ELSEVIER

Pattern Recognition Letters 19 (1998) 255–259

Pattern Recognition Letters

# A graph distance metric based on the maximal common subgraph

Horst Bunke [a,*], Kim Shearer [b]

[a] *Institut für Informatik und angewandte Mathematik, University of Bern, Bern, Switzerland*
[b] *Department of Computer Science, Curtin University of Technology, Perth, WA, Australia*

Received 22 July 1997; revised 12 November 1997

**Abstract**

Error-tolerant graph matching is a powerful concept that has various applications in pattern recognition and machine vision. In the present paper, a new distance measure on graphs is proposed. It is based on the maximal common subgraph of two graphs. The new measure is superior to edit distance based measures in that no particular edit operations together with their costs need to be defined. It is formally shown that the new distance measure is a metric. Potential algorithms for the efficient computation of the new measure are discussed. 

*Keywords:* Error-tolerant graph matching; Distance measure; Maximal common subgraph; Graph edit distance; Metric

## 1. Introduction

One of the most general and powerful data structures useful in a variety of applications are graphs. For example, in computer vision and pattern recognition, graphs are often used to represent unknown objects, which are to be recognized, and known models, which are stored in a database. Thus, the recognition problem turns into a graph matching problem. Applications of graph matching in pattern recognition and machine vision include character recognition (Lu et al., 1991; Cordella et al., 1997), schematic diagram interpretation (Lee et al., 1990; Messmer and Bunke, 1996), shape analysis (Pearce et al., 1994), image registration (Christmas et al., 1995), 3-D object recognition (Cho and Kim, 1992; Wong, 1992) and video indexing (Shearer et al., 1997).

Classical algorithms of graph matching include graph and subgraph isomorphism (Read and Corneil, 1977; Ullman, 1976). However, due to errors and distortions in the input data and the models, approximate, or error-tolerant, graph matching methods are needed in many applications. One way to cope with errors and distortions is graph edit distance (Shapiro and Haralick, 1981; Bunke, 1997). Here one introduces a set of edit operations, for example, the deletion, insertion and substitution of nodes and edges, and defines the similarity of two graphs in terms of the shortest (or least cost) sequence of edit operations that transforms one graph into the other. Another approach to error-tolerant graph matching is based on the maximal common subgraph of two graphs (Horaud and Skordas, 1989; Levinson, 1992).

When defining distance or similarity measures, certain properties are desirable. For example, one may wish that the distance from object $A$ to $B$ is the same as the distance from $B$ to $A$ (symmetry).

* Corresponding author. E-mail: bunke@iam.unibe.ch.

0167-8655/98/$19.00 
*PII* S0167-8655(97)00179-7

Speaking more generally, it is often desired that the distance measure $d$ fulfills the properties of a metric:
1. $d(A,B) = 0 \Leftrightarrow A = B$,
2. $d(A,B) = d(B,A)$,
3. $d(A,B) + d(B,C) \geqslant d(A,C)$.

Usually edit distance measures are metrics. Only if the costs of the underlying edit operations satisfy certain conditions, the properties listed above will hold. But these conditions are sometimes too restrictive, or incompatible with the considered problem domain.

In the present paper, we propose a new graph distance measure that is based on the maximal common subgraph of two graphs. The main contribution of the paper is the formal proof that the new distance measure is a metric. An advantage of the new distance measure over graph edit distance is the fact that it does not depend on edit costs. It is well known that any edit distance measure critically depends on the costs of the underlying edit operations. But the problem how these edit costs are obtained is still unsolved. Using the new distance measure, this problem can be avoided.

In the next section of this paper we will present basis definitions. The following section will first define the maximal common subgraph based distance measure. Then it will be shown that the measure is a metric. Concluding remarks will make up the final section, including a discussion of potential algorithms for the computation of the new distance measure.

## 2. Basic definitions

In this paper, we consider graphs with labeled nodes and edges. Let $L_V$ and $L_E$ denote the finite sets of node and edge labels, respectively. (Unlabeled graphs are obtained as a special case if $|L_V| = |L_E| = 1$.)

**Definition 1.** A *graph* is a 4-tuple $G = (V, E, \mu, \nu)$, where
- $V$ is a set of finite vertices,
- $E \subseteqq V \times V$ is the set of edges,
- $\mu : V \to L_V$ is a function assigning labels to the vertices,
- $\nu : E \to L_E$ is a function assigning labels to the edges.

If $V = \emptyset$ then $G$ is called the empty graph.

**Definition 2.** Given a graph $G = (V, E, \mu, \nu)$, a *subgraph* of $G$ is a graph $S = (V_S, E_S, \mu_S, \nu_S)$ such that
- $V_S \subseteqq V$,
- $E_S = E \cap (V_S \times V_S)$,
- $\mu_S$ and $\nu_S$ are the restrictions of $\mu$ and $\nu$ to $V_S$ and $E_S$, respectively, i.e.,

$$\mu_S(v) = \begin{cases} \mu(v) & \text{if } v \in V_S, \\ \text{undefined} & \text{otherwise,} \end{cases}$$

$$\nu_S(e) = \begin{cases} \nu(e) & \text{if } e \in E_S, \\ \text{undefined} & \text{otherwise.} \end{cases}$$

The notation $S \subseteqq G$ is used to indicate that $S$ is a subgraph of $G$.

**Definition 3.** A bijective function $f : V \to V'$ is a *graph isomorphism* from a graph $G = (V, E, \mu, \nu)$ to a graph $G' = (V', E', \mu', \nu')$ if
- $\mu(v) = \mu'(f(v))$ for all $v \in V$,
- for any edge $e = (v_1, v_2) \in E$ there exists an edge $e' = (f(v_1), f(v_2)) \in E'$ such that $\nu(e) = \nu(e')$, and for any $e' = (v_1', v_2') \in E'$ there exists an edge $e = (f^{-1}(v_1'), f^{-1}(v_2')) \in E$ such that $\nu(e') = \nu(e)$.

**Definition 4.** An injective function $f : V \to V'$ is a *subgraph isomorphism* from $G$ to $G'$ if there exists a subgraph $S \subseteqq G'$ such that $f$ is a graph isomorphism from $G$ to $S$.

Note that finding a subgraph isomorphism from $G$ to $G'$ implies finding a subgraph of $G'$ isomorphic to the whole of $G$. This distinction becomes important in later discussion.

**Definition 5.** Let $G$, $G_1$, and $G_2$ be graphs. $G$ is a *common subgraph* of $G_1$ and $G_2$ if there exists subgraph isomorphisms from $G$ to $G_1$ and from $G$ to $G_2$.

**Definition 6.** A common subgraph $G$ of $G_1$ and $G_2$ is *maximal* if there exists no other common subgraph $G'$ of $G_1$ and $G_2$ that has more nodes than $G$.

The maximal common subgraph of two graphs $G_1$ and $G_2$ will be denoted by $\mathrm{mcs}(G_1,G_2)$. Notice that $\mathrm{mcs}(G_1,G_2)$ is not necessarily unique for two given graphs, $G_1$ and $G_2$. The number of nodes of a graph $G=(V,E,\mu,\nu)$ is given by $|V|$. For the purpose of notational convenience, we also denote the number of nodes of $G$ by $|G|$.

## 3. Graph distance measure

**Definition 7.** The *distance* of two non-empty graphs $G_1$ and $G_2$ is defined as

$$d(G_1,G_2)=1-\frac{|\mathrm{mcs}(G_1,G_2)|}{\max(|G_1|,|G_2|)}.$$

An example is shown in Fig. 1. Here we have $|G_1|=5$, $|G_2|=4$ and $|\mathrm{mcs}(G_1,G_2)|=3$. Hence, $d(G_1,G_2)=0.4$.

**Theorem 1.** *For any graphs $G_1$, $G_2$ and $G_3$, the following properties hold true*:

1. $0\leqslant d(G_1,G_2)\leqslant 1$,
2. $d(G_1,G_2)=0 \Leftrightarrow G_1$ *and* $G_2$ *are isomorphic to each other*,
3. $d(G_1,G_2)=d(G_2,G_1)$,
4. $d(G_1,G_3)\leqslant d(G_1,G_2)+d(G_2,G_3)$.

**Proof.** Properties 1–3 follow directly from Definition 7. In the following proof of the triangle inequality we distinguish two cases:

*Case* A. The graphs $\mathrm{mcs}(G_1,G_2)$ and $\mathrm{mcs}(G_2,G_3)$ are disjoint, or speaking more strictly, the maximal common subgraph of $\mathrm{mcs}(G_1,G_2)$ and $\mathrm{mcs}(G_2,G_3)$ is empty. For a Venn diagram illustration see Fig. 2(a).

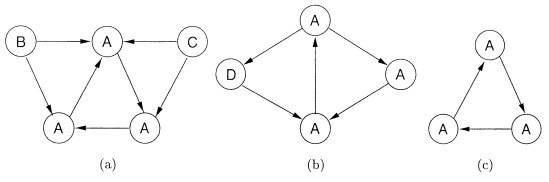


Fig. 1. An example of Definition 7: (a) a graph $G_1$; (b) a graph $G_2$; (c) the maximal common subgraph, $\mathrm{mcs}(G_1,G_2)$, of $G_1$ and $G_2$. Here we have $d(G_1,G_2)=0.4$.

Let $m_{12}=|\mathrm{mcs}(G_1,G_2)|$, $m_{23}=|\mathrm{mcs}(G_2,G_3)|$, and $m_{13}=|\mathrm{mcs}(G_1,G_3)|$. Then the following relation holds true:

$$m_{12}+m_{23}\leqslant|G_2|. \tag{1}$$

Property 4 in Theorem 1 is equivalent to the following inequality:

$$1-\frac{m_{12}}{\max(|G_1|,|G_2|)}+1-\frac{m_{23}}{\max(|G_2|,|G_3|)} \geqslant 1-\frac{m_{13}}{\max(|G_1|,|G_3|)}. \tag{2}$$

We will show that the left-hand side of this inequality is always greater than or equal to 1, which is equivalent to

$$\max(|G_1|,|G_2|)\max(|G_2|,|G_3|) \geqslant m_{12}\max(|G_2|,|G_3|)+m_{23}\max(|G_1|,|G_2|). \tag{3}$$

We proceed by a simple case analysis.

*Case* A.1: $|G_1|\geqslant|G_2|\geqslant|G_3|$. Here Eq. (3) is equivalent to

$$|G_1|\cdot|G_2|\geqslant m_{12}|G_2|+m_{23}|G_1|. \tag{4}$$

From Eq. (1) we conclude that

$$|G_1||G_2|\geqslant m_{12}|G_1|+m_{23}|G_1|\geqslant m_{12}|G_2|+m_{23}|G_1|.$$

*Case* A.2: $|G_1|\geqslant|G_3|\geqslant|G_2|$. Here Eq. (3) becomes

$$|G_1|\cdot|G_3|\geqslant m_{12}\cdot|G_3|+m_{23}\cdot|G_1|. \tag{5}$$

Using Eq. (1) again we conclude

$$|G_1||G_3|\geqslant|G_1||G_2| \geqslant m_{12}|G_1|+m_{23}|G_1|\geqslant m_{12}|G_3|+m_{23}|G_1|.$$

The remaining four cases $|G_2|\geqslant|G_1|\geqslant|G_3|$, $|G_2|\geqslant|G_3|\geqslant|G_1|$, $|G_3|\geqslant|G_1|\geqslant|G_2|$ and $|G_3|\geqslant|G_2|\geqslant|G_1|$ can be shown similarly.

*Case* B. Here we assume that the maximal common subgraph of $\mathrm{mcs}(G_1,G_2)$ and $\mathrm{mcs}(G_2,G_3)$ is not empty (see Fig. 2(b)).

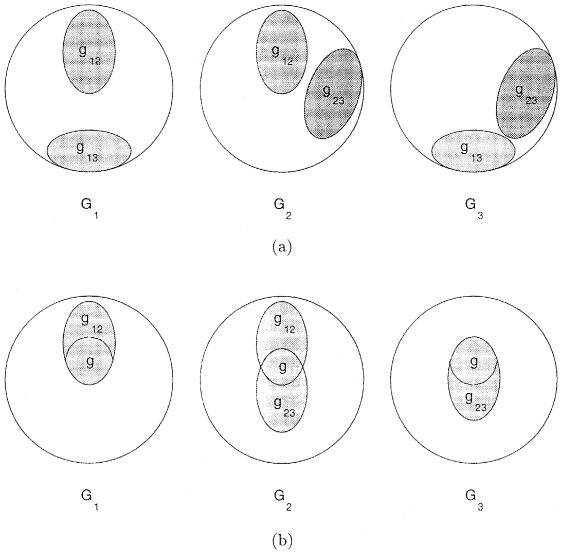


Fig. 2. Illustration of disjoint and overlapping common subgraphs: (a) the maximal common subgraphs $\mathrm{mcs}(G_1,G_2)=g_{12}$ and $\mathrm{mcs}(G_2,G_3)=g_{23}$ are disjoint; (b) $\mathrm{mcs}(G_1,G_2)$ and $\mathrm{mcs}(G_2,G_3)$ share a common subgraph $g$, i.e., $g=\mathrm{mcs}(\mathrm{mcs}(G_1,G_2),\mathrm{mcs}(G_2,G_3))$.

Let $m=|\mathrm{mcs}(\mathrm{mcs}(G_1,G_2),\mathrm{mcs}(G_2,G_3))|>0$. It follows that there exists a maximal common subgraph of $G_1$ and $G_3$ with size greater than or equal to $m$. Furthermore it follows that

$$m_{12}+m_{23}-m\leqslant|G_2|,\quad m\leqslant m_{12},\quad m\leqslant m_{23}. \tag{6}$$

We will show that

$$1-\frac{m_{12}}{\max(|G_1|,|G_2|)}+1-\frac{m_{23}}{\max(|G_2|,|G_3|)}\geqslant 1-\frac{m}{\max(|G_1|,|G_3|)} \tag{7}$$

which implies property 4 of Theorem 1. Obviously inequality (7) is equivalent to

$$\begin{aligned}&\max(|G_1|,|G_2|)\max(|G_2|,|G_3|)\max(|G_1|,|G_3|)\\&\quad\geqslant m_{12}\max(|G_2|,|G_3|)\max(|G_1|,|G_3|)\\&\qquad+m_{23}\max(|G_1|,|G_2|)\max(|G_1|,|G_3|)\\&\qquad-m\max(|G_1|,|G_2|)\max(|G_2|,|G_3|).\end{aligned} \tag{8}$$

Again we proceed by case analysis.

*Case* B.1: $|G_1|\geqslant|G_2|\geqslant|G_3|$. Here Eq. (8) is equivalent to

$$\begin{aligned}&|G_1||G_1||G_2|\\&\quad\geqslant m_{12}|G_1||G_2|+m_{23}|G_1||G_1|-m|G_1||G_2|\end{aligned}$$

which can be simplified to

$$\begin{aligned}|G_1||G_2|&\geqslant m_{12}|G_2|+m_{23}|G_1|-m|G_2|\\&=(m_{12}-m)|G_2|+m_{23}|G_1|.\end{aligned} \tag{9}$$

From Eq. (6) it follows that

$$\begin{aligned}|G_1||G_2|&\geqslant m_{12}|G_1|+m_{23}|G_1|-m|G_1|\\&=(m_{12}-m)|G_1|+m_{23}|G_1|\end{aligned}$$

from which we get Eq. (9) due to $m_{12}\geqslant m$.

*Case* B.2: $|G_1|\geqslant|G_3|\geqslant|G_2|$. Here Eq. (8) becomes

$$\begin{aligned}&|G_1||G_1||G_3|\\&\quad\geqslant m_{12}|G_1||G_3|+m_{23}|G_1||G_1|-m|G_1||G_3|\end{aligned}$$

which can be simplified to

$$|G_1||G_3|\geqslant m_{12}|G_3|+m_{23}|G_1|-m|G_3|. \tag{10}$$

We proceed analogously to Case B.1.

$$\begin{aligned}|G_1||G_3|&\geqslant|G_1||G_2|\geqslant m_{12}|G_1|+m_{23}|G_1|-m|G_1|\\&\geqslant m_{12}|G_3|+m_{23}|G_1|-m|G_3|.\end{aligned} \tag{11}$$

The remaining cases can be shown similarly. □

From Theorem 1 it follows in particular that our proposed distance measure is a metric. [1]

## 4. Discussion and conclusion

We have shown that the graph distance measure of Definition 7 is in fact a metric. As discussed earlier it is often difficult to form a metric from edit distance measures. Therefore in applications where the properties of a metric are important, the largest common subgraph metric could be used.

One application where this is important is information retrieval from images and video databases (Chang et al., 1987; Lee and Hsu, 1992; Shearer et al., 1997). This area relies heavily on browsing to locate required database elements. Thus it is necessary for the distance measure chosen to be ‘‘well

[1] Strictly speaking, this statement is only true if isomorphic graphs are regarded equal. But this assumption is certainly justified in most applications.

behaved'' to allow sensible navigation of the database. The use of a metric, such as that proposed, for the distance measure ensures that the behaviour of the similarity retrieval will be consistent and comprehensible, aiding the user in their search task.

Classical algorithms for computing the maximal common subgraph of two graphs are based on maximal clique detection (Levi, 1972) or backtracking (McGregor, 1982). These algorithms are conceptually simple, but have a high computational complexity. For example, the worst case time complexity of the method described by Levi (1972) is $O((nm)^n)$, where $n$ and $m$ denote the number of nodes of the two graphs under consideration. Recently, however, a new algorithm has been developed which uses preprocessing of a database of model graphs to detect the maximal common subgraph from an input graph to the models in the database with worst case time complexity of $O(2^n)$ (Shearer et al., 1997). This algorithm has demonstrated near real-time behaviour in a video indexing application.

In a recent paper, it has been shown that maximal common subgraph computation can be regarded a special case of graph edit distance computation under a particular cost function (Bunke, 1997). An immediate consequence is that any algorithm for graph edit distance computation can be used to compute the maximal common subgraph if it is run under the cost function given by Bunke (1997). This opens up additional possibilities for the computation of the distance measure proposed in this paper, particularly with respect to an efficient algorithm for graph edit distance computation reported by Bunke and Messmer (1997).

## References


Bunke, H., 1997. On a relation between graph edit distance and maximum common subgraph. Pattern Recognition Lett. 18 (8), 689–694.

Bunke, H., Messmer, B., 1997. Recent advances in graph matching. Internat. J. Pattern Recognition Artif. Intell. 11 (1), 169–203.

Chang, S., Shi, Q., Yan, C., 1987. Iconic indexing by 2D strings. IEEE Trans. Pattern Anal. Machine Intell. 9 (3), 413–428.

Cho, C.J., Kim, J.J., 1992. Recognizing 3-D objects by forward checking constrained tree search. Pattern Recognition Lett. 13 (8), 587–597.

Christmas, W.J., Kittler, J., Petrou, M., 1995. Structural matching in computer vision using probabilistic relaxation. IEEE Trans. Pattern Anal. Machine Intell. 17 (8), 749–764.

Cordella, L., Foggia, P., Sansone, C., Vento, M., 1997. Subgraph transformations for the inexact matching of attributed relational graphs. In: Jolion, J.-M., Kropatsch, W. (Eds.), Preproceeding GbR'97: IAPR Workshop on Graph Based Representations, Lyon.

Horaud, R., Skordas, T., 1989. Stereo correspondence through feature grouping and maximal cliques. IEEE Trans. Pattern Anal. Machine Intell. 11 (11), 1168–1180.

Lee, S., Hsu, F., 1992. Spatial reasoning and similarity retrieval of images using 2D C-string knowledge representation. Pattern Recognition 25 (3), 305–318.

Lee, S.W., Kim, J.H., Groen, F.C.A., 1990. Translation-, rotation-, and scale invariant recognition of hand-drawn symbols in schematic diagrams. Internat. J. Pattern Recognition Artif. Intell. 4 (1), 1–15.

Levi, G., 1972. A note on the derivation of maximal common subgraphs of two directed or undirected graphs. Calcols 9, 341–354.

Levinson, R., 1992. Pattern associativity and the retrieval of semantic networks. Comput. Math. Appl. 23, 573–600.

Lu, S.W., Ren, Y., Suen, C.Y., 1991. Hierarchical attributed graph representation and recognition of handwritten Chinese characters. Pattern Recognition 24, 617–632.

McGregor, J.J., 1982. Backtrack search algorithms and the maximal common subgraph problem. Software Practice and Experience 12, 23–34.

Messmer, B., Bunke, H., 1996. Automatic learning and recognition of graphical symbols in engineering drawing. In: Kasturi, R., Tombre, K. (Eds.), Graphics Recognition, Lecture Notes in Computer Science, vol. 1072. Springer, Berlin, 1996, pp. 123–134.

Pearce, A., Caelli, T., Bischof, W.F., 1994. Rulegraphs for graph matching in pattern recognition. Pattern Recognition 27 (9), 1231–1246.

Read, R.C., Corneil, D.G., 1977. The graph isomorphism disease. J. Graph Theory 1, 339–363.

Shapiro, L.G., Haralick, R.M., 1981. Structural descriptions and inexact matching. IEEE Trans. Pattern Anal. Machine Intell. 3, 504–519.

Shearer, K., Bunke, H., Venkatesh, S., Kieronska, D., 1997. Efficient graph matching for video indexing. In: Jolion, J.-M., Kropatsch, W. (Eds.), Preproceeding GbR'97: IAPR Workshop on Graph based Representations, Lyon.

Ullman, J.R., 1976. An algorithm for subgraph isomorphism. J. ACM 23 (1), 31–42.

Wong, E.K., 1992. Model matching in robot vision by subgraph isomorphism. Pattern Recognition 25 (3), 287–304.