International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 22, No. 6 (2008) 1141–1169

World Scientific
www.worldscientific.com

FLEXIBLE 3D OBJECT RECOGNITION FRAMEWORK USING 2D VIEWS VIA A SIMILARITY-BASED ASPECT-GRAPH APPROACH

JWU-SHENG HU* and TZUNG-MIN SU†

Department of Electrical and Control Engineering
National Chiao-Tung University
Hsinchu, Taiwan, R.O.C.
**jshu@cn.nctu.edu.tw*
†*linux.ece89g@nctu.edu.tw*

This work presents a flexible framework for recognizing 3D objects from 2D views. Similarity-based aspect-graph, which contains a set of aspects and prototypes for these aspects, is employed to represent the database of 3D objects. An incremental database construction method that maximizes the similarity of views in the same aspect and minimizes the similarity of prototypes is proposed as the core of the framework to build and update the aspect-graph using 2D views randomly sampled from a viewing sphere. The proposed framework is evaluated on various object recognition problems, including 3D object recognition, human posture recognition and scene recognition. Shape and color features are employed in different applications with the proposed framework and the top three matching rates show the efficiency of the proposed method.

Keywords: Aspect-graph; object representation; object recognition; human posture recognition; scene recognition.

1. Introduction

Object recognition is an important topic in computer vision where various approaches have been developed.[7,27,32,44] However, numerous technical issues require further investigation, especially for 3D object recognition. Variations in viewing direction and angle,[7,34] illumination changes,[13,19] and scene clutter and occlusion[28,36] are the main challenges for object recognition. In recent years, many researches were presented for solving these issues. For example, a generic object class detection system[38] that combines the Implicit Shape Model and multiview specific object recognition is presented to detect object instances from arbitrary viewpoints. A new framework[35] that combines a visual-cortex-like hierarchical structure and an increasingly complex and invariant feature was proposed for robust object recognition. Furthermore, a new object representation, Multicolored Region Descriptor (M-CORN),[29] was proposed to describe the color and local shape information of objects. Moreover, some low-level visual features, such as object shading, surface

texture and an object's contour or binocular disparity, have recently been proposed to describe 3D object representation.[8,12] However, 3D object recognition is primarily influenced by position variations and illumination source type, and the relative positions of an observer and object.

Some advanced theorems of 3D object perception have been investigated to solve these issues and enhance the 3D object recognition task.[3] Existing theorems for high-level 3D object perception can be categorized as object- and viewer-centered representations based on a coordinate system,[32] and as volume- (or model-based) and view-based representations based on the constituent elements.[40] Viewer-centered representation describes portions of an object relative to a coordinate system based on an observer. A view-based representation characterizes a 3D object using a set of object views. Both viewer-centered and view-based frameworks conform to the intuition of human perception, during which a person memorizes an object using several primary views without requiring an exhaustive 3D object model. Moreover, Kim *et al.*[18] proposed a combined model-based method to recognize 3D objects using a combination of a bottom-up process (model parameter initialization) and a top-down process (model parameter optimization).

1.1. *Human posture recognition*

Human posture recognition is an important example of 3D object recognition. A considerable number of studies have been made on this field over the past ten years.[1,17] Existing approaches[6] for human posture recognition are classified as direct and indirect approaches based on the human body model. The model has either a 2D or 3D representation based on the dimensionality of features. The direct approach typically consists of a detailed human body model. For example, Ghost[14] developed a silhouette-based body model, incorporating hierarchical body pose estimation, a convex hull analysis of the silhouette and a partial mapping from body parts to silhouette segments. Furthermore, Pfinder[42] utilized color information to develop a multiclass statistical model and identified human body parts using shape detection. However, occlusions and perspective distortion lead to the unreliable results. The indirect approach extracts features about the human body instead of a detailed human body model, and combines classifiers to estimate human posture. For example, Ozer*et al.*[31] utilized the AC-coefficients as the features and adopted principal component analysis as the classifier. A recent work[30] used color, edge and shape as the features and the hidden Markov model as the classifier. Furthermore, complex 3D models utilize different equipment to solve problems associated with the angle from which human postures are observed. For instance, Delamarre *et al.*[9] proposed a method for building a 3D human body via three or more cameras, and then calculated the projection of the silhouette for comparison with 2D projections in a database. Additionally, 3D laser scanners[41] or thermal cameras[16] have also been adopted to build a 3D human body model. However, these 3-D solutions require enormous computing time and high device costs.

1.2. *Scene recognition*

Recognizing scene can be addressed as a problem of 3D object recognition, where the scene represents variations due to changing the viewer location or camera pose. Scene recognition is a fundamental element in the topological representation of environment,[23,39] where the graph node of the adjacency graph describes the robot's location. Moreover, scene recognition can also be employed to memorize and detect visual landmarks in geometrical representation of environments.[11,21] In Ref. 2, a series of experiments were presented to show that only the overall geometry and a few key features are required to perform scene recognition. For capturing the key features, Kröse *et al.*[22] proposed a method for appearance-based modeling of an environment by extracting scene features using principal component analysis (PCA). Moreover, a framework combined with a supervised method for recognizing the door and an unsupervised method for learning door-reaching behavior has been proposed in Ref. 5.

1.3. *Aspect-graph representation*

The common challenge in 3D object recognition, human posture recognition, and scene recognition is the variation in orientations. The simplest method for solving this problem is to characterize an object with a densely sampled collection of independent views. The object can be described in detail by constructing an object model with numerous 2D views; however, this approach significantly increases computing time due to the expansive search space. Thus, several approaches have been developed to extract a minimal set of object views. Appearance-based methods focus on changes in intensity of each view. However, changes to object lighting, rotation, deformation and occlusion affect object recognition results when using the appearance-based method. Aspect-graph representations focus on shape changes to an object's projection. Koenderink *et al.* developed the underlying theory that describes 3D objects using aspect-graph representation.[20] Moreover, the traditional aspect-graph method[37] assumes that an object belongs to a limited class of shapes, and that characteristic views can be extracted using prior knowledge of the object. Aspect-graph vertices represent the characteristic views extracted from points on a transparent viewing sphere with an object in the object center. These characteristic views are extracted as prototypes of an object from a densely sampled collection of object views.

1.4. *Motivation for the proposed method*

Cyr and Kimia[7] presented a similarity-based aspect-graph method to extract the characteristic views using shape similarity between views. The viewing sphere is sampled at regular (five-degree) intervals and two similarity metrics, which one based on curve matching and the other based on shock matching, are applied to

combine views into aspects. Let there be N objects $\{O_1, O_2, \ldots, O_n, \ldots, O_{N-1}, O_N\}$, which comprise an object database. Each object is composed of M views sampling the viewing sphere giving rise to a set of views $\{V_1^1, \ldots, V_m^n, \ldots, V_M^N\}$ where V_m^n denotes the mth view of object O_n. The aspect p of object n is defined as A_p^n, which is a collection of views ranging from $V_{m-k^-}^n$ to $V_{m+k^+}^n$ and represented by the characteristic view V_m^n. Moreover, the dis-similarity of two views is represented as $d(V_m^n, V_j^i)$, which is the distance between the mth view of object n and the ith view of object j. The goal is to minimize the set of views required to represent each object O_n. Two criteria are imposed to maintain successful object recognition while forming aspects representation by characteristic views. The first criterion (local monotonicity) supposes that the dis-similarity of two views increases as their relative viewing angle between them increases. The second criterion describes that the distance of each view V_i^n in an aspect A_m^n and the characteristic view of that aspect V_m^n is smaller than the distance between any non-aspect view V_j^n and the characteristic view V_m^n.

The training views of an object in Ref. 7, which are sampled at five-degree increments and sorted by order, are collected in advanced. When additional views of an object are collected to improve object representation in the work of Ref. 7, the total views of an object must be resorted in order of view angles. The first criterion, local monotonicity, is not suitable when the object is symmetrical in the feature space. It is inconvenient to update the aspect-graph representation while collecting more new 2D views. To improve the flexibility of an update mechanism, this work presents an incremental database construction method for building and updating the aspect-graph with object views sampled at random intervals. Object representation becomes increasingly detailed using additional captured and characteristic views without recalculating similarity measures by resorting total views. Moreover, the first criterion in the work of Ref. 7, local monotonicity, is not utilized, thereby improving flexibility of extracting aspects of symmetrical objects. Although the proposed approach cannot confirm the view angle of a test view using a specific object view, the proposed approach improves flexibility when building an aspect-graph representation, and reduces computing time when updating object aspects. Additionally, the accuracy of the object representation increases with minimal growth of search space while collecting additional new object views.

The remainder of this paper is organized as follows. Section 2 presents an overview of the proposed method. Section 3 describes the procedure for extracting features and the similarity measures for building database and object matching. Section 4 describes the incremental database construction method for extracting the aspects and characteristic views of objects. Furthermore, the object matching procedure is described with a weighting combination between different similarity measures. Section 5 presents experimental results that demonstrate the performance of the proposed method for 3D rigid objects, human postures and scene recognition. Conclusions are discussed in Sec. 6.

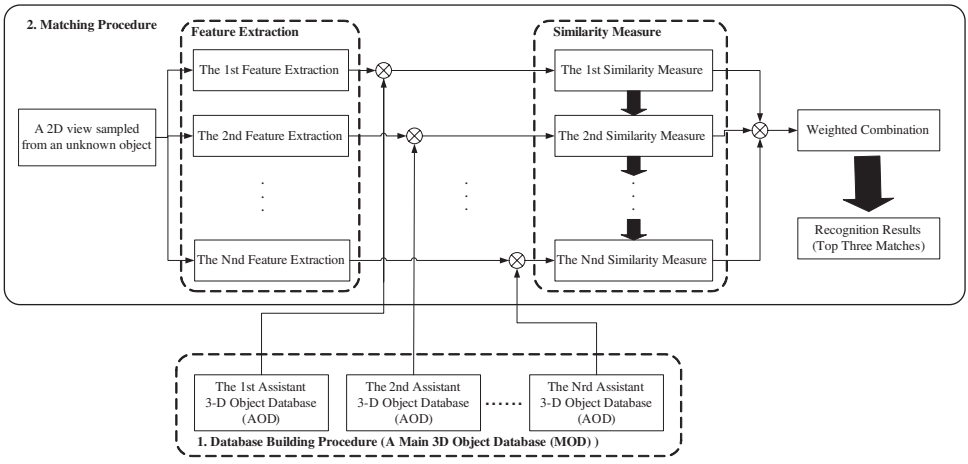


Fig. 1. The system architecture of the proposed framework. A main 3D object database (MOD) comprises of total AODs.

2. The Overview of the Proposed Method

The proposed framework (Fig. 1) contains two parts, which are called the database building procedure and the matching procedure. Suppose an object database contains T_0 objects, and T_1 2D views of each object are randomly sampled from a viewing sphere. In the database building procedure, a proposed incremental database construction method (Fig. 2) is applied to extract the aspects of each object using T_1 2D views. The main 3D database contains a set of assistant 3D object databases (AOD). Furthermore, an AOD comprises the aspects of each object, where the aspects are represented by their characteristic views. Figure 3 illustrates the inner structure of an AOD. In Fig. 3, a set of aspects is employed to represent the database of a 3D object in the aspect level. The prototypes for these aspects, called the characteristic views, are utilized to represent an object for object matching. The passage from one characteristic view to another is defined with only the similarity measure. The proposed similarity-based aspect-graph focuses on an efficient learning method with associated features and similarity functions. While the object features are sufficient to discriminate the similarity between each two 2D training views, the

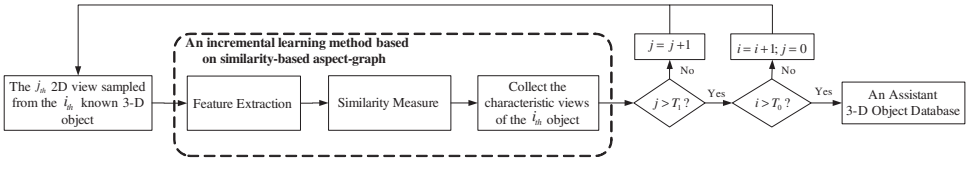


Fig. 2. The database building procedure, where T_0 is the number of objects in the database and T_1 is the number of sampled views required to build the aspect-graph representation of an object.

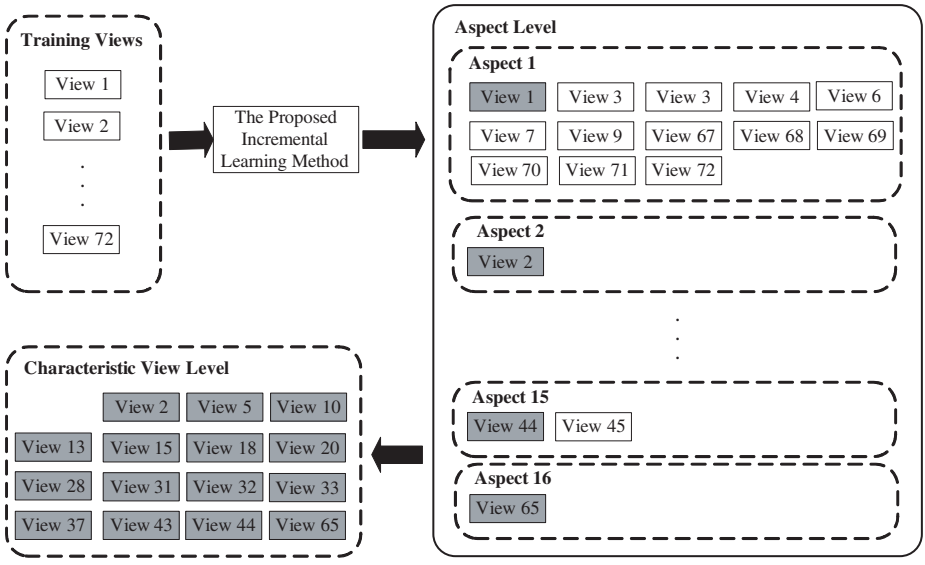


Fig. 3. The inner structure of an AOD.

aspects and characteristic views can be extracted using associated similarity measures. Even if the objects are complex, the characteristic views can be extracted in the feature space.

In the matching procedure, a similarity measure is applied between a 2D view sampled from an unknown object and all the characteristic views of the 3D object database. After the weighted combination of all similarity measures, the first three characteristic views that have the highest similarity with the testing 2D view are regarded as the recognition results (the top three matches).

3. Object Representation

In this work, shape and color features are utilized to measure similarity between two object views. To extract shape information, a robust background subtraction framework from previous works[15,24] is utilized to extract foreground regions while considering shadows and highlights. Foreground detection provides flexibility when constructing the object database, even in an out-of-control environment. Canny edge detection[4] is then applied to extract shape edge, and the Gradient Vector Flow Snake (GVF)[43] is applied to extract the contour information. Assume that the contour information is included in a set $\mathbf{Z}$, which is composed of N points z_i, where z_i is a complex form given by Eq. (1). Two kinds of shape features, which are called the Fourier descriptor (FD) and the point-to-point length (PPL), are extracted from $\mathbf{Z}$.

$$\mathbf{Z} = \{z(i)\} = \{x_i + jy_i\}, \quad 0 \le i < N. \tag{1}$$

3.1. *Shape features*

The points inside the set $\mathbf{Z}$ are resampled using Eq. (2) to eliminate variations in shift and scale.

$$\tilde{\mathbf{Z}} = \{\tilde{z}(i)\} = \{L_c[(x_i - x_c) + j(y_i - y_c)]/L\} \tag{2}$$

where $0 \leq i < N$; L denotes contour length of $\mathbf{Z}$, L_c is expected contour length, and (x_c, y_c) is the location of the contour center of $\mathbf{Z}$. Then, the Fourier transform is applied to $\tilde{\mathbf{Z}}$ to compute FD using Eq. (3).

$$\mathrm{FD}(k) = \sum_{n=0}^{N-1} \tilde{z}(n) \exp(-j2\pi kn/N), \quad 0 \leq \mathrm{k} < N. \tag{3}$$

The low-frequency parts of FD are extracted with the consideration of decreasing the variations of high-frequency noises, and are defined as MAG. Notably, MAG is composed of $2T_2$ magnitude values of frequency information selected among $2N$ frequencies. The method for extracting MAG is given by Eq. (4).

$$\mathrm{MAG} = \{\,|\,\tilde{Z}(k)\,|\,,\ |\,\tilde{Z}(N-k)\,|\,,\ 1 \leq k \leq T_2\}. \tag{4}$$

Intuitively speaking, MAG only characterizes the shape and not the orientation of human posture. Therefore, MAG cannot discriminate between similar shapes oriented differently. To solve this problem, phase information for FD must be used for memorizing an object. The work in Ref. 25 proposes that memorizing the phase value at low frequency is sufficient. Suppose the phase information is θ_z, then θ_z can be calculated using FD(1) and FD($N-1$), as described in Eqs. (5) and (6).

$$\mathrm{FD}(1) = |\,\mathrm{FD}(1)\,| \cdot \exp(j\theta_1) = R_1 + jI_1 \tag{5}$$

$$\mathrm{FD}(N-1) = |\,\mathrm{FD}(N-1)\,| \cdot \exp(j\theta_{N-1}) = R_{N-1} + jI_{N-1}. \tag{6}$$

Furthermore, θ_z can be calculated using Eq. (7).

$$\theta_z = (\theta_1 + \theta_{N-1})/2 = (\arctan(I_1/R_1) + \arctan(I_{N-1}/R_{N-1}))/2 \tag{7}$$

where R_1 and R_{N-1} denote the real parts of FD(1) and FD($N-1$), I_1 and I_{N-1} denote the imaginary parts of FD(1) and FD($N-1$), and θ_1 and θ_{N-1} are the phases of FD(1) and FD($N-1$).

Moreover, the lengths between each pair of points in $\mathbf{Z}$ are defined as PPL, which is suitable for describing shape details. To calculate PPL is time consuming since each point is considered as a start point. Equations (8) and (9) describe the calculating process of PPL.

$$\begin{aligned}\mathrm{PPL}(k) &= \{l_i\} = \{\|\tilde{z}(i) - \tilde{z}(i-1)\|\} \\ &= \{\sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}\}, \\ &\qquad 1 \leq k \leq N, \quad k \leq i \leq N + k\end{aligned} \tag{8}$$

$$\tilde{z}(k) = \tilde{z}(N+k). \tag{9}$$

3.2. *Color features*

Numerous features, such as edge, corner, texture, color and shape, have been utilized to extract useful information from an image. Among these features, color involves the intuitive information to represent the conceptual idea of an image. Therefore, pixel color and pixel position are utilized in this work to extract the conceptual idea of an image. The color space used is RGB color space, a format common to most video devices. To enhance the regional information of an image, the position (x, y) feature is combined with RGB color information as the feature vector. That is, each pixel contains a 5D feature vector (R, G, B, x, y), which is shown in Fig. 4.

This work applies Gaussian mixture model (GMM) to model region information in a scene image as a blob model, which is defined as BM, using 5D feature vectors (R, G, B, x, y). We assume that the density function of color and position features have Gaussian distributions. First, each pixel x is defined as a five-dimensional vector at time t. Moreover, N Gaussian distributions are used to construct the GMM, which is described in Eq. (10).

$$f(x \mid \lambda) = \sum_{i=1}^{N} w_i \frac{1}{\sqrt{(2\pi)^d \mid \sum_i \mid}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \sum_i^{-1}(x-\mu_i)\right) \tag{10}$$

λ represents the parameters of GMM,

$$\lambda = \left\{ w_{\mathrm{i}}, \mu_{\mathrm{i}}, \sum_{\mathrm{i}} \right\}, \quad i = 1, 2, \ldots, N \quad \text{and} \quad \sum_{\mathrm{i=1}}^{\mathrm{N}} w_i = 1.$$

Next, parameters λ of GMM are calculated to enable the GMM to match the feature vector distribution with least errors. The most common method for calculating parameters λ is maximum likelihood (ML) estimation. The objective of ML estimation is to identify model parameters by maximizing the likelihood function of GMM obtained from training feature vectors X. The ML parameters are

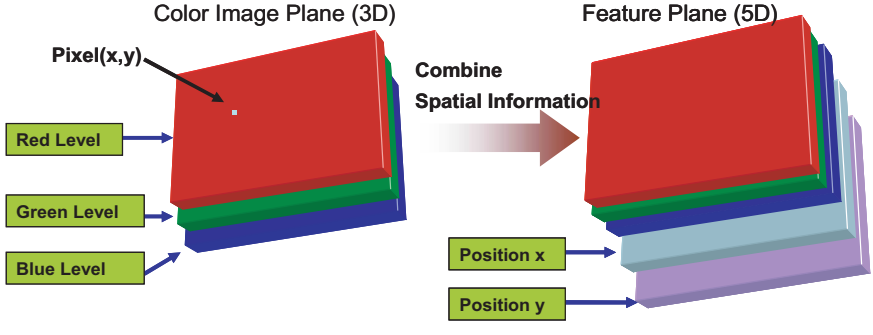


Fig. 4. 5D feature vector construction.

derived iteratively using the expectation maximization (EM) algorithm.[10] Supposing there are s feature vectors $x_1, x_2, \ldots, x_s$ (in this work s is defined as image size, $320 \times 240 = 76,800$), then the ML estimation of λ can be calculated using Eq. (11).

$$\lambda_{\mathrm{ML}} = \arg\max_{\lambda} \sum_{j=1}^{s} \log f(x_j \mid \lambda). \tag{11}$$

Furthermore, unsupervised data clustering is used before the EM algorithm iterations to accelerate convergence. This study uses the K-means algorithm[26] for clustering. The number of clusters is defined, and then the initial center of each cluster is obtained randomly. The appropriate center and variance of each cluster can be estimated iteratively using the K-means algorithm and applied as the initial mean and variance of each Gaussian component of the GMM.

3.3. *Similarity functions*

To determine the similarity between two objects when building databases and recognizing objects, a similarity measurement $D(U, V)$ is applied to the extracted features. We assume that the features extracted from two contours are $U = \{u_0, \ldots, u_i, \ldots, u_{I-1}\}$ and $V = \{v_0, \ldots, v_i, \ldots, v_{I-1}\}$, respectively, where I denotes the feature size. Two similarity measures are applied using 1-norm distance [Eq. (12)] and K-L distance[33] [Eq. (13)], where c denotes the number of points on an extracted contour and s denotes image size. In this work, c is defined as 256 and s is defined as 76,800.

$$D_{1-\mathrm{norm}}(U, V) = \sum_{i=0}^{I-1} |u_i - v_i|, \quad I = c \tag{12}$$

$$D_{KL}(U, V) \approx \sum_{t=0}^{I-1} \left(p_1(t) \cdot \log\left(\frac{p_1(t)}{p(t)}\right) + p_0(t) \cdot \log\left(\frac{p_0(t)}{p(t)}\right)\right), \quad I = s \tag{13}$$

where

$$p_0(t) = \frac{u_t}{u_{\mathrm{sum}}}, \quad p_1(t) = \frac{v_t}{v_{\mathrm{sum}}}, \quad u_{\mathrm{sum}} = \sum_{i=0}^{s-1} u_i,$$

$$v_{\mathrm{sum}} = \sum_{i=0}^{s-1} v_i, \quad p(t) = \frac{p_0(t) + p_1(t)}{2}.$$

3.4. *Similarity measures*

Suppose V^n_{new} represents a new sampled view of the nth object and C^n_m represents the mth characteristic view of the nth object. Moreover, $A_{m^{\min}}$ denotes the aspects that have the minimum distance from V^n_{new} and $C^n_{m^{\min}}$ represents the minimal distance, where $m^{\min}$ is the index of $A_{m^{\min}}$. $C^n_{m^{\min}-1}$ and $C^n_{m^{\min}+1}$ denote the neighboring views of $C^n_{m^{\min}}$. Let $d^1_M(V^n_{\mathrm{new}}, C^n_m)$ [Eq. (14)] denote the similarity

measure using MAG and 1-norm distance, $d_M^2(V_{\text{new}}^n, C_m^n)$ [Eq. (15)] denote the similarity measure using PPL and 1-norm distance, $d_M^3(V_{\text{new}}^n, C_m^n)$ [Eq. (16)] denote the similarity measure using BM and K-L distance, and $d_a^1(V_{\text{new}}^n, C_m^n)$ [Eq. (17)] denote the similarity measure using θ_z and 1-norm distance.

$$d_M^1(V_{\text{new}}^n, C_m^n) = \sum_{k=1}^{T_2} |\text{MAG}^{V_{\text{new}}^n}(k) - \text{MAG}^{C_m^n}(k)| + |\text{MAG}^{V_{\text{new}}^n}(N-k) - \text{MAG}^{C_m^n}(N-k)| \tag{14}$$

$$d_M^2(V_{\text{new}}^n, C_m^n) = \sum_{k=1}^{N} |\text{PPL}^{V_{\text{new}}^n}(k) - \text{PPL}^{C_m^n}(k)| \tag{15}$$

$$d_M^3(V_{\text{new}}^n, C_m^n) = \sum_{t=0}^{s-1} \left(p_1(t) \cdot \log\left(\frac{p_1(t)}{p(t)}\right) + p_0(t) \cdot \log\left(\frac{p_0(t)}{p(t)}\right)\right) \tag{16}$$

where

$$p_0(t) = \frac{u_t}{u_{\text{sum}}}, \quad p_1(t) = \frac{v_t}{v_{\text{sum}}}, \quad u_{\text{sum}} = \sum_{i=0}^{s-1} u_i,$$

$$v_{\text{sum}} = \sum_{i=0}^{s-1} v_i, \quad p(t) = \frac{p_0(t) + p_1(t)}{2}$$

$$d_a^1(V_{\text{new}}^n, C_m^n) = |\theta_z^T(k) - \theta_z^D(k)|. \tag{17}$$

4. A Flexible 3D Object Recognition Framework

A flexible framework using the proposed incremental database construction method is described in this section. In the framework, a MOD is composed of one or more AODs. Each AOD is built using one main feature or using one main feature with one assistant feature. Moreover, each feature has its similarity function, such as Eqs. (14)–(17).

4.1. *Generation of aspects and characteristic views*

The proposed incremental database construction method is a four-step procedure and is illustrated as in Fig. 5. Steps A-1 to A-4 are applied to extract aspects and characteristic views. Those aspects comprise the object database and the characteristic views are used for object matching with a new view V_{new}^n.

Step A-1. Initialize the number of aspects to be zero. 2D views of the nth object are randomly sampled from a viewing sphere and each 2D view is regarded as V_{new}^n.

Step A-2. When the number of existing aspects of the nth object equals zero, V_{new}^n is regarded as a characteristic view of a new aspect.

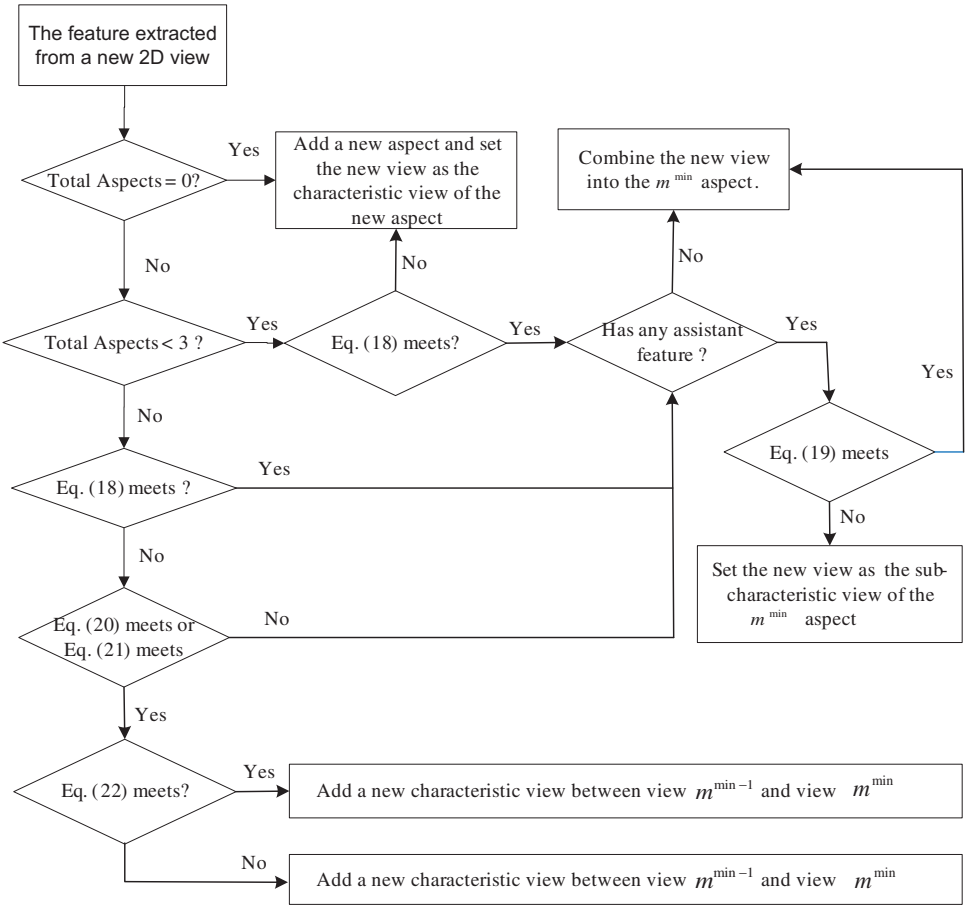


Fig. 5. The procedure of the proposed incremental database construction method.

Step A-3. When the number of existing aspects of the nth object equals one or two,

(A-3.1) When Eqs. (18) and (19) are both satisfied, V_{new}^{n} is combined into the $m^{\min}$ aspect, and the characteristic view of the $m^{\min}$ aspect remains the same,

$$\min_{\text{allC}_{\text{m}}^{\text{n}} \in A_{m^{\min}}} d_M(V_{\text{new}}^{n}, C_m^n) < T_3 \tag{18}$$

$$\min_{\text{allC}_{\text{m}}^{\text{n}} \in A_{m^{\min}}} d_a(V_{\text{new}}^{n}, C_m^n) < T_5 \tag{19}$$

where T_3 and T_5 are both predefined threshold values.

(A-3.2) Otherwise, if Eq. (18) is satisfied and Eq. (19) is not, V_{new}^{n} is combined into the $m^{\min}$ aspect, and is regarded as a new characteristic view of the $m^{\min}$ aspect.

(A-3.3) Otherwise, if Eqs. (18) and (19) are both unsatisfied, a new aspect of the nth object is established, and V_{new}^n is regarded as the new characteristic view of the new aspect.

Step A-4. When the number of existing aspects of the nth object is ≥ 3,

(A-4.1) If either Eq. (20) or Eq. (21) is true, a new aspect is constructed and V_{new}^n is considered the characteristic view of the new aspect. When a new aspect is established, the aspect order can be determined to let similar aspects be close to each other using Eq. (22). If the similarity distance between V_{new}^n and $C_{m^{\min}+1}^n$ exceeds that between V_{new}^n and $C_{m^{\min}-1}^n$, then the new aspect is inserted between aspect $m^{\min}$ and aspect $m^{\min}-1$; otherwise, the new aspect is inserted between aspects $m^{\min}$ and $m^{\min}+1$.

$$\min_{\text{allC}_m^n \in A_{m^{\min}}} d_M\left(V_{\text{new}}^n, C_m^n\right) > T_4 \tag{20}$$

$$T_3 < \min_{\text{allC}_m^n \in A_{m^{\min}}} d_M(V_{\text{new}}^n, \text{C}_m^n) < T_4 \quad \text{and} \quad d_M(V_{\text{new}}^n, \text{C}_{m^{\min}\pm 1}^n) > T_4 \tag{21}$$

$$d_M(V_{\text{new}}^n, C_{m^{\min}+1}^n) > d_M(V_{\text{new}}^n, C_{m^{\min}-1}^n). \tag{22}$$

(A-4.2) Otherwise, if Eqs. (20) and (21) are both unsatisfied and Eq. (19) is true, V_{new}^n is combined into the $m^{\min}$ aspect and the characteristic view of the $m^{\min}$ aspect remains the same.

(A-4.3) Otherwise, if Eqs. (20) and (21) are both unsatisfied and Eq. (19) is not true, V_{new}^n is combined into the $m^{\min}$ aspect and is regarded as a new characteristic view of the $m^{\min}$ aspect.

Terms T_3 and T_4 are two predefined threshold values, where $T_4 \geq T_3$. The criterion for selecting T_3 and T_4 depends on the precise level for describing the object. If T_3 and T_4 are both small, then the criterion of combining 2D views becomes strict and, thus, the number of aspects increases. Furthermore, if the difference between T_3 and T_4 decreases, the tolerance for the difference between 2D views inside an aspect decreases, thereby increasing the number of aspects. Additionally, T_3 and T_4 should be initialized manually and modified iteratively until the final number of aspect reaches an acceptable number, which is determined based on the degree of object symmetry. In this work, $T_3^{d_M^1}$ and $T_4^{d_M^1}$ are defined as T_3 and T_4 while adopting MAG as the feature; $T_3^{d_M^2}$ and $T_4^{d_M^2}$ are defined as T_3 and T_4 while adopting PPL as the feature; $T_3^{d_M^3}$ and $T_4^{d_M^3}$ are defined as the T_3 and T_4 while adopting BM as the feature, and $T_5^{d_a^1}$ is defined T_5 while adopting θ_z as the feature. Section 5 presents the values of $T_3^{d_M^1}$, $T_3^{d_M^2}$, $T_3^{d_M^3}$, $T_4^{d_M^1}$, $T_4^{d_M^2}$, $T_4^{d_M^3}$, and $T_5^{d_a^1}$.

4.2. *Object recognition using 2D characteristic views*

After constructing the aspect-graph representation of each object, a test view of an unknown object is recognized by matching itself with all the characteristic views of each AOD. If multiple AODs are utilized in the framework, a hierarchical matching process is applied to calculate the final recognition results with a weighting combination of all similarity measures. Suppose the candidate objects in the kth AOD are included in a set of N^k, and $n(N^k)$ denotes the number of candidate objects in N^k. The number of candidate objects reduces after each object matching procedure, which is described in Eq. (23).

$$n(N^{k+1}) \leq n(N^k), \quad k \geq 1. \tag{23}$$

In the kth AOD, the main feature and the assistant feature of the test view are extracted to match with all the characteristic views. Suppose V_j^i denotes a test view of an unknown object, the object matching in the proposed framework is described as Eqs. (24a) and (24b).

$$d^k(V_j^i, C_m^n) = d_m^k(V_j^i, C_m^n) + \omega_1^k \cdot d_a^k(V_j^i, C_m^n), \quad n \in N^k, \quad k = 1 \tag{24a}$$

$$d^k(V_j^i, C_m^n) = d_{\min}^{k-1}(n) + \omega_2^k \cdot (d_m^k(V_j^i, C_m^n) + \omega_1^k \cdot d_a^k(V_j^i, C_m^n)), \quad n \in N^k, \quad k \geq 2. \tag{24b}$$

Let $d_m^k(V_j^i, C_m^n)$ and $d_a^k(V_j^i, C_m^{n'(k)})$ denote the main and assistant similarity distances between the unknown object and C_m^n, where n denotes the nth object in the set of candidate objects N^k. If the framework comprises only one AOD, the characteristic views of the first three smallest similarity distances in $d^1(V_j^i, C_m^n)$ [Eq. (24a)] are regarded as the top-three matches. In Eq. (24a), ω_1^k is a weighting parameter for combining different similarity measures. When no assistant feature is utilized, ω_1^k is set to zero. Otherwise, the objects included in the first half smallest similarity distances of $d^k(V_j^i, C_m^n)$ are defined as a set N^{k+1}. The objects in N^{k+1} are preserved for further recognition in the $(k+1)$th AOD.

If the framework comprises two or more AODs, the characteristic views of the first three smallest similarity distances in $d^k(V_j^i, C_m^n)$ [Eq. (24b)] are regarded as the top-three matches. In Eq. (24b), $d_{\min}^{k-1}(n)$ denotes the minimum similarity distance between the unknown object and the nth candidate object in the $(k-1)$th database. Moreover, ω_2^k is a weighting parameter for combining the similarity measure between the kth and $(k-1)$th AODs.

4.3. *Applications*

The proposed framework is evaluated on various object recognition problems, including 3D object recognition, human posture recognition and scene recognition. The features and similarity measures described in Sec. 3 are employed in the three applications.

In the 3D rigid object recognition, two AODs are utilized with two main features MAG and PPL. The weighting combination of the similarity measures is described

in Eqs. (25) and (26).

$$d^1(V_j^i, C_m^n) = d_m^1(V_j^i, C_m^n), \quad n \in N^1 \tag{25}$$

$$d^2(V_j^i, C_m^n) = d_{\min}^1(n) + \omega_2^2 \cdot (d_m^2(V_j^i, C_m^n)), \quad n \in N^2. \tag{26}$$

Moreover, $d_m^1(V_j^i, C_m^n)$ is calculated with MAG using Eq. (14), and $d_m^2(V_j^i, C_m^n)$ is calculated with PPL using Eq. (15). Furthermore, the weighting parameters ω_1^1 and ω_1^2 are both set to zero and the weighting parameters ω_2^2 is defined as Eq. (27). $T_4^{d_M^1}$ and $T_4^{d_M^2}$ are the threshold values applied on the incremental database construction method, and are defined in Sec. 4.1.

$$\omega_2^2 = T_4^{d_M^2} / T_4^{d_M^1}. \tag{27}$$

In the human posture recognition, only one AOD is utilized with one main feature MAG and one assistant feature θ_z. The weighting combination of the similarity measures is described in Eq. (28).

$$d^1(V_j^i, C_m^n) = d_m^1(V_j^i, C_m^n) + \omega_1^1 \cdot d_a^1(V_j^i, C_m^n), \quad n \in N^1. \tag{28}$$

In Eq. (28), $d_m^1(V_j^i, C_m^n)$ is calculated using Eq. (14) and $d_a^1(V_j^i, C_m^{n'})$ is calculated using Eq. (17). Furthermore, the weighting parameter ω_1^1 is defined as Eq. (29), where $T_5^{d_a^1}$ is the threshold value applied on the incremental database construction method, and is defined in Sec. 4.1.

$$\omega = 1/T_5^{d_a^1}. \tag{29}$$

In the scene recognition, only one AOD is utilized with one main feature BM. The weighting combination of the similarity measures is described in Eq. (30).

$$d^1(V_j^i, C_m^n) = d_m^1(V_j^i, C_m^n), \quad n \in N^1. \tag{30}$$

In Eq. (30), $d_m(V_j^i, C_m^n)$ is calculated using Eq. (16). Furthermore, the weighting parameter ω_1^1 is defined as zero.

5. Experimental Results

This section describes several experiments demonstrating the effectiveness of the proposed method. A SONY EVI-D30 PTZ camera was employed to capture object views. The following three databases were built to test the proposed method: Fig. 6 contains 12 3D rigid objects, Fig. 7 contains six 3D human postures, and Fig. 8 contains 11 scenes. The notation $\mathbf{V}_d^{1,j}$ and $\mathbf{V}_d^{2,j}$ denote the sets of training views captured at five-degree intervals, where $\mathbf{V}_d^{1,j}$ is employed during rigid object recognition, and $\mathbf{V}_d^{2,j}$ is employed during human posture recognition. The notation $\mathbf{V}_d^{3,j}$, which denotes the set of training views captured at each location at a one-degree increment, is utilized during scene recognition. Moreover, $\mathbf{V}_t^{1,j}$ and $\mathbf{V}_t^{2,j}$ denote the set of testing views captured from trisection points between each pair of points separated by five-degree, where $\mathbf{V}_t^{1,j}$ is utilized during rigid object recognition,

Fig. 6. The first database containing 12 3D rigid objects.

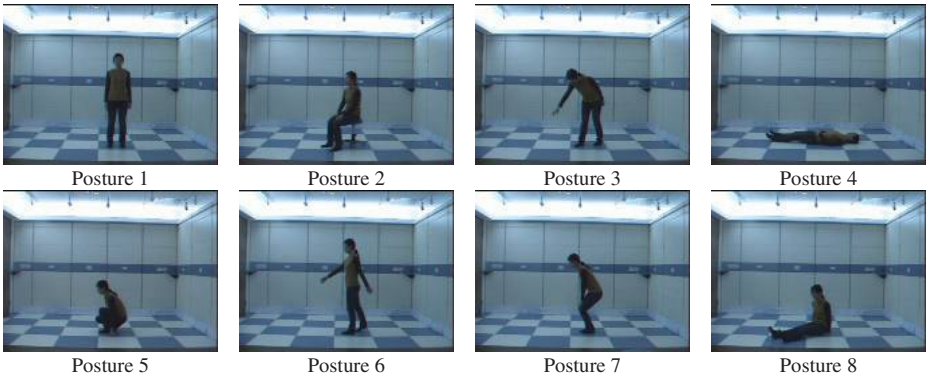


Fig. 7. The second image database containing eight 3D human postures.

and $\mathbf{V}_t^{2,j}$ is utilized during human posture recognition. Moreover, $\mathbf{V}_t^{3,j}$ denotes the set of testing views captured at locations away from the original locations in four directions (forward, backward, left and right), five distances (5 cm, 10 cm, 15 cm, 20 cm and 50 cm) and five covering rates (5%, 10%, 15%, 20% and 50%). $\mathbf{V}_t^{3,j}$ is utilized during scene recognition. The descriptions of the captured views are given in Eqs. (31)–(36).

$$\mathbf{V}_d^{1,j} = \{V_d^{1,j}(i)\}, \quad \text{where } 1 \le j \le 12,\ 1 \le i \le 72 \tag{31}$$

$$\mathbf{V}_d^{2,j} = \{V_d^{j}(i)\}, \quad \text{where } 1 \le j \le 8,\ 1 \le i \le 72 \tag{32}$$

Fig. 8. The third image database containing 11 scenes.

$$\mathbf{V}_d^{3,j} = \{V_d^j(i)\}, \quad \text{where } 1 \le j \le 11,\ 1 \le i \le 61 \tag{33}$$

$$\mathbf{V}_t^{1,j} = \{V_t^j(i)\}, \quad \text{where } 1 \le j \le 12,\ 1 \le i \le 216 \tag{34}$$

$$\mathbf{V}_t^{2,j} = \{V_t^j(i)\}, \quad \text{where } 1 \le j \le 8,\ 1 \le i \le 216 \tag{35}$$

$$\mathbf{V}_t^{3,j} = \{V_t^j(i)\}, \quad \text{where } 1 \le j \le 11,\ 1 \le i \le 6100. \tag{36}$$

In the following experiments, T_0 denotes the number of objects, and is 12 for the rigid object recognition, 8 during human posture recognition, and 11 during scene recognition; T_1 denotes the number of training views, and is 72 during rigid object recognition and human posture recognition, and 61 during scene recognition. T_2 denotes the number of low frequency information in FD, and is 40 in the following experiment. Moreover, the threshold values used in the proposed incremental database construction method are listed as in Table 1.

Table 1. The threshold values for the proposed incremental database construction method.

	The First AOD			The Second AOD		
	Main Feature		Assistant Feature	Main Feature		Assistant Feature
Experiments	T_3	T_4	T_5	T_3	T_4	T_5
3D Object Recognition	$T_3^{d_M^1} = 640$	$T_4^{d_M^1} = 1.25 * T_3^{d_M^1}$	N/A	$T_3^{d_M^2} = 336$	$T_4^{d_M^2} = 1.25 * T_3^{d_M^2}$	N/A
Human Posture Recognition	$T_3^{d_M^1} = 1450$	$T_4^{d_M^1} = 1.25 * T_3^{d_M^1}$	$T_5^{d_a^1} = 10$	N/A		
Scene Recognition	$T_3^{d_M^3} = 1100$	$T_4^{d_M^3} = 1.25 * T_3^{d_M^3}$	N/A	N/A		

Computing time for calculating similarity between a test view and a view in the database was approximately 0.006 s for rigid object recognition, 0.004 s for human posture recognition and 0.01 s for scene recognition on a P4 3.2 G CPU with 1 GB RAM.

5.1. *3-D rigid object recognition*

In the first experiment, the efficiency of the proposed framework was assessed using 2-D views captured at random intervals with the first database (Fig. 6). To determine average performance of the proposed method, training views were generated by sampling views in $V_d^{1,j}$ in 200 different random orders. Background subtraction was first performed on training 2D views to extract foreground objects. After that, Canny edge detection and GVF were performed on the extracted foreground objects to extract the object contour. Two features, called the MAG and PPL, were then extracted from the object contour and used for building the AODs with the proposed incremental database construction method (Fig. 5). The characteristic views of aspects in each AOD were utilized for object matching. A recognition result is calculated with a weighted combination of the similarity measures from both AODs. Figure 9 illustrates the system architecture of the proposed framework for the 3D rigid object recognition.

Table 2 presents statistical information on the aspect numbers using MAG and PPL. Furthermore, symmetrical objects, such as objects 2, 5, 6 and 7, had few aspects, thereby reducing computing time for recognizing objects. The views in $\mathbf{V}_t^{1,j}$ were adopted as unknowns, and tested whenever aspect-graph representations were built each time (200 times). The proposed aspect-graph generation is efficient due to its high recognition rate in the Top 1 to Top 3 matches in Table 3.

The proposed method, which constructs an aspect-graph representation using sampled views at random intervals, generates a practical updating mechanism that integrates the database using new collected views. In this experiment, 18 random views sampled from $\mathbf{V}_d^{1,j}$ are first utilized to construct a coarse aspect-graph

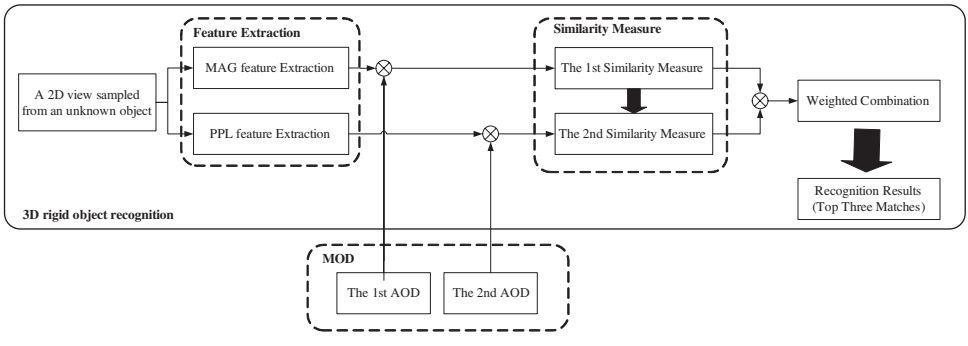


Fig. 9. The system architecture of the proposed framework applied during the first experiment (3D rigid object recognition).

Table 2. The result of rigid object recognition using 2D views via MAG and PPL.

Recognition Results	The Index of the Objects in the First Database Listed in Fig. 6												
	1	2	3	4	5	6	7	8	9	10	11	12	Avg.
Numbers of Aspect of *MAG*	34.66	3.84	27.83	24.75	6.87	9.47	2.04	25.62	17.14	16.16	16.62	28.75	17.81
Numbers of Aspect *PPL*	38.72	14.08	14.32	22.84	10.98	20.12	8.41	31.07	25.79	17.68	23.61	19.88	20.63
Top 1 Match (%)	98.25	99.97	97.71	97.39	100	99.81	99.79	99.35	99.90	97.97	98.44	96.83	98.78
Top 2 Match (%)	99.21	100	98.96	98.73	100	99.96	99.86	99.67	99.97	98.68	99.47	98.17	99.39
Top 3 Match (%)	99.61	100	99.39	99.34	100	99.98	99.89	99.78	99.99	98.98	99.77	98.64	99.62

Table 3. Results for the aspect numbers using MAG and PPL after updating by additional training views.

Numbers of Aspect	The Index of the Objects in the First Database Listed in Fig. 6											
	1	2	3	4	5	6	7	8	9	10	11	12
D_{18}	14.11	3.40	11.98	10.13	5.32	6.36	1.58	12.64	8.80	8.43	8.73	11.10
D_{36}	22.86	3.60	18.83	16.15	6.23	8.01	1.80	19.16	12.53	12.17	12.48	18.29
D_{54}	29.52	3.74	23.95	20.96	6.65	8.94	1.92	23.24	15.20	14.53	15.06	24.05
D_{72}	34.66	3.84	27.83	24.75	6.87	9.47	2.04	25.62	17.14	16.16	16.62	28.75
D_{90}	39.28	3.99	28.68	25.99	7.14	9.83	2.14	27.32	18.04	17.37	17.86	30.99
D_{108}	43.28	4.07	29.50	27.14	7.36	10.12	2.26	28.67	18.90	18.50	19.06	33.12

representation of each object, called D_{18}. Eighteen additional random views are then adopted from the remaining views in $\mathbf{V}_d^{1,j}$ to increase the accuracy of the database D_{18}, called D_{36}. Similarly, D_{54} and D_{72} are constructed using views in remaining $\mathbf{V}_d^{1,j}$. Additionally, D_{90} and D_{108} are further constructed with extra random views sampled from $\mathbf{V}_t^{1,j}$. Table 3 presents the average aspect numbers for each rigid object from 200 iterations. Although the aspect numbers increase when new views are employed to update the coarse database, the number of stored views remains significantly smaller than the number of original views. Figure 10 presents the recognition rate results obtained when using coarse to fine databases. Figure 11 presents the standard deviations for recognition rates. The recognition rate increases when aspect-graph representations are trained using additional object views. Moreover, stability increases based on decreasing standard deviation. Therefore, the proposed method is demonstrated as effective for updating aspect-graph representations without resorting the overall collected views, or recalculating overall similarity measures.

5.2. *Human posture recognition*

The efficiency of the proposed method is demonstrated using the second image database (Fig. 7). As the same preprocessing in the first experiment, object contour (the contour of human posture) was extracted for further utilization. In the second

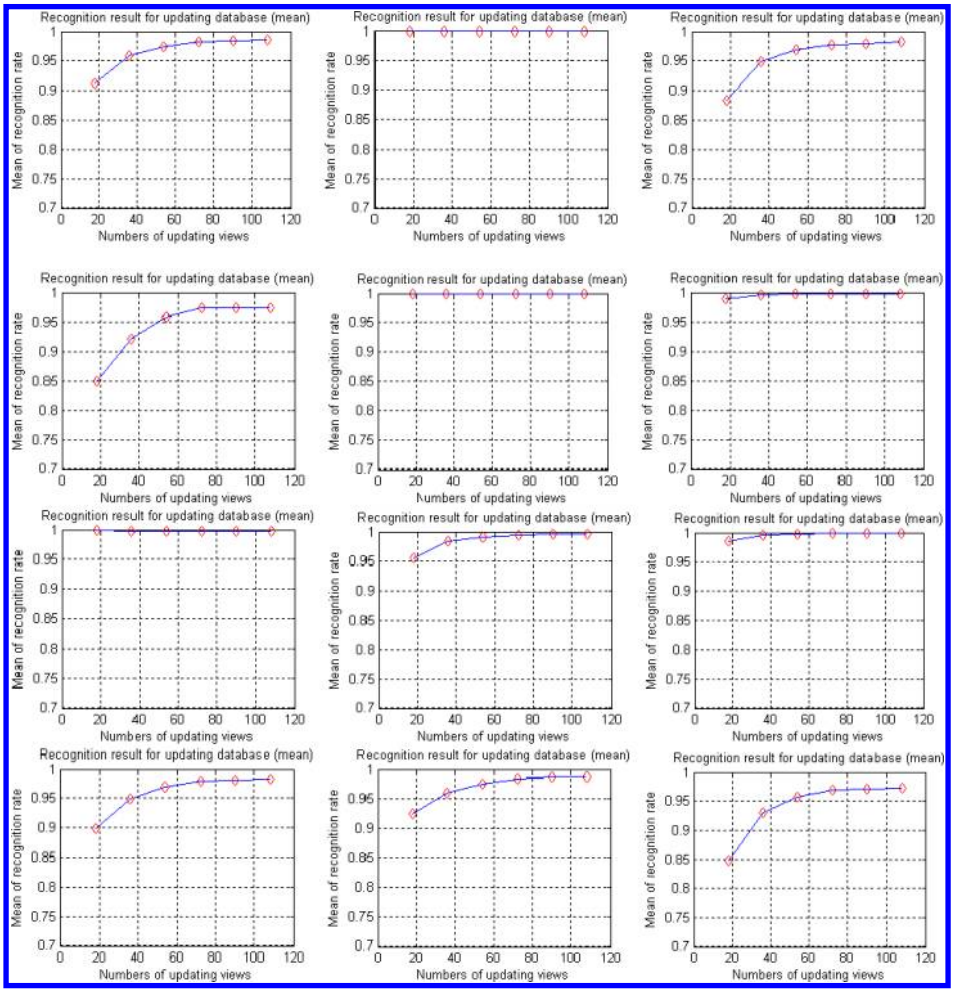


Fig. 10. Recognition rates of coarse and fine databases (D_{18}, D_{36}, D_{54}, D_{72}, D_{90} and D_{108}), calculated using 200 results.

experiment, two features, called the MAG and θ_z, are extracted from the object contour and used for building the AODs (Fig. 5). The characteristic views of aspects in each AOD are utilized for human posture recognition. Figure 12 illustrates the system architecture of the proposed framework for human posture recognition. Table 4 shows the efficiency of the proposed method with a high recognition rate.

The proposed method decreases the number of aspects for each human posture, and, thus, reducing the computing time for recognizing objects. Furthermore, adopting θ_z instead of PPL reduces the computing time. The similarity measure, which is based on posture contour with N points between an unknown posture and the posture in the database, requires computing N similarity distances while adopting PPL as the feature, but the similarity is computed only once while adopting θ_z.

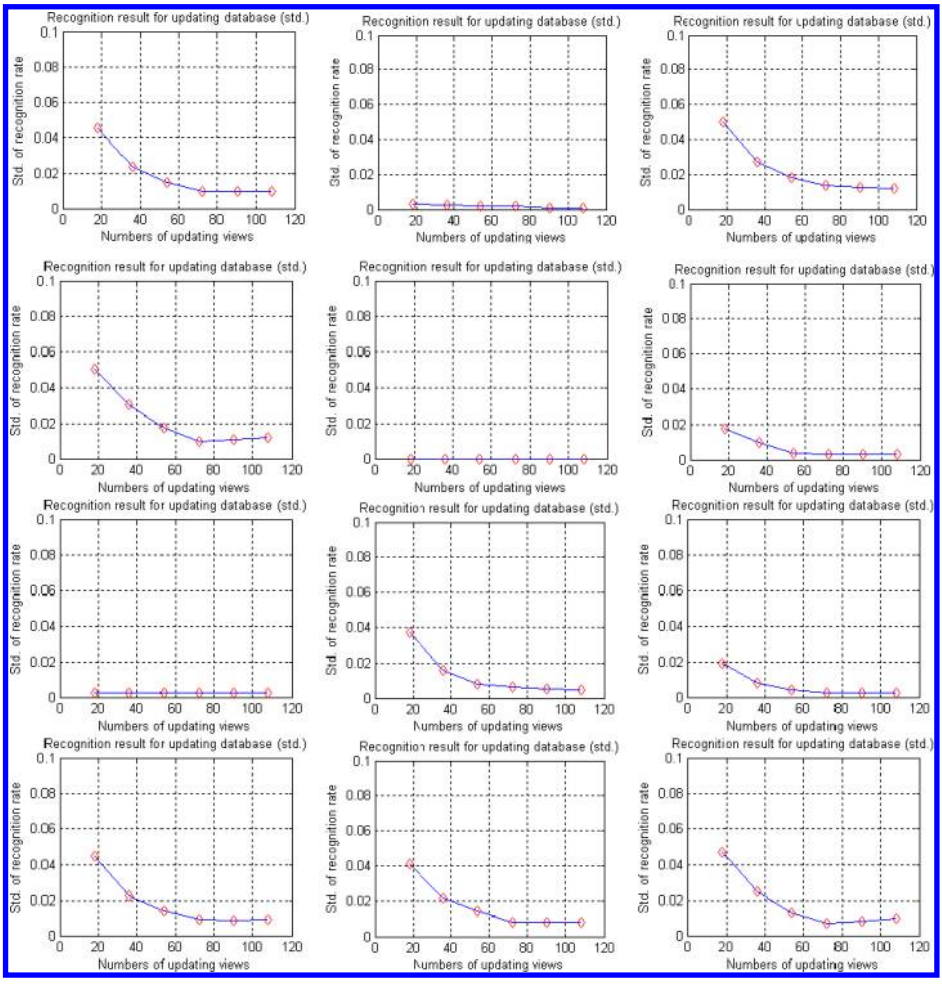


Fig. 11. Standard deviations of recognition rates using coarse to fine databases (D_{18}, D_{36}, D_{54}, D_{72}, D_{90} and D_{108}), calculated using 200 results.

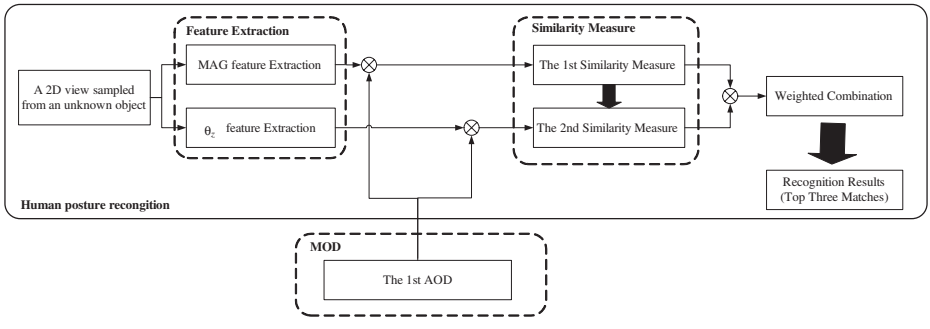


Fig. 12. The system architecture of the proposed framework applied during the second experiment (human posture recognition).

Table 4. Results of human posture recognition using 2D views via MAG and θ_z.

	The Index of the Postures in the Second Database Listed in Fig. 7								
Recognition Results	1	2	3	4	5	6	7	8	Avg.
Number of Aspects	8	25	37	41	42	38	8	38	29.63
Top 1 Match (%)	94.91	99.07	98.15	100	99.07	96.30	99.54	100	98.38
Top 2 Match (%)	99.07	99.54	100	100	100	99.54	100	100	99.77
Top 3 Match (%)	100	99.54	100	100	100	99.54	100	100	99.88

5.3. *Scene recognition*

In the third experiment, training images of 11 locations (Fig. 8) in an environment (Fig. 13) are obtained by rotating the PTZ camera from $-30°$ to $30°$ using one-degree increments at each location, thereby generating 61 images for each position. Furthermore, 12 Gaussian distributions are adopted in this work to build the blob model (BM feature). The number of aspects of each scene is below 13 after the combination processes. Figure 14 presents the sample training images, blob models and conceptual descriptions for each scene. Figure 15 illustrates the system architecture of the proposed framework for the scene recognition. Additionally, for the sake of illustration, the set of characteristic views at the sixth position in the indoor environment is cited as an instance of scene (Fig. 16).

To test the efficiency of the proposed method, test images are captured by a mobile robot moving in four directions (forward, backward, left and right) at five different distances (5 cm, 10 cm, 15 cm, 20 cm and 50 cm) and five different levels of occlusion (5%, 10%, 15%, 20% and 50%). Sixty-one images are captured at each position by rotating the camera from $-30°$ to $30°$ at one-degree increments with no occlusion. Figure 17 presents the test image samples captured at the sixth position. Figure 17(a) shows the test images captured in the forward and backward directions; Fig. 17(b) presents the test images captured in the left and right directions.

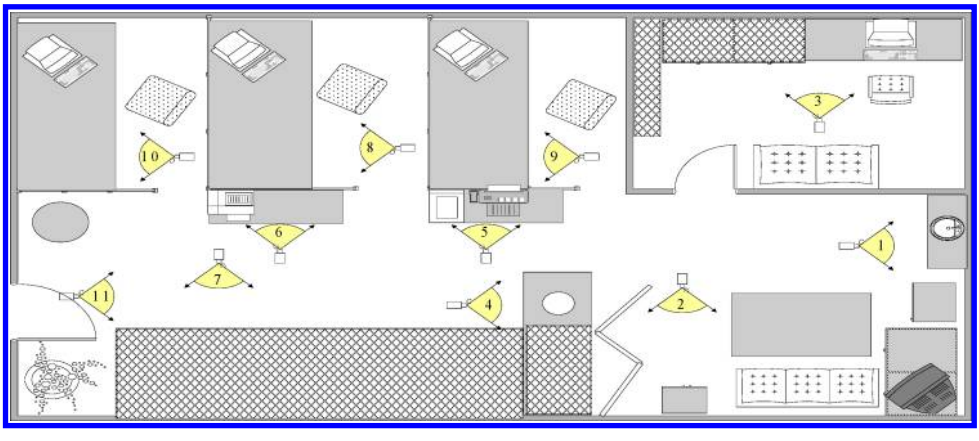


Fig. 13. The indoor environment from which scenes in the third database are obtained.

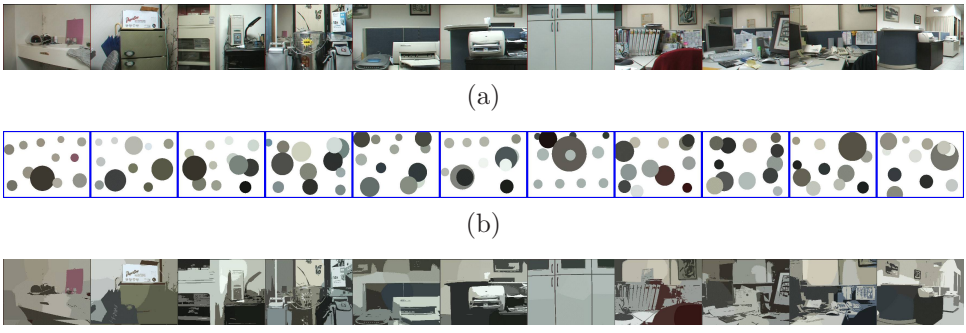


Fig. 14. The sample training image, blob model and conceptual description of each scene captured in the indoor environment (Fig. 13). (a) The sample image captured at each location in the indoor environment (from left to right are positions 1, 2, ..., 11). (b) The blob model of each sample captured image in (a) with 12 Gaussian distributions. (c) The conceptual description of each sample captured image in (a), which are calculated by comparing the original pixel values of each captured image with its blob model.

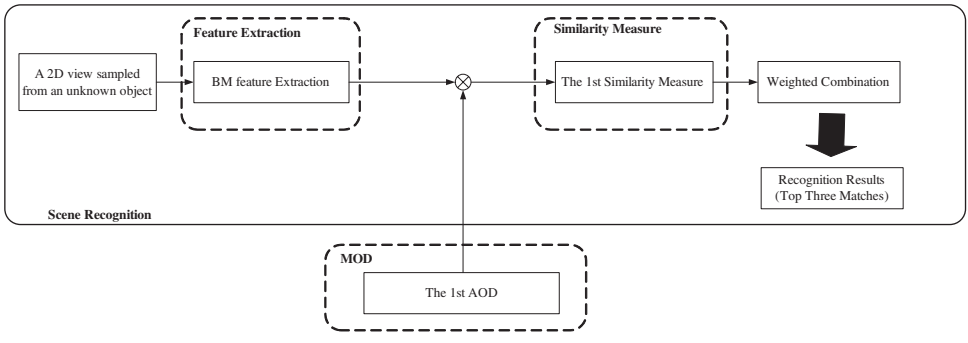
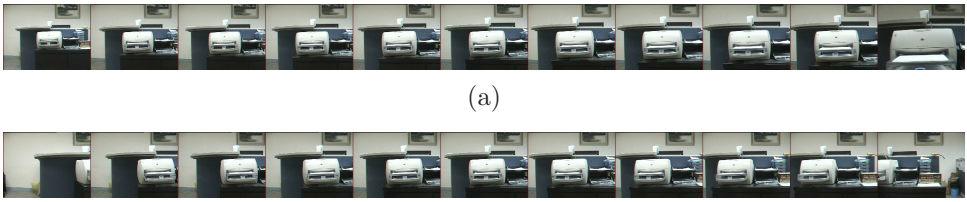


Fig. 15. The system architecture of the proposed framework applied during the second experiment (human posture recognition).

Fig. 16. The 11 characteristic views at the sixth position in the indoor environment.

To increase the robustness of scene cognition, multiple-view recognition is appropriate for testing. In this experiment, three arbitrary images I_i $(1 \leq i \leq 3)$ obtained with different rotating angles of the PTZ camera are utilized for scene recognition. The first three recognized results that have the first three minimum similarity measures are adopted as candidates for further processing. Suppose O is the set of recognized results defined as follows:

$$O = \{o_{ij}\}, \quad 1 \leq i \leq 3, \quad 1 \leq j \leq 3, \quad 1 \leq o_{ij} \leq 11$$



(a)

(b)

Fig. 17. The test images captured from the sixth position in the indoor environment. (a) The test images captured in the forward and backward directions; the shifted distances are as follows: backward 50 cm, backward 20 cm, backward 15 cm, backward 10 cm, backward 5 cm, 0 cm, forward 5 cm, forward 10 cm, forward 15 cm, forward 20 cm and forward 50 cm. (b) The test images captured in the left and right directions; the shifted distances are left 50 cm, left 20 cm, left 15 cm, left 10 cm, left 5 cm, 0 cm, right 5 cm, right 10 cm, right 15 cm, right 20 cm and right 50 cm.

where i is the index of the test image, and j is the index of the order of recognition result.

Three methods are proposed for estimating the final scene cognition result. The first result, R_1, is estimated using only one recognition result with only one captured image. The second result, R_2, uses the first three recognition result with only one captured image. The third result, R_3, uses all combinations of the first three recognition results with three captured images. The descriptions of R_1, R_2 and R_3 are derived by Eq. (31).

$$R_k = \begin{cases} o_{11}, & k = 1 \\ o_{11} \cdot \bar{D}_1 + r_1 \cdot D_1, & k = 2 \\ o_{11} \cdot (\bar{D}_1\, \bar{D}_2\, \bar{D}_3) + r_1 \cdot (D_1) + \bar{D}_1[r_2 \cdot (D_2) + \bar{D}_2(r_3 \cdot D_3)], & k = 3 \end{cases} \quad (37)$$

where

$$r_p = \arg\max(F_p), \quad 1 \le p \le 3$$

$$D_p = \begin{cases} 1, & \text{if argmax}(\mathrm{F}_p) \text{ exists} \\ 0, & \text{if argmax}(\mathrm{F}_p) \text{ does not exist} \end{cases} \quad 1 \le p \le 3$$

$$F_p = \{f_{pq}, 1 \le q \le 11\}, \quad f_{pq} = \sum_{j=1}^{3} \delta(q - v_{pj}).$$

Based on recognition results (Table 5), the recognition rates of the three methods are all above 95% when the level of occlusion is less than 20% and variation positions are below 20 cm. Although the level of occlusion is 50% and variation positions are 50 cm, recognition rates are still above 50%. Moreover, the third method, R_3, performs best and is reasonable based on the human vision used for localization. When a person enters an unknown place, multidirectional views are captured by the eyes to assist recall of past experiences of the unknown place. In this work, the same strategy is adopted to increase scene cognition robustness.

Table 5. Human posture recognition results using 2D views via BM with position variations and different levels of occlusion.

Shift Distance (cm) and Direction		Covering Rate (1.000 = 100%)														
		5%			10%			15%			20%			50%		
		R_1	R_2	R_3	R_1	R_2	R_3	R_1	R_2	R_3	R_1	R_2	R_3	R_1	R_2	R_3
	0 cm	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.800	0.769	0.817
5	Forward	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.797	0.775	0.809
	Backward	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	0.794	0.763	0.809
	Left	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.794	0.779	0.806
	Right	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.791	0.754	0.818
10	Forward	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.796	0.784	0.802
	Backward	0.997	0.979	1.000	1.000	0.997	1.000	1.000	0.997	1.000	0.997	0.996	1.000	0.785	0.738	0.802
	Left	1.000	0.993	1.000	1.000	0.996	1.000	1.000	0.996	1.000	1.000	0.993	1.000	0.796	0.772	0.805
	Right	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	1.000	1.000	0.999	1.000	0.770	0.747	0.817
15	Forward	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.784	0.769	0.797
	Backward	1.000	0.996	1.000	1.000	0.993	1.000	0.997	0.988	1.000	0.994	0.987	1.000	0.763	0.726	0.781
	Left	0.997	0.979	1.000	0.997	0.979	1.000	0.997	0.979	1.000	0.994	0.975	1.000	0.779	0.736	0.794
	Right	0.992	0.982	0.997	0.992	0.977	0.997	0.992	0.979	0.997	0.992	0.977	0.997	0.726	0.705	0.757
20	Forward	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.765	0.751	0.782
	Backward	0.999	0.987	1.000	0.994	0.982	1.000	0.991	0.979	1.000	0.988	0.976	1.000	0.733	0.711	0.748
	Left	0.976	0.960	0.987	0.979	0.970	0.993	0.975	0.961	0.979	0.975	0.963	0.979	0.748	0.699	0.776
	Right	0.975	0.955	0.984	0.979	0.970	0.993	0.976	0.951	0.984	0.975	0.951	0.984	0.714	0.694	0.744
50	Forward	0.845	0.838	0.854	0.845	0.844	0.849	0.842	0.829	0.845	0.832	0.815	0.839	0.508	0.503	0.523
	Backward	0.881	0.839	0.925	0.848	0.821	0.896	0.830	0.809	0.872	0.796	0.785	0.833	0.525	0.508	0.553
	Left	0.750	0.741	0.775	0.748	0.742	0.768	0.733	0.729	0.754	0.723	0.711	0.735	0.502	0.501	0.531
	Right	0.811	0.794	0.841	0.799	0.784	0.827	0.781	0.770	0.817	0.753	0.763	0.778	0.532	0.502	0.531

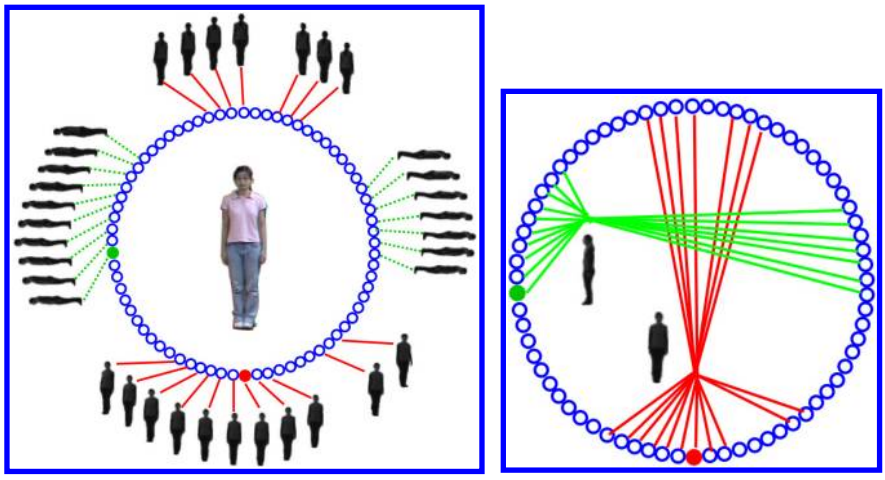
6. Conclusions

This study presents a flexible framework for recognizing 3-D objects by building aspect-graph representations using 2D views sampled at random intervals. The proposed framework comprises an incremental database construction method and a hierarchal weighting combination structure. A robust database, called a MOD, is composed of AODs. Each AOD is built using the incremental database construction method with one main feature or one main feature and one assistant feature. The final recognition result can be estimated by combining the results calculated from each AOD. To demonstrate the efficiency of the proposed framework, three various object recognition problems, including 3D object recognition, human posture recognition, and scene recognition are performed in the experiments.

Although the threshold values (T_3 and T_4) applied in the proposed incremental database construction method are determined manually case by case, the criteria for selecting T_3 and T_4 are described in Sec. 4.1. The selection of T_3 and T_4, which is a trade-off in this work, affects the number of aspects and thus affects the computing time and the error performance. Moreover, the feature selection plays an important role while applying the proposed method in different applications. Although the recognition rate decreases while the number of objects in the database increases in most applications, the proposed framework provides a hierarchical structure to combine more features to maintain the robustness of the recognition system.

Moreover, the proposed incremental database construction method is practical for extracting aspects when features of an object conflict with the first criterion, namely, local monotonicity, as indicated by Cyr and Kimia.[7] For instance, the combinational algorithm developed by Cyr and Kimia[7] cannot efficiently combine 2D views of a human posture with MAG. However, the proposed incremental database construction method overcomes this problem, and efficiently decreases the aspect number. In Fig. 18(a), the blue circles represent 2D views of a human posture, and the black human postures with the red and green lines connected to the blue circle represent the 2D views belonging to the same aspect. In Fig. 18(b), the black human postures are the characteristic views of the aspects of human postures. The two aspects in Fig. 18(b) clearly contain two clusters of 2D views that are opposites.

The proposed method decreases computing time when updating the aspects with new 2D views. Using the method proposed by Cyr and Kimia,[7] an object with N collected views requires a computing time of $N(N+1)/2$ to calculate the mutual similarity distances between the $(N+1)$ 2D views and to extract the aspects and characteristic views. However, the proposed method requires only a computing time of N times to calculate the similarity distance between new incoming views and N existing views via the proposed method. However, as the proposed method has a high computation requirement, improving its efficiency is a topic for future works.



(a) (b)

Fig. 18. The aspect-graph representation of the first human posture listed in Fig. 6 via MAG only. (a) The similar 2D views of two aspects. (b) The characteristic views of two aspects.

Acknowledgments

This work was supported by National Science Council of the R.O.C. under grant no. NSC94-2218-E009064 and DOIT TDPA Program under the project number 95-EC-17-A-04-S1-054.

References

1. K. Akita, Image sequence analysis of real world human motion, *Patt. Recogn.* **17**(4) (1984) 73–83.
2. M. Bessa, A. Coelho, J. B. Cruz and A. Chalmers, Selective presentation of perceptually important information to aid orientation and navigation in an urban environment, *Int. J. Patt. Recogn. Artif. Intell.* **20**(4) (2006) 467–482.
3. V. Blanz, M. J. Tarr and H. H. Bultho, What object attributes determine canonical views? *Perception* **28** (1999) 575–599.
4. J. Canny, A computational approach to edge detection, *IEEE Trans. Patt. Anal. Mach. Intell.* **8**(6) (1986) 679–698.
5. G. Cicirelli, T. D'Orazio and A. Distante, Different learning methodologies for vision-based navigation behaviors, *Int. J. Patt. Recogn. Artif. Intell.* **19**(8) (2005) 949–975.
6. R. Cucchiara, C. Grana, A. Prati and R. Vezzani, Probabilistic posture classification for human-behavior analysis, *IEEE Trans. Syst. Man Cybern.* **35**(1) (2005) 42–54.
7. C. M. Cyr and B. Kimia, A similarity-based aspect-graph approach to 3D object recognition, *Int. J. Comput. Vis.* **57**(1) (2004) 5–22.
8. C. de Trazegnies, C. Urdiales, A. Bandera and F. Sandoval, 3D object recognition based on curvature information of planar views, *Patt. Recogn.* **36**(11) (2003) 2571–2584.
9. Q. Delamarre and O. Faugeras, 3-D articulated models and multi-view tracking with silhouettes, *IEEE Conf. Comput. Vis.* (September 1999), pp. 716–721.
10. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Stat. Soc.* **39**(1) (1977) 1–38.

11. G. N. Desouza and A. C. Kak, Vision for mobile robot navigation: a survey, *IEEE Trans. Patt. Anal. Mach. Intell.* **24** (2002) 237–267.
12. A. Diplaros, T. Gevers and I. Patras, Color-shape context for object recognition, *IEEE Workshop on Color and Photometric Methods in Computer Vision (in conjunction with ICCV 2003)*, Nice, France (2003).
13. A. Diplaros, T. Gevers and I. Patras, Combining color and shape information for illumination-viewpoint invariant object recognition, *IEEE Trans. Imag. Process.* **15**(1) (2006) 1–11.
14. I. Haritaoglu, D. Harwood and L. S. Davis, Ghost: a human body part labeling system using silhouettes, *Proc. Int. Conf. Patt. Recogn.* **1** (1998) 77–82.
15. J. S. Hu, T. M. Su and S. C. Jen, Robust background subtraction with shadow removal for indoor environment surveillance, *Proc. IEEE IROS*, China (October 2006).
16. S. Iwasawa, K. Ebihara, J. Ohya and S. Morishima, Real-time human posture estimation using monocular thermal images, *IEEE Int. Conf. Automatic Face and Gesture Recognition* (April 1998), pp. 492–497.
17. H. Jiang, Z. N. Li and M. S. Drew, Recognizing posture in pictures with successive convexification and linear programming, *IEEE. Trans. Multimed.* **14**(6) (2007) 26–37.
18. S. Kim, G. J. Jang, W. H. Lee and I. S. Kweon, Combined model-based 3D object recognition, *Int. J. Patt. Recogn. Artif. Intell.* **19**(7) (2005) 839–852.
19. T. K. Kim, J. Kittler and R. Cipolla, Discriminative learning and recognition of image set classes using canonical correlations, *IEEE Trans. Patt. Anal. Mach. Intell.* **29**(6) (2007) 1005–1018.
20. J. J. Koenderink and A. J. van Doorn, The singularities of the visual mapping, *Biol. Cybern.* **24** (1976) 51–59.
21. A. Kosaka and A. C. Kak, Fast vision-guided mobile robot navigation using model-based reasoning and prediction of uncertainties, *Comput. Vis. Graph. Imag. Process. — Imag. Underst.* **56**(3) (1992) 271–329.
22. B. J. A. Kröse, N. Vlassis, R. Bunschoten and Y. Motomura, A probabilistic model for appearance-based robot localization, *Imag. Vis. Comput.* **19** (2001) 381–391.
23. P. Lamon, A. Tapus, E. Glauser, N. Tomatis and R. Siegwart, Environmental modeling with fingerprint sequences for topological global localization, *IEEE Int. Conf. Intell. Robots and Systems* (October 2003), pp. 3781–3786.
24. C. C. Lin, Shape memorization and recognition of 3-D objects using a similarity-based aspect-graph approach, Department of Electrical and Control Engineering, National Chiao-Tung University, Hsinchu, Taiwan, R.O.C., Master thesis (June 2005).
25. P. C. Lin, Human posture recognition system using 2-D shape features, Department of Electrical and Control Engineering, National Chiao-Tung University, Hsinchu, Taiwan, R.O.C., Master thesis (June 2006).
26. B. MacQueen, Some methods for classification and analysis of multivariate observations, *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, Berkeley (University of California Press, 1967), Vol. 1, pp. 281–297.
27. G. Mamic and M. Bennamoun, Representation and recognition of 3D free-form objects, *Dig. Sign. Process.* **12** (2002) 47–76.
28. S. Mian, M. Bennamoun and R. A. Owens, Three-dimensional model-based object recognition and segmentation in cluttered scenes, *IEEE Trans. Patt. Anal. Mach. Intell.* **28**(10) (2006) 1584–1601.
29. S. K. Naik and C. A. Murthy, Distinct multi-colored region descriptor for object recognition, *IEEE Trans. Patt. Anal. Mach. Intell.* **29**(7) (2007) 1291–1296.

30. L. B. Ozer, T. Lu and W. Wolf, Design of a real-time gesture recognition system: high performance through algorithms and software, *IEEE Sign. Process. Mag.* **22** (2005) 57–64.
31. L. B. Ozer and W. Wolf, Real-time posture and activity recognition, *Proc. IEEE Workshop, Motion and Video Computing* (2002), pp. 133–138.
32. G. Peters, Theories of three-dimensional object perception — a survey, *Recent Research Developments in Pattern Recognition*, Transworld Research Network (2000).
33. Y. Rubner, C. Tomasi and L. J. Guibas, The earth mover's distance as a metric for image retrieval, *Int. J. Comput. Vis.* **40**(2) (2000) 99–121.
34. H. Schneiderm and T. Kanade, Object detection using the statistics of parts, *Int. J. Comput. Vis.* **56**(3) (2004) 151–177.
35. T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber and T. Poggio, Robust object recognition with cortex-like mechanisms, *IEEE Trans. Patt. Anal. Mach. Intell.* **29**(3) (2007) 411–426.
36. Y. Shan, H. S. Sawhney, B. Matei and R. Kumar, Shapeme histogram projection and matching for partial object recognition, *IEEE Trans. Patt. Anal. Mach. Intell.* **28**(4) (2006) 568–577.
37. I. Shimshoni and J. Ponce, Finite-resolution aspect graphs of polyhedral objects, *IEEE Trans. Patt. Anal. Mach. Intell.* **19**(4) (1997) 315–327.
38. A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele and L. Van Gool, Towards multi-view object class detection, *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, New York, USA (June, 2006).
39. I. Ulrich and I. Nourbakhsh, Appearance based place recognition for topological localization, *IEEE Conf. Robotics and Automation* (November 2000), pp. 1023–1029.
40. I. Weiss and M. Ray, Model-based recognition of 3D objects from single images, *IEEE Trans. Patt. Anal. Mach. Intell.* **23**(2) (2001) 116–128.
41. N. Werghi and Y. Xiao, Recognition of human body posture from a cloud of 3-D data points using wavelet transform coefficients, *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition* (2002), pp. 70–75.
42. C. Wren, A. Azarbayejani, T. Darrell and A. Pentland, Pfinder: real-time tracking of the human body, *IEEE Trans. Patt. Anal. Mach. Intell.* **19**(7) (2007) 780–785.
43. C. Xu and J. L. Prince, Snakes, shapes, and gradient vector flow, *IEEE Trans. Imag. Process.* **7**(3) (1998) 359–369.
44. P. Yan, S. M. Khan and M. Shah, 3D model based object class detection in an arbitrary view, *IEEE Int. Conf. Computer Vision (ICCV)*, Rio de Janeiro, Brazil (October 2007).

Jwu-Sheng Hu received the B.S. degree from the Department of Mechanical Engineering, National Taiwan University, Taiwan, in 1984, and the M.S. and Ph.D. degrees from the Department of Mechanical Engineering, University of California at Berkeley, in 1988 and 1990, respectively. He is currently a Professor in the Department of Electrical and Control Engineering, National Chiao-Tung University, Taiwan, R.O.C.

His current research interests include microphone array signal processing, active noise control, intelligent mobile robots, embedded systems and applications.

Tzung-Min Su received the B.S. degree in electrical and control engineering from National Chiao Tung University, Taiwan, R.O.C in 2000. He is currently a Ph.D. candidate in the Department of Electrical and Control Engineering at National Chiao Tung University, Taiwan, R.O.C. He was awarded the championship at the national competition held by the Ministry of Education Advisor Office in 2001.

His research interests include background subtraction, 3D object recognition, and home-care surveillance.