Inexact Pattern Matching using Genetic Algorithm

Surapong Auwatanamongkol
Department of Computer Science
School of Applied Statistics
National Institute of Development Administration
Bangkok, Thailand 10240
Email: surapong@as.nida.ac.th

ABSTRACT

A Genetic Algorithm for graphical pattern matching based on angle matching had been proposed. It has proven quite effective in matching simple patterns. However, the algorithm needs some modifications to enhance its accuracy on pattern matching when there are some differences between two patterns in terms of numbers of nodes, shapes and rotations. This paper presents the modifications, such as the introduction of node exemption, inexact matching between straight lines and curves in the patterns, and consideration of rotational degrees of the patterns. Each angle is also given with a weight to indicate the significant degree of the angle. A multi-objective function is used to reflect the similarity between two patterns. The experiments designed to evaluate the algorithm have shown very promising results. It is highly accurate on patterns matching with dissimilarities in shapes, numbers of nodes and rotational degrees.

Categories and Subject Descriptors

I.5.2 [Pattern Recognition]: Design Methodology – *Classifier design and evaluation*

General Terms

Algorithms

Keywords

Inexact pattern matching, Genetic algorithm, Graph matching

1. INEXACT PATTERN MATCHING

We assume a graphical pattern has been divided into straight line segments and is represented as a graph. The pattern is also augmented with indirect links [1]. The indirect links cannot crossover any original or direct links and represents the lines of sight from one node to another. Any angle formed between two adjacent links is assigned with an angle weight that represents the significant degree of the angle that helps identify the pattern. The sum of all angle weights in a pattern is set to be equal to 1.

Our objective for the inexact pattern matching is finding the best 1-to-1 matching between nodes in a given input pattern and nodes in a known pattern, which yields the maximum degree of matching. Also, we would like to match straight line segments between the two patterns as many segments as possible. The candidate matching solution can simply be encoded as a vector of integer numbers. The vector length is n where n is the number of nodes in the input pattern. Each element in the vector has a value ranging between 0 and m, where m is the number of nodes in the known pattern. The i^th^ element of the vector represents the node id of the known pattern that the i^th^ node of the input pattern is matched with. Since the number of nodes in the two patterns can be unequal, some nodes in both patterns will be left unmatched. The zero element of the vector designates null matching or unmatched condition for the corresponding node in the input pattern.

Consider the two patterns in figure 1. If nodes A, D and E are matched with node P, Q and R, respectively, while nodes B and C are exempted, the indirect link AD and the curve ABCD must be replaced by the direct link AD. All indirect links involved with exempted nodes, i.e. AC, AD, BE and CE, must be now exempted or discarded.

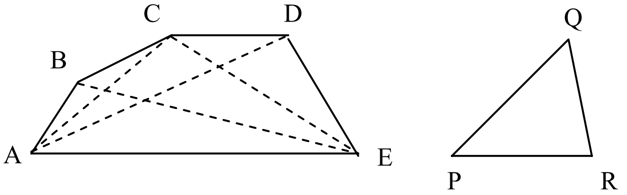

Figure 1 A pattern with some exempted nodes i.e. node B and C and its counterpart pattern

From figure 1, we can see that the curve ABCD can be matched with direct link PQ. Direct link AE and DE can be matched with direct link PR and QR, respectively. The matching degree for the angle BAE, ϕ_{BAE}, can be defined as follows:

$$\phi_{BAE} = W_{BAE} * \mu_{ABCD} * \mu_{AE} * \text{sim}(\Delta)$$

where

W_{BAE} is the weight of the angle BAE,

μ_{ABCD} and μ_{AE} the straightness degrees of ABCD and AE, respectively. The straightness degree of a curve can be computed as the square (concentration) of the ratio of the length of straight

GECCO'05, June 25-29, 2005, Washington, DC, USA.
ACM 1-59593-010-8/05/0006.

line connected the two end points of the curve and the length of the curve itself,

Δ is the absolute different value between the magnitudes of the angles DAE and QPR,

Sim is a similarity function that defined on Δ. The similarity function value should decrease as the difference Δ increases. The function value peaks at 1.0 when Δ is equal to 0 and decreases to a constant when Δ approaches 2π or 360 degrees.

The angle matching degrees for the other angles can be computed in the same way except for those of angles incident to exempted nodes, e.g. ABC and BCD. Because of the limitation on page length of this paper the scheme to compute the degrees will not be discussed in this paper.

Let S_a be the sum for angle matching degrees ϕ of all angles in the pattern. Let S_t be the sum of the maximum values of angle matching degrees for all angles in the pattern, which is actually the sum of weights of all angles in the pattern. Hence, S_t is equal to 1. Therefore, the matching degree that the pattern matches with the other can be computed a

$$\phi\,(\text{pattern}) = S_a / S_t = S_a$$

2. STRUCTURAL FITNESS FUNCTION

Let G_1 and A_1 be the input pattern and its corresponding augmented pattern, respectively. Let G_2 and A_2 be the known pattern and its corresponding augmented pattern, respectively.

Hence, $\phi(G_1)$, $\phi(G_2)$, $\phi(A_1)$ and $\phi(A_2)$ represent the matching degrees for patterns G1, G2, A_1 and A_2, respectively. We will use a simple weighting approach to combine all the matching degrees into one fitness value that represents the structural matching degree between the two patterns. The structural fitness function can be expressed as follows:

$w_1 * (w_2 * \phi(G_1) + (1- w_2) * \phi(A_1)) + (1-w_1) * (w_2 * \phi(G_2) + (1- w_2) * \phi(A_2))$

where w_1 is the weight used to combine the matching degrees for the input pattern and the known pattern, while w_2 is the weight used to combine the matching degrees of original pattern and its augmented pattern.

3. ROTATIONAL FITNESS FUNCTION

Let vector AD and PQ represent the link AB(CD) and PQ of the patterns in figure 1 respectively. We can superimpose node A on node P since they match with each other. Therefore, the rotation degree difference Δ between the two links can be measured in term of degree of the angle ADQ (or APQ). The difference can then be converted into rotational similarity (RS) ranging between 0 and 1 using a half bell shape transformation function (intensification). Hence, the rotational fitness for link AB(CD) and PQ can be defined as follows:

$$\mu_{ABCD} * \mu_{PQ} * \text{RS of ABCD and PQ}$$

Notice that the rotational fitness of links BC and CD are approximated to be equal to the rotational fitness of link AB(CD). The rotational fitness for a given pattern G can be computed as the average of rotational fitness of all direct links in the pattern.

Finally, the rotational fitness function can then be computed using the weighting scheme as follows

w_1 * Rotational fitness of G_1 + $(1-w_1)$ * Rotational fitness of G

The two fitness functions, namely structural and rotational fitness functions, can now be combined to form the total fitness function as follows

w_3 * Structural Fitness Function + (1- w_3) * Rotational Fitness Function

where w_3 is the weight that represents the preference on structural fitness.

4. EXPERIMENTS ON THE GENETIC ALGORITHM

Patterns of alphabet characters, A to Z, are chosen to be the known patterns in our experiments designed to evaluate the effectiveness of the proposed algorithm. Each pattern is plotted on 18 x 24 matrix. The patterns are segmented and then augmented with indirect links as described in section 1. The testing sets consist of five suits of input patterns of characters A-Z. Hence, 130 input patterns are tested. Each of the five suits contains twenty six character patterns of A-Z which shapes and the numbers of nodes are different from those of the known patterns and those from the other suits. They are also input in XY coordinates then segmented.

Each testing pattern in each suit is tried twice and the best matching among the trials is selected. A tournament selection scheme is used in the experiments. A crossover rate of 0.6 and a mutation rate of 0.05 are used for the best results. The values of w1, w2 and w3 used in the experiments are 0.5, 0.5 and 0.7 respectively. Using both structural and rotational fitness functions, 99.23 % accuracy in classifying the patterns is achieved while 86.92 % accuracy is achieved if only structural fitness function is used. The rotation fitness function can help distinguish similar patterns such as N and Z, M and W etc.

5. CONCLUSIONS

The inexact pattern matching algorithm based on genetic algorithm is designed to take into account the dissimilarities of patterns on both shapes, numbers of nodes and rotational degrees. The algorithm achieves very high accuracy on pattern matching in our experiments. It can be used for not only pattern recognition but also image retrieval applications. The algorithm is also suitable for parallel implementation, a prerequisite for large complex pattern matching.

6. REFERENCES

[1] Auwatanamongkol Surapong, *Pattern Recognition using Genetic Algorithm*, 2000 Congress of Evolutionary Computation, p. 822-828.

[2] Coughlan James, and Shen Huiying, *Shape Matching with Belief Propagation: using dynamic quantization to accommodate occlusion and clutter*, Proceedings of the 2004 IEEE conference of Computer Vision and Pattern Recognition Workshop (CVPRW'04).

[3] Milan Sonka, Vaclav Hlavac and Roger Boyle, *Image processing, Analysis and Machine Vision*, Second Edition, 1999, PWS Publishing.