradient value and hence accumulate in the same histogram bin. As a consequence, the total accumulation for a particular histogram bin represents a measure of the boundary length between two homogeneously painted surface patches. Because each non zero bin indicates the presence of a distinct boundary, the histogram is indicative for the boundary variety in view.

For a measure of match quality, let rank r^{Q_i} denote the position of the correct match for test image Q_i, $i = 1, ..., N_2$, 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, the average ranking percentile is defined by:

$$\overline{r} = (\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{N_1 - r^{Q_i}}{N_1 - 1}) 100\% \qquad (24)$$

5.1 The Datasets

For comparison reasons, we have selected two different datasets: Amsterdam and Columbia - COIL-100, which are publicly available and often used in the context of object recognition [10], [21].

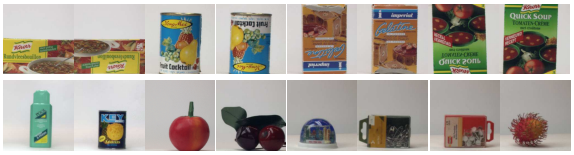

Figure 1. *Various images which are included in the Amsterdam image dataset of 500 images. The images are representative for the images in the dataset. Objects were recorded in isolation (one per image).*

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 256x256 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: In order to test the performance of our algorithm against variability in appearance, the COIL-100 has been selected which have been collected at the Columbia University. For these experiments, we aim at multi-view object recognition i.e. there are various images taken from the object (i.e. back, front, from aside etc.), see Fig. 2. However, a number of objects in the database are single-colored and therefore hard to recognize.

5.2 Viewpoint Robustness

To test the effect of change in viewpoint we used the COIL-100 dataset. This dataset consist of 7200 images from 100 objects which have been put perpendicularly in front of the camera and in total 72 recordings were generated by varying the angle between the camera with 5 degrees with respect to the object, see Fig. 2. We conducted our experiment as following, we gave as a query a view of an object ranging from 0 to 70 degrees (15 different views) and the method had to recognize the corresponding object from the 0 degrees views of all 100 objects. This was done for all 100 objects and for all 15 views of each object. We

Figure 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.*

concentrate on the quality of the recognition rate with respect to varying viewpoint differentiated by color information, shape information, and the integration of both. Therefore, we first study to what extent the proposed framework is viewpoint independent. In the mean time, we research on whether the combination of color and shape invariant information will outperform the matching scheme based on only color or shape. To this end, we have constructed three different color-shape context histograms. Firstly, we have included both color and shape invariant information denoted

by $\mathcal{H}_{\mathrm{CS}}$. Secondly, only color is considered which will be given by $\mathcal{H}_{\mathrm{C}}$. Thirdly, we used only shape information denoted by $\mathcal{H}_{\mathrm{S}}$. Also, we have included in our experiment the well-known, color-based method of Histogram Intersection [17] denoted by $\mathcal{HI}_{RGB}$. With $\mathcal{HI}_{inv-rgb}$ we denote the results of the Histogram Intersection when the images have been initially transformed to the normalized rgb color space. The performance of the recognition scheme is given

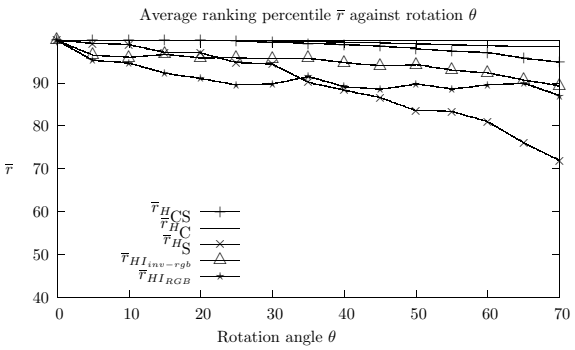

Figure 3. *The discriminative power of the color-shape matching process under varying viewpoint differentiated by color information, shape information, and the integration of both. The average ranking percentile of color-shape, color, and shape contexts are denoted by $\overline{r}_{\mathcal{H}_{CS}}$,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

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

In conclusion, recognition based on our framework using color invariant and both shape and color invariant information produces the highest discriminative power. The small performance gain in using only color denotes the lack of robustness of the shape invariant part of our framework to viewpoint change. Our method always outperform the Histogram Intersection method even when the images have been transformed to the normalized rgb color space. 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.

5.3 Illumination Robustness

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 $\alpha \in \{ 0.3, 0.5, 0.7, 0.8, 0.9,\ 1.0, 1.1, 1.2, 1.3, 1.5, 1.7\}$. The discriminative power of the histogram matching process differentiated plotted against illumination intensity is shown in Fig. 4. The color-shape context is to a large degree robust

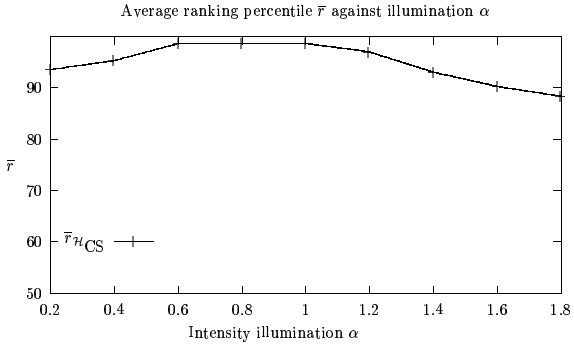

Figure 4. *The discriminative power of the color-shape matching process plotted against the varying illumination intensity.*

to illumination intensity changes even if the shape based recognition is not. To measure the sensitivity of our method

Figure 5. *2 objects under varying illumination intensity generating each 4 images with SNR $\in$ $\{ 24, 12, 6, 3\}$.*

with respect to varying SNR, 10 objects were randomly chosen from the Amsterdam image dataset. Then, each object has been recorded again under a global change in illumination intensity (i.e. dimming the light source) generating images with 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 $3 <$ SNR < 12, the results show a rapid decrease in the performance. For these SNR's, the color-shape based recognition scheme still outperforms the shape based recog-

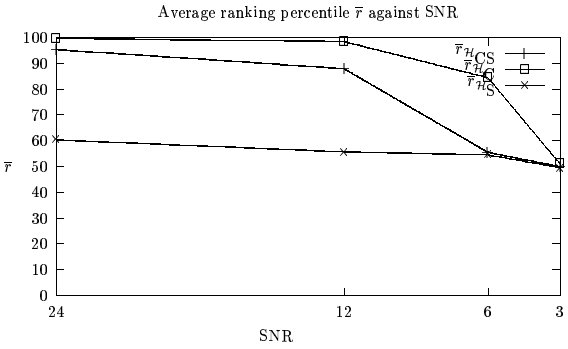

Figure 6. *The discriminative power of the color-shape matching process plotted against the varying SNR.*

nition. For SNR < 3, the performance of all methods incline. 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 signal-to-noise ratios but can not outperform the color based recognition since the other part of its components does not performs well. Even when the object is nearly visible, object identification is still sufficient. Again, the matching based on both shape and color invariants produces good discriminative power but not as good as the color-based approach.

5.4 Occlusion Cluttering Robustness

To test our method for complex scenes, we used a subset of the Amsterdam dataset. An image which contained multiple objects and occlusion was used as the query. The dataset was 100 randomly selected images, including two instances of an occluded object. The average ranking percentile for these two queries was 99%. Note that despite the cluttering and the quite large amount of occlusion the method was capable to identify the object.

Figure 7. *Left: Image containing cluttering and occlusion. Right: The objects that were retrieved from a dataset of 100 objects.*

6 Conclusions

In this paper, we proposed computational models and techniques to merge color and shape *invariant* information to recognize objects. A vector-based framework is proposed to index images on the basis of illumination (color) invariants and viewpoint (shape) invariants. From the experimental results it is shown that the method is able to recognize rigid objects in 3-D complex scenes robust to illumination, viewpoint and noise.

References

[1] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, April 2002.

[2] G. Carpaneto and Toth P. Solution of the assignment problem (algorithm 548). *ACM Transactions on Mathematical Software*, 6:104–111, 1980.

[3] L. van Gool, T. Moons, and D. Ungureanu, Geometric/Photometric Invariants for Planar Intensity Patterns, ECCV, pp. 642-651, 1996.

[4] A. Pentland, R. Picard, and S. Sclaroff. Photobook: content-based manipulation of image databases. *International Journal of Computer Vision*, 18(3):233–254, June 1996.

[5] S.M. Smith and J.M. Brady. Susan - a new approach to low level image processing. *Int. Journal of Computer Vision*, 23(1):45–78, May 1997.

[6] H.A.L. van Dijck. *Object Recognition with Stereo Vision and Geometric Hashing*. PhD thesis, University of Twente, Enschede, February 1999.

[7] Finlayson, G.D., Drew, M.S., and Funt, B.V., Spectral Sharpening: Sensor Transformation for Improved Color Constancy, JOSA, 11, pp. 1553-1563, May, 1994.

[8] Funt, B. V. and Finlayson, G. D., Color Constant Color Indexing, IEEE PAMI, 17(5), pp. 522-529, 1995.

[9] Th. Gevers and Arnold W.M. Smeulders, Color Based Object Recognition, Pattern Recognition, 32, pp. 453-464, March, 1999.

[10] Th. Gevers and A. W. M. Smeulders, Image Indexing using Composite Color and Shape Invariant Features, Int. Conf on Computer Vision, pp. 234-238, Bombay, India, 1998.

[11] Th. Gevers and H. Stokman, Classification of Color Edges in Video into Shadow-Geometry, Highlight, or Material Transitions, IEEE Trans. on Multimedia, 2003.

[12] B. W. Mel, SEEMORE: Combining Color, Shape, and Texture Histogramming in a Neurally Inspired Approach to Visual Object recognition, Neural Computation, 9, pp. 777-804, 1997.

[13] F. Mindru, T. Moons, and L. van Gool, Recognizing Color Patterns Irrespectively of Viewpoint and Illumination, IEEE CVPR, pp. 368-373, 1999.

[14] S. K. Nayar, and R. M. Bolle, Reflectance Based Object Recognition, International Journal of Computer Vision, Vol. 17, No. 3, pp. 219-240, 1996

[15] A.P. Reeves *et. al.*, Three-Dimensional Shape Analysis Using Moments and Fourier Descriptors, IEEE trans. PAMI, vol. 10, no. 6, 1988.

[16] S.A. Shafer, Using Color to Separate Reflection Components, COLOR Res. Appl., 10(4), pp 210-218, 1985.

[17] Swain, M. J. and Ballard, D. H., *Color Indexing*, International Journal of Computer Vision, Vol. 7, No. 1, pp. 11-32, 1991.

[18] J. R. Taylor, *An Introduction to Error Analysis*, University Science Books, 1982.

[19] T.H. Reis, Recognizing Planar Objects using Invariant Image Features, Springer- Verlag, Berlin, 1993.

[20] Rothwell, C. A., Zisserman, A., Forsyth, D. A. and Mundy, J. L., *Planar Object Recognition Using Projective Shape Representation*, International Journal of Computer Vision, Vol. 16, pp. 57-99, 1995.

[21] N. Sebe and M.S. Lew and D.P. Huijsmans, *Toward Improved Ranking Metrics*, IEEE Trans. on PAMI, 22(10), pp. 1132-1143, 2000.

[22] Veiblen, O. and Young. J. W., *Projective Geometry*, Ginn. Boston, 1910.

[23] I. Weiss, Geometric Invariants and Object Recognition, International Journal of Computer Vision, 10(3), pp. 207-231, 1993.