COMPUTER VISION AND IMAGE UNDERSTANDING
Vol. 71, No. 1, July, pp. 74–93, 1998
ARTICLE NO. IV970603

Using Photometric Invariants for 3D Object Recognition*

Kenji Nagao

Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts and Multimedia Systems Research Laboratory, Matsushita Electric Industrial Co., Ltd.

and

W. Eric. L. Grimson

Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts

Received May 8, 1995; accepted January 24, 1997

In this paper we describe a new efficient algorithm for recognizing 3D objects by combining photometric and geometric invariants. We derive some photometric properties that are invariant to the changes of illumination and to relative object motion with respect to the camera and/or the lighting source in 3D space. We show that recognition does not require a full constancy of colors; rather, it only needs something that remains unchanged under the varying light conditions and poses of the objects. Combining the derived color invariants and spatial constraints on the object surfaces, we identify corresponding positions in the model and the data space coordinates, using centroid invariance of corresponding groups of feature positions. Tests are given to show the stability and efficiency of our approach to 3D object recognition. © 1998 Academic Press

1. INTRODUCTION

In a typical approach to model-based object recognition [16], geometric models are matched against features extracted from an image, where the features are typically localized geometric events, such as vertices. Objects are considered to have undergone a transformation in space to yield a novel view for the image. To solve for this transformation explicitly, recognition methods use matches of features to hypothesize a transformation, which is used to align the model with the image and select the best-fit pair of transformation and model. While this approach to recognition has achieved considerable success, there still remain practical problems to be solved.

One such problem is the computational complexity of the method. For example, even with popular algorithms (e.g., [22, 36]) to recognize an object with m features from an image with n features, we must examine m^3n^3 combinations of hypotheses where m and n can be easily on the order of several hundreds in natural pictures. A second problem is the tolerance of the algorithm to scene clutter. To verify the hypothesized transformation, object recognition algorithms have to collect evidence of actual correspondences characterized by that transformation. This is usually done by looking for nearest image features around the transformed model features, or equivalently by casting votes to a hash table of parameters, such as affine invariant parameters, leading to a correspondence (e.g., [27]). In either case, when features are extracted from the image with perturbations, and if the image is cluttered so that the feature distribution is too dense, it is difficult to tell whether an image feature thus detected is the one actually corresponding to the model feature or if it just happened to fall close to the transformed model feature. This issue has been extensively analyzed, both theoretically and empirically, giving arguments about the limitations of geometric feature based approaches to recognition (e.g., [1, 16, 17]).

Given the limitations of conventional approaches to recognition which depend solely on local geometrical features, it is natural to consider cues other than simple local geometric features. One such candidate is photometric information like color, because we know that color often characterizes objects well and it is almost invariant to change of view and lighting conditions. In parallel with geometry, color properties of the object surface should be a strong key to the perception of the surface. However, most authors who have exploited color in recognition used it simply for segmentation, e.g., [5, 18, 33], mostly because color is considered to be more contributive in building up salient features on the object surface than in giving precise information on the location and the poses of the objects. Exceptions include Swain [31, 32], Funt and Finlayson [10], and Nayar and Bolle [30] who have used photometric information more directly for recognition, the first two authors for indexing, and the third for matching processes.

At the same time, however, they abandoned the use of local geometric features, which are still very useful in predicting the locations and the poses of the objects. Swain used only a color

* This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research Contract N00014-91-J-4038. W.E.L.G. was also supported by NSF Grant IRI-8900267.

histogram for representing objects and matched it over the image to identify the object and localize its presence in the image. Funt and Finlayson improved this method by using their new color invariant as the input to the histogram process, instead of using the color values directly. Nayar and Bolle proposed a photometric invariant and used it for matching regions with consistent colors given the partitioned model and image derived by some other color properties. Therefore, it requires a preliminary segmentation of the image into regions having consistent colors.

In this paper, we attempt to exploit both geometric and photometric cues to recognize 3D objects, by combining them more tightly. Our goal is to develop an efficient and reliable algorithm for recognition by taking advantage of the merits of both geometric and color cues: the ability of color to generate larger and thus more salient features reliably, as well as of adding more selectivity to features, which enables more efficient and reliable object recognition, and the rich information carried by the set of local geometric features that is useful in accurately recovering the transformation that generated the image from the model. To realize this, we have developed new photometric invariants which are suitable for this approach. Then, we combine the proposed photometric properties with the centroid alignment approach of matching geometric feature groups in the model and the image that we have recently proposed [28]. This strategy gives an efficient and reliable algorithm for recognizing 3D objects. In our testing, it took only 0.2 s to derive corresponding positions in the model and the image for natural pictures. A preliminary version of this work was presented in [29].

2. NOVEL PHOTOMETRIC INVARIANTS

In this section, we develop some photometric invariants that can be used as strong cues in the recognition of 3D objects. The motivation is to find simple cues that can be used reliably to isolate portions of an object under a range of image acquisition variations.

2.1. A Related Issue: Color Constancy

The invariants we develop here are related to the notion of *color constancy*, that is—whether in human or machine vision —the perceptual ability to determine the surface reflectance property of the target object given the reflected light from the object surface in the receptive field. If a color constancy algorithm could perform sufficiently well, we could use it for object recognition because it would provide a unique property of the object itself. Unfortunately, however, color constancy is generally difficult to compute in practice. Actually, as far as we have checked almost all authors have addressed problems in a strongly constrained world like Mondrian space [9, 14, 15, 21, 35, 37]: a 2D space composed of several matte patches overlapping each other. Then, based on the assumption that both the ambient light and the surface reflectance for planar surfaces can be approximated by linear combinations of a small number of fixed basis functions [7, 24], the problem becomes manageable [9, 12, 14, 15, 35, 37]. Finlayson recently removed the assumption of linear model of the ambient light by instead requiring surface observations under two different illuminants [9]. However, all of those works do not, at least explicitly, address the problem for 3D surfaces. Thus, we tentatively conclude that conventional color constancy algorithms cannot be used for recognizing a 3D world as presented. Contrastively, the invariant property to be presented here is effectively computed from the images at the same time as geometrical features are extracted.

2.2. Novel Color Invariants

Let $S(\mathbf{x}, \lambda)$ be the spectral reflectance function of the object surface at $\mathbf{x}$, that is the property one has to recover in color constancy, let $E(\mathbf{x}, \lambda)$ be the spectral power distribution of the ambient light, and let $R_k(\lambda)$ be the spectral sensitivity of the kth sensor, then $\rho_k(\mathbf{x})$, the scalar response of the kth sensor channel to be observed, is described as

$$\rho_k(\mathbf{x}) = \int S(\mathbf{x}, \lambda) E(\mathbf{x}, \lambda) R_k(\lambda)\, d\lambda, \tag{1}$$

where, generally, S can be an arbitrary function describing geometric and spectral properties of the surface at $\mathbf{x}$ and E could also be an arbitrary function of $\mathbf{x}$ and λ. The integral is taken over the visible spectrum (usually from 380 to 800 nm). The geometric factor of the object surface, that is usually considered to include the surface normal and the relative angle of the incident and reflecting light direction with respect to the surface normal, is very crucial in the 3D world [20], which makes the color constancy more confounding for 3D surfaces. Since it is known that a spectrum distribution of the surface reflectance of many materials depends very little on the surface geometry [26], we may break up the surface reflectance function $S(\mathbf{x}, \lambda)$ into the product of geometry $G(\mathbf{x})$ and spectrum property $L(\mathbf{x}, \lambda)$ such that $S(\mathbf{x}, \lambda) = G(\mathbf{x}) L(\mathbf{x}, \lambda)$. Then, Eq. (1) becomes

$$\begin{aligned}\rho_k(\mathbf{x}) &= \int G(\mathbf{x}) L(\mathbf{x}, \lambda) E(\mathbf{x}, \lambda) R_k(\lambda)\, d\lambda \\ &= G(\mathbf{x}) \int L(\mathbf{x}, \lambda) E(\mathbf{x}, \lambda) R_k(\lambda)\, d\lambda. \end{aligned} \tag{2}$$

Assumption AE1: Constant ambient light assumption over the entire surface. If we assume that the ambient light spectrum distribution is constant over the entire surface of the objects, E becomes simply a function of wavelength λ. This assumption is justified when the lighting source is sufficiently far away from the object relative to the size of the object surface, and mutual illumination and shadowing are not significant. This yields

$$\rho_k(\mathbf{x}) = G(\mathbf{x}) \int L(\mathbf{x}, \lambda) E(\lambda) R_k(\lambda)\, d\lambda. \tag{3}$$

Taking the ratios between two (i, j) channel responses eliminates the geometric factor $G(\mathbf{x})$ which depends on the relative orientation of the object surface with respect to the camera and/or the lighting source,

$$\gamma_{ij} \equiv \frac{\rho_i(\mathbf{x})}{\rho_j(\mathbf{x})} = \frac{\int L(\mathbf{x},\lambda)E(\lambda)R_i(\lambda)\,d\lambda}{\int L(\mathbf{x},\lambda)E(\lambda)R_j(\lambda)\,d\lambda}. \tag{4}$$

By the same reasoning, we have a similar form after the motion of the object with respect to the camera and/or the lighting source,

$$\gamma'_{ij} \equiv \frac{\rho'_i(\mathbf{x}')}{\rho'_j(\mathbf{x}')} = \frac{\int L'(\mathbf{x}',\lambda)E'(\lambda)R_i(\lambda)\,d\lambda}{\int L'(\mathbf{x}',\lambda)E'(\lambda)R_j(\lambda)\,d\lambda}, \tag{5}$$

where primes show the function after the motion, and this prime notation applies to any symbol expressing some quantity after the motion of the object in the rest of this paper unless otherwise described. In the following, we show that under some assumptions on camera sensors or ambient light properties we can derive photometric invariants from γ's defined above that can be used for recognizing 3D objects.

Assumption AS1: Frequency-selecting sensor assumption. When we approximate the spectral absorption functions R by frequency-selective filters R' such that

$$R'_i \equiv \int_{\alpha_i}^{\beta_i} \delta(\lambda - u)R_i(u)\,du, \tag{6}$$

that is centered around the peak wavelength λ_i of the sensor R_i, where α_i and β_i ($\alpha_i < \lambda_i < \beta_i$) are lower and upper cut-off wavelengths, respectively, we have

$$\rho_i(\mathbf{x}) \approx \int L(\mathbf{x},\lambda)E(\lambda)R'_i(\lambda)\,d\lambda \tag{7}$$

$$= \int L(\mathbf{x},\lambda)E(\lambda)\left[\int_{\alpha_i}^{\beta_i} \delta(\lambda - u)R_i(u)\,du\right]d\lambda \tag{8}$$

$$= \int_{\alpha_i}^{\beta_i}\left[\int L(\mathbf{x},\lambda)E(\lambda)\delta(\lambda - u)\,d\lambda\right]R_i(u)\,du, \tag{9}$$

where the geometrical term G has been omitted, and in the last step the order of integrals has been interchanged.

Assumption AE2: Slow spectral variation of ambient light assumption. Here, if we can assume that the spectral variation of the ambient light is slow enough with respect to the effective range of wavelength, i.e., $\alpha_i \le \lambda \le \beta_i$ such that $E(\lambda) \approx E(\lambda_i)$ for $\alpha_i \le \lambda \le \beta_i$, we have

$$\rho_i(\mathbf{x}) = E(\lambda_i)\int_{\alpha_i}^{\beta_i} L(\mathbf{x},u)R_i(u)\,du. \tag{10}$$

Therefore, we obtain

$$\gamma_{ij}(\mathbf{x}) \approx \frac{E(\lambda_i)\int_{\alpha_i}^{\beta_i} L(\mathbf{x},u)R_i(u)\,du}{E(\lambda_j)\int_{\alpha_j}^{\beta_j} L(\mathbf{x},u)R_j(u)\,du} \tag{11}$$

$$\gamma'_{ij}(\mathbf{x}') \approx \frac{E'(\lambda_i)\int_{\alpha_i}^{\beta_i} L'(\mathbf{x}',u)R_i(u)\,du}{E'(\lambda_j)\int_{\alpha_j}^{\beta_j} L'(\mathbf{x}',u)R_j(u)\,du}. \tag{12}$$

Note that $L(\mathbf{x},\lambda) = L'(\mathbf{x}',\lambda)$, because the spectrum property of the surface reflectance would not be affected by the object motion. Taking the ratio of γ's before and after the motion and/or the change of lighting conditions yields

$$\frac{\gamma_{ij}(\mathbf{x})}{\gamma'_{ij}(\mathbf{x}')} \approx \epsilon_{ij}, \tag{13}$$

where

$$\epsilon_{ij} = \frac{E(\lambda_i)}{E(\lambda_j)}\bigg/\frac{E'(\lambda_i)}{E'(\lambda_j)}. \tag{14}$$

As ϵ's are independent of the position on the surface, Eqs. (13) and (14) show that under the assumptions AE1, AE2, and AS1, we obtain a photometric property γ that is invariant within a consistent scale ϵ to the changes of the spectral property of the ambient light and the orientations of the object surfaces. It is very important to note that the assumption of slow spectral variation of the ambient light must be used in conjunction with that of the frequency selective filter. This is because as demonstrated above "slow variation" is required only for the range of the wavelength for which the sensors are effective. For instance, when we look at the spectral components of the sun's radiation observed on the ground, the power is fairly constant above the wavelength 500 nm that is the ranges covered by Green and Red channels in standard cameras [23].

Assumption AS2: Narrow band sensor assumption. Instead of assuming slow spectral variation of the ambient light and approximating the sensor by frequency-selective sensor, if we place a narrow band filter in front of the camera sensor or when we approximate the spectral absorption functions R by narrow band filters such that $R_i(\lambda) \approx s_i\delta(\lambda_i - \lambda)$, where s_i is the channel sensitivity and the λ_i is the peak of the spectral sensitivity of the ith channel, we obtain ratios from (4) and (5):

$$\gamma_{ij}(\mathbf{x}) \equiv \frac{\rho_i(\mathbf{x})}{\rho_j(\mathbf{x})} \approx \frac{s_i L(\mathbf{x},\lambda_i)E(\lambda_i)}{s_j L(\mathbf{x},\lambda_j)E(\lambda_j)} \tag{15}$$

$$\gamma'_{ij}(\mathbf{x}') \equiv \frac{\rho'_i(\mathbf{x}')}{\rho'_j(\mathbf{x}')} \approx \frac{s_i L(\mathbf{x},\lambda_i)E'(\lambda_i)}{s_j L(\mathbf{x},\lambda_j)E'(\lambda_j)}. \tag{16}$$

Since the bandwidth over which a real camera sensor responds varies from camera to camera, and the standard ones may not be too narrow, this is only an approximation if we do not actually use ones. Taking the ratio of γ's before and after the motion and/or the change of lighting conditions yields

$$\frac{\gamma_{ij}(\mathbf{x})}{\gamma'_{ij}(\mathbf{x}')} \approx \epsilon_{ij}, \tag{17}$$

where

$$\epsilon_{ij} = \frac{E(\lambda_i)}{E(\lambda_j)} \bigg/ \frac{E'(\lambda_i)}{E'(\lambda_j)}. \tag{18}$$

Again, ϵ_{ij} is independent of the position on the surface and depends only on the incident light. Therefore, under the assumptions AE1 and AS2, $\gamma_{ij}(\mathbf{x})$ can be regarded as approximately invariant to the changes of illuminant conditions and to the motions of the object within a consistent scale factor over the object surface.

In order to use γ's thus derived in Eqs. (11), (12) and (15), (16) for object recognition, we might need to normalize its distribution because generally it is invariant only within a scale factor ϵ. As we describe in the next section, considering also the correlative property between γ_{ij}'s of natural objects, obtained using different channel pairs (i, j), we normalize γ's by decorrelating their joint distributions. This allows us to make full use of the information from γ's as well as to remove the scale factor thus deriving an invariant.

Assumption AE3: Only locally constant ambient light assumption. Now, let us assume only a locally constant ambient light spectrum distribution, instead of the globally constant one over the object surface: $E(\mathbf{x}_l, \lambda) = E(\mathbf{x}_m, \lambda)$ for nearby positions $\mathbf{x}_l, \mathbf{x}_m$. Then, accordingly, Eqs. (11) and (12) must be modified, respectively, as

$$\gamma_{ij}(\mathbf{x}) \approx \frac{E(\mathbf{x}, \lambda_i)\int_{\alpha_i}^{\beta_i} L(\mathbf{x}, u)R_i(u)\,du}{E(\mathbf{x}, \lambda_j)\int_{\alpha_j}^{\beta_j} L(\mathbf{x}, u)R_j(u)\,du} \tag{19}$$

$$\gamma'_{ij}(\mathbf{x}') \approx \frac{E'(\mathbf{x}', \lambda_i)\int_{\alpha_i}^{\beta_i} L'(\mathbf{x}', u)R_i(u)\,du}{E'(\mathbf{x}', \lambda_j)\int_{\alpha_j}^{\beta_j} L'(\mathbf{x}', u)R_j(u)\,du}. \tag{20}$$

By the same reason, Eqs. (15) and (16) should be

$$\gamma_{ij}(\mathbf{x}) \equiv \frac{\rho_i(\mathbf{x})}{\rho_j(\mathbf{x})} \approx \frac{s_i L(\mathbf{x}, \lambda_i)E(\mathbf{x}, \lambda_i)}{s_j L(\mathbf{x}, \lambda_j)E(\mathbf{x}, \lambda_j)} \tag{21}$$

$$\gamma'_{ij}(\mathbf{x}') \equiv \frac{\rho'_i(\mathbf{x}')}{\rho'_j(\mathbf{x}')} \approx \frac{s_i L(\mathbf{x}, \lambda_i)E'(\mathbf{x}', \lambda_i)}{s_j L(\mathbf{x}, \lambda_j)E'(\mathbf{x}', \lambda_j)}. \tag{22}$$

Incorporating the assumption, that is, $E(\mathbf{x}_l, \lambda) = E(\mathbf{x}_m, \lambda)$ and $E'(\mathbf{x}'_l, \lambda) = E'(\mathbf{x}'_m, \lambda)$, we again have the following invariant ψ_{ij}^{lm} for both cases; i.e., under the assumptions AE1, AE2, and AS1 (Eqs. (11), (12)),

$$\psi_{ij}^{lm} \equiv \frac{\gamma_{ij}(\mathbf{x}_l)}{\gamma_{ij}(\mathbf{x}_m)} \approx \frac{\int_{\alpha_i}^{\beta_i} L(\mathbf{x}_l, u)R_i(u)\,du}{\int_{\alpha_j}^{\beta_j} L(\mathbf{x}_l, u)R_j(u)\,du} \bigg/ \frac{\int_{\alpha_i}^{\beta_i} L(\mathbf{x}_m, u)R_i(u)\,du}{\int_{\alpha_j}^{\beta_j} L(\mathbf{x}_m, u)R_j(u)\,du}, \tag{23}$$

and under AE1 and AS2 (Eqs. (15), (16)),

$$\psi_{ij}^{lm} \equiv \frac{\gamma_{ij}(\mathbf{x}_l)}{\gamma_{ij}(\mathbf{x}_m)} \approx \frac{L(\mathbf{x}_l, \lambda_i)}{L(\mathbf{x}_l, \lambda_j)} \bigg/ \frac{L(\mathbf{x}_m, \lambda_i)}{L(\mathbf{x}_m, \lambda_j)}. \tag{24}$$

The quantity ψ is invariant, that is $\psi_{ij}^{lm} \approx \psi_{ij}^{lm'}$, because it depends only on surface properties that remain unchanged: $L(\mathbf{x}, \lambda) = L'(\mathbf{x}', \lambda)$. However, ψ_{ij}^{lm} is obviously sensitive to perturbations contained in the image signals, especially when one makes the values of $\gamma_{ij}(\mathbf{x}_m)$ (the denominator in (23) or (24)) close to zero. To stabilize this, we adopt a normalized measure in place of ψ itself:

$$\varphi_{ij}^{lm} \equiv \frac{\gamma_{ij}(\mathbf{x}_l)}{\gamma_{ij}(\mathbf{x}_m) + \gamma_{ij}(\mathbf{x}_l)}. \tag{25}$$

It is easy to see $\varphi \approx \varphi'$, that is, φ is approximately invariant to the change of illumination conditions and of orientations of the object surfaces. Note that for γ_{ij} we cannot derive this kind of normalized invariant formula.

An important thing to remember here is that in order to make φ useful, the surface reflectance properties associated with two nearby positions $\mathbf{x}_l, \mathbf{x}_m$ to be picked must be sufficiently different from each other. Otherwise, even if an invariant of φ in (25) holds true, as the γ's tend to have the same value for $\mathbf{x}_l, \mathbf{x}_m$, the φ's always return values that are close to 0.5, so that it does not provide any useful information involved in their color properties. Fortunately, as we describe later, when color properties are picked from different sides of brightness boundaries, this situation may often be avoided.

In summary, we have derived two ratios γ, φ which are general values that are distinctive to the objects from which they arise and are not influenced by environmental factors such as lighting. They both correspond to simple ratios, which are easily extracted from imagery. It may be claimed that we could show a way to derive them using only frequency selective sensor assumption for the ambient light of slow spectral variation, without nessesarily using narrow band sensors as are often assumed in deriving conventional invariants. This combination of the assumptions can also be used for those other invariants which are described in the next section [11, 30], allowing them to derive invariants without using narrow band filters. We will actually demonstrate the performance of the proposed invariant computed using the RGB outputs of the usual camera sensors in the experiments.

2.3. Related Photometric Invariants

A related invariant to our photometric invariants was proposed earlier based on an *opponent color model* by Faugeras for image processing applications [8]. The opponent color model was first introduced by Hering [19] to describe the mechanism of human color sensation. He advocated that the three pairs Red–Green, Blue–Yellow, White–Black form the basis of human color perception. A simple mathematical formulation of this [3], which is a linear transformation of R, G, B, was used as a color invariant in [31, 32] for indexing 3D objects: $[\text{R-G, Bl-Y, W-Bk}]^T = L[\text{R, G, B}]^T$, where L is a linear transformation. A similar formalization of an opponent color model was also used for the correspondence process in color stereopsis [5]. However, there are no theoretical explanations of the linear transformation model for the full 3D object surfaces, because, as we noted in the derivation of our invariants, the surface orientation in 3D space with respect to the lighting source and the camera is an unignorable factor (see also [20]) in deriving invariants for a 3D world, and it is never removed by any linear transformation.

Unlike this linear transformation case, Faugeras' form is the logarithm of the ratios between different channel responses for a chromatic model, so is similar to ours, and is the logarithm of the products of three R, G, B responses but with low-pass filtering accounting for lateral inhibition for achromatic responses.

In [4] a unique illuminant-invariant was proposed which, assuming the existence of at least four local distinct color surfaces, uses the volumetric ratio invariant of the parallel pipe generated by the responses of the three receptors. It seems to us, however, that the assumption of four local distinct color surfaces is demanding too much in practice.

Recently, a new photometric invariant was proposed for object recognition by Nayar and Bolle [30]. When its application was limited to only geometrically continuous smooth surfaces, it used as an invariant the ratio between the channel intensities of two adjacent points through a narrow band filter. This invariant ratio is almost identical to our invariant ψ that is actually the ratio of theirs in two channels. The effect of taking the ratio in different color channels is that it removes the geometric factor of the surface reflectance, so the surface smoothness is not required. Another new invariant was proposed by Funt and Finlayson for use in indexing [11]. This invariant is defined either as the first derivative or the Laplacian of the logarithm of a single color channel, which in essence is the color ratios between neighboring points, so it is basically the same as Nayar and Bolles'.

2.4. Experiments

Experiments were conducted to examine the accuracy of the proposed photometric invariants. The goal here is to show that the ratios we described remain invariant over changes in image acquisition and thus are indicative of properties of the object, not the environment.

Figure 1 shows pictures of a man-made convex polyhedron composed of six planar surfaces each with a different surface orientation. The left picture is a front view of the polyhedron, hereafter pose P_A, while in the right picture the object is rotated around the vertical axis (y-axis) by about 30°, hereafter pose P_B. On each side of the boundary of adjacent surfaces, several matte patches with different colors were pasted. Then, we picked corresponding positions manually within each colored patch in the pictures for the poses (P_A, P_B). The selected positions within patches are depicted by crosses in the pictures. We used manually selected points because we want to verify that our invariants perform correctly given a correct point correspondence. Later, we will worry about the performance when there is error or uncertainty in the correspondence.

To test the accuracy of the proposed invariants γ, φ under varying illuminant conditions and surface orientations of the object with respect to the illuminant and the camera, we took three pictures: the first at the pose P_A under the usual lighting conditions (P_A and L_U), the second at the pose P_B under a

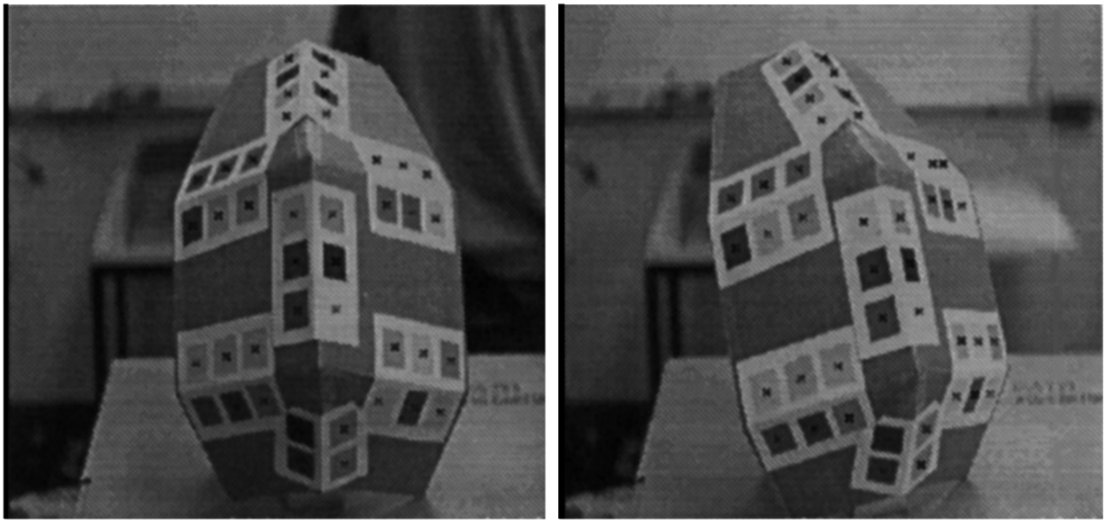

FIG. 1. Tests of invariant on convex polyhedron. The pictures show the convex polyhedron in different poses: left pose P_A, right pose P_B. This object is composed of 6 planar surface patches each with different surface orientation. On each side of the boundary of adjacent surfaces, several matte patches with different colors were pasted. Then, we picked corresponding positions manually within each colored patch in both pictures. The selected positions within patches are depicted by crosses.

greenish light (P_B and L_G), and the third at the pose P_B but under a bluish light (P_B and L_B). To change the source light spectrum, i.e., to get greenish or bluish light, we covered a tungsten halogen lamp with cellophane of colors green and blue.

For φ, the surface positions within planar patches facing over the boundaries of planar surfaces were used as neighboring positions to satisfy the requirement of (locally) constant ambient light. To compute the invariants in practice, we used the ratios G/R, B/R for γ and $\varphi_1 = (G^1/R^1)/(G^1/R^1 + G^2/R^2)$, $\varphi_2 = (B^1/R^1)/(B^1/R^1 + B^2/R^2)$ for φ, where R, G, B are the outputs from the sensor channels, respectively, of Red, Green, Blue, and the indices attached to R, G, B show the sides of the surfaces used for computing φ's with respect to their boundaries.

As described previously, in our theory, when we use the RGB channel outputs to compute invariants, instead of outputs through the exact narrow band filters, the assumption of slow spectral variation of the ambient light together with frequency selective filter of the sensor must be satisfied or an extreme approximation of the sensors by narrow band filters must be taken. The following results show that the values of γ and φ computed using RGB are fairly invariant to the changes of the illumination conditions as well as the surface orientations and demonstrate that some of those requirement we have described may be satisfied.

In Table 1, the correlation coefficients between the sets of values for each invariant measure computed at corresponding positions in different pictures are given, that are measured by the formula

$$\sqrt{\frac{C^2_{\alpha\alpha'}}{C_{\alpha\alpha}C_{\alpha'\alpha'}}}, \tag{26}$$

where the C_{ab}'s ($a, b \in \{\alpha, \alpha'\}$) are the covariances between the sets of the values of the measure α (e.g., γ) before (α) and after (α') the motion of the objects or the changes of the lighting conditions, which is defined by

$$C_{ab} = \sum P(a, b)(a - \bar{a})(b - \bar{b}), \tag{27}$$

where $\bar{x}$ is the average of the measure x, $P(a, b)$ is the probability density function, and the sum is taken over all corresponding values of the measures a, b. A high correlation, that gives a value close to 1, shows that the proposed invariant measures remained unchanged within a consistent scale over the set of positions between the two pictures, while a low correlation, that gives a value close to 0, means that the values of the measures changed in an irregular manner. For comparison, other color properties including raw (R, G, B), a linear-transformation implementation of the opponent color model [3], and the invariant ϖ proposed by Nayar and Bolle [30], which is defined by $\varpi = (\rho_1 - \rho_2)/(\rho_1 + \rho_2)$, where ρ_i is ideally a narrow band filter output at position i, are also included. In this experiment, to implement ϖ, we simply used the RGB output from the same pair of positions used to compute our invariant φ, instead of the narrow band filter output: $\varpi_1 = (R_1 - R_2)/(R_1 + R_2)$, $\varpi_2 = (G_1 - G_2)/(G_1 + G_2)$, $\varpi_3 = (B_1 - B_2)/(B_1 + B_2)$. Since the two faces of the convex polyhedrom used for computing ϖ have some jumps of the orientation on their boundaries though the invariant ϖ is designed to be used for locally smooth surfaces, this is only a test of how it can tolerate against the rough surfaces. From the table, we note between the first and the second pictures the invariants φ and ϖ performed perfectly, though other properties R, G, B, $R-G$, $B-Y$, $\gamma = G/R$, B/R, were also good. This means those properties may have been changed but only within a consistent scale between the different pictures (recall the property of γ being invariant within a scale factor). Looking at the results using the first and third pictures, however, we note that the property ϖ degraded, while φ was almost perfect. To see how far the color properties remained unchanged in addition to the correlative relation, in Fig. 2 the actual distribution of the color properties are plotted, where the horizontal axes are the values for the pose P_A, while the vertical axes are those for the pose P_B. If the color measures remained unchanged between the two pictures before and after the motions of the object and/or the changes of the light conditions, the distributions should present linear shapes, and their slopes should be close to 1. Indeed, the measure φ is certainly found to remain almost unchanged under varying light conditions, while the property $\gamma = G/R$, B/R is found to change. The property ϖ also remains almost unchanged. The biases of the slopes of γ either toward the horizontal or vertical axes indicate that the light spectrum has been changed between the two compared pictures.

TABLE 1
Correlation Coefficients between the Sets of the Values of the Color Properties from Different Pictures of Test-Object

	P_A and L_U – P_B and L_G	P_A and L_U – P_B and L_B
R	0.988368	0.989877
G	0.967951	0.974081
B	0.946251	0.882816
$R - G$	0.985398	0.985687
$B - Y$	0.935039	0.908867
$\varpi_1 = (R_1 - R_2)/(R_1 + R_2)$	0.991229	0.958538
$\varpi_2 = (G_1 - G_2)/(G_1 + G_2)$	0.988581	0.961559
$\varpi_3 = (B_1 - B_2)/(B_1 + B_2)$	0.979000	0.945734
$\gamma = G/R$	0.978163	0.988289
$\gamma = B/R$	0.962186	0.907126
$\varphi_1 = (G^1/R^1)/(G^1/R^1 + G^2/R^2)$	0.997766	0.997532
$\varphi_2 = (B^1/R^1)/(B^1/R^1 + B^2/R^2)$	0.991843	0.988893

Note. The correlation coefficients between the sets of values of the proposed invariants from pictures taken under different light conditions and at the different poses of the object are given to show how much they remain unchanged within a consistent scale. For comparison, other color properties including raw (R, G, B), a linear-transformation implementation of the opponent color model [3], and the invariant ϖ proposed by Nayar and Bolle [30], which was computed using RGB output as $\varpi_1 = (R_1 - R_2)/(R_1 + R_2)$, $\varpi_2 = (G_1 - G_2)/(G_1 + G_2)$, $\varpi_3 = (B_1 - B_2)/(B_1 + B_2)$, are also included.

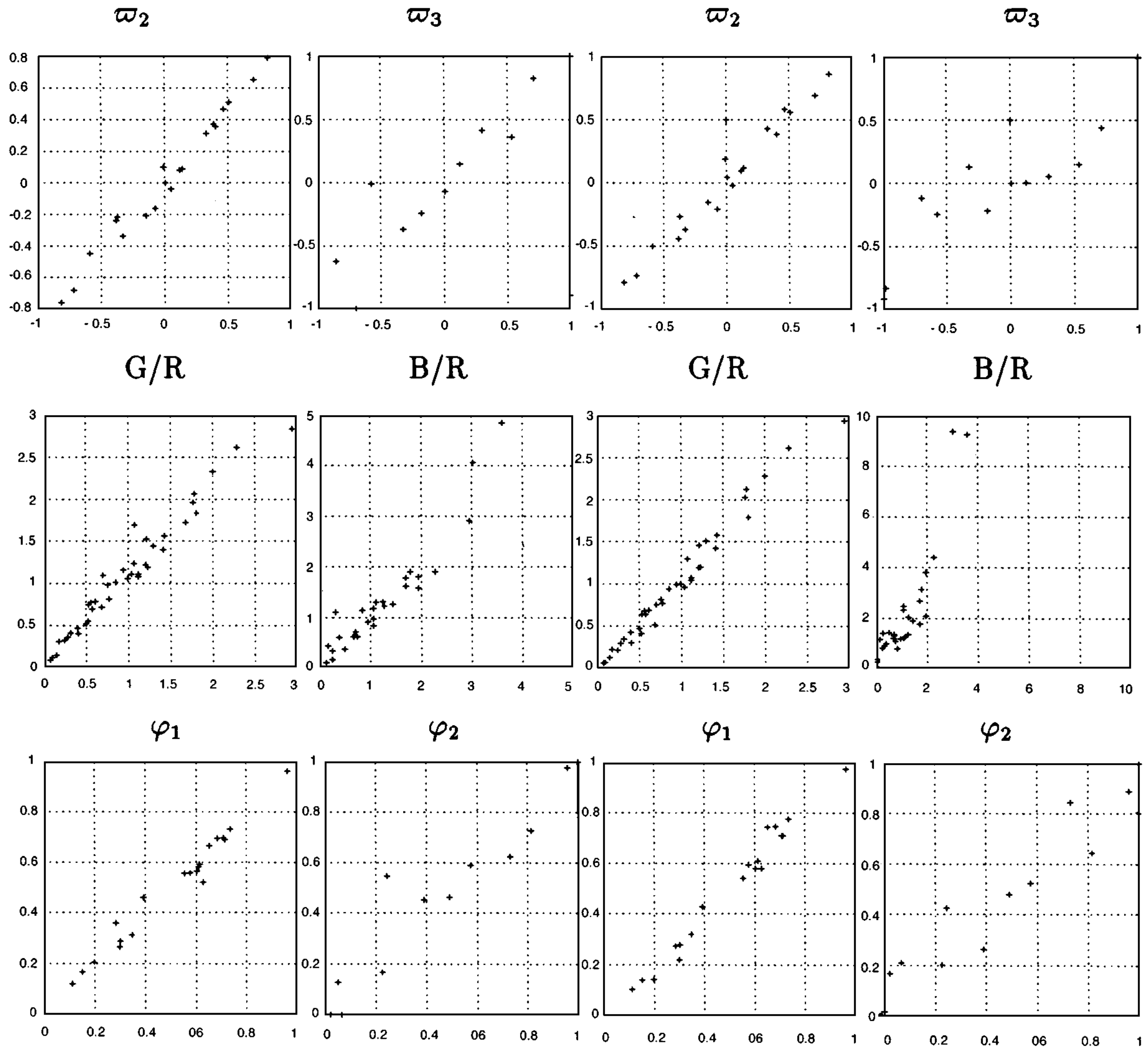

FIG. 2. Distributions of invariants on convex polyhedron. The left two columns are from pictures taken under P_A and L_U (horizontal axis) and P_B and L_G (vertical axis), and the right two columns are from pictures taken under P_A and L_U (horizontal axis) and P_B and L_B (vertical axis). The rows in each of the two columns are, respectively, top left and right, $\varpi_2 = (G_1 - G_2)/(G_1 + G_2)$ and $\varpi_3 = (B_1 - B_2)/(B_1 + B_2)$; middle left and right, G/R and B/R; bottom left and right, $\varphi_1 = (G^1/R^1)/(G^1/R^1 + G^2/R^2)$ and $\varphi_2 = (B^1/R^1)/(B^1/R^1 + B^2/R^2)$.

Figure 3 shows the performance of γ constancy against the change of the object pose, under the same lighting conditions. In other words, unlike in the last experiments, this time the ambient light has not been changed for both of the two pictures, and only the object pose has been changed. For comparison, the performance of $B - Y$ (linear-trans implementation for blue vs yellow, the second figure from the left) as well as raw B (blue, the first one) are also shown. Note that what should be observed here is how the slopes of the distributions are close to 1. Except for the two samples in the upper area in the figure (the fourth picture), $\gamma = B/R$ is found to be almost unchanged between the two pictures. The two exceptional samples were from patches with almost saturated blue channel in the picture at pose P_B. The performance of $\gamma = G/R$ (the third figure) is almost perfect. On the other hand, $B - Y$ and B are perturbed around the slope of 1, which is probably caused by the perturbed orientations of the patches. This suggests that γ may be used for object recognition without applying any normalization process, so that extracting object regions might not be a prerequisite, as long as the lighting conditions are not changed.

Similarly, in Table 2 the results of the similar tests as above but on a natural object, a doll which is shown in Fig. 4, are given, for which both the ambient light and the object pose were changed. We refer to the pose of the doll similarly to the above tests on the test-object: left pose P_A, right pose P_B. The first picture was taken under a usual lighting conditions from the oblique angle (P_A and L_U), the second and third were taken, respectively, under a greenish and a bluish light from the front angle (P_B and L_G, P_B and L_B). Corresponding positions were picked manually as done in the previous tests. As the surface colors varied

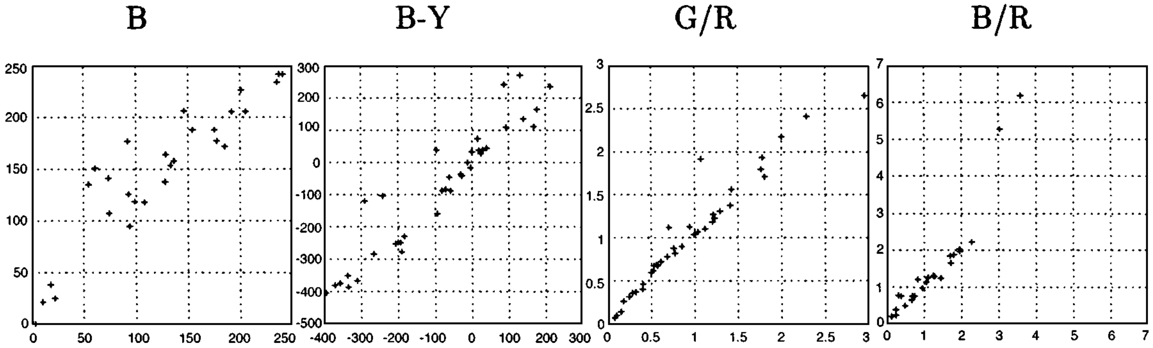

FIG. 3. Tests of γ at different poses of object but under the same illuminant conditions. The first from the left, distribution of Blue; the second, $B - Y$ (Blue vs Yellow); the third, G/R; the fourth, B/R. The horizontal axis is for the pose P_A and the vertical axis is for the pose P_B. Except for the two samples in the upper right area of the distribution, $\gamma = B/R$ is found to be almost unchanged in both of the pictures because the slope is almost 1, while $B - Y$ and B are perturbed around the slope of 1. Those two exceptional samples were from patches with almost saturated blue channel in the picture at pose P_B. The distribution of $\gamma = G/R$ is almost perfect. This gives the evidence that γ may be used for object recognition without applying any normalization process, so that extracting object regions might not be a prerequisite as long as the lighting conditions are not changed.

smoothly, we cannot expect that we could pick up corresponding points accurately. Thus, unwanted errors could be introduced in this operation. This time for φ, two positions which are closest to each other among the selected points are used. The property ϖ is not included in this test, since it was very hard to find a pair of positions where surface normals do not have a big change but have a change of colors (*Note*. The values of invariant ϖ such as φ will be close to 0 in places where colors do not change). In these tests, R, G, B performed poorly. The linear model $R-G$, $B-Y$, and $\gamma = G/R$, B/R performed well again, though γ was better. The measure φ is quite stable again. Unlike the results on the test-object, however, since the surface of the doll, especially in the body parts, had similar surface colors in nearby positions, the distribution of φ (i.e., $\varphi_1 = (G^1/R^1)/(G^1/R^1 + G^2/R^2)$, $\varphi_2 = (B^1/R^1)/(B^1/R^1 + B^2/R^2)$) did not spread very well, thus having a weak selectivity photometrically, as seen in Fig. 5. Therefore, when picking two nearby positions for φ for object recognition, it is important that they have different spectral reflectance. This will also be true in using the invariant proposed by Nayar and Bolle. We should note that the property γ does not require this condition, though in terms of invariance it is not superior to ϖ or φ. For comparison, the values of γ are also plotted in Fig. 5.

TABLE 2
Correlation Coefficients between the Sets of the Values of the Color Properties from Different Pictures of the Doll

	P_A and L_U- P_B and L_G	P_A and L_U- P_B and L_B
R	0.764343	0.819267
G	0.588161	0.881416
B	0.936572	0.843604
$R - G$	0.764240	0.939152
$B - Y$	0.948642	0.877519
G/R	0.779377	0.944164
B/R	0.962186	0.895180
$\varphi_1 = (G^1/R^1)/(G^1/R^1 + G^2/R^2)$	0.996245	0.998781
$\varphi_2 = (B^1/R^1)/(B^1/R^1 + B^2/R^2)$	0.988840	0.983675

Note. The results on a natural object, a doll, are given. The first picture was taken under usual lighting conditions from the oblique angle (P_A and L_U), the second and third were taken, respectively, under a greenish and a bluish light from the front angle (P_B and L_G, P_B and L_B). This time for φ (i.e., $\varphi_1 = (G^1/R^1)/(G^1/R^1 + G^2/R^2)$, $\varphi_2 = (B^1/R^1)/(B^1/R^1 + B^2/R^2)$) two positions which are closest to each other are used. In this test, R, G, B were very unstable. The linear model $R - G$, $B - Y$, $\gamma = G/R$, B/R did perform well again, though γ was better. The measure φ is quite stable again.

2.5. *Sensing Limitations*

As we note in the examination above, the invariant properties γ, φ are sometimes perturbed around the ideal values which support our theories. This is caused mainly by the limited dynamic range of the sensors of the camera. These effects include *color clipping* and *blooming* as argued carefully in [26]. When the incident light is too strong and exceeds the dynamic range of the sensor, the sensor cannot respond to that much input and thus clips the upper level beyond the range. This means the sensor no longer correctly reflects the intensity of the light. Note that this is very serious for our invariants, because both γ and φ are ratio invariant, and a basis of their theory is, whether locally or globally, the consistency of the amount of light falling onto the concerning positions on the object surfaces. Here, our natural and important assumption is that this consistency is correctly reflected in the responses of the sensors. Therefore, if the sensor response does not meet this assumption, our theory no longer holds. The same arguments also hold for the blooming effect. When the incoming light is too strong to be received by the sensor element of the CCD camera, the overloaded charge will travel to the nearby pixels, thus crippling the responses of such pixels.

2.6. *Summary*

In this section, we have developed two different photometric invariants and demonstrated their accuracy on a set of natural

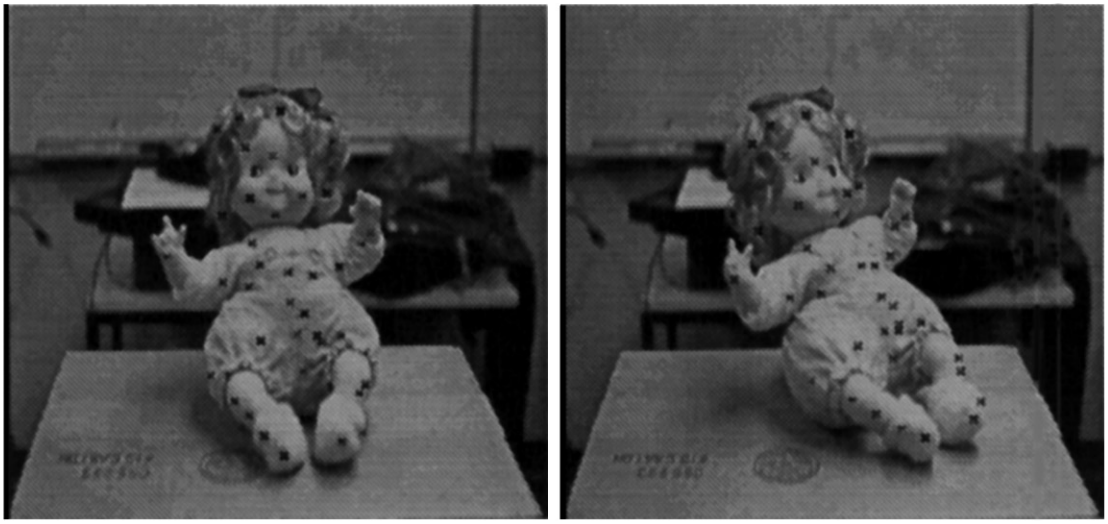

FIG. 4. Tests of invariant on natural pictures. The pictures show a doll at different poses: left pose A, right pose B. We picked corresponding positions in both views. The selected positions are depicted by crosses.

images. In the next section, we turn to the question of how to utilize these photometric constraints in conjunction with geometric constraints, for use in recognition.

3. COMBINING PHOTOMETRIC AND GEOMETRIC CONSTRAINTS FOR 3D OBJECT RECOGNITION

In this section, we describe how we can exploit the photometric invariant developed in the preceding section for recognizing 3D objects. The basic idea is to combine it with the centroid alignment approach we recently proposed in [28].

3.1. *Centroid Invariant of Geometric Feature Groups*

We argued in [28] that when an object undergoes a linear transformation caused by its motion, the centroid of a group of 3D surface points is transformed by the same linear transformation. Thus, it was shown that under an orthographic projection model,

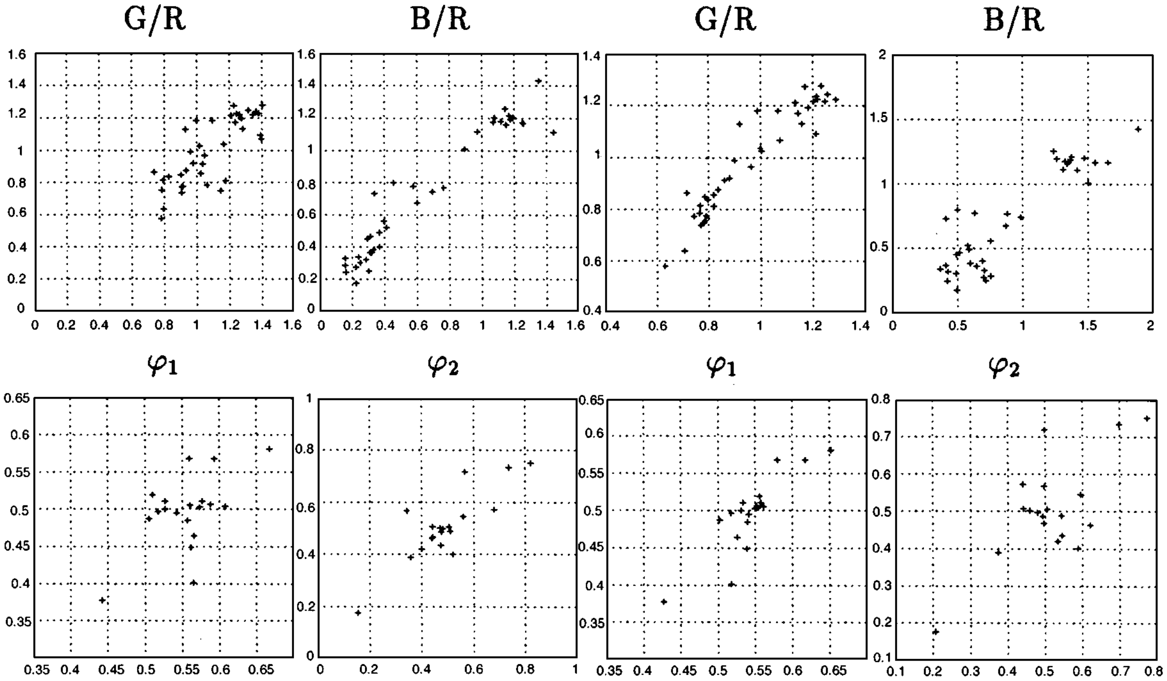

FIG. 5. The distributions of invariant measures on Doll pictures. The left two columns are from pictures taken under P_A and L_U (horizontal axis) and P_B and L_G (vertical axis), and the right two columns are from pictures taken under P_A and L_U (horizontal axis) and P_B and L_B (vertical axis). The rows in each two columns are, respectively, top left and right, H and S; middle left and right, G/R and B/R; bottom left and right, $\varphi_1 = (G^1/R^1)/(G^1/R^1 + G^2/R^2)$ and $\varphi_2 = (B^1/R^1)/(B^1/R^1 + B^2/R^2)$.

centroids of 2D image geometric features always correspond over different views regardless of the 3D pose of the 3D object in space. This is true for any object surfaces (without self-occlusion). Note that this property is very useful, because if we have some way to obtain corresponding feature groups over different views, we can replace simple local features used for defining alignment in conventional methods by those groups, thereby reducing computational cost. We demonstrated the effectiveness of this approach to object recognition on natural as well as simulation data [28].

3.2. Grouping by Photometric and Geometric Constraints

While the idea of using centroids of similar groups dramatically reduces the complexity of alignment-style recognition schemes, we need to find reliable ways of extracting such groups. One set of methods was described in [28]. Here we focus on amplifying this approach by utilizing our photometric invariants to obtain corresponding groups of 2D geometric features.

In [28], to obtain corresponding geometric feature groups, a clustering operation, in which the clustering criterion was rotationally invariant, was applied in the coordinates which had been normalized up to a rotation prior to a clustering. This time, we again use a clustering technique to obtain corresponding geometric feature groups in different views. Our intention is to yield corresponding cluster configurations using a criterion incorporating spatial proximity constraints of geometric features and the invariance of their associated photometric invariants. Therefore, we assume that surface colors (surface spectral reflectance) vary mostly from place to place. In other words, within some local areas surface colors are almost consistent. Note that this assumption should be justified for most object surfaces, because otherwise we must always be seeing diffused colors over the surface and thus always having difficulty in trying to distinguish surfaces.

When we are provided with geometric and photometric feature sets from almost corresponding model and data points, we normalize those feature distributions using the decor-relating transformation presented in [28]. This operation performs invariance between the model and data features and brings stability in the results of clustering. As we will see in the experiments, however, the requirement of correspondences of feature sets proves not so strict. The distribution of geometric features are normalized using the same transformation used in [28]. This operation has been confirmed, both mathematically and empirically, to generate a unique distribution up to a rotation, for feature sets from a planar surface on the object, regardless of the surface orientation in 3D space. We note that even 3D object surfaces often tend to become planar in their visible surfaces, thus justifying the use of our transformation for 3D object surface. We also decorrelate the joint distributions of the different photometric invariants by a transformation defined similarly to that for geometric features. The reason of this is, as we will see in the experiments, that the distribution of different photometric invariants may be jointly highly correlative (for instance, see the second row of Fig. 6), which, from the information theoretic point of view, means that if they are used directly they are less informative than they could be, as a result of reducing the dimension of the distributions.

Thus, we define the extended feature vector f for clustering as

$$f = \left[f_g^T, s f_p^T \right]^T, \tag{28}$$

where f_g is the 2D geometric feature composed of spatial coordinates $f_g = (x, y)^T$ of a feature point in the xy image plane, f_p is the vector of photometric invariant properties we proposed in the preceding sections, and s is a balancing parameter. Then, the distribution of this feature f is transformed (normalized) by a decorrelating transformation $D = diag\{T_g, T_p\}$ where T_g, T_p are the matrices defined as follows and, respectively, decorrelate the distributions of f_g and f_p,

$$T = \Lambda^{-1/2} \Phi^T, \tag{29}$$

where Φ and Λ are eigenvector and eigenvalue matrices of the covariance matrix of f_g or f_p, $[\cdot]^{-\frac{1}{2}}$ denotes the square root matrix of a positive definite matrix, and $[\cdot]^T$ is the matrix transpose. It should be noted that the invariance of the distribution of f_p, which is composed of γ or φ, up to a rotation is never damaged through this operation. Moreover, by this normalizing operation all the physical dimensions included in the geometric and photometric features, such as image resolution and power of the light energy, are removed. Therefore, in constructing a high dimensional feature vector for clustering, we can provide a consistency between the physical dimensions of the components of features. So, in theory, even when the physical properties of the image change, the balancing parameter s may not have to be readjusted.

3.3. Implementation

We employ the *Kmean* clustering algorithm, in which the criterion is rotationally invariant, to obtain corresponding feature groups in the feature set from different views. Note that what we ultimately need here is simply the configuration of geometric features, that is f_g, in the clustering results, and the photometric invariant is used only as a cue in performing clustering.

After the clustering, an alignment process starts by using centroids of clusters so derived to recover the transformation which generated a novel view, the image data, from the model. It is known that only 3-point correspondences suffice to recover the transformation either by using linear combination of the models [36] or a full 3D object model [22]. Therefore, we examine every possible combination of triples of cluster centroids of models and data that are generated by clustering and select the best-fit transformation to generate the data from the model in terms of their match. In our testing, which we will see later, this number of clusters could be suppressed to less than ten. Further, we should note that we only need to consider the combination of model and data cluster centroids which have compatible values of γ or φ. This means that adding photometric properties contributes not only to the clustering but also to the selectivity of the features (cluster centroids). Therefore, considering the computational complexity of conventional alignment approach to recognition, this should bring a noticeable computational improvement.

4. EMPIRICAL RESULTS

In this section, we show experimental results of our algorithm for identifying corresponding positions in different views. Tests were conducted on natural pictures including 3D unoccluded/occluded objects to be recognized, which are taken under varying light conditions and poses of objects.

4.1. Preliminaries

Geometric features used for our algorithm can be extracted as follows:

1. Use an edge detector [6] after preliminary smoothing to obtain edge points from the original gray level images.
2. Link individual edge points to form edge curve contours.
3. Using local curvatures along the contours, identify features as corners and inflection points, respectively, by detecting high curvature points and zero crossings based on the method described in [22]. Before actually detecting such features, we smooth the curvatures along the curves [2].

In obtaining color attributes from corresponding positions we should note that the positions of the geometric features thus extracted in different views do not always correspond exactly in discrete image coordinate space. This is not only due to quantization error, but also because edges detected to derive feature points can shift to the other side of the surface beyond the boundary under an object rotation within an image plane. Note that this is serious because the occurrences of gray level edges often tend to coincide with color edges [5]. So, we cannot simply use the color attributes of the geometrical feature points derived from gray level edges. To solve this problem, we picked color values from two positions over the gray level boundary, which are away from the geometric feature positions in the opposite directions along the local normals of the contours. Then, we used two color values from both of the two positions. As we do not know which side of an edge in one picture corresponds to which side in another, the distance metric between the photometric invariant vectors associated with two different feature positions should be independent of the correspondences of those sides of the surfaces. Thus, the actual measure used for photometric invariant vector f_p and the distance metric for two of those (that are used for computing the values for clustering criterion) are designed such that they support the symmetry on the sides of the surfaces over the boundaries $f_p = [f_p^{1^T}, f_p^{2^T}]^T$, where

$$f_p^i = (G^i/R^i, B^i/R^i)$$

for γ,

$$f_p^i = \left(\frac{(G^i/R^i)}{(G^i/R^i + G^j/R^j)}, \frac{(G^j/R^j)}{(G^i/R^i + G^j/R^j)}, \frac{(B^i/R^i)}{(B^i/R^i + B^j/R^j)}, \frac{(B^j/R^j)}{(B^i/R^i + B^j/R^j)} \right)$$

for φ, indices $(i, j) \in \{(1, 2), (2, 1)\}$ show the sides of the surfaces with respect to their boundaries, and the distance metric between f_{p1} and f_{p2} for geometric feature positions 1, 2 is

$$|f_{p1} - f_{p2}|^2 = \min\left\{ \left\| f_{p1}^1 - f_{p2}^1 \right\|^2 + \left\| f_{p1}^2 - f_{p2}^2 \right\|^2, \left\| f_{p1}^1 - f_{p2}^2 \right\|^2 + \left\| f_{p1}^2 - f_{p2}^1 \right\|^2 \right\}, \quad (30)$$

where $\|\cdot\|$ denotes Euclidean distance. This apparently supports the symmetry on the sides of the surfaces over the boundaries of the gray level and is invariant to the rotation of the objects within an image plane. The following experiments test our algorithm with both of the proposed invariants γ, φ. For each feature position, the associated invariant φ was computed using color attributes of those two points mentioned above, that is, two points a little away from the geometrical feature points along the contour normals in the opposite directions. As described earlier, since gray level edges tend to coincide with color edges, the color values collected from those two positions facing across the gray level edges are usually quite different, thereby producing φ distributions that spread over the feature space.

4.2. Tests on Images without Occlusion

The first experiment tests our algorithm on feature sets from almost corresponding model and data regions. The region extraction was done manually though we expect that this could be done automatically using several cues such as motion, color, and texture (see, e.g., [31–34]). Then, through the normalization process of the distribution of γ as well as geometric features as described, γ becomes a complete invariant. Note that, however, in using φ these processes, i.e., region extraction and normalization, are not necessarily required, as long as the background in the picture happened to have different colors than the object. This is because φ is a complete invariant, unlike γ which needs normalization to remove scale factors. This is also true for γ when the ambient light has not been changed before and after the motion of the objects.

It would not be hard to see that identifying corresponding positions perfectly is not an easy task, because in doing that we must fight against two different kinds of instabilities: one in extracting geometric features, the most serious of which is missing features and the other substantially contained in photometric properties of the image, such as the ones described in the arguments for sensing limitations. Remember that, however, for our ultimate objective, which is recognizing objects using the identified positions, only three correspondences are sufficient under

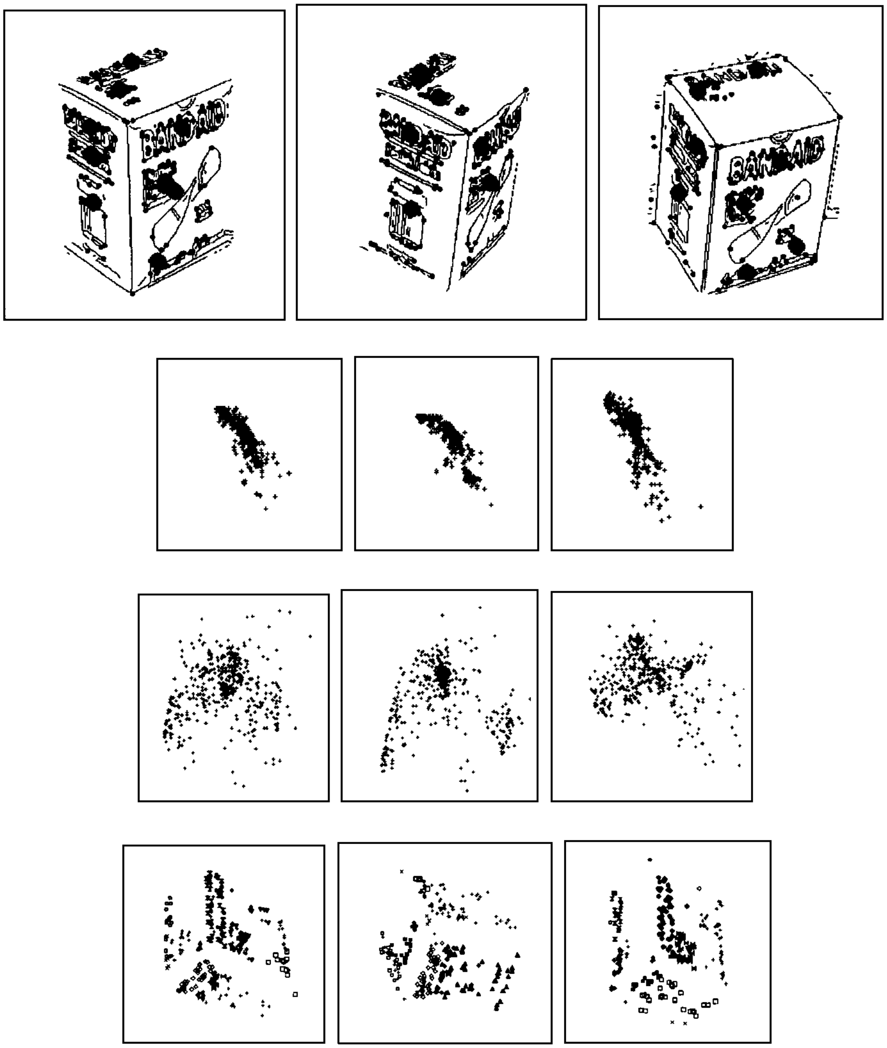

FIG. 6. Tests with γ on Bandage-box pictures. Edge maps are shown with extracted geometric features superimposed on them in the first row. The first picture (from the left) was taken under usual light conditions. The second and third pictures were taken, respectively, under a greenish and a bluish light at a different pose. Identified corresponding positions using our algorithm are also superimposed by large closed circles. The figures in the second and third rows show the respective original and normalized distributions of γ. The intermediate results of clustering are shown in the fourth row figures in their normalized coordinate of the geometric features.

orthographic projection model [36] or weak perspective projection model [22]. Therefore, what must be observed in the following results is whether our algorithm could identify at least this minimum number of correspondences or not. First, the results of using γ as the photometric invariant are shown.

Using γ for photometric invariant. Figure 6 shows the results of obtaining feature group centroids on Bandage-box pictures, which includes characters of some different colors on a white base on the surface. All the pictures were taken to involve the same three surfaces of the box, which are to be used for the recognition. The figures in the first row from the top show the edge maps, with extracted geometric features superimposed on them with small closed circles. The first from the left (hereafter, first) picture was taken under usual light conditions. The second from the left (hereafter, second) and third from the left (hereafter, third) pictures were taken, respectively, under a greenish and a bluish light at a different pose from the first one. Throughout the rest of the paper, we refer to the figures by the order they are presented from the left as above. The lighting conditions were changed by the same way as in the experiments presented in Section 2.4. The figures in the second and the third rows show the respective original and normalized distributions of γ. In the second row, the horizontal axes of the figures are for G/R while the vertical axes are for B/R. These figures show how the invariant property γ remained unchanged between the different pictures. When it performs well, the original distributions of γ should show similar shape over different views except for some scale change along the axes. Then, those scale distortions (e.g., dilation) should be corrected by the decorrelating process of the distribution, thus ideally showing the same distribution within rotations. Note that even if the shape of the γ distributions are distorted in addition to the dilation, we cannot conclude that the proposed invariants performed poorly. This is because unstable results of the geometrical feature extraction will also distort the shape of the distribution of the photometric properties.

This intermediate results of clustering are shown in the fourth row in their normalized coordinate of the geometric features. In the figures of the first row, identified corresponding positions using our algorithm are superimposed by large closed circles. Therein, the accuracy of our algorithm is found to be fairly good. Apparently perturbations of identified positions were caused partly by the unstable results of feature extraction, e.g., missing features, rather than by clustering errors or incompleteness of the proposed photometric invariant. Note again that what is required is that at least three of these features correspond between pairs of views. As seen in Fig. 6, there are five common features between the first two views, seven common features between the first and third view, and four common features between the second and third views. Thus an alignment method will easily correctly identify the pose base on the pairings of common triples of features and will reject poses based on other pairings. More importantly, the number of pairings to be tested has been drastically reduced, without losing the correct answer.

In Fig. 7, results on Spaghetti-box pictures taken in the same way as the Bandage-box pictures are given. The surfaces of this box include some textures including large/small characters. This is a lightly cluttered texture compared with the Bandage-box surface. The first row shows the edges with extracted geometric features superimposed on them. The first picture was taken under usual light conditions. The second and the third pictures were taken, respectively, under a greenish and a bluish light at different poses. The second and the third row figures show the respective original and normalized distribution of γ. The algorithm could perform identification of the corresponding positions fairly accurately as we see in the top figures. Similar to the previous case, between any pair of views there are either five, six, or seven common features, so that an alignment method will correctly find the true pose.

Similarly, in Fig. 8 the results on Doll (the same one used in Section 2.4) pictures are presented. Unlike the last two examples, the surface of this doll does not have man-made texture such as characters, but only has color/brightness changes partly due to the change of materials and partly due to depth variations. The surface is mostly smooth except for some parts including hair, face, and finger parts. The pictures in the first row show the edges with extracted geometric features superimposed on them. The first and second pictures were taken under usual light conditions, but at different poses of the doll. The third picture was taken under a moderate greenish light plus usual room light. For the fourth picture, we used an extremely strong tungsten halogen lamp with a bluish cellophane covering it. The second and the third row figures show the respective original and normalized distributions of γ. Comparing the shapes of original and normalized distributions of γ for the first and the second pictures, we can confirm that when the light conditions have not been changed the distributions of γ are not affected by the change of pose of the object. The algorithm could perform identification of the corresponding positions fairly accurately as we see in the pictures.

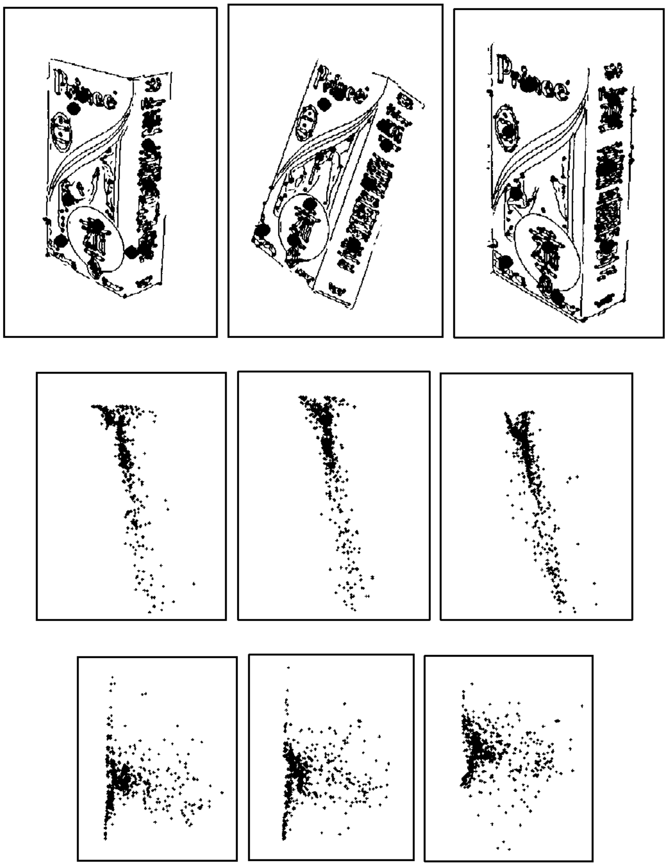

FIG. 7. Tests with γ on Spaghetti-box pictures. The surface of this box include some colored textures including large/small characters. The pictures in the first row show the edges with extracted geometric features superimposed on it. The first picture (from the left) was taken under usual light conditions. The second and third pictures were taken, respectively, under a greenish and a bluish light at a different pose from the first one. The second and third rows show the respective orignal and normalized distributions of γ. The identified positions are depicted by large closed circles in the figures of the first row. The algorithm could perform identification of the corresponding positions fairly accurately as we see in the upper figures.

Using φ. The results of using φ as a photometric invariant on the same pictures used for γ are shown. Figure 9 presents the results on Bandage-box pictures. The first row shows the edge maps with extracted geometric features superimposed on them with closed circles. In the second row, respective distributions of φ are shown. The horizontal axes are for $(G^i/R^i)/(G^i/R^i + G^j/R^j)$, while the vertical axes are for $(B^i/R^i)/(B^i/R^i + B^j/R^j)$, where $(i, j) \in \{(1, 2), (2, 1)\}$. As described already, since we do not know the correspondences of the sides of the surface over the edges (contours), we included properties from both sides of the edges. Consequently, we had 2-fold symmetric distributions of φ around its centroid as noted in the second row figures (see Eq. (25)). When φ performs well as an invariant, this

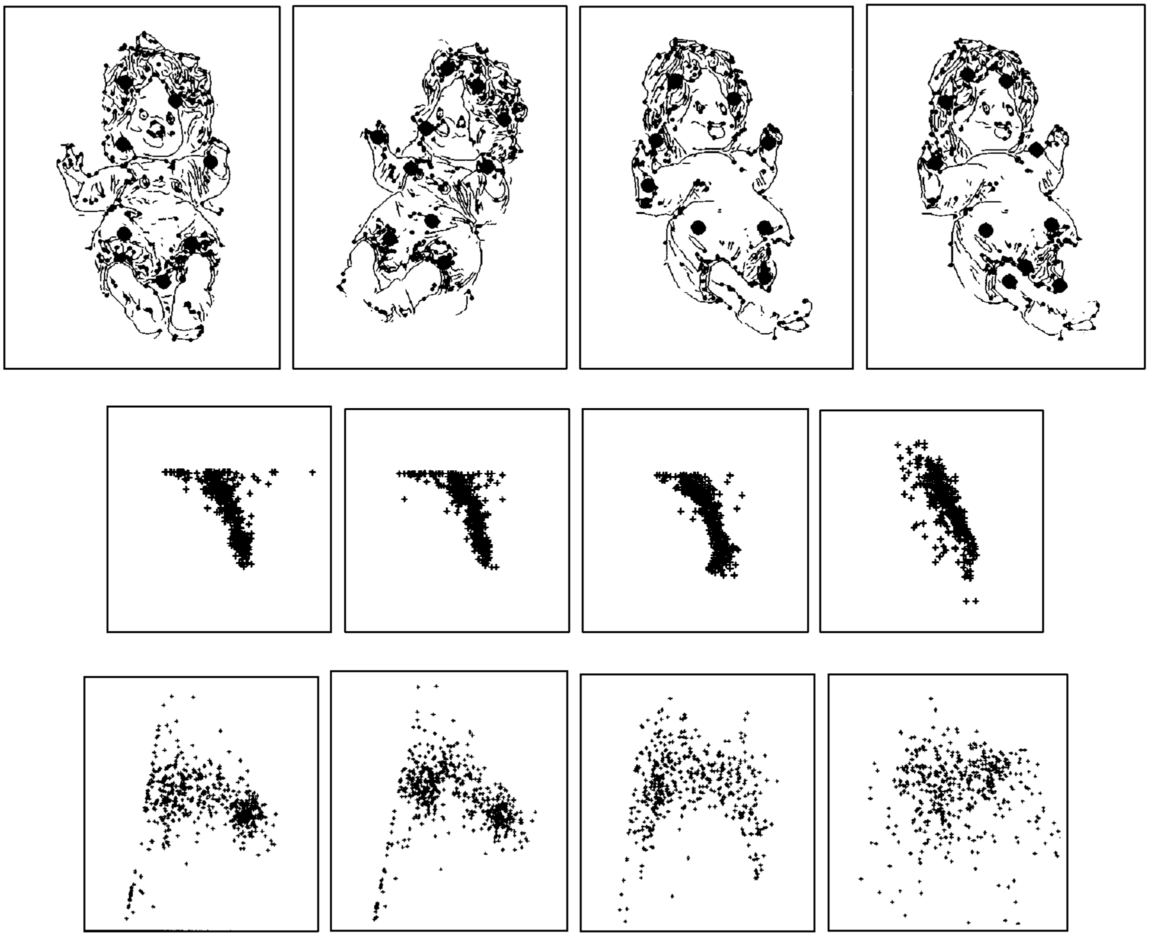

FIG. 8. Tests with γ on Doll pictures. The surface of this doll does not have man-made texture like characters, but only has color/brightness variation partly due to the changes of materials and partly due to depth variations. The surface is mostly smooth except for some parts including hairs, face, and finger parts. The first row shows the edge maps with the extracted geometrical features superimposed on it with small closed circles. The first and second pictures (from the left) were taken under usual light conditions, but at different poses of the doll. The third picture was taken under a moderate greenish light plus usual room light. For the fourth picture, we used an extremely strong tungsten halogen lamp with bluish cellophane covering it. The second and the third rows show the respective original and normalized distributions of γ. The identified positions are depicted by large closed circles in the figures of the first row. The algorithm could perform identification of the corresponding positions fairly accurately as we see in the figures.

distribution should remain unchanged over different pictures. Thus, the second row figures demonstrate a fairly good performance for this picture. The third row shows their decorrelated distributions. The intermediate results of clustering are given in the fourth row figures in their normalized coordinate of the geometric features. In the figures of the first row, identified corresponding positions using our algorithm are also superimposed by large closed circles. Thus, the accuracy of our algorithm is found to be fairly good.

In Fig. 10 the results with φ on Spaghetti-box are given. The first row shows the extracted geometric features. The second and the third rows show the original and the decorrelated distributions of φ. The performance of φ is almost perfect. As we see in the pictures, the algorithm with φ could perform identification of the corresponding positions very well.

Figure 11 presents the results on Doll pictures. In the first row, the edge maps with extracted geometric features superimposed on them are shown. The second and the third rows show the the respective original and the decorrelated distributions of φ. Since for the fourth picture we used extremely intensive blue light, the blue channel of many pixels were saturated. As a consequence, the distribution of φ was shrunk in the vertical direction as noted in the fourth picture of the second row. For these doll pictures, generally, the results of identifying corresponding positions with φ were not as good as those with γ, though they were not very bad. This is probably because as the surface colors of the doll vary quite smoothly in most parts, the distribution of φ did not spread well, so that it did not work so well to separate clusters in terms of colors.

4.3. *Tests on Images with Occlusion*

In the following experiment we examine the tolerance of the algorithm against occlusions of local data parts. In this test, occlusion was produced by manually removing nearly 20 to 35%

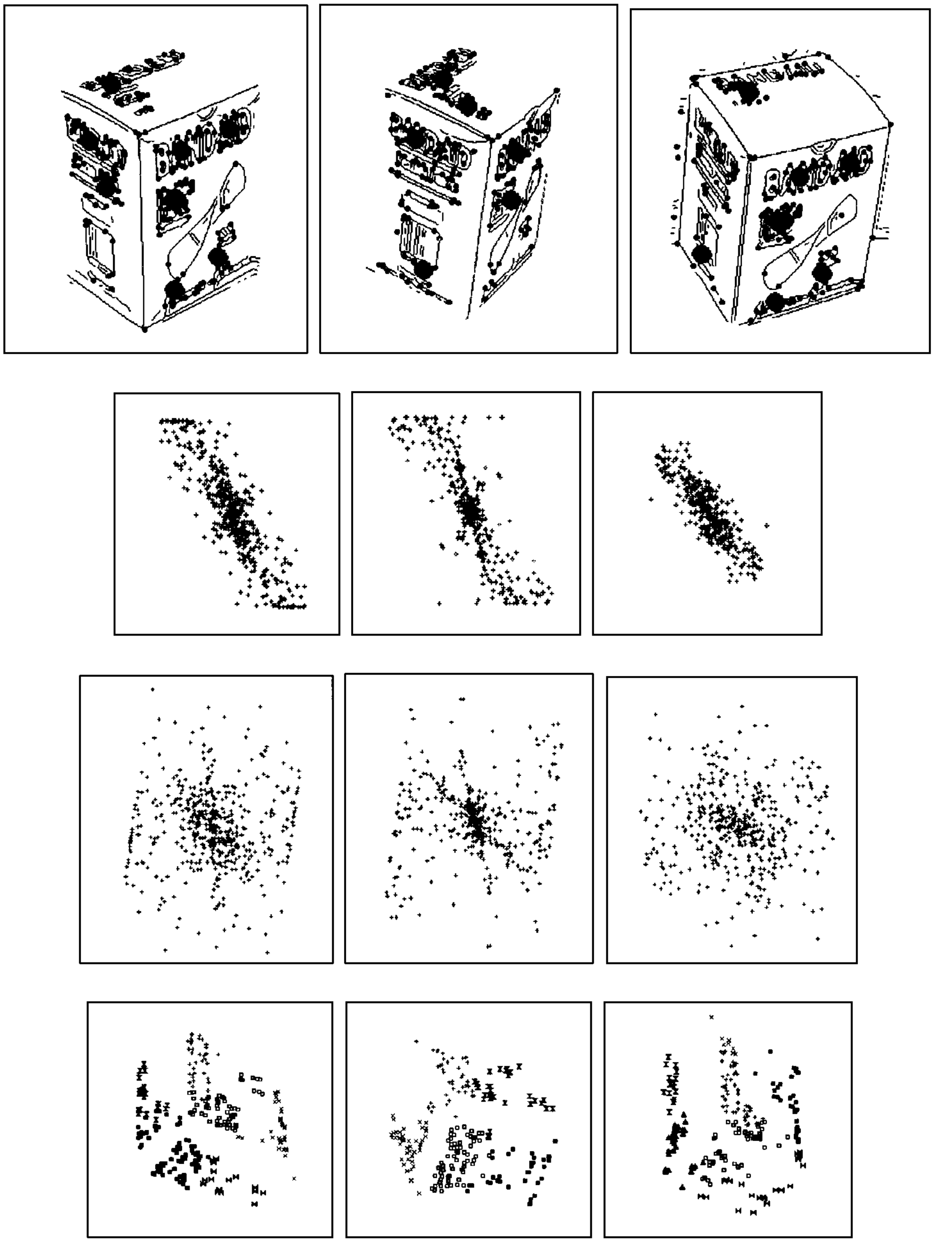

FIG. 9. Tests with φ on Bandage-box pictures. The pictures in the top show the edge maps with extracted geometric features superimposed on them. The first picture (from the left) was taken under usual light conditions. The second and third pictures were taken, respectively, under a greenish and a bluish light at a different pose from the first one. The second and the third row figures show the respective original and decorrelated distributions of φ. The fourth row shows the intermediate results of the clustering. The identified positions are depicted by large closed circles in the figures of the first row. The algorithm could perform identification of the corresponding positions fairly accurately as we see in the upper figures.

of the whole object region. After extracing the photometric and geometric features out of remaining regions, those features were decorrelated in the same way as in the tests on almost complete data sets. Theoretically, in this case the invariance of the features between the different views no longer hold if we decorrelate the distributions, due to the collapse of the correspondences. However, we will see that this does not have a significant effect on the results of clustering, as long as the percentage of the dropped region was not so large, e.g., up to around 35%, and the object surface has enough variety of colors.

Using γ for photometric invariant. Figure 12 shows the results on the Bandage-box pictures which are the same as those used for the first experiments, except that input image data have drops of the local regions. In the top, for our convenience, we again include the result of a picture without any dropping of

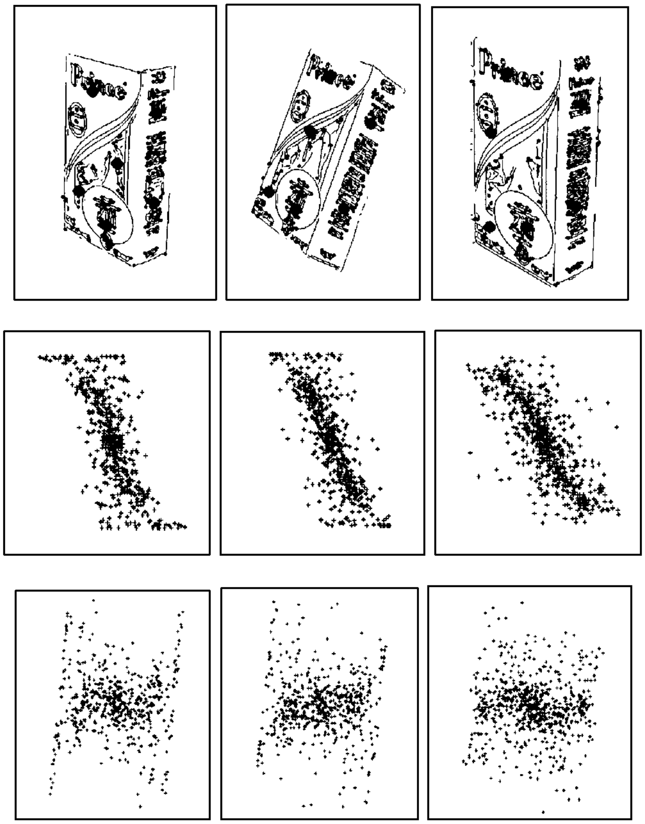

FIG. 10. Tests with φ on Spaghetti-box pictures. The surface of this box includes some colored textures including large/small characters. Top pictures show the edges with extracted geometric features superimposed on it. The first picture was taken under usual light conditions. The second and third pictures were taken, respectively, under a greenish and a bluish light and at a different pose. The second and the third figures show the respective original and decorrelated distributions of φ. The identified positions are depicted by large closed circles in the figures of the upper row. The algorithm could perform identification of the corresponding positions fairly accurately as we see in the upper figures.

regions. In the bottom left picture, the lower right corner of the object region was dropped, which was nearly 35% of the whole object area, which included about 20% of the feature points. The picture in the bottom right has a drop of the upper left corner which was almost 20% of the whole region and included 23% feature points. Comparing with the results on the almost complete data sets presented in Fig. 6, we note the accuracy of detecting the salient feature positions, i.e., cluster centroids, in the remaining feature sets is almost the same, providing corresponding salient features sufficient to subsequent matching process. The reason for this stability in spite of the break of the invariance of the features is explained as follows: As argued in our previous work [28], since the clustering algorithm *Kmean* we employed tries to detect local parts in which features are concentrated. As long as those concentrations are not damaged by occlusions, it can still be detected no matter how other remote local areas are devastated. In terms of this, looking at the surface of the object Bandage-box, we note it has some local regions having dense feature distributions with consistent colors coming from the man-made textures. Thus, those local parts are stably detected even in the presence of occlusions of other remote areas.

Similarly, Fig. 13 shows the results on occluded Doll pictures. The top figure is the result on the complete object view, while in the bottom figures results are given in which almost 20% of the whole object regions are occluded: in the first view upper right part is dropped losing about 32% of the whole features, in the second view the lower left is dropped losing 12% features, and in the third the upper left is removed which included 27% features. Although the object Doll's surface has almost regular feature patterns rather than locally dense ones, the results of extracting the salient points correspond well over the different views.

From those results, we conclude that by using the invariant γ we can provide an algorithm that can still tolerate occlusions of the object surfaces despite the fact that we can no longer support the complete invariance of the features.

Using φ. Similar tests are conducted using φ as invariant. We use the same set of pictures including the occluded objects as used in the tests for γ. In Fig. 14, we note the results of detecting the salient features are fairly good, providing still enough commonality: four common features between the first (top figure) and the second (bottom left figure), three between second and third (bottom right), and six between the first and the third.

In Fig. 15, however, the accuracy of the correspondences of the extracted features degrades. Specifically, between the second (bottom left) and the third (bottom middle) views only one or two plausible correspondences were obtained which is not enough for alignment style recognition algorithm, though in other pairs of the views at least three correspondences were obtained. This decrease in accuracy will be due to the fact that the surface of the object Doll has almost regular geometric feature distributions in the image space and that the color varies only in small areas on the surface: the color changes only on the boundaries of hair and face, face and body, and body and arms and legs. As we argued in the derivation of φ since it provides no selectivity in places where color does not change, always returning the same value 0.5, it does not contribute to clustering in such places.

Thus, from this experiment for φ, it is not robust against occlusions when the surface of the object does not have enough color variations from place to place.

5. DISCUSSION AND CONCLUSION

We argued that by combining the proposed photometric invariants with geometric constraints, we can realize efficient and reliable recognition of 3D objects. Specifically, we conducted experiments of identifying the corresponding feature positions

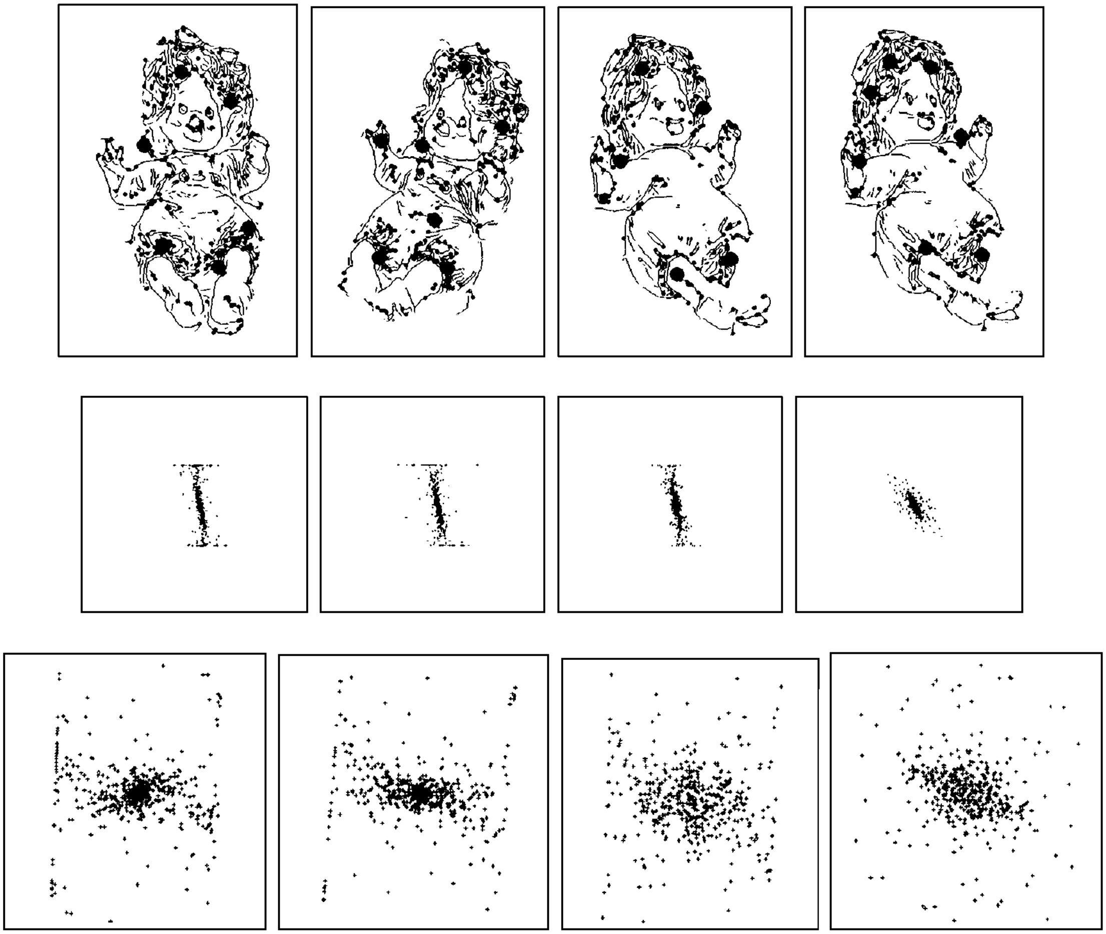

FIG. 11. Tests with φ on Doll pictures. The surface of this doll does not have man-made texture like characters, but only has color/brightness variation due to the change of material. The surface is mostly smooth except for some parts including hairs, face, and finger parts. The pictures in the top row show the edges with extracted geometric features superimposed on it. The first and second pictures were taken under usual light conditions, but at different poses of the doll. The third picture was taken under a moderate greenish light and the fourth picture was taken under an extremely bright bluish light. The figures in the second and the third row show the respective original and decorrelated distributions of φ. The identified positions are depicted by large closed circles in the figures of the upper row. The algorithm could perform identification of the corresponding positions fairly well as we see in the pictures.

over the different views taken under different conditions. In our method, we apply a geometric and photometric normalization to bring features into a coordinate frame in which they are invariant up to a rotation in the feature space, and we use these invariant properties to yield the same cluster configurations in the clustering results. The centroids of those groups can then be used as input to an alignment style recognition system, such as [22] or the linear combination of the model [36]. We note that the feature groups obtained (as shown by the large circles in the figures) are not perfect but in each case there was sufficient commonality of the extracted feature groups so that an alignment technique would correctly identify the pose of the object. Of course, this assumes that alignment will also be able to use verification of the full model to distinguish correct from incorrect index sets, as was demonstrated in [28].

In the experiments, we showed that our methods could tolerate considerable occlusions in addition to the perturbations of color and geometric properties and could provide at least a minimum number of correspondences of positions necessary for object recognitions. Although generally it might be better to extract object regions prior to feature detection and clustering processes, we stress again that, as demonstrated, our method does not require the accuracy of those preliminary processes so strictly. Moreover, as long as the background has different colors from the object, we can use φ without any preliminary processing for region extraction. This also holds true for γ when the ambient light has remained unchanged. The weakness of φ comes out when the discontinuities of gray level do not coincide with the discontinuities of colors. In this case, the distribution of φ does not spread very well. This emerged in the body parts of the doll. Compared with the conventional approaches of matching local features of which the number is of the order of several hundreds, the computational cost of our approach for recognizing 3D objects should be very small. The time for identifying (about

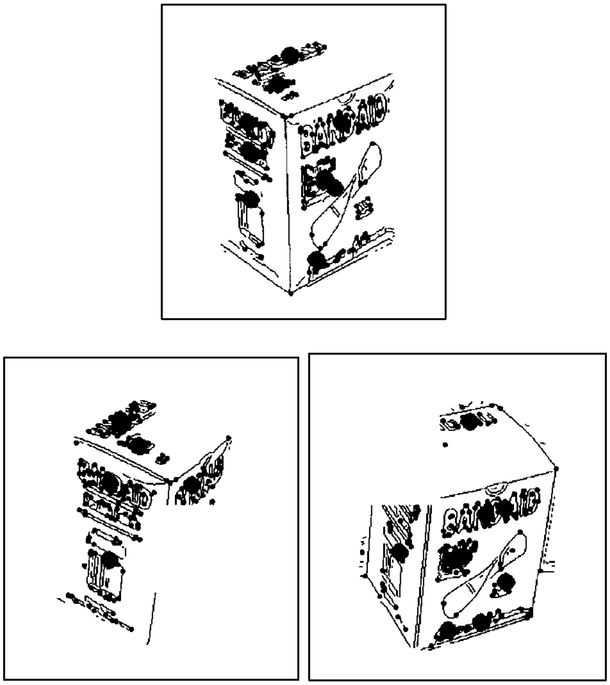

FIG. 12. Tests with γ on occluded Bandage-box pictures. The result of a picture without any drop of regions is given on the top. In the bottom left picture, the lower right corner of the object region was dropped, which was nearly 35% of the whole object area, which included about 20% of the feature points. The picture in the bottom right has a drop of the upper left corner which was almost 20% of the whole region and included 23% of the feature points.

10) corresponding feature positions, i.e., cluster centroids, was around 0.2 s for pictures with several hundreds of features. In addition, we can use the invariant photometric values in searching for the correspondences between the derived feature points in the model and the image, so that needless searches could be further suppressed. As for the stability of our algorithm against the change of the parameters included in the algorithm, we actually noted that when we changed the weight *s* which balances geometry and photometry in forming the extended features, the clustering configuration perturbed slightly. This kind of instability always accompanies when one includes a clustering process in the algorithm. However, in alignment-style recognition this can also be handled in its consistent framework by simply treating those perturbed candidates as just another candidate, thus increasing the search space by just a few scale.

The differences and similarities of our approach and Nayar's are as follows. Their method used invariant photometric properties designed for neighboring points for regions each with a consistent and a different color, so that the color segmentation is a prerequisite. In our view, this color segmentation is an essential process to reduce the size of the search space for correspondences, and the photometric invariant was used only for further limiting possible matches between the model and the data regions. Unfortunately, however, achieving complete color segmentation is often quite hard and time consuming [33]. Of course, it can still contribute to reducing the computational cost, since in general the number of color regions included in the entire image could still be on the order of some tens. But, it appears to be less of a contribution than color segmentation to the reduction of computational cost. On the other hand, our method in some cases does not require color segmentation and in others requires only rough extraction of the object region. As far as we have experienced, the feature detection that is not required of Nayar's method is not a hard task, and is not time consuming, as long as we do not require high accuracy. After those preliminary processes, since the color invariant properties are passed to the following clustering plus feature centroid alignment process, our method can tolerate many confounding factors, such as inaccuracies of region and/or feature extraction, happening in the application to the real world. The clustering plus feature centroid alignment process is very suitable for compensating those uncertainties. We should also point out that, to be theoretical, the region centroids which they used for matching cannot be used for 3D surfaces, while our feature centroids can.

The weakness of both our and Nayar's methods will be against large occlusions, especially on objects having small color variations. Since both try to produce corresponding partitions, whether

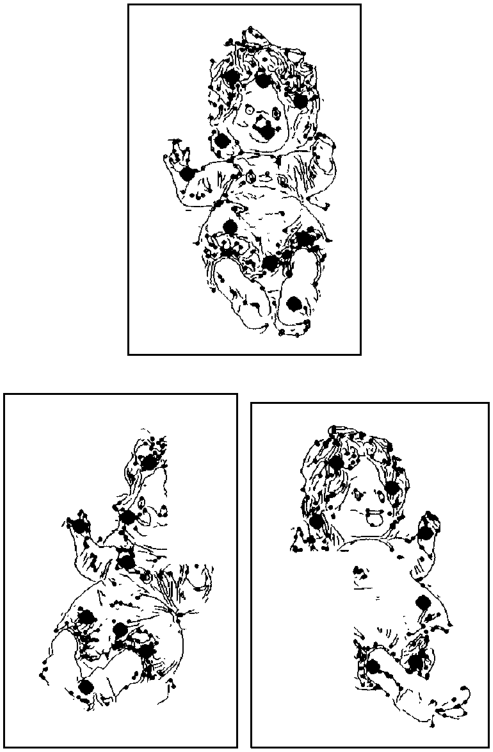

FIG. 13. Tests with γ on occluded Doll pictures. The top figure is the result on the complete object view, while in the bottom figures results are given in which almost 20% of the whole object regions are occluded: in the first view the upper right part is dropped losing about 32% of the whole features, in the second view the lower left is dropped losing 12% features, and in the third the upper left is removed which included 27% of the features.

in feature space or in image space, if the whole object area is damaged significantly we will never be able to obtain corresponding groupings. Also, their method and ours using the invariant φ will not work well on objects which do not have enough variety of colors on the surfaces, as those invariants would provide no information except on the color bounraries. In our experiments using φ on occluded Doll pictures, this happened exactly and the accuracy of the correspondences of extracted features degraded.

An alternative way of using the proposed photometric invariant in recognition is just to incorporate it into the conventional framework of recognition. For example, in selecting features to form hypothesized corresponding triples of features between the model and the data, photometric properties can be used to limit the possible matches between the model and the data features, trimming a bunch of needless combinations in the search space, thereby effectively reducing the computational cost. This kind of idea has been used in [30] for matching corresponding regions.

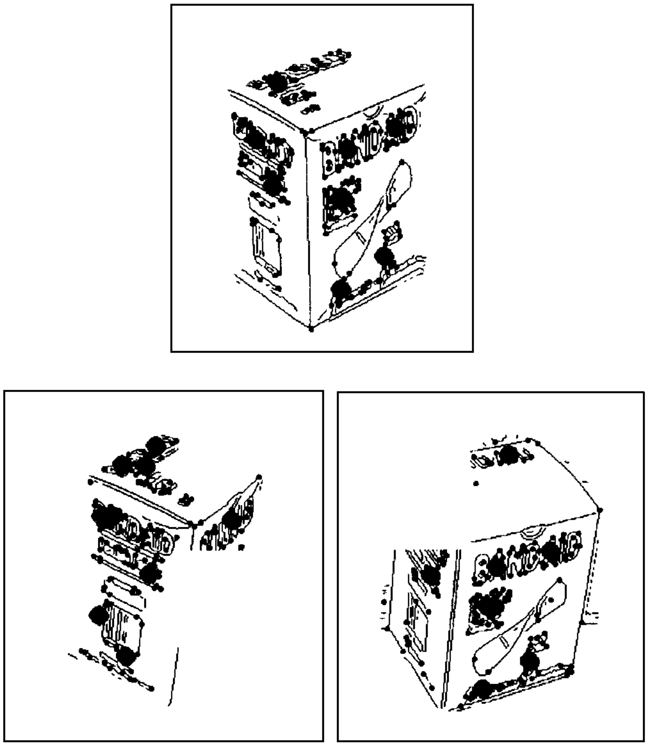

FIG. 14. Tests with φ on occluded Bandage-box pictures. The result of a picture without any drop of regions is given on the top. In the bottom left picture, the lower right corner of the object region was dropped which was nearly 35% of the whole object area, which included about 20% of the feature points. The picture in the bottom right has a drop of the upper left corner which was almost 20% of the whole region and included 23% of the feature points. We note that the results of detecting the salient features are fairly good, providing still enough commonality: four common features between the first (top) figure and the second (bottom left) figure, three between the second and third (bottom right), and six between the first and the third.

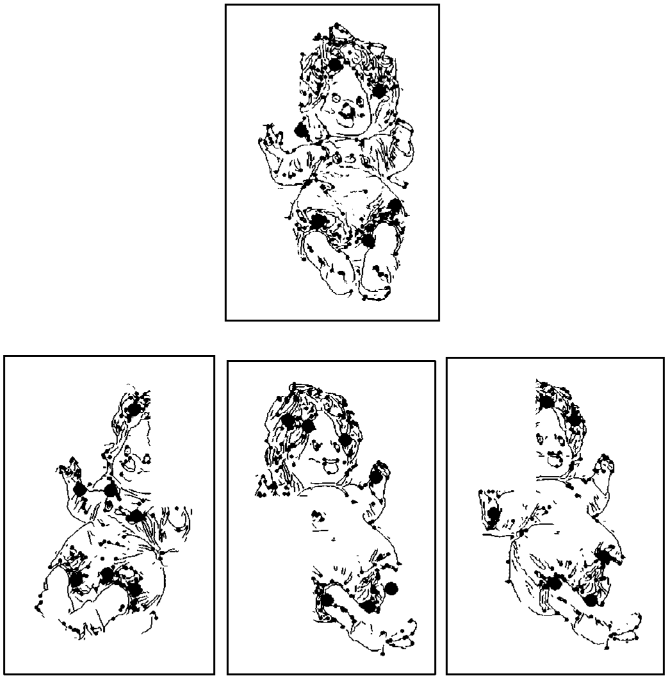

FIG. 15. Tests with φ on occluded Doll pictures. The top figure is the result on the complete object view, while in the bottom figures results are given in which almost 20% of the whole object regions are occluded: in the first view the upper right part is dropped losing about 32% of the whole features, in the second view the lower left is dropped losing 12% of the features, and in the third the upper left is removed which included 27% of the features. The accuracy of the correspondences of the extracted features degrades: only one or two correspondences were obtained between the second (bottom left) and the third (bottom middle) views.

ACKNOWLEDGMENTS

Kenji Nagao is thankful to Berthold Horn for his useful comments on this research. He also thanks many other colleagues in the MIT AI LAB for the discussions on his research.

REFERENCES

1. T. Alter and W. Eric L. Grimson, Fast and robust 3D recognition by alignment, in *Proc. ICCV 93, 1993*, pp. 113–120.
2. H. Asada and M. Brady, Curvature primal sketch, *IEEE Trans. Pattern Anal. Mach. Intell.* **8**, 1986, 2–14.
3. D. H. Ballard and C. M. Brown, *Computer Vision*, pp. 31–35, Prentice Hall Englewood Cliffs, NJ, 1982.
4. M. H. Brill, A device performing illuminant-invariant assessment of chromatic relations, *J. Theoret. Biol.* 1978, 473–478.
5. D. C. Brockelbank and Y. H. Yang, An experimental investigation in the use of color in computational stereopsis, *IEEE Trans. Systems Man Cybernet.* **19**(6), 1989, 1365–1383.
6. J. F. Canny, A computational approach to edge detection, *IEEE Trans. Pattern Anal. Mach. Intell.* **8**, 1986, 34–43.
7. J. Cohen, Dependency of the spectral reflectance curves of the Munsell color chips, *Psychon. Sci.* **1**, 1964, 369–370.
8. O. D. Faugeras, Digital color image processing within the framework of a

human visual model, *IEEE Trans. Acoustics Speech Signal Process.* **27**(4), 1979, 380–393.

9. G. D. Finlayson, M. S. Drew, and B. V. Funt, Diagonal transforms suffice for color constancy, in *Proc. ICCV, 1993*, pp. 164–170.
10. B. V. Funt and G. D. Finlayson, Color constant color indexing, *IEEE Trans. Pattern Anal. Mach. Intell.* **17**(5), 1995, 522–529.
11. G. D. Finlayson, B. V. Funt, and K. Barnard, Color constancy under varying illumination, in *Proc. ICCV95, 1995*, pp. 720–725.
12. D. A. Forsyth, A novel algorithm for color constancy, *Internat. J. Comput. Vision* **5**(1), 1990, 5–36.
13. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, San Diego, 1972.
14. B. V. Funt and M. S. Drew, Color constancy computation in near-Mondrian scenes using a finite dimensional linear model, in *Proc. CVPR, 1988*, pp. 544–549.
15. R. Gershon, A. D. Jepson, and J. K. Tsotsos, From [R,G,B] to surface reflectance: computing color constraint descriptors in images, in *Proc. 10th Int. Jt. Conf. on Artificial Intelligence, 1987*, pp. 755–758.
16. W. E. L. Grimson, *Object Recognition by Computer*, MIT Press, Cambridge, MA, 1991.
17. W. E. L. Grimson, Affine matching with bounded sensor error: a study of geometric hashing and alignment, A.I. Memo 1250, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, August 1991.
18. W. E. L. Grimson, A. Lakshmi Ratan, P. A. O'Donnell, and G. Klanderman, An active visual attention system to "Play Where's Waldo," *IEEE CVPR Workshop on Visual Behaviors, Seattle*, June 1994.
19. E. Hering, *Outlines of a Theory of the Light Senses*, translated by Leo M. Hurvich and Dorothea. Harvard Univ. Press, Cambridge, MA, 1964.
20. B. K. P. Horn, *Robot Vision*, MIT Press, Cambridge, MA, 1986, pp. 185–277.
21. B. K. P. Horn, Determining lightness from an image, *Comput. Graphics Image Process.* **3**, 1974, 277–299.
22. D. P. Huttenlocher and S. Ullman, Recognizing solid objects by alignment with an image, *Internat. J. Comput. Vision* **5**(2), 1990, 195–212.
23. J. S. Lim, *Two-Dimensional Signal and Image Processing*, Prentice Hall, Englewood Cliffs, NJ, 1990, pp. 413–423.
24. D. B. Judd, D. L. MacAdam, and G. Wyszecki, Spectral distribution of typical daylight as a function of the correlated color temperature, *J. Opt. Soc. Am.* **54**, 1964, 1031–1040.
25. G. J. Klinker, *Physical Approach to Color Image Understanding*, Ph.D. thesis, Carnegie Mellon University, 1988.
26. G. J. Klinker, Steven A. Shafer, and Takeo Kanade, The measurement of highlights in color images, *Internat. J. Comput. Vision* 1988, 7–32.
27. Y. Lamdan, J. T. Schwartz, and H. J. Wolfson, Affine invariant model based object recognition, *IEEE Trans. Robotics Automation* **6**, 1988, 238–249.
28. K. Nagao and W. E. L. Grimson, Object recognition by alignment using invariant projections of planar surfaces, *Comput. Vision Image Understanding*, in press. [Also in *Proc. 12th ICPR, 1994*, pp. 861–864, and in A.I. Memo 1463, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, February 1994.]
29. K. Nagao, Recognizing 3D object using photometric invariant, in *Proc. ICCV95, 1995*, pp. 480–487. [Also in A.I. Memo 1523, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, February 1995.]
30. S. K. Nayar and R. M. Bolle, Reflectance ratio: a photometric invariant for object recognition, in *Proc. Fourth International Conference on Computer Vision, 1993*, pp. 280–285.
31. M. J. Swain, *Color Indexing*, Ph.D. thesis, Chapter 3, University of Rochester Technical Report 360, November 1990.
32. M. J. Swain and D. H. Ballard, Color indexing, *Internat. J. Comput. Vision*, **7**(1), 1991, 11–32.
33. T. F. Syeda-Mahmood, Data and model-driven selection using color regions, in *Proc. European Conference on Computer Vision, 1992*, pp. 321–327.
34. W. B. Thompson, K. M. Mutch, and V. A. Berzins, Dynamic occlusion analysis in optical flow fields, *IEEE Trans. Pattern Anal. Mach. Intell.* **7**, 1985, 374–383.
35. M. Tsukada and Y. Ohta, An approach to color constancy using multiple images, in *Proc. ICCV, 1990*, pp. 385–389.
36. S. Ullman and R. Basri, Recognition by linear combinations of models, *IEEE Trans. Pattern Anal. Mach. Intell.* **13**(10), 1991, 992–1006.
37. B. A. Wandell, The synthesis and analysis of color images, *IEEE Trans. Pattern Anal. Mach. Intell.* **9**(1), 1987, 2–13.