WEIGHTED MINIMUM COMMON SUPERGRAPH FOR CLUSTER REPRESENTATION

H. Bunke[1], C. Guidobaldi[2] and M. Vento[3]

[1] Institut für Informatik und angwandte Mathematik,
Universität Bern, Neubrückstrasse 10, CH-3012 Bern, Switzerland
bunke@iam.unibe.ch
[2] Dipartimento di Informatica e Sistemistica,
Università di Napoli "Federico II", Via Claudio, 21 I-80125 Napoli (Italy)
cguidoba@unina.it
[3] Dipartimento di Ingegneria dell'Informazione ed Ingegneria Elettrica,
Università di Salerno, via P.te Don Melillo, I-84084 Fisciano (SA), Italy
mvento@unisa.it

ABSTRACT

Graphs are a powerful and versatile tool useful for representing patterns in various fields of science and engineering. In many applications, for example, in image processing and pattern recognition, it is required to measure the similarity of objects for clustering similar patterns. In this paper a new structural method for representing a cluster of graphs is proposed. Using this method it becomes easy to extract the common information shared in the patterns of a cluster, make evident this information and separate it from noise and distortions that usually affect graph representation of real images.

1. INTRODUCTION

Attributed graphs (*graphs*, in the following) are a powerful and versatile tool useful in various fields of science and engineering. There are applications, for example, in image processing,pattern recognition, in which a structural representation of the patterns is particularly suitable. In those cases the representation of patterns by graph is extremely convenient [1,11]. A central problem in artificial vision is to perform a *clustering*, of a given set of object, i.e. to identify *natural groupings* in sets of the input patterns. If graphs are used for the representation of structured objects, then grouping similar objects becomes equivalent to find those graphs that are similar to each other in a set of graphs. Many graph clustering methods are known from the literature [5,7,9,11]. In many artificial vision applications, a cluster is composed of many distorted versions of the same object. In these cases the information contained in the cluster is the original object itself, while the noise of the cluster are all the deformations of the distorted objects. When graph clustering is performed and a grouping is achieved, a very interesting question is how to represent the information contained in each cluster. If patterns are represented by vectors, cluster representation is a well known problem and many possibilities have been exploited[6]. The representation of a cluster of graphs is an open problem; it has been studied since a few year [7,9]. In this paper a new method for representing clusters of graphs is proposed. A cluster of graphs will be described by a graph that is, in general, different from all the graphs of the given cluster. This graph, called the Weighted Minimum Common Supergraph, *WMCS*, of the cluster, is a graph summarizing all (and only) the properties of all the graphs belonging to the cluster. Moreover weights indicating the frequency of each property in the cluster are added during the construction process, obtaining a weighted graph in which the frequency of each property is known. The main advantage of using *WMCS* is that the structural information contained into the cluster is preserved and it is also easy to separate this information from noise. The main contribution of this paper is a formal definition of WMCS and an approximate procedure for its computation.

The remainder of the paper is organized as follows. In Sec.2, definitions are introduced. In Sec.3 a new approach for representing a cluster of graphs is proposed. In Sec.4 approximate methods are discussed, while experimental results are reported in Sec.5. Finally future work is discussed and some conclusions are drawn in Sec.6.

2. DEFINITIONS

A *graph isomorphism* is a bijective mapping between the nodes of two graphs that have the same number of nodes, identical labels and identical edge structure. Similarly, a *subgraph isomorphism* between two graphs g_1 and g_2 is an isomorphism between g_1 and a subgraph of g_2.

The *maximum common subgraph* of two graphs g_1 and g_2, $mcs(g_1,g_2)$, is a subgraph of both g_1 and g_2 and has, among all those subgraphs, the maximum number of nodes. The

Minimum Common Supergraph of two graphs g_1 and g_2, $MCS(g_1,g_2)$, is a supergraph of both g_1 and g_2 and has, among all those supergraphs, the minimum number of nodes. Notice that both $mcs(g_1,g_2)$ and $MCS(g_1,g_2)$ are not necessarily unique for two given graphs. The *difference* between g_2 and a subgraph g_1, $g_2 - g_1$ is obtained by removing g_1 from g_2, including edges connecting g_1 with the rest of the graph. These edges are called the *embedding* of g_1 in g_2, $E = emb(g_1,g_2)$. The *union* of g_1 and g_2 including E, $g_1 \cup_E g_2$ is a graph composed of the graphs g_1 and g_2, joined by the edges in E.

In [4] more details are given above these definitions and formal proof of the following theorem has been given.

Theorem 2.1: Let g_1 and g_2 be graphs. Then

$$MCS(g_1,g_2)= mcs(g_1,g_2)\cup_{E1}(g_1-mcs(g_1,g_2))\cup_{E2}(g_2-mcs(g_1,g_2)) \qquad (1)$$

where $E_1=emb(mcs(g_1,g_2),g_1)$ and $E_2=emb(mcs(g_1,g_2),g_2)$

The computation of *mcs* is a NP-complete problem. Several algorithms for *mcs* are known from literature. In this paper we will use a version of McGregor algorithm [10]. Theorem 2.1 shows a way how *MCS* can be actually computed for two given graphs g_1 and g_2.

Def 2.1 A *weighted graph* is a 6-tuple $g=(V,\lambda,E,\varepsilon,\alpha,\beta)$, where

- V is the finite set of vertices (also called nodes)
- λ: $V \rightarrow N^+$ is a function assigning positive weights to the nodes
- $E \subseteq V \times V$ is the set of edges
- ε: $V \rightarrow N^+$ is a function assigning positive weights to the edges
- $\alpha : V \rightarrow L$ is a function assigning labels to the vertices
- $\beta : E \rightarrow L$ is a function assigning labels to the edges

Def. 2.2: Let $g_1=(V_1,\lambda_1,E_1,\varepsilon_1,\alpha_1,\beta_1)$ and $g_2=(V_2,\lambda_2,E_2,\varepsilon_2\,\alpha_2,\beta_2)$ be two weighted graphs. A *weighted common subgraph* of g_1 and g_2 $wcs(g_1,g_2)$ is a weighted graph $g_2=(V_2,\lambda_2,E_2,\varepsilon_2,\alpha_2,\beta_2)$ such there exist subgraph isomorphisms from g to g_1 and g from g_2. We call g a *weighted maximum common subgraph* of g_1 and g_2, $wmcs(g_1,g_2)$, if there exists no other common subgraph of g_1 and g_2 that has more nodes than g. Let m be the size of $wcs(g_1,g_2)$, and let $V = \{v_1,\ldots,v_m\}$ the set of nodes of g. Furthermore let $V_i = \{v'_{i1},\ldots,v'_{im}\} \subseteq V$, $i \in \{1,2\}$ be the subsets of nodes corresponding to $wcs(g_1,g_2)$. There exist subgraph isomorphisms $f_i(v_j)$: $V \rightarrow V_i'$, $\forall i \in \{1,2\}$, $\forall j \in \{1,\ldots,m\}$. The weight of each node v_j is $\lambda(v_j)=\lambda_1(f_1(v_{1i}))+\lambda_2(f_2(v_{2i}))$. Let $f_i(e_{jk})=(f_i(v_j), f_i(v_k))$, $i \in \{1,2\}$ and $\forall j$, $\forall k \in \{1,\ldots,m\}$. The weight of each edge e_{jk} is $\varepsilon(e_{jk})=\varepsilon_1(f_1(e_{1jk}))+\varepsilon_2(f_2(e_{2jk}))$.

Def. 2.3: Let g_1 and g_2 be weighted graphs. A *Weighted Common Supergraph* of g_1 and g_2, $WCS(g_1,g_2)$, is a weighted graph g such that there exist subgraph isomorphisms from g_1 to g and from g_2 to g. We call g a *Weighted Minimum Common Supergraph* of g_1 and g_2, $WMCS(g_1,g_2)$, if there exists no other weighted common supergraph of g_1 and g_2 that has less nodes than g.

It is possible to extend the Theorem 2.1 also in case of weighted graphs, obtaining a procedure for constructing $WMCS(g_1,g_2)$. The weights of $WMCS(g_1,g_2)$ are defined according to the construction procedure.

```
input: $g_1,g_2$; output: WMCS($g_1,g_2$);
procedure WMCS($g_1,g_2$)
begin
   $E_1 = emb(wmcs(g_1,g_2),g_1)$; $E_2 = emb(wmcs(g_1,g_2),g_2)$
  $WMCS(g_1,g_2) =$
   $wmcs(g_1,g_2)\cup_{E1}(g_1-wmcs(g_1,g_2))\cup_{E2}(g_2-wmcs(g_1,g_2))$;
end procedure
```

Notice that, according to Def. 2.2 and 2.3, $wmcs(g_1,g_2)$ and $WMCS(g_1,g_2)$ are not necessarily unique for two given graphs. In the following, a graph g and a weighted graph g' in which $\lambda(v) = 1\ \forall v \in V$ and $\varepsilon(e) = 1\ \forall e \in E$, will be considered the same.

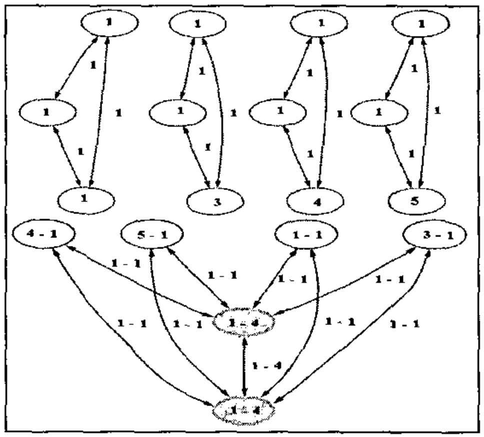

Fig.1: A set of graphs G and a $WMCS(G)$. On each node (edge) of $WMCS(G)$ there is a second label for the weight. This weight is the frequency of the node (edge) in the cluster.

Def 2.4: Let $G=\{g_1,\ldots,g_n\}$ be a set of graphs; a *Weighted Common Supergraph* of G, $WCS(G)$, is a weighted graph $g=(V,\lambda,E,\varepsilon\,\alpha,\beta)$ such there exist subgraph isomorphisms from g_i to g $\forall i \in \{1,\ldots,n\}$. We call g a *Weighted Minimum Common Supergraph* of the set of graphs G, $WMCS(G)$, if there exists no other weighted common supergraph of G that has fewer nodes than g.

Let $f_i : V_i \rightarrow V$ be a bijective mapping between g_i and a subgraph of g. The weight of each node v of g is k if there exist exactly k distinct isomorphisms f_i such that $f_i(v_i) = v$. Let (u,v) be a pair of nodes of V. The weight of each edge $e = (u,v)$ is h if there exist exactly h distinct isomorphisms $f_i(e_{il}) = (f_i(u_i), f_i(v_i))$ such that $f_i(e_{il}) = e$.

Def 2.5: Let g be a $WMCS(G)$. A $WMCS_p(G)$, is a subgraph of g containing all the nodes of g whose weight is at least p and all edges of g having weight at least p, that connect those nodes in V having weight at least p.

3. OPTIMAL CLUSTER REPRESENTATION

Let G be a set of n graphs. We suppose that each graph of G is extracted from the same image with different distortions, due, for instance, to a not optimal system for graph extraction. The graphs of this set are different instances of the same pattern and we call this set a *cluster*. A central problem is to represent a cluster of graphs using a new graph, different from each graph of the cluster.

The *WMCS*(G) is a suitable instrument for representing a cluster of graphs. It is the smallest graph containing, as subgraph, all the graphs of the cluster. As a consequence it is the smallest graph representing each property appearing in the cluster. According to Def. 2.4, the weight of each node (edge) of *WMCS*(G) is can be interpreted as the frequency of representation of a node (edge) in the cluster. Another very interesting problem is to extract from the graphs of the cluster those properties that belong to the original image and to separate them from the noise affecting the graphs. The properties of an image represented by a graph can be represented by subgraphs of that graph. In the rest of this section the terms *subgraph* and *property* will be used synonymously. The *WMCS*(G) is a suitable instrument for finding the common properties of a cluster of graphs for several reasons.

- Nodes (edges) of properties present in many different graphs, i.e. information of the cluster, will correspond to nodes (edges) of *WMCS*(G) having large weights, conversely nodes (edges) of properties present in a few graphs of the cluster, i.e. the noise of the cluster, will correspond to nodes (edges) of having small weights.
- A possibility to separate information from noise is the computation of $WMCS_p(G)$, where p is a threshold for noise-rejection. The selection of an optimal value for the threshold p is useful to reconstruct the information contained in the cluster.
- The construction of $WMCS_p(G)$ can be viewed as a process of generalization, i.e. a learning by example. Indeed only those subgraphs present at least in p graphs of the cluster, are also present in $WMCS_p(G)$: it summarizes all and only those properties appearing at least in p different graphs of the cluster. Thus, if the graphs represent distorted replicas of the same image, $WMCS_p(G)$ is useful for retrieving the original content of the image.
- The common properties summarized in $WMCS_p(G)$ are also expressed by a weighted graph that can be easily displayed and interpreted by an expert of the underlying domain. The knowledge summarized in $WMCS_p(G)$ is not hidden, but is interpretable by an expert.

4. WMCS APPROXIMATIONS

The proposed definition of *WMCS* is very expensive to compute. Indeed the computation of a *WMCS* of two graphs is a NP-hard problem. Moreover, the complexity of computing a *WMCS* of a set of graphs G is also exponential in the number of graphs. Consequently, we propose an approximated procedure compute a *WMCS* (G). Firstly, each graph of the set G is considered a weighted graph, assigning to each node (edge) the weight 1. Then an ordering θ is chosen for the graphs of the set G: θ = ord(G). Now the procedure *WMCS* (g_1,g_2) is iterated for computing *WMCS* (θ).

```
procedure WMCS(θ)
input: g_1,...,g_n, θ output: WMCS(θ);
begin
  assign to each node and edge the weight 1;
  order the graphs according to θ; W = g_1 ;
  for i = 2 to n
    W = WMCS( W , g_i );
  end for
  WMCS(W , g_i ) = W;
end procedure
```

Using this procedure, the complexity becomes linear in the number of graphs. The computed graph remain a common supergraph, but in most cases it is not minimal, i.e. it will usually include some extra nodes. It is also possible that the weight of some node doesn't have the proper value, but it is smaller. Indeed, it can happen that a node already present in the graph under construction is not recognized and thus inserted once more. The weight is updated, for each iteration of the procedure, only on one of the two copies of the node, thus some node could have a weight smaller than the real frequency in the cluster. A possibility to reduce the approximation generated by the procedure, is to realize a *random search* algorithm. The first step is to realize a number n of different orderings $\theta_1,\ldots,\theta_n$ of the graphs of the cluster randomly selected (*shuffles*). For each shuffle θ_j the procedure *WMCS*(θ_j) is applied and then the best approximation is chosen.

5. EXPERIMENTAL RESULTS

Let us suppose that the graphs of a cluster G represent an image with different distortions, due, for instance, to a not optimal system used for graph extraction. We want to evaluate the quality of the approximated WMCS for representing the cluster G. The experiments are organized as follows. Firstly the image graph is generated. This graph is a *mesh* [3,8]. Meshes are suitable for the representation of 3D image surfaces. In our meshes the number of nodes is fixed to 15, the number of edges is fixed to 22 and the number M of attributes is fixed to 50. The method for generating meshes is explained in details in [2]. A degree of distortion is added to the image graph, obtaining a different graph. This operation is repeated 10 times, thus a cluster G of 10 distorted graphs is obtained. A *distortion* is defined as the substitution of an attribute of a node (edge) with another attribute of the set, randomly chosen. Let p be the probability of distortion of a node (edge); the *graph distortion* is the fraction of distorted nodes (edge). Two experiments have been conducted. In both of them the distortion p has been varied in the range 0.1 – 0.8; the aim of the first experiment is to quantify the size of the prototype (*wmcs*) describing the cluster, whilst in our second experiment the maximum weight of the prototype is evaluated. Results are in average, each experiment has been repeated 50 times.

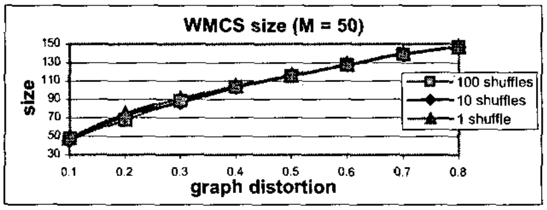

Fig. 2: The approximated WMCS is increasing if the distortion of the graphs increases. It is worth notice that a small effort can be obtained increasing the shuffles in the cluster of graphs.

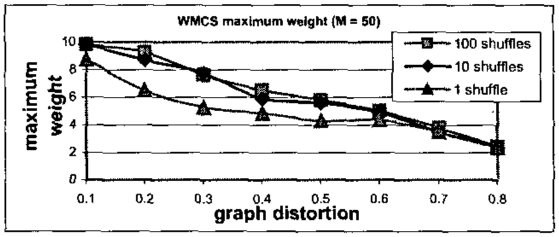

Fig. 3: The maximum weight of the approximated WMCS decreases on the increasing distortion of the graphs, due a smaller overlap among the graphs. Morcover, if more than 10 shuffles are produced, a weak effort is obtained in the approximation.

The approx *WMCS* is computed choosing a number n of shuffles $\theta_1,\ldots,\theta_n$, computing $WMCS(\theta_1),\ldots,WMCS(\theta_2)$ and choosing the best. The best $WMCS(\theta)$ is defined according to this criterion: $WMCS(\theta_j)$ with the smallest number of nodes is chosen; if there is a tie, then *WMCS* representing more properly the most frequent information (i.e. the one maximizing the highest weight) is selected.

In Fig.2 the size of the approximated *WMCS* is shown for an increasing distortion. When the graph distortion increases, also the *WMCS* size increases, due the less overlap of the graphs. It is worth notice that there is no significant improvement in the quality of the approximation when the number of shuffles of the graphs is raised. Moreover, it is interesting to observe in Fig.3, that the maximum weight of the approximated *WMCS*, is increasing when the number of shuffles increases form 1 to 10, thus a better result is obtained. On the contrary, when the number of shuffles increases from 10 to 100, there is only a small further improvement. It follows that a good approximated *WMCS* can be obtained with a small number of shuffles, and further shuffles, that are very time consuming, do not cause very significant improvements.

6. CONCLUSIONS AND PERSPECTIVES

In this paper a new method, the *WMCS* for representing a cluster of graph has been proposed. The main advantage the method is that the structural information contained into the graphs is preserved and that is also easy to separate this information from noise. Due the high complexity of the method, an approximation is proposed. Experiments have been conducted to quantify the behavior of the approximated *WMCS* on the variation of the noise of the cluster. Preliminary tests show that a good approximated *WMCS* can be used for representing a cluster of graphs and this behavior encourage to produce a deeper analysis. A more precise measure of the performance could be obtained introducing a more complete model of distortions on the graphs.

REFERENCES

[1] 1 L.Brun, M. Mokhtari, Graph-based Representation in Different Application Domains. 3rd IAPR TC-15 GbR pp. 115-124, 2001.

[2] H. Bunke, P. Foggia, C. Guidobaldi, C. Sansone, M. Vento: A Comparison of Algorithms for Maximum Common Subgraph on Randomly Connected Graphs SSPR/SPR2002, pp 123-132. 2002.

[3] H. Bunke, M. Vento, Benchmarking of Graph Matching Algorithms,Proc.2nd GbR, Haindorf,pp.109-114, 1999

[4] H. Bunke, X.Jiang, A. Kandel, On the Minimum Common Supergraph of two Graphs, Computing 65, No. 1, pp. 13-25. 2000.

[5] C.De Mauro,M. Diligenti, M. Gori, M. Maggini, Similarity Learning for Graph-Based Image Representations. Proc. of the 3rd IAPR-TC15 GbR, pp.250-259. 2001

[6] R.O. Duda, P.E. Hart, D.G. Stork. Pattern Classification, II Ed., Wiley-Interscience Publication. 2000.

[7] S. Günter, H. Bunke, Self-organizing Map for Clustering in the Graph Domain, PRLe 23, pp401-417.2002.

[8] P. Foggia, C. Sansone, M. Vento, A Database of Graphs for Isomorphism and Sub-Graph Isomorphism Benchmarking, 3rd IAPR TC-15 GbR pp. 176-187, 2001.

[9] B. Luo, R. Wilson, R. Hancock, Spectral Feature Vectors for Graph Clustering. SSPR/SPR,pp.83-93.2002.

[10] J.J. McGregor, Backtrack Search Algorithms and the Maximal Common Subgraph Problem, Software Practice and Experience, Vol. 12, pp. 23-34, 1982.

[11] J.Vergés-Lahí, A.Sanfeliu, F.Serratosa, R.Alquézar, Face recognition: Graph matching versus neural techniques. NSPRIA'99,Spain,Vol.I, pp.259-266, 1999A.

[12] A.K.C. Wong, J. Constant, M.L. You, Random Graphs, in Syntactic and Structural Pattern Recognition - Theory and Applications 4:197-234. 1990.