Recognition of 3-D Objects in Multiple Statuses Based on Markov Random Field Models

Ying Huang *Xiaoqing Ding* *Shengjin Wang*
Department of Electronic Engineering, Tsinghua University, Beijing, China
hy@ocrserv.ee.tsinghua.edu.cn

ABSTRACT

A general framework is presented to realize 3-D object recognition invariant to object scaling, deformation, rotation, occlusion, and viewpoint change. This framework utilizes densely sampled grids with different resolutions to represent the local information of the input image. A Markov random field (MRF) model is then created to model the geometric distribution of the object key nodes. Flexible matching, which is aim to find the accurate correspondence map between the key points of two images, is performed by combining the local similarities and the geometric relations together using the highest confidence first (HCF) method. Afterwards, a global similarity is calculated for object recognition. Experimental results on Coil-100 object database are presented. The excellent recognition rates achieved in all the experiments indicate that our approach is well-suited for appearance-based recognition.

1. INTRODUCTION

In real-world scenes, the representation of a 3-D object may be modified due to multiple factors such as: i) object scale, viewpoint and illumination variations; ii) partial occlusion; iii) noise data; iv) object deformation. Human can distinguish different objects easily without considering these variations. However, this is a quite difficult task for computer vision systems. Most object recognition approaches aim to find object features and then match these features between the observed data and the object databases. Generally, this problem can be traced back to establish the relation between two images. In the following we will present a novel method to solve this visual correspondence problem.

There are numerous research efforts dealing with the object recognition problem and the existing approaches can mainly be classified into two distinct categories: global feature based approaches and local feature based approaches. The global feature based approaches extract global image features such as color histograms [1], and receptive field histograms [2]. Global features are robust to scale and viewpoint changes, but it has been difficult to extend them to partially occluded and noise images. Methods based on support vector machines [3, 4], SNoW (Sparse Network of Winnows) [9], and eigenspace matching [5] can handle images corrupted by noise and partially occlusions successfully, but fail to recognize viewpoint variant or heavily occluded objects. The local feature based approaches identify local feature points, and then create a local image descriptor at each interest point [6, 7]. Object matching is performed by finding similar point pairs between test images and training models. Advantages of the local image features are that they are only partially affected by object occlusion, viewpoint modification, and deformation. However, their recognition results depend on the accuracy of the point detection. Points detected in one object may be missed in another image of the same object. Furthermore, no geometric constraint between interest points is utilized to rectify the matching results.

Our method uses dense local features that sampled at a large number of repeatable locations to represent objects. Markov random field models are established to model the geometric constraints between object key points. The matching program is composed of two procedures: local matching and global matching. Local matching procedure calculates the similarities between the key points of two images. In the global matching procedure, the highest confidence first method is introduced to reach a local minimum of the MRF model. This final result decides which pairs of key points correspond to the same point and which points have no corresponding partner in the other image. Our object recognition method is evaluated using the Coil-100 object database containing 7200 image of 100 objects. Different numbers of images are selected as training examples, and the remaining as test images. We also test our method on the images corrupted by synthetically generated rotation, scaling and occlusions. The remarkable recognition rates show the potential of our approach in the problem of recognizing 3D objects.

The remainder of the paper is organized as follows: The following section presents our approach in detail. Section 3 describes the local and global matching procedures. Some experiments and comparisons are given in Section 4. We conclude this paper in Section 5.

2. GENERAL FRAMEWORK

2.1. Motivation

As mentioned above, objects can be represented by local features. These local features are computed at some interest points, which can be extracted using corner detectors. Object matching is performed by comparing the features of the key points between two images. However, the detected positions of these key points will be influenced by image transformations. Thus for two images of the same objects, some points in the first image cannot find their accurate partners in the second one. In addition, the geometric positions of the key points have some relations even after complex transformations. For example, considering a key point in the first image, the accurate partners of its neighbor key points are also located in the neighborhood of its partner in the second image. This information should be considered.

To solve these problems, this paper introduces two kinds of topological grids. The first kind of grid is utilized to extract local image features, while the other grid combine these local features with the geometric constraints between grid nodes. Unlike previous approaches, our method does not identify key locations. Thus we do not have to consider the accuracy of point detection, and the algorithm complexity is decreased. In our framework, the

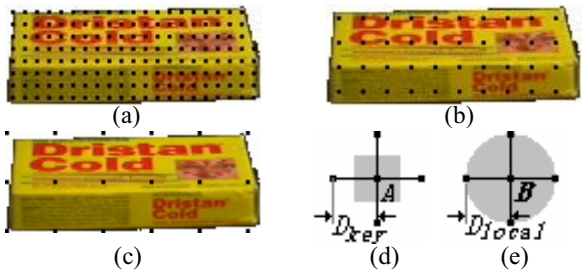


Fig.1 (a), (b): two level local grid. (c): key grid. (d): neighbourhood of the key node *A* (gray regions). (e): neighbourhood of the local node *B* (gray regions).

interest points are equidistantly distributed on different levels, and these points are sampled densely enough to cover most of the object details.

2.2. MRF-based Framework

We call the first kind of grid "local grid", and the second kind of grid "key grid". The nodes of these grids are equidistantly sampled. The local grid is composed of several levels of local nodes (Fig. 1a-b). The distance between two adjacent local nodes of a sampling level k is defined as $D_{local}(k)$ (Fig. 1e). The key grid is introduced to perform the match between two images (Fig. 1c). D_{key} is the distance between two adjacent key nodes (Fig. 1d):

$$D_{key} = F(image)/V_{stad} \quad (1)$$

V_{stad} is a standard value that decides the resolution of the key grid, and $F(image)$ is calculated from some physical parameters of the input image, such as the image pixel number, or the image bounding box. $D_{local}(k)$ is formulated as:

$$D_{local}(k) = D_{key} \times 0.5 \times k / K \quad (2)$$

where K is the maximum level number. For example, K is equal to 2 in Fig. 1.

Since D_{key} is much larger than $D_{local}(k)$, in the neighbourhood of each key node (Fig. 1d), there are several local nodes. These local nodes are used to calculate the similarity between two key nodes in the matching procedure.

Given two images of the same objects, assume that the neighbourhood of two key nodes A and B correspond to the same region on the physical object. The corresponding partners of the neighbour key nodes of A should also be adjacent to the node B. On the other hand, if we have already determined the partners of the neighbour nodes of A, then the coordinate of B depends mainly on two factors: one is the similarity between A and B, and the other is the positions of these partners. Therefore, this issue can be modeled by a Markov random field model. As shown in Fig.2, a 12-node neighbourhood and its associated clique are defined in our MRF model. More details of the potential calculation and the relaxation algorithm will be described in Section 3.2.

2.3. Color Histograms

Researchers have developed many local image descriptions, such as color histograms, Gabor features, differential invariants, and scale invariant features. To accelerate the test experiments, we take use of the color histogram description.

Color histograms are popular used in many applications because they are trivial to compute, and robustly tolerate image transformations and changes in camera viewpoint. We quantize colors into a set of K representative colors $C = \{c_1, c_2, \ldots, c_K\}$. For each local node, we compute the color histogram feature using its neighbor pixels.

Two color histograms H_1 and H_2 are compared by computing

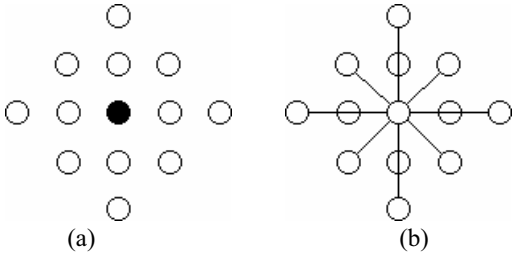


Fig.2 12-node neighbourhood (a) and its associated clique (b).

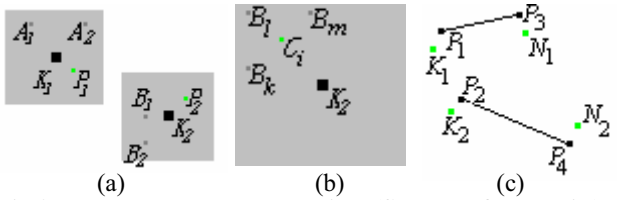


Fig.3 Local and global matching (See text for details). (a), (b): matching between two key nodes K_1 and K_2. (c): geometric potential calculation.

their intersection, an idea introduced by Swain and Ballard [1]. The intersection is

$$I(H_1, H_2) = \sum_{i=1}^{K} \min[H_1(i), H_2(i)] \quad (3)$$

This intersection is equal to a value between 0 and 1. We define this data as the similarity S_{local} between two local nodes.

3. TEMPLATE MATCHING

3.1. Local Matching

The objective of the local matching procedure is to find the similarity S_{key} between two key nodes. This data is calculated from the similarities between the local nodes that located in the neighbourhood of the two key nodes.

In Fig.3a, two key nodes K_1 and K_2 are displayed. Local nodes located in their neighbourhood are signed as (A_1, A_2, …, A_N) and (B_1, B_2, …, B_M), respectively. $S_{local}(i, j)$ is defined as the similarity between the local node A_i and B_j. To obtain S_{key}, we first extract data S_{Ai}, i = 1,2,…, N for A_1, A_2, …, A_N, which describes the similarities between these nodes and their optimum matching points in K_2's neighbourhood. Since A_1, A_2, …, A_N are uniformly distributed, S_{key} can then be estimated from these data.

The largest value of $S_{local}(i, j)$, j = 1, 2, …, M can be selected as S_{Ai}. However, this method will incur matching errors. Since the local nodes are equidistantly sampled, usually the best matching point of A_i is not sampled as a local node in the second image. For example, in Fig.3b, C_i is the best point. B_l, B_k, and B_m are its neighbour nodes. The three nodes can only serve as A_i's approximate partners. Nevertheless, since the local nodes of our framework are densely distributed, the local features of B_l, B_k, and B_m are similar to the feature of C_i. Hence we use not only the largest similarity $S_{local}(i, m_1)$, but also the second and the third largest value $S_{local}(i, m_2)$, $S_{local}(i, m_3)$ to calculate S_{Ai}:

$$S_{Ai} = \min[S_{local}(i, m_1) + S_{local}(i, m_2)/4 + S_{local}(i, m_3)/6, \quad 1.0] \quad (4)$$

The coordinate of C_i are also reconstructed:

$$\mathbf{X}_{Ci} = S_{local}(i, m_1)\mathbf{X}_{B\,m1} + S_{local}(i, m_2)\mathbf{X}_{B\,m2} + S_{local}(i, m_3)\mathbf{X}_{B\,m3} \quad (5)$$

Based on the same principle, the coordinates of the two key nodes K_1 and K_2 may be incorrectly matched. Thus we integrate the coordinates of C_i (i = 1,2,…, N) with S_{Ai} to calculate the positions of the two matching point P_1 and P_2 (Fig.3a):

$$X_{P1} = \sum_{i=1}^{N} S_{Ai} X_{Ai} / \sum_{i=1}^{N} S_{Ai}; X_{P2} = \sum_{i=1}^{N} S_{Ai} X_{Ci} / \sum_{i=1}^{N} S_{Ai} \quad (6)$$

To compute the local similarity S_{key}, we have to consider all the neighbour pixels of K_1. Each point may be located in the neighbour regions of multiple local nodes. Among the similarity of these nodes, we can obtain a maximum data. Then S_{key} is formulated as:

$$S_{key} = \sum_{D \in \mathbf{N}(K_1)} \max_{D \in \mathbf{N}(Ai), i=1,2,..N} (S_{Ai}) / N_{po\,\mathrm{int}} \quad (7)$$

Here D is a neighbour pixel, $\mathbf{N}(K_1)$ denotes the neighbour points of K_1, $\mathbf{N}(A_i)$ denotes the neighbour points of A_i. $N_{po\,\mathrm{int}}$ is the number of pixels in $\mathbf{N}(K_1)$. Note that the matching point of K_1 may be another one if the matching destination changes form K_2 to another key node. In addition, if S_{key} between two key nodes is lower than a threshold, we judge that the two key nodes are dissimilar. In this case, the two key nodes are directly set as the best matching points.

3.2. Global Matching

Given two images $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$, the local matching procedure has calculated the similarity between every two key nodes of the two images. The global matching procedure combines these results with $\boldsymbol{I}_1$'s MRF model to decide the corresponding map between the key nodes of $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$.

Our MRF model is composed of:

- a set of sites $S = \{s_1, s_2, \ldots, s_n\}$. Each site corresponds to a key node of $\boldsymbol{I}_1$.
- a set of possible labels for each site $\Omega_i \subset \Omega = \{l_1, l_2, \ldots, l_m\}, i = s_1, s_2, \ldots s_n$. These labels correspond to the key nodes of the second image $\boldsymbol{I}_2$.

Then we have the conditional posterior potential for each site:

$$E_i(\Omega_i \mid s_i) = V(\Omega_i \mid s_i) + \sum_{c:i \in c} V_c(s_i \mid s_{Ni}) \quad (8)$$

where $c : i \in c$ means any clique c containing the site s_i. Assume that $\Omega_i = l_j$, then $V(\Omega_i \mid s_i)$ is the negative of the similarity between the key nodes s_i and l_j of the two images. The other item gives the geometric potential of the current node. The item contains only one component because only one clique is defined (Fig.2b). s_{Ni} defines the twelve neighbour nodes of the site s_i.

Assume that in Fig.3c, the key node K_1 corresponds to the site s_i, K_2 corresponds to the label l_j. P_1 and P_2 are their matching point. N_1 is a neighbour node of K_1. Its accurate partner is N_2. P_3 and P_4 are their matching points. Two parameters can be obtained from the lengths and directions of the two lines P_1P_3 and P_2P_4:

$$r(P_1P_3, P_2P_4) = \frac{Len(P_1P_3) \times D_{key}(K_2)}{Len(P_2P_4) \times D_{key}(K_1)}$$
$$ddir(P_1P_3, P_2P_4) = dir(P_1P_3) - dir(P_2P_4) \quad (9)$$

where D_{key} is defined by (1). The two data describe the length ratio and direction difference between the two lines. If K_2 is K_1's accurate partner, for each one of the twelve neighbour nodes of K_1, the two parameters should remain approximately unchanged. Furthermore, the length ratio is close to 1.0. Hence the potential $V_c(s_i \mid s_{Ni})$ is formulated as:

$$V_c(s_i \mid s_{Ni}) = -\exp[-\mathrm{var}\_ddir \times w1 - \mathrm{var}\_r \times w2 - (avg\_r - 1) \times (avg\_r - 1) \times w3] \quad (10)$$

Here $w1$, $w2$ and $w3$ are three weights, avg_r is the average of the twelve ratios, var_r is their variance, and var_ddir is the variance of the twelve direction differences.

The highest confidence first (HCF) algorithm [8] is a deterministic relaxation technique which can be converged to a local minimum of the MRF, but induce drastically less computational cost than a stochastic relaxation scheme. The HCF algorithm classifies the sites into two classes: "committed" and "uncommitted". Initially all sites are set uncommitted, and a site has no effect on its neighbours unless it has committed. Thus in Equation (9, 10), only committed sites are used to compute V_c.

A stability measure is calculated for each site based on the local conditional posterior potential defined in (8). This measure determines the order in which the sites are to be visited. Details of this algorithm have been described in [8]. The procedure terminates when the criterion function can no longer be decreased by reassignment of the labels.

After global matching, we obtain a posterior potential for each key node. The negative of the potential describe the local and geometric similarity between the current node and its partner. Therefore, the average value of these data is used to define a global similarity for the following object recognition experiments.

4. EXPERIMENTS

Our object recognition algorithm is test on the COIL (Columbia Object Image Library) [5] database. The COIL database consists of 7,200 images of 100 objects (72 views for each of the 100 objects). The images are color images of 128×128 pixels. The objects are positioned in the center of a turntable and observed from a fixed viewpoint. For each object, the turntable is rotated of 5° per image (Fig.4a-b).

We test the proposed system using all the 100 objects. Different numbers of views are selected as the training set. For each object, the numbers of training images vary from 4 (one every 90°), 8 (one every 45°), 18 (one every 20°) to 36 (one every 10°). The remaining images compose the test sets. The image pixel number is selected to calculate $F(image)$ in (1). In addition, since the resolution of the key grid is decided by the standard value V_{stad}, performance of the framework with different V_{stad} is tested. Comparisons between our method and other approaches are list in Table 1 and 2.

Table 1 Recognition rates using different training sets

Training images	400	800	1800	3600
Test images	6800	6400	5400	3600
Our framework (color&V_{stad} = 25)	**95.75%**	**99.30%**	**100.0%**	**100.0%**
Our framework (color&V_{stad} = 9)	93.66%	97.89%	99.41%	99.98%
SNoW (edges)[9]	88.28%	89.23%	94.13%	96.25%
SNoW (intensity)[9]	81.46%	85.13%	92.31%	95.81%
Nearest neighbor (color)	79.74%	93.08%	98.91%	99.89%
Linear SVM (color) [3]	83.99%	95.36%	99.31%	100.0%

(a) (b) (c) (d)

Fig.4 Coil-100 object database. (a), (b): two images. (c): a synthesized image under scaling and rotation. (d): a synthesized image occluded by a randomly placed $k \times k$ window of uniformly distributed random noise (k = 48).

Table 2 Comparisons between our framework and other methods [4] (Training images: 400)

Representation	Our Framework (V_{stad} =25)	Columbia [5]	Roobaert [4]
Color only	**95.75%**	77.5%	82.3%
Shape&color	---	87.6%	86.9%

In Table 1, the image representations of the associated methods are displayed in the parentheses. Results of SNoW are cited from [9]. Results of the methods in Table 2 are cited from [4]. For the nearest neighbor classifier and support vector machines [3], we reduce the image spatial resolution from 128×128 to 32×32 and then transform each COIL image into an eight-bit vector of 32×32×3 = 3072 components.

Our method performs recognition with excellent percentages of success even in the presence of very similar objects. The ability of handling viewpoint changing is much higher than the other sited methods. The average match time between two images is no more than 30ms for the low-resolution framework (V_{stad} = 9), and 200ms for the high-resolution framework (V_{stad} = 25) on a P4 1.7GHz computer. However, it still took us quite a long time to finish the test of the high-resolution framework. Therefore, the low-resolution framework is adopted to continue the following test, and eight images for each object are selected as the training set.

Our method is invariant to image shifting, scaling and rotation. This ability is evaluated using synthetically generated images (Fig. 4c). The 7200 synthesized images are tested on the training set, and the recognition rate is 97.94%.

In order to verify the systems against occlusion, test images corrupted by generated occlusions are also synthesized (Fig. 4d). Table 3 lists the results of our method and the SVM-based method. From the obtained experimental results, we conclude that our method achieves very good rates even half pixels of the images are occluded.

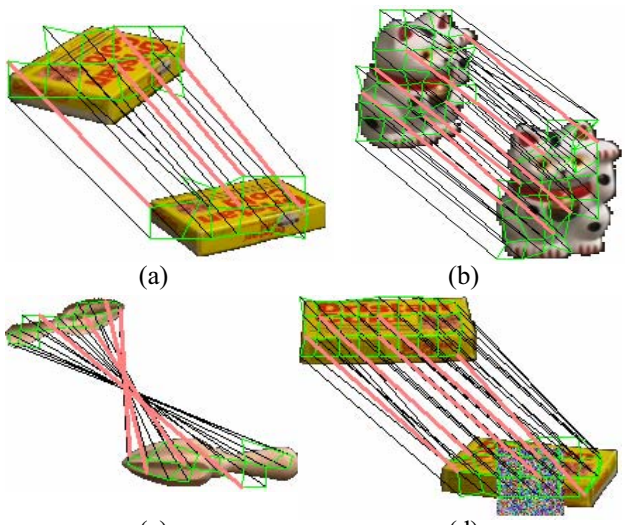
Table 3 Recognition rates for COIL images occluded by a randomly placed $k \times k$ window of uniformly distributed random noise

k	24	48
Our method	**97.93%**	**96.72%**
Linear SVM [3]	95.32%	84.99%

A shortcoming of our object recognition algorithm is the ability against noise corruption. However, this drawback is mainly caused by the color histograms. Gabor features, which are robust to additive Gaussian noise, will be used to further enhance our system.

Finally we show some correspondence maps between the key nodes of different COIL images in Fig.5.

5. CONCLUSIONS



(a) (b) (c) (d)

Fig.5 Key node correspondence maps (in dark lines) between different images of the same objects. The matching points of two adjacent key nodes are linked by green lines. Some mapping lines are emphasized in red color. (a), (c): V_{stad} = 9. (b), (d): V_{stad} = 25.

The main contribution of this paper is to introduce a general framework, which can combine the local image descriptions with the geometric structure of an object together to establish point correspondence maps between different images. Markov random field is introduced to model this framework. Object recognition is then performed from these maps. This method is successfully tested on the Coil-100 image database. The remarkable experimental results indicate that this approach is well-suited for 3-D object recognition robust under viewpoints changing, occlusion, rotation, and scaling. More efforts will be done to further extend this method to detect objects in complex environments.

REFERENCES

[1] M. J. Swain, and D. H. Ballard. "Color Inexing." In *IJCV*, 1991, 7(10):11-32.

[2] B. Schiele, and J. L. Crowley. "Object Recognition using multidimensional receptive field histograms." In *Proc. ECCV*, 1996.

[3] M. Pontil, and A. Verri. "Support vector machines for 3D object recognition." In *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1998, 20(6):637-646.

[4] D. Roobaert, and M. Van Hulle. "View-based 3D object recognition with support vector machines." In IEEE *international workshop on neural networks for signal processing*, 1999.

[5] Murase, Hiroshi, and S. K. Nayar. "Visual learning and recognition of 3-D objects from appearance." In *IJCV*, 1995, 14 (1): 5-24,.

[6] D. G. Lowe. "Object recognition from local scale-invariant features." In *Proc. ICCV*, 1999.

[7] D. Jugessur, and G. Dudek. "Local appearance for robust object recognition." In *Proc. CVPR*, 2000.

[8] P. B. Chou, and C. M. Brown. "The theory and pratice of Bayesian image labeling." In *IJCV*, 1990, 4(3): 185-210.

[9] M. H. Yang, D. Roth, and N. Ahuja. "Learning to Recognize 3D Objects with SNoW." In *Proc. ECCV*, 2000.