21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Hierarchical Multilevel Object Recognition Using Markov Model

Muhammad Attamimi, Tomoaki Nakamura, and Takayuki Nagai
Dept. of Mechanical Eng. and Intelligent Systems, The University of Electro-Communications
1-5-1 Chofugaoka Chofu-shi, Tokyo, 182-8585 Japan
{*m_att, naka_t*}*@apple.ee.uec.ac.jp, tnagai@ee.uec.ac.jp*

Abstract

In this study, we address the issue on multilevel object recognition. The multilevel object recognition is object recognition in various levels, that is, simultaneous recognition of its instance, category, material, etc. At each level, many recognition methods have been proposed in the literature. Therefore it is straightforward to design a multilevel object recognition system using conventional methods independently. However, these "levels" are related each other and form hierarchical structure. Hence the recognition performance can be improved by considering consistency of the recognition results at all levels. To model the consistency, we formulate the problem as finding the Viterbi path in a Markov model, since the consistent recognition results can be thought of as the most likely sequence of the states. We implemented the proposed multilevel object recognition system and evaluated it to show validity.

1. Introduction

With developments in technology, robots have been used in a variety of environments for various purposes in recent years. The domestic service robots can carry out various tasks by employing a visual recognition system in various levels, such as instances, categories, and materials. In the cleaning task, for instance, the robot should recognize material of the object as well as its category and/or the instance in order to separate garbage. Of course, a number of research have been made on object recognition [1]. However, few researchers address the multilevel object recognition. In fact, the recognition result at each level is not independent but deeply related each other. Moreover, the relationship is in a form of hierarchical structure.

This paper presents a system which enables object recognition in various levels (i.e. *instances*, *categories*, and *materials*) consistently. Multiple cues, such as colors, textures, 3D point clouds, and NIR reflection intensities are adaptively incorporated to construct the recognition system. A probabilistic representation of the hierarchical object structure is introduced using a Markov model (MM) since the consistent recognition results are considered as the most likely sequence of the states in the MM framework. Each state of the proposed MM corresponds to an object class to be recognized and has the emission probability representing the likelihood of the state. Moreover, the emission probability, representing the tendency of false classifications, is modeled by a Gaussian distribution that improves the recognition performance as a whole.

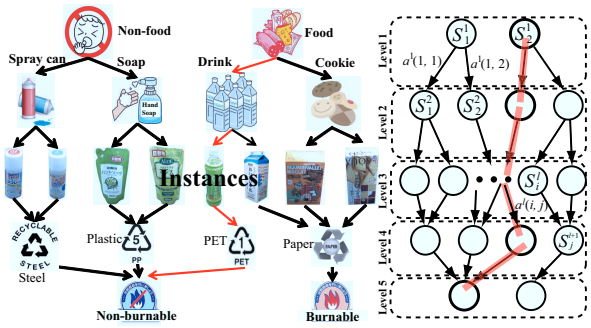


Figure 1: Hierarchical structure of objects and corresponding Markov model for multilevel object recognition.

Hierarchical object recognition has been proposed in [2] and [3]. However, [2] only deals with a single feature, and the consistency of recognition is not taken into consideration. In [3], a tree-based object recognition which enables instance, category, and pose recognition has been proposed. However, the main issue of [3] is the pose recognition rather than the multilevel object recognition.

2. Proposed Method

2.1. Overview

Figure 1 illustrates the idea of multilevel object recognition. From the figure one can see analogy between the multilevel object recognition and the Markov model (MM). Hence the problem can be formulated as finding the most likely path. In the figure, each circle (state s_i^ℓ) represents a class to be recognized. Each level

has a classifier whose results, i.e. scores, are used for calculating the emission probability of each state. The emission probability is modeled by a Gaussian distribution so that the tendency of classification errors is taken into consideration. Although any classifier is applicable to this model, k-NN based classifier, which will be explained below, is involved in this paper. Each arc has a transition probability, which encodes the hierarchical relationship among object classes.

2.2. Vision Sensor

In this paper, a 3D visual sensor [4] is used. This sensor consists of a TOF and two CCD cameras that can capture color information and 3D point clouds in real time. Moreover, NIR reflection intensities can also be acquired from the TOF camera. Hence, colors, textures, 3D point clouds, and NIR reflection intensities, are used for implementing the object recognition system as shown in Fig. 2.

2.3. Object Extraction

Object extraction is required as the first step in the learning and recognition phases. Here we assume two different methods, which use the information acquired from the 3D visual sensor. *The first method* is motion-attention based object extraction [5]. This method uses a motion detector for extracting an initial object region, and then the object region is refined using color and depth information of the initial region. *The second method* is plane detection based object extraction. If objects are on the table, the plane detection technique is beneficial to detect the objects. The 3D randomized Hough transform is utilized for fast and accurate plane detection.

2.4. Feature Extraction

Color information is calculated as a color histogram of hue and saturation in HSV color space. *Texture information* is represented by the Bag of Keypoints (BoK). We utilize dense scale invariant feature transform (DSIFT) for better result. Before taking a histogram, DSIFT is vector quantized using a predetermined 500 dimensional codebook, which is generated from many indoor images by k-means clustering. *Shape information* is represented by shape distribution (SD). SD represents characteristics of the object's shape by calculating a metric among vertices. We use distances between all combinations of two vertices in the object region, followed by taking a histogram of these distances as the object feature. *Material information* is calculated as a histogram of NIR reflection intensities. Since the NIR reflection intensity is varied according to the distances, the compensated value is used for calculating the histogram.

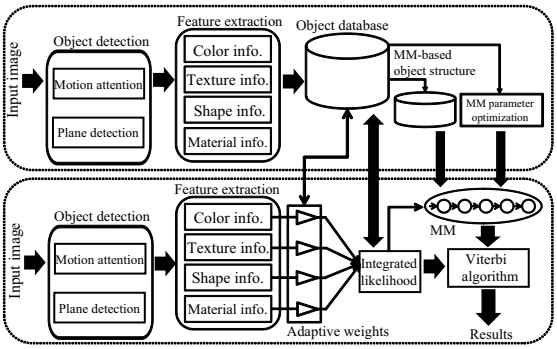


Figure 2: Hierarchical object recognition system.

2.5. Hierarchical Multilevel Object Recognition

The proposed system is divided into learning and recognition phases as shown in Fig. 2. In the learning phase, an object database, consisting of multiple feature vectors in various views, is generated. The consistency model based on MM is also learned in this phase. In the recognition phase, the extracted target object is classified at each level independently, and then the scores are used for finding a Viterbi path, which corresponds to a final recognition result. This process allows the system to correct the false classification at each level.

The MM has parameters such as, initial, transition, and emission probabilities. Learning of these parameters is carried out by the following procedures.

Learning of Emission Probability Now, let the classification results, i.e. scores, at level ℓ be $\boldsymbol{x}_c^\ell = \{x_1^\ell, x_2^\ell, \cdots, x_{C_\ell}^\ell\}$, where C_ℓ is the number of classes at level ℓ. x_c^ℓ represents a score for the class c at level ℓ, which can be obtained by $x_c^\ell \propto \exp\{-\lambda(\bar{d}_c^\ell)^2\}$. λ is a predetermined coefficient for adjusting the variance of the distances. $\bar{d}_c^\ell$ represents an average of top-k smallest distances between the input data and the reference of class c at level ℓ:

$$d_c^\ell = \sum_f \frac{w^f \{D_f(\boldsymbol{h}_c^f, \boldsymbol{h}_{\rm in}^f)\}^2}{\sigma_f^2}, \tag{1}$$

where f represents a type of features $f \in \{\mathrm{color}, \mathrm{texture}, \mathrm{shape}, \mathrm{material}\}$. w^f is a weight for the feature f, which is determined adaptively considering the environments and similarity among objects in the database. w^f is normalized as $\sum_f w^f = 1$. σ_f^2 represents a variance of distances for the feature f, which is calculated by the cross validation over the database. $\boldsymbol{h}_{\rm in}^f$ is a histogram of the feature f for a given target object, while $\boldsymbol{h}_c^f$ represents a reference histogram, which belongs to the class c in the object database. Here, $D_f(\boldsymbol{h}_c^f, \boldsymbol{h}_{\rm in}^f)$ is the Bhattacharrya distance between these histograms.

The emission probability of each state is modeled as a Gaussian distribution over the classification scores.

The parameters of the Gaussian distribution of the state c at level ℓ, $\boldsymbol{\theta}_c^\ell = (\boldsymbol{\mu}_c^\ell, \boldsymbol{\Sigma}_c^\ell)$ can be estimated by

$$\boldsymbol{\mu}_c^\ell = \frac{1}{M_c} \sum_{m=1}^{M_c} \boldsymbol{x}_{(m,c)}^\ell, \tag{2}$$

$$\boldsymbol{\Sigma}_c^\ell = \frac{1}{M_c} \sum_{m=1}^{M_c} (\boldsymbol{x}_{(m,c)}^\ell - \boldsymbol{\mu}_c^\ell)(\boldsymbol{x}_{(m,c)}^\ell - \boldsymbol{\mu}_c^\ell)^\top, \tag{3}$$

where M_c is the number of objects that belong to the class c at level ℓ. $\boldsymbol{x}_{(m,c)}^\ell$ represents the score distribution of the m-th object. Thus, for the given $\boldsymbol{x}^\ell$, the emission probability of state c at level ℓ can be obtained by

$$b_c^\ell = \frac{\exp\{-\frac{1}{2}(\boldsymbol{x}^\ell - \boldsymbol{\mu}_c^\ell)^\top (\boldsymbol{\Sigma}_c^\ell)^{-1} (\boldsymbol{x}^\ell - \boldsymbol{\mu}_c^\ell)\}}{(2\pi)^{C_\ell/2} |\boldsymbol{\Sigma}_c^\ell|^{1/2}}. \tag{4}$$

Optimization of Transition Probability The EM algorithm is not required in this particular case, since all state transitions are known for the training data. Intuitively, the transition probabilities encodes object structure, which can be easily calculated by counting the number of objects in each class. However, we can further optimize the transition probabilities by minimizing the likelihood of all paths except for the correct one considering the emission probabilities so that the probability of the Viterbi path is maximized. For given emission probabilities of m-th input $\boldsymbol{o}_m = \{b_{(m,c_1)}^1, b_{(m,c_2)}^2, ..., b_{(m,c_T)}^T\}$, the likelihood of the correct path $P_m^\ell(i,j)$ transitioning from state i to state j at level ℓ can be written as

$$\begin{aligned} P_m^\ell(i,j) &= \pi_{(m,c_1)} b_{(m,c_1)}^1 a^1(c_1,c_2) b_{(m,c_2)}^2 \cdots \\ &\quad \times a^\ell(i,j) \cdots b_{(m,c_T)}^T = A_m^\ell(i,j) a^\ell(i,j). \end{aligned} \tag{5}$$

The total likelihood P_m^{all} can be calculated using the forward coefficient $\alpha_m^\ell(i)$ and the backward coefficient $\beta_m^{\ell+1}(j)$ as follows:

$$P_m^{\text{all}} = \sum_{i,j} \alpha_m^\ell(i) \beta_m^{\ell+1}(j) b_{(m,j)}^{\ell+1} a^\ell(i,j). \tag{6}$$

Here, the likelihood of the correct path can be maximized by minimizing the following difference $\xi_m^\ell(i,j)$

$$\xi_m^\ell(i,j) = \left\| P_m^{\text{all}} - P_m^\ell(i,j) \right\|^2. \tag{7}$$

Applying (7) to M training data, the cost function can be written as

$$\begin{aligned} \xi_{\text{all}}^\ell &= \sum_{m=1}^{M} \xi_m^\ell(i,j) = \sum_{m=1}^{M} \left\| \boldsymbol{\gamma}_m^{\ell\top}(i,j) \boldsymbol{a}^\ell \right\|^2 \\ &= \boldsymbol{a}^{\ell\top} \boldsymbol{\Gamma}^\ell \boldsymbol{a}^\ell, \end{aligned} \tag{8}$$

where $\boldsymbol{a}^\ell$ represents a transition vector, whose elements are nonnegative and normalized to $\sum_k a^\ell(i,k) = 1$. $\boldsymbol{\gamma}_m^\ell$ is a vector of length $C_\ell \times C_{\ell+1}$. These vectors are defined as

$$\boldsymbol{a}^\ell = (a^\ell(1,1) \cdots a^\ell(1,C_{\ell+1}) \cdots a^\ell(C_\ell, C_{\ell+1}))^\top,$$

$$\boldsymbol{\gamma}_m^\ell(i,j) = \begin{pmatrix} \alpha_m^\ell(1) \beta_m^{\ell+1}(1) b_{(m,1)}^{\ell+1} \\ \vdots \\ \alpha_m^\ell(i) \beta_m^{\ell+1}(j) b_{(m,j)}^{\ell+1} - A_m^\ell(i,j) \\ \vdots \\ \alpha_m^\ell(C_\ell) \beta_m^{\ell+1}(C_{\ell+1}) b_{(m,C_{\ell+1})}^{\ell+1} \end{pmatrix}. \tag{9}$$

Therefore, the problem can be formulated as follows:

$$\begin{aligned} \min \boldsymbol{a}^{\ell\top} \boldsymbol{\Gamma}^\ell \boldsymbol{a}^\ell, \ \text{s.t.}\ \boldsymbol{y}^{(\ell,i)\top} \boldsymbol{a}^\ell = 1, \ a^\ell(i,j) \geq 0, \\ 1 \leq i \leq C_\ell, \ 1 \leq j \leq C_{\ell+1}, \end{aligned} \tag{10}$$

where $\boldsymbol{y}^{(\ell,i)}$ is a vector consisting of zero vectors $\mathbf{0}^{\ell+1}$ and $\mathbf{1}^{\ell+1}$ (all elements are 1) of length $C_{\ell+1}$,

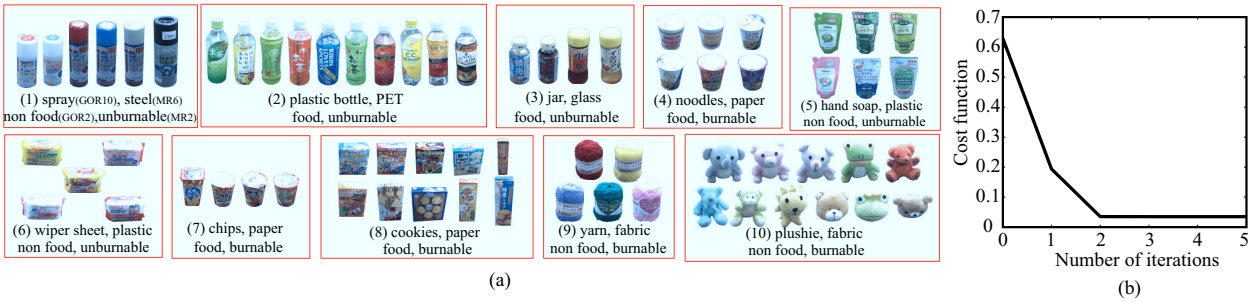
$$\boldsymbol{y}^{(\ell,i)} = (\mathbf{0}_1^{\ell+1\top} \ \mathbf{0}_2^{\ell+1\top} \ \cdots \ \mathbf{1}_i^{\ell+1\top} \ \cdots \ \mathbf{0}_{C_\ell}^{\ell+1\top})^\top. \tag{11}$$

The problem can be solved by a quadratic programming solver. The above procedure is for a fixed level ℓ. Therefore the optimization is carried out by iterating the above procedure with respect to ℓ until convergence. It should be noted that the algorithm converges to a local minimum depending on the initial values. We use the initial values that are calculated by counting the number of objects in each class.

Recognition Phase In the recognition phase, the feature vectors of input data are compared to the object database to perform classification at each level. The classification results are used to form the score distribution $\boldsymbol{x}^\ell$ as in the learning phase. Then, we can calculate the emission probability in state c at level ℓ, i.e. b_c^ℓ, according to (4).

The final result of hierarchical multilevel object recognition can be obtained by finding the most likely path for a given data $\boldsymbol{X} = \{\boldsymbol{x}^1, \boldsymbol{x}^2, \cdots, \boldsymbol{x}^T\}$. The emission probabilities are calculated from (4). The Viterbi algorithm is involved in obtaining the most likely path.

Unknown Objects So far we assume that the target object is known at all levels. In this situation, the proposed model works for improving the recognition performance by considering consistency of all levels. On the other hand, for an unknown object instance, it is important to recognize its category and/or material. However, direct application of the model to this situation suffers from negative effects on final recognition due to



(a) (b)

Figure 3: (a) Objects used in the experiment with hierarchical categories, and (b) value of the cost function vs. number of iterations.

unreliable classification results at the instance level. To cope with this problem, all emission probabilities are set to 1 in all states of the level if all classification scores are less than a predefined threshold at the level. This procedure avoids the negative effects and allows to pursue the consistency among reliable classification results.

3. Experiment

3.1. Experimental Setup

An experiment was conducted using 67 objects in Fig.3(a). Five levels, i.e. categories with 2 and 10 classes, materials with 2 and 6 classes, and instances with 67 classes are considered in this experiment. Each object instance has 36 images from different view points. Leave-one-instance-out was used to evaluate the proposed algorithm. Therefore, the target object is always unknown for the system at the instance level. For comparison, the independent recognition (independent method) was also carried out using k-NN at each level.

3.2. Experimental Result

Values of the cost function (8) against the number of iterations are plotted on Fig. 3 (b). From the figure, one can see that the cost decreased monotonously and the optimization converged within a few iteration.

Recognition results are given in Fig. 4. Please note that the instance classification gives very low score, since the target object is unknown for the system at the instance level. Hence all emission probabilities of the states belonging to the instance level are set to 1 as mentioned in the previous section. From Fig. 4, it can be seen that the proposed method outperforms the independent method at all levels except for the material recognition with 2 classes. This is because the proposed method considers consistency of all classification results. Moreover, distributions of classification scores, which encode the tendency of false classification at each level, work reasonably well in this experiment. The average recognition rates over all levels are 81.8% (independent) and 90.8% (proposed) respectively.

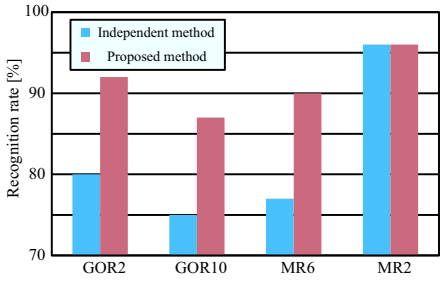


Figure 4: Results of multilevel object recognition. "GORxx" and "MRxx" stand for category and material recognition, followed by the number of classes.

4. Conclusion

In this paper, we proposed a novel multilevel object recognition system (instances, categories, and materials) based on Markov model with integrated features. We showed that the recognition performance can be improved by considering consistency of the classification results at all levels. The results given in this paper indicates that the proposed system improves the multilevel object recognition performance compared to the independent recognition. The pursuit of optimal combination of classifiers and features is left for the future research.

References

[1] K. Lai *et. al.*, "A Large-Scale Hierarchical Multi-View RGB-D Object Dataset", in Proc. of ICRA, pp.1817–1824, May 2011.

[2] A. Dhua *et. al.*, "Hierarchical, Generic to Specific Multi-class Object Recognition", in Proc. of ICPR, pp.783–788, Aug. 2006.

[3] K. Lai *et. al.*, "A Scalable Tree-based Approach for Joint Object and Pose Recognition", in Proc. of AAAI, Aug. 2011.

[4] M. Attamimi *et. al.*, "Real-Time 3D Visual Sensor for Robust Object Recognition," in Proc. of IROS, pp.4560–4565, Oct. 2010.

[5] M. Attamimi *et. al.*, "Learning Novel Objects Using Out-of-Vocabulary Word Segmentation and Object Extraction for Home Assistant Robots", in Proc. of ICRA, pp.745–750, May 2010.