

Edge Orientation-based Multi-view Object Recognition

Weiyu Zhu

Dept. of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign
weiyuzhu@ifp.uiuc.edu

Stephen Levinson

sel@ifp.uiuc.edu

Abstract

An edge orientation-based algorithm for multi-view object recognition is presented in this paper. The distribution of edge point orientations, combined with the normalized second moments, is taken as a feature vector to describe and index each object instance. For each unknown testing object, a set of likelihood weights for all the possible candidate objects is obtained by computing the Euclidean distances between the unknown feature set and all the available template feature vectors. A convincing coefficient is introduced to evaluate the confidence of the best match. New views (photo shots) will be automatically taken if the best match is thought to be insufficiently convincing. In the experiments, our algorithm has achieved an average of 91.5% correct recognition rate under the 5-view scheme for 320 testing images taken from eight natural objects.

1. Introduction

3-D object recognition from images is a challenging research topic in the field of computer vision. Over years, many approaches have been developed to solve this problem but at least so far, a general method is still not available. Each specific method is tailored to deal with some certain kind of problems and no single method will work in all situations. Some approaches are based on the study of the 3-D structure of objects, such as [3, 8], in which a complete 3-D model for each object is built. Upon recognition, each candidate object model is examined to match with the unknown object image and the decision is made based on some probability hierarchy structure, such as a decision tree or graph. Appearance feature-based object recognition [5, 6] is another popular recognition methodology, in which each object is directly mapped to some feature sets corresponding to their possible projective images. The biggest issue with the feature-based recognition methods is that the projective images of each object may change greatly under different viewing angles and illumination conditions. Therefore,

finding a feature set invariant to the change of viewing conditions is a “hot” research topic in recent years [4, 7].

For both 3-D structure-based and appearance feature-based object recognition, a pre-built object database is necessary for most recognition approaches. Such a requirement may impose a great limitation in real applications. For example, in our autonomous robot learning project, one of our goals is to teach the robot to learn the objects around him, as if teaching a child (or infant) to understand the world. In this case, no prior knowledge is available and the robot should learn, or in other words, memorize, the objects introduced to him simply by a series of interactions with his supervisor. The more the robot is taught, the more he would learn and the more sophisticated his database would become.

This paper focuses on the study of single object recognition under simple backgrounds (slightly noisy edges are allowed). The system takes the distribution of edge point orientations and the normalized second moments of objects as feature vectors to build up the object database on-line through the interaction between human and computer (robot). For the recognition, each candidate object is assigned a likelihood weight according to the minimum distance between the unknown feature vector and its possible instance. The best match is evaluated with a convincing coefficient, which is defined to reflect the confidence of the correctness in the best match. A new observation from another random view angle will be taken if the best match can not provide sufficient confidence, which is our so-called multi-view decision scheme.

The remaining part of this paper is organized as follows: Section 2 will give a detailed description of our algorithm, including the feature extraction and the multi-view recognition scheme. Section 3 describes the experiments we have done with a set of eight natural objects. Finally, a brief conclusion of our work will be given in the last section.

2. Algorithm Description

Our object recognition system consists of feature extraction and multi-view recognition two parts.

2.1. Feature Extraction

The feature extraction procedure is shown in Figure 1.

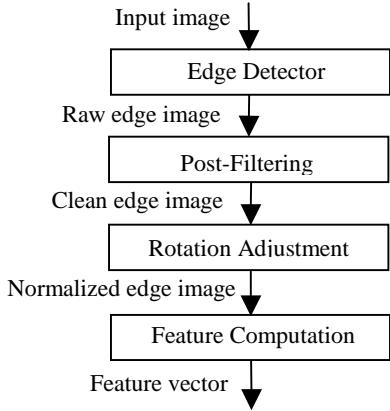


Figure 1. Procedure of feature extraction

Canny edge detection [2] is applied to generate an edge image for each object instance. A post-filtering procedure is introduced to remove all the unrelated background edges that do not belong to the object of interest. The post-filtering algorithm works under two assumptions about the “noisy edge image”: (1) Most edge points in the noisy edge image belong to the object. (2) The noisy edges do not occur at the center the object and they do not intersect with object edges. Object orientation is computed [2] based on the obtained clean edge image and then, the object would be rotated (normalized) to a certain predefined orientation, e.g., along the vertical axis. Two features are extracted from the normalized object edge image:

- **Distribution of edge point orientations.** Each edge point is assigned an orientation, which might be one of the four values: 0° , 45° , 90° and 135° . The distribution of edge point orientations is defined as:

$$D_\alpha = \frac{N_\alpha}{N} \quad (1)$$

Where N_α is the number of edge point with the orientation of α and N is the total number of edge points in the object image. The values of D_α imply the information of edge orientations of the object.

- **Normalized second moments in the horizontal and vertical directions.** Let \bar{x} and \bar{y} be the center coordinates of the object and $x' = x - \bar{x}$, $y' = y - \bar{y}$ be the relative coordinates of each edge point. Define the normalized second moments a , b [1] as:

$$a = \frac{\iint (x')^2 dx' dy'}{(\Delta x)^2}, \quad b = \frac{\iint (y')^2 dx' dy'}{(\Delta y)^2} \quad (2)$$

Where Δx and Δy are the width and height of the object contour, respectively. The values of a and b may provide us with some shape information about the objects. For example, larger values of a indicate more edge points distributed far away from the object center in the horizontal direction.

The distribution of edge point orientations is actually a histogram, in which four bins, corresponding to the four directions, are contained, and the normalized second moments only have two values, a and b . If we consider the object as a whole, then six floating point numbers are sufficient for the description. However, in order to provide more information for the recognition, we further divide each image into four regions: *top-left*, *top-right*, *bottom-left* and *bottom-right*, relative to the geometric center of the object. Independent feature sets for every regions are computed and combined together to form the ultimate feature vector of 24 floating point values for each object instance. A weight ratio of the two features is assigned when computing the Euclidean distance in the recognition stage.

2.2. Multi-view Object Recognition and Learning

Each object instance, represented by a feature vector and its corresponding object identity index, is a record in our object database. The database is initially set to be empty. Afterward, for each test object sample, a procedure of recognition (Figure 3), followed by the database adjustment (Figure 4), is carried out.

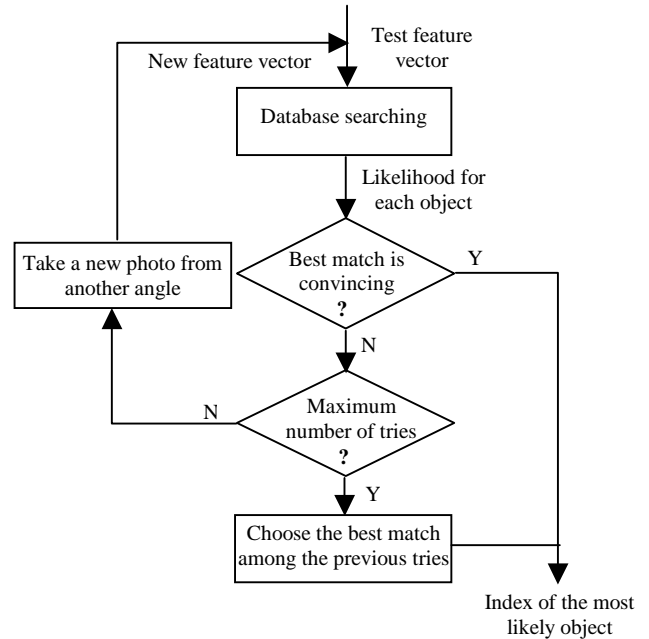


Figure 2. Flow chart of multi-view object recognition

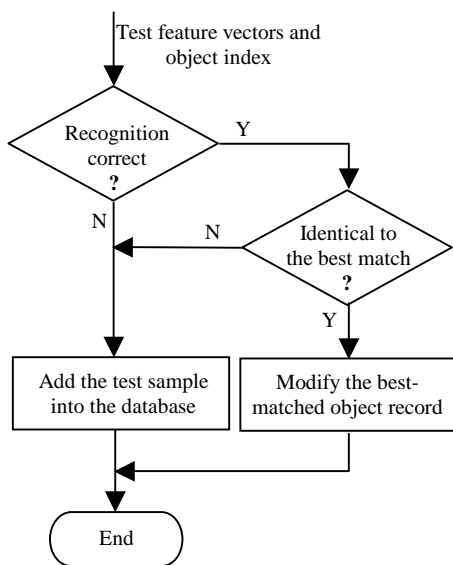


Figure 3. Flow chart of object database adjustment

In the object recognition process, feature vectors of unknown objects are compared with all the records in the database and for each object class i , a best match with the smallest Euclidean distance, denoted as d_i , is obtained. Define the likelihood weight w_i for each object class as:

$$w_i = \frac{d_{\min}^2}{d_i^2} \quad (3)$$

where d_{\min} is the smallest distance among all the candidate objects. Obviously, we have $w_i \leq 1$ where the best match has a weight equal to 1.

For the best-matched object, a convincing coefficient c is computed as:

$$c_i = \frac{w_{\text{best}}^{(i)}}{w_{\text{second}}^{(i)}} \quad (4)$$

where i indicates the sequential number of the try and w_{second} is the second largest likelihood weight among all the objects for the current view. If the value of c_i exceeds a pre-set threshold, e.g., 2.0, the best match is thought to be convincing and the corresponding object identity is accepted. Otherwise, the recognition decision can not be made right away and a new photo from another random view angle will be taken and another round of matching would occur. The matching and evaluating processes will be repeated until a convincing best match is found or the pre-determined maximum number of tries is reached. In case that no decision can be made after a certain number of views, the best match with a highest convincing coefficient among the previous tries will be accepted. Obviously, the multi-view decision procedure is somewhat similar to the decision process of human being. That is, if

the information provided by the current view is too weak to make a decision, the human observer may try again from another view angle. In case that none of the best matches in several tries can reach the decision threshold, the most convincing one would be chosen.

Following the recognition stage, a database adjustment process is performed, where the testing feature vector may be linked into the database as a new object instance. However, if the recognition is correct and the testing sample is “identical” to the corresponding best match (the distance between them is small enough), no new record will be built. In this case, we simply replace the record of the best-matched instance by the mean value of the two feature vectors. With such a scheme, the size of the database will not unlimitedly increase even if a large amount of recognition tasks is done, as long as the size of the object set is limited.

3. Experiments

In our experiment, eight items: a bag, a cup, a stapler, a joystick, a remote controller, a pair of sunglasses, a wallet and a data link are selected as the set of objects (Figure 4).



Figure 4. Training object set

Each object is randomly picked and observed from a random view angle. The goal of the experiment is to “teach” the computer (robot) to learn these objects simply based on what it observes. At the beginning of the experiment, the system database is empty. After that, for each testing image, the system tries to decide which object it contains. If the best match is not convincing, a new photo will be automatically taken. The system will adjust its knowledge base correspondingly based on the correctness of the recognition.

320 images of the eight objects are picked as unknown test samples in the experiment. Under the multi-view schemes, not necessarily all of the 320 object samples would result in a decision, because some of them may act as extra samples in case that the best matches in the previous tries are not sufficiently convincing. The feature weight ratio of the distribution of edge point orientations over the normalized second moments is set to be 2:1. Figure 5 depicts the curve of the incorrect recognition rates (averaged over 12 experiments) of the 320 test object

samples versus the different pre-defined maximum number of views.

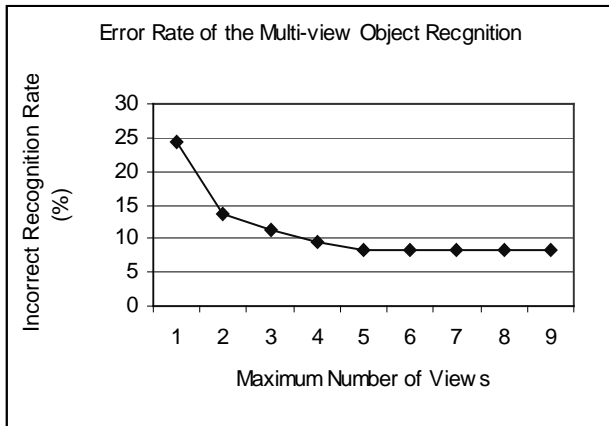


Figure 5. Error rate of the object recognition algorithm

The figure shows that the recognition error rate decreases with the changes of the maximum number of views from 1 to 5, and then keeps stable at a rate of about 8.5% for 5-9 views. In addition to this figure, we have other two observations regarding the experimental results:

- The typical number of decisions made under 5-view scheme is about 180, which means that the total number of errors in this case is about 15 (with 8.5% error rate). Since the object database is empty at the beginning of the experiment, an error must occur for the first test sample of each object. We call these errors as “compulsory errors”, which mean that the error is inevitable and independent of the performance of the recognition algorithm. For our case, if we do not take the compulsory errors into account, the number of errors may decrease to about 7 in 170 recognition decisions.
- Most recognition errors occur on the first several test samples and few errors are made after about 15 samples of each object under the 5-view scheme and 25 samples under the single view scheme. This means that, for each object, an average of 15 training samples under 5-view scheme is nearly sufficient to distinguish the object from others. That is, $15 \times 24 = 360$ floating point numbers is almost enough to describe any of the eight candidate objects.

4. Conclusions

In this paper, we present an edge orientation-based multi-view object recognition algorithm. The distribution of edge point orientations and the normalized second moments are introduced to describe object shapes. A feature vector of 24 floating point numbers is used to index object instance and an average correct recognition rate of 91.5% is achieved under the ≥ 5 -view schemes in the experiment. The features of our approach can be briefly summed as follows:

- No prior object knowledge required
- Immune to the changes of illumination, camera rotations and translations
- Small size of feature set (24 floats/record) and real-time processing
- Multi-view decision scheme similar to the decision process of human being

5. References

- [1] B.K.P.Horn, “Binary Images” in *Robot Vision*, MIT Press, 1986, Chapters 3
- [2] Emanuele trucco, Alessandro verri, *Introductory techniques for 3-D computer vision*, pp. 71-78, Prentice Hall, Inc. New Jersey 1998
- [3] S. J. Dickinson, H. I. Christensen, J. K. Tsotsos, “Active Object Recognition Integrating Attention and Viewpoint control”, *Computer Vision and Image Understanding*, Vol. 67, No. 3, September, pp. 239-260, 1997
- [4] E. Rivlin, I. Weiss, “Deformation Invariants in Object Recognition”, *Computer Vision and Image Understanding*, Vol. 65, No. 1, January, pp. 95-108, 1997
- [5] H. Murase, S.K.Nayar, “Visual Learning and Recognition of 3-D Objects from Appearance”, *International Journal of Computer Vision*, Vol. 14, No. 1, pp. 5-24, 1995
- [6] W. M. Wells, “Statistical Approaches to Feature-based Object Recognition”, *International Journals of Computer Vision*, Vol. 21, No. 1/2, pp. 63-98, 1997
- [7] T.K.Leung, M.C.Burl, P.Perona, “Probabilistic Affine Invariants for Recognition”. *CVPR*, pp. 678-684, 1998.
- [8] S. J. Dickinson, A. P. Pentland, A. Rosenfeld, “From Volumes to Views: An Approach to 3-D Object Recognition”, *CVGIP*, March, pp. 130-154, 1992.