1994 International Symposium on Speech, Image Processing and Neural Networks, 13-16 April 1994, Hong Kong

A Two-Level Model-Based Object Recognition Technique

Yip-San Wong and Andrew Choi
Dept. of Comp. Sci., University of Hong Kong, Pokfulam Road, Hong Kong
Email:yswong@csd.hku.hk, choi@csd.hku.hk

Abstract

Object recognition is the problem of detecting the presence and determining the pose of a set of known objects in given images. For some applications, the known objects may be composed of identical components. The relations among these components can be exploited to improve recognition accuracy. This paper introduces a two-level model for object recognition which reduces redundant work due to objects with identical components by explicitly specifying the components and the relations among them. Using automatic analysis of music scores as example, an empirical study is presented, demonstrating the effectiveness and properties of the technique.

1 Introduction

When a model-based object recognition technique [1] is applied directly to locate instances of objects in a scene, a pass must be made through the entire scene for each known object. For example, with three different objects, represented by the object models *a*, *b*, and *c*, all instances of *a* in the scene must be identified, then those of *b*, and then those of *c*. However, if the objects contain components with identical shapes, some operations among the different passes are redundant and should be avoided. Examples for such objects can be found in music scores, in which quarter notes, eighth notes, sixteenth notes, and so on are composed of noteheads, stems, flags, and beams. The effort in locating noteheads, for example, will be duplicated if each type of notes is identified in a separate pass.

Furthermore, in some cases, the presence and the location of one object may be used as a clue in detecting and locating another object. For example, in music scores, a dot, if it appears, must be to the right of a note, to form a dotted note. If a recognition algorithm is to claim that a certain object is a dot, the presence of a note to its left provides additional evidence in support of the claim. On the other hand, if a note is not detected to the left, the claim should be made with less certainty.

To prevent redundant operations and exploit the relations among objects in the recognition process, we developed and studied a technique that uses a two-level object model. The model specifies known objects in terms of their components, or *sub-objects*, and captures joint occurrence frequencies and relative positions among them. For example, in modeling objects in music scores, an eighth note is divided into three parts: a notehead, a stem, and a flag (figure 1). An object recognition technique based on this model will be described next.

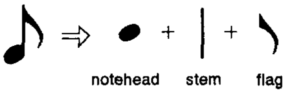


Figure 1

2 Overview of the Technique

The recognition process is divided into three steps. Step one attempts to identify all instances of each sub-object in the scene. Step two refines these recognition results by making use of the relations among sub-objects. Step three applies a clustering algorithm to locate the object instances.

Step one: Hough transform. The Hough transform is a common method for determining an object's pose [2, 3]. It does so by accumulating counts in a Hough table, which represent the likelihood of different possible coordinate transformations. Large clusters in the Hough table of values greater than a certain threshold are taken to be poses of the objects' instances. A modified version of the Hough transform is employed in this step, which not only makes use of foreground information contained in each object model (or template), but also makes use of background information. Negative values are added to the boundary of template of each sub-object in order to improve the accuracy of the method. Step one applies this modified Hough transform to the input scene to generate a Hough table for each sub-object template.

Step two: Merging Hough tables. In detecting a sub-object, say a notehead, the presence of stems, flags, dots, etc., should be treated as evidence in support of its detection. We use the term *main sub-object* to refer to the sub-object being located and the term *supporting sub-objects* to refer to the sub-objects used as supporting evidence. Figure 2 shows an example of an input scene and a one-dimensional projection along the vertical direction of the corresponding Hough table of the notehead (introducing the projection simplifies the graphical

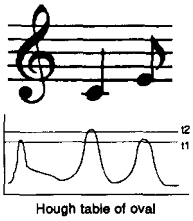


Figure 2

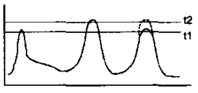
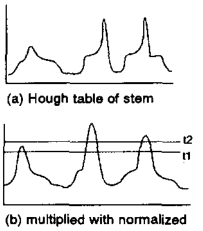


Figure 3



(a) Hough table of stem

(b) multiplied with normalized supporting table of stem

Figure 4

presentation). When the threshold *t1* is applied in figure 2, an incorrect cluster results. The threshold *t2*, although not resulting in any incorrect cluster, causes a correct one to be ignored. If the values of the correct clusters can be raised above *t2* without changing the incorrect ones, both noteheads will be detected successfully (figure 3). This is achieved by normalizing a supporting Hough table to the range [1, *t2/t1*] and multiply it (element by element) to the main table. Note that the values of *t1* and *t2* must be determined empirically. For example, consider the use of the stem as supporting sub-object. Figure 4a shows the corresponding Hough table of the stem and figure 4b shows the Hough table of the notehead after it is multiplied with the normalized Hough table for the stem.

Step three: Clustering and Peak Detection. In this step, the peaks of regions (centers of clusters) with values above the threshold are identified. They will correspond to the centers of the detected main sub-objects. Each entry in the Hough table with value above the threshold is placed in a bucket *B* if it is next to one of the entries in bucket *B*. A new bucket is created for it if no such bucket exists. Then, pairs of buckets are merged if they represent adjacent regions. Buckets that correspond to compound clusters are split. Clusters whose width and height are approximately multiples of those of the main sub-object are considered compound clusters. Each bucket now corresponds to a single instance of the main sub-object. Its geometric center can be calculated.

3 Experiments

We implemented and tested the two-level model-based object recognition system described above. The test data were two sets of music scores. The first set consisted of 3 well-printed scores, containing a total of about 500 noteheads. The second set consisted of 15 poorly-printed scores, containing a total of about 2,750 noteheads. All images were scanned by an HP ScanJet IIc scanner at 300 D.P.I. in black-and-white drawing mode.

Figure 5 shows all the supporting sub-objects used in the recognition of the notehead. Note that a staff line can be both a middle staff line and boundary staff line (with respect to different notes). When a staff line is used as a middle staff line to refine the Hough table of noteheads, its relative vertical position from the notehead is zero. However, when it is used as a boundary staff line to refine the Hough table of noteheads, its relative vertical position is offset by a value which is ±1/2 the height of a notehead.

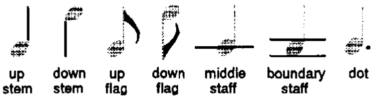


Figure 5

To assess the performance of the system, define the *recognition rate* (RR) to be the ratio of number of instances of the main sub-object correctly located to the total number of such instances in the scene. If there are *n* instances in the scene of which *x* are correctly located, the recognition rate is *x/n*. The *false detection rate* for an object is defined to be the ratio of incorrect claims to the total number of instances claimed by the system to be that object. The *reliability* (Re) of recognizing a certain object is then one minus the false detection rate. Table 1 shows the recognition rate and reliability of using traditional Hough transform to locate noteheads in the second set of music scores. The *occurrence frequency* of object *a* relative to object *b* is defined to be the percentage of instances of object *b* with corresponding instances of object *a* at a set of predefined relative positions. Table 2 shows occurrence frequency of different supporting sub-objects relative to the notehead (the main sub-object) in the second set of music scores.

	At threshold 98%	At threshold 90%	At threshold 80%
RR	63.7%	99.3%	99.9%
Re	98.5%	89.1%	74.1%

Table 1

s. object	up stem	down stem	up flag	down flag
frequency	0.57	0.43	0.23	0.19
s.object	mid. staff	b. staff	dot	-
frequency	0.50	0.50	0.21	

Table 2

3.1 Comparison with a Single-Level Method

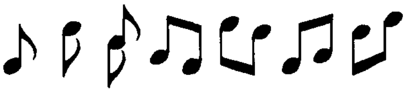
In the first set of experiments, the improvement in performance due to the use of the relations among objects in the recognition process is assessed. In one experiment, all the dots in the set of well-printed images were to be located. The performances of two recognition systems were compared. The first is the two-level recognition system, S_t, which uses the dot as the main sub-object and the notehead as the supporting sub-object. The other is a traditional Hough transform recognition system, S_h, which uses the dot as template. It is found that to achieve the same reliability of close to 100%, S_t has a recognition rate of 96% while S_h has only a recognition rate of 80%.

Then, the effect of dividing objects into sub-objects on recognition performance is tested. The same set of data is presented to the two systems again to locate all quarter-notes and eighth-notes. S_t uses the notehead as the main sub-object and the stem and flag as the supporting sub-objects. S_h uses the entire quarter-note and eighth-note as templates. Table 3 shows the results of the experiment.

Object	Two-level object recognition system		Hough transform recognition system	
	RR	Re	RR	Re
quarter-note	100	99.4	99.8	99.6
eighth-note	99.4	99.4	99.2	99.2

Table 3

From this experiment, we note that the performance of the two systems is quite similar. But as mentioned before, quarter-notes have two orientations. So, in S_h, two Hough transform are needed (one for each orientation). And in S_t, one Hough transform using the notehead template, one Hough transform using the stem template, and two Hough table mergings are needed. The running time of Hough transform is proportional to the size of the image times the size of the template. It should be noticed that the size of a whole quarter-note is roughly equal to the size of a notehead plus the size of a stem. Therefore the running time of one Hough transform of S_h is roughly equal to the two Hough transforms of S_t. Also, since each Hough table merging in S_t can be done in time linear to the size of the image, its running time is negligible compared to that of a Hough transform. So, S_t is about twice as fast as S_h for locating quarter notes. In practice, the Hough tables of the supporting sub-objects can be reused for locating other types of notes (figure 6 shows the different configurations of sub-objects in eight notes). The savings in running time is then even greater.



For eighth note, there are total 11 orientations

Figure 6

3.2 Factors Affecting Performance

In this set of experiments, the notehead is used as the main sub-object and the stem, flag, dot, and staff-line are used as the supporting sub-objects. Each Hough table of a supporting sub-object obtained in step one is normalized to the range [1.0, 1.225]. The value 1.225 is obtained as follows. Recall from section 2 that our goal is to raise the values of correct clusters from a threshold which achieves high recognition rate, say a threshold of 90%, to another threshold that achieves high reliability, say a threshold of 98% (refer to table 1). We call this threshold the *desired threshold*, which is chosen to be 98% for all experiments below. Thus we normalize the values in the Hough table of the supporting sub-object to between 1.0 and $98/90 = 1.225$. From the experiments, we found that a number of factors affect the performance of the technique. These are described next.

Occurrence frequency vs. Recognition rate. In this experiment, the effect of different occurrence frequencies of supporting sub-objects on the recognition rate of the main sub-object is studied. The Hough table of each supporting sub-object obtained in step one is multiplied to the main Hough table and the recognition rate is then recorded at the desired threshold. Table 4 shows the results of the experiment on the set of poorly-printed music score.

s. object	up stem	down stem	up flag	down flag
RR	83.5%	77.6%	70.5%	68.4%
s.object	mid. staff	b. staff	dot	-
RR	71.2%	84.2%	70.5%	

All above RR values are taken at the desired threshold. Note that without using supporting sub-objects, the corresponding RR=63.7%

Table 4

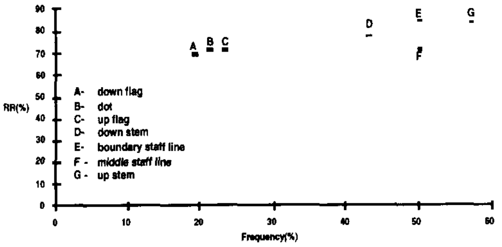


Figure 7

Figure 7 plots the recognition rates against occurrence frequencies of the supporting sub-objects. The graph shows that the occurrence frequency is roughly proportional to the improvement in recognition rate. This suggests that a sub-object that occurs frequently together with the main sub-object will serve as a better supporting sub-object.

No. of supporting sub-objects used vs. Recognition rate. The effect of using a different number of supporting sub-objects on the recognition rate is studied. We use the term *combined occurrence frequency* to refer to the frequency of occurrence of the union of supporting sub-objects. Table 5 shows the recognition rate and combined occurrence frequency of difference set of supporting sub-objects.

Supporting sub-objects used	RR(%)	Combined Occurrence Frequency(%)
up stem	83.5	57
up stem + up flag	84.2	57
up stem + dot	85.7	62
up stem + down flag	92.5	76
up stem + down stem	97.2	100
up stem + down stem + up flag	97.5	100
up stem + down stem + down flag	97.5	100
up stem + down stem + dot	97.3	100
up stem + down stem + up flag + down flag	97.8	100

Table 5

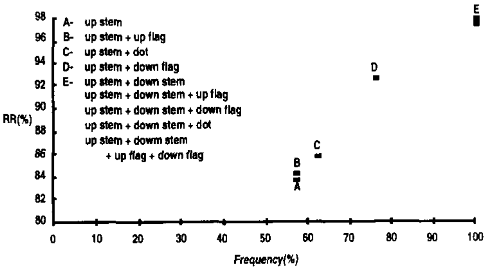


Figure 8

Figure 8 plots recognition rate against combined occurrence frequency, which shows that an improvement in the former is roughly proportional to the latter. Therefore, to achieve a high recognition rate, the combined occurrence frequency also needs to be high. In other words, the introduction of an additional supporting sub-object improves the recognition rate only if the combined occurrence frequency is higher than the occurrence frequency of the original supporting sub-object.

Reliability. The two results above provide guidelines for increasing the recognition rate in an implementation of the technique. However, for an object recognition technique to be successful, a high reliability must also be maintained. Table 6 shows the effect of different supporting sub-objects on reliability for the same set of music score.

s. object	up stem	down stem	up flag	down flag
Re	97.8%	70.5%	99.1%	96.7%
s.object	mid. staff	b. staff	dot	-
Re	54.2%	54.5%	94.4%	

All above Re values are taken at the desired threshold. Note that *without using supporting sub-objects, the corresponding* Re at threshold 90% is 89.1%, at threshold 98% is 98.5%

Table 6

Note that the reliabilities of using staff lines and down stems as supporting sub-objects are quite low. The poor performance of the staff lines is explained by their frequent appearances near sub-objects that are easily mistaken for noteheads. Their use as supporting sub-objects thus increases the number of incorrect clusters. Cases that noteheads are incorrectly recognized due to the use of down stems as supporting sub-objects are shown in figure 9.

The reliabilities with which the individual sub-objects can be recognized are then found to be as follows.

Re of up stem, Re of up flag > Re of down flag > Re of dot >
Re of down stem > Re of middle staff line, Re of boundary staff line
———Eq. (1)

Comparing Eq. (1) with the values of table 6, we note that the higher the reliability of the supporting sub-objects, the higher the reliability of the recognition results using that sub-object as supporting sub-object.

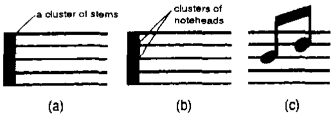


Figure 9

The effect of different sets of the two supporting sub-objects were also studied. The results of one of these experiments are tabulated in table 7, which show that the two-level technique will maintain a high reliability of recognition if the both supporting sub-objects used have high individual reliability of recognition. An analytical study of and further empirical results for the two-level technique appear in [4].

Supporting sub-objects used	Re(%)	Supporting sub-objects used	Re(%)
up stem + up flag	97.3	down stem + up flag	90.2
up stem + down flag	95.2	down stem + down flag	86.5
up stem + dot	94.1	down stem + dot	86.7
up stem + down stem	87.9	-	-

Table 7

Summary

This paper describes a two-level model-based object recognition technique and experimental study of its performance and properties. The experiments show that a high recognition rate is achieved using supporting sub-objects with high occurrence frequencies. To maintain a high reliability, the supporting sub-objects must also be recognized reliably individually.

References

[1] W. Eric L. Grimson and Tomas Lozano-Perez. Model-based recognition and localization from sparse range or tactile data. *International Journal of robotics Research, vol. 3,* No. 3, pp. 382-414, Fall 1984.

[2] Jerry L. Turney, Trevor N. Mudge and Richard A. Volz. Recognizing partially occluded parts. *IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI-7,* No. 4, pp. 410-421, July 1985.

[3] W. Eric L. Grimson and Daniel P. Huttenlocher. On the sensitivity of the Hough transform for object recognition. *IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI-12,* No. 3, pp. 255-274, March 1990.

[4] Y.S. Wong and A. Choi. Empirical and analytical study of a two-level model-based object recognition technique. Technical Report, Department of Computer Science, University of Hong Kong.