Pattern Recognition 120 (2021) 108132

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Learning graph edit distance by graph neural networks

Pau Riba[a,*], Andreas Fischer[b,c], Josep Lladós[a], Alicia Fornés[a]

[a] Computer Vision Center, Universitat Autònoma de Barcelona, Spain
[b] Department of Informatics, DIVA Group, University of Fribourg, Fribourg, Switzerland
[c] Institute of Complex Systems, University of Applied Sciences and Arts Western Switzerland, Fribourg, Switzerland

ARTICLE INFO

Article history:
Received 3 August 2020
Revised 19 April 2021
Accepted 23 June 2021
Available online 2 July 2021

Keywords:
Graph neural networks
Graph edit distance
Geometric deep learning
Keyword spotting
Document image analysis

ABSTRACT

The emergence of geometric deep learning as a novel framework to deal with graph-based representations has faded away traditional approaches in favor of completely new methodologies. In this paper, we propose a new framework able to combine the advances on deep metric learning with traditional approximations of the graph edit distance. Hence, we propose an efficient graph distance based on the novel field of geometric deep learning. Our method employs a message passing neural network to capture the graph structure, and thus, leveraging this information for its use on a distance computation. The performance of the proposed graph distance is validated on two different scenarios. On the one hand, in a graph retrieval of handwritten words i.e. keyword spotting, showing its superior performance when compared with (approximate) graph edit distance benchmarks. On the other hand, demonstrating competitive results for graph similarity learning when compared with the current state-of-the-art on a recent benchmark dataset.

1. Introduction

Graph-based representation have been widely used in several application domains such as computer vision [1], bioinformatics [2], computer graphics [3] or pattern recognition [4,5]. Graphs are powerful and flexible representations able to describe shapes, images, knowledge, etc. in terms of relationships between constituent parts or primitives. In the core of any pattern recognition application, there is the ability to compare two objects. This operation, which is trivial when considering feature vectors defined in $\mathbb{R}^n$, is not properly defined in the graph domain [6,7]. Due to the inherent graph flexibility, it forces us to adopt some definitions of dissimilarity (similarity) ad hoc to particular purposes. Borgwardt [8] formally defines such problem as follows :

Definition 1.1 (Graph Comparison Problem)**.** Let $g_1 = (V_1, E_1, \mu_1, \nu_1)$ and $g_2 = (V_2, E_2, \mu_2, \nu_2)$ be two graphs from $\mathcal{G}$, the graph comparison problem is to find a function

$$d : \mathcal{G} \times \mathcal{G} \rightarrow \mathbb{R}$$

such that $d(g_1, g_2)$ quantifies the dissimilarity (similarity) of g_1 and g_2.

Lots of efforts have been made in this direction. In the literature, error-tolerant or inexact graph matching algorithms have been proposed. For instance, Zhou and de la Torre [9] proposed to factorize the large pairwise affinity matrix into smaller matrices that encode, on the one hand, the local node structure of each graph, and on the other hand, the pairwise affinity between both nodes and edges. Moreover, graph embeddings and kernels have been also proposed as mechanisms to compare two graphs. Graph embedding refers to those techniques that aim to explicitly map graphs to vector spaces [10,11]. Similarly, implicit graph embedding or graph kernel aims at finding a function able to map the input graph into a *Hilbert space* which basically defines a way to compute the similarity between two graphs in terms of a dot product [8,12]. One of the most popular error-tolerant graph matching methods is the graph edit distance (GED) [13–15]. Now, the graph comparison problem is formulated in terms of finding the minimum transformation cost in such a way that an isomorphism exists between the transformed graph g_1 and the second one g_2. In addition, GED algorithms, unlike most embedding and kernel methods, are able to cope with any type of labeled graph and any type of labels on nodes and edges.

The main drawback of GED techniques is that the time complexity is exponential in terms of the number of nodes of the input graphs. Hence, GED is unfeasible in a real scenario, where there may be no constraints in terms of the graph size. Therefore, several algorithms have been proposed to cope with this complexity [16,17]. However, these approximate algorithms only consider

* Corresponding author.
E-mail addresses: priba@cvc.uab.cat (P. Riba), andreas.fischer@unifr.ch (A. Fischer), josep@cvc.uab.cat (J. Lladós), afornes@cvc.uab.cat (A. Fornés).

https://doi.org/10.1016/j.patcog.2021.108132
0031-3203/

very local node structures in their computation and they do not adapt their costs according to the problem being addressed.

In Euclidean domains such as vectors, images or sequences, deep learning has been proposed as a solution to perform a huge variety of tasks. In the last decade, the particular case of deep neural networks have supposed a breakthrough in computer vision, pattern recognition and artificial intelligence [18]. However, until recent years, deep learning advances were not able to process non-Euclidean data in its framework. Lately, *geometric deep learning*[1] has emerged as a generalization of deep learning methods to non-Euclidean domains such as manifolds and graphs [19]. This field has arisen in the recent years allowing the developed models to encode structural and relational data. Several fields have benefited from this new paradigm, for instance computer vision [1], quantum chemistry [2] and computer graphics [3] among others.

Inspired by the efficient GED approximations and the powerful framework provided by the new advances in geometric deep learning, we propose to leverage its effectiveness as a learning framework to enhance a graph distance computation. Therefore, we are facing a graph metric learning problem. It can be formulated as a contrastive learning problem that finds contrast between similar and dissimilar objects. A siamese architecture is suitable for this problem. Bromley et al. [20] proposed it for signature verification. Siamese networks make use of the same model and weights on two separated branches in order to learn a representation where distances can be computed. Later, several approaches have extended this idea, being the triplet loss [21] one of the most successful methods. Recently, novel approaches have focused on extending this concept in order to exploit groups of samples instead of pairs or triplets [22]. Moreover, contrastive learning has been used, not only as a metric learning framework but it has also raised some attention due to its astonishing improvement on unsupervised learning tasks [23].

In this work, we define a triplet learning framework for the graph metric learning problem. Our proposed distance is inspired by the Hausdorff Edit Distance introduced by Fischer et al. [17] as an efficient approximation of the real graph edit distance with key differences. In comparison, our framework has the ability to enrich the initial graph representation by means of message passing operations. These learned features allows to use features distances as node substitution costs. Furthermore, insertion and deletion costs, are dynamically learned according to nodes local contexts. Therefore, we avoid a costly manual process on setting the edit cost operations per each specific problem. Moreover, a key difference is that our model does not require any distance at the edge level as it is already encoded at node level during the message passing phase. Thus, we present the first learnable extension of the Hausdorff edit distance. In addition, the proposed model is able to provide an interpretation in terms of the classical edit operation without explicit ground-truth.

The proposed approach is validated using standard graph datasets for keyword spotting and object classification. In this application scenario, the proposed approach based on a message passing neural network shows competitive results demonstrating the efficacy of our learning framework.

This article supposes a significantly extended version of our previous conference paper [24]. To the best of our knowledge, it was the first work that introduced the idea of learning a graph metric by means of message passing architectures. In this work, we have enhanced our previous graph neural network architecture. Specifically, the main changes from our previous work are (i) a new graph neural network architecture to obtain a better node context representation; (ii) a novel graph similarity which makes use of learned insertions, deletions and substitutions as edit operations; (iii) a learning approach making use of triplets of graphs with in-triplet hard negative mining. Finally, a thorough analysis and evaluation of the involved parameters as well as a performance comparison with the recent state-of-the-art literature is presented.

The rest of this paper is organized as follows. Section 2 introduces the related work on graph neural networks. Section 3 introduce the graph edit distance algorithm along with relevant approximations. Sections 4 and 5 proposes our learning framework and learning strategy. Section 6 evaluates the proposed framework. Finally, Section 7 draws the conclusions and future work.

2. Related work on geometric deep learning

In the following, some basic notations, definitions and previous works are presented in the context of geometric deep learning. As mentioned in the introduction, *Geometric deep learning* has emerged as a generalization of deep learning methods to non-Euclidean domains such as graphs and manifolds [19]. In this paper we will specifically focus on its application to graphs.

A graph is as a symbolic data structure describing relations (*edges*) between a finite set of objects (*nodes*). Graphs are formally defined as:

Definition 2.1 (Graph)**.** Let L_V and L_E be a finite or infinite label sets for nodes and edges, respectively. A *graph* g is a $4-$tuple $g=(V,E,\mu,\nu)$ where, V is the finite set of *nodes*, also known as *vertices*; $E\subseteq V\times V$ is the set of *edges*; $\mu: V\rightarrow L_V$ is the node labelling function, and; $\nu: E\rightarrow L_E$ is the edge labelling function.

We denote $|V|$ and $|E|$ as the *order* and *size* of a graph, namely, the number of nodes and edges respectively. Moreover, the *neighborhood* of a given node $v\in V$ in a graph g is defined as the set of nodes $\{v_i\}$ adjacent to v. We denote the neighborhood as $\mathcal{N}(v)$.

Graph Neural Networks were first proposed by Gori and Scarselli [25,26] as the first attempt to generalize neural networks to graphs. Later, Bruna et al. [27] proposed the first formulation of CNNs on graphs taking advantage of the graph spectral domain. However, the ability to process graph data came with a huge time complexity becoming inappropriate for real scenarios. Later, some works addressed these computatuinal drawbacks [28–30]. In its simplest form, a GNN layer is defined as

$$h^{(k+1)}=G_c(h^{(k)})=\rho\left(\sum_{B\in\mathcal{A}^{(k)}}Bh^{(k)}\Theta_B^{(k)}\right), \tag{1}$$

where $h^{(k)}$ is the node hidden state at the kth layer, ρ is a nonliniarity such as ReLU$(\cdot)$, $\mathcal{A}$ is a set of graph intrinsic linear operators that act locally on the graph signal and Θ are learnable parameters. The set of graph intrinsic linear operators can handle multi-relational graphs, however, in most cases, $\mathcal{A}$ only contains the adjacency matrix.

Recently, Gilmer et al. [2] proposed an approach named *Message Passing Neural Networks* (MPNNs) as a general supervised learning framework for graphs. This approach is able to generalize several GNN layers to a common pipeline. They propose to define each layer by means of two differentiable functions, on the one hand, a message function $M^{(k)}(\cdot)$ which collects the information from the neighboring nodes and edges according to

$$m_v^{(k+1)}=\sum_{u\in\mathcal{N}(v)}M^{(k+1)}(h_v^{(k)},h_u^{(k)},e_{vu}), \tag{2}$$

where $h_u^{(k)}$ and $h_v^{(k)}$ are the hidden states of nodes v and u at iteration k and $\mathcal{N}(v)$ denotes the neighbours of v in the graph g. On the other hand, an update function which updates the hidden state of the central node v according to the message $m_v^{(k+1)}$. The update function is formally defined as $h_v^{(k+1)}=U^{(k+1)}(h_v^{(k)},m_v^{(k+1)})$.

[1] http://geometricdeeplearning.com/

Several important GNN layers have been proposed along the last years, the most relevants for the scope of this work are the Graph Attention Networks (GAT) [31] and the Gated Graph Neural Networks (GG-NN) [32].

The literature on graph neural networks and their applications is quite large, so we refer the interested readers to recent survey papers for a comprehensive overview of these methodologies [19,33–35]. Very recently, Dwivedi et al. [36] presented a reproducible benchmarking framework.

In the context of similarity learning, several papers have explored the siamese neural network architecture. In its origins, Baldi et al. [37] makes use of a siamese model for fingerprint recognition. Siamese neural networks use a pair of samples to train with positive and negative examples i.e. being similar or not. Later, several approaches have extended this idea in order to take always into account positives and a negatives examples. These approaches are known as triplet networks [21]. In this case, three branches with shared weights are used to bring similar examples together and dissimilar examples to be far apart.

In the graph domain, Li et al. [38] presented two different models to solve the graph similarity problem, namely a graph embedding model, and a graph matching network. Both models can be trained with pairs or triplets. The **Graph embedding model** takes advantage of siamese GNN's to embed the given graphs into a vectorial space. Then, given the pair of vectorial representations, a similarity metric in the vector space can be computed by means of the Euclidean, cosine or Hamming similarities. Very similar approaches have also been presented in other works, for instance Chaudhuri et al. [39] who trained a similar approach with contrastive loss or the work introduced by Zhang et al. [40] which uses the L_2 loss to mimic the real similarity score. On the other hand, the **Graph matching networks (GMN)** uses two GNNs with shared weights to process the input graphs. However, in this case, the authors propose to modify the node update module to take into account not only the aggregated messages on the edges of each graph, but a cross-graph message which measures how well the nodes match from one graph to the other. Finally, following the same idea as the previous model, each graph is finally converted into a vectorial representation which is later used in a similarity metric.

SimGNN is another interesting approach proposed by Bai et al. [41]. In this work, the authors proposed to combine graph-level embeddings and node-node similarity scores by taking their histogram of features. However, as the histogram function is not differentiable, this methodology still relies on the graph-level embedding for computing the final similarity score. Their model is trained according to the real graph edit distance for small datasets whereas the smallest distance computed by three well-known approximate algorithms is taken to handle large datasets. The authors extended this work by proposing a new model named GraphSim [42]. In this architecture, only three node-node similarities scores are used corresponding to node embeddings at different scales. After that, the similarity matrices are treated as images and a CNN is used to process them to discover the optimal node matching pattern. However, to deal with the permutation invariant ordering of graph nodes, they propose a BFS ordering. Moreover, the similarity matrices are padded to $\max(|V_1|, |V_2|)$, where V_1 and V_2 are the node sets of the graphs involved and resized to meet the expected size. Finally, a precomputed similarity score is used to guide the training.

Compared to these works, our model makes use of a node-node distance matrix to obtain a global graph distance metric. Therefore, we are not obtaining a global vectorial representation of our graphs nor applying cross-convolution layers in our graph neural network architecture. This allows us to make use of any differentiable graph or set distance, which preserves the permutation invariance property, while avoiding the computational overhead of the cross-convolution layers. Moreover, we avoid the loss of structural information of other approaches when obtaining a vectorial graph representation by explicitly dealing with the structure in the distance itself.

3. Related work on graph edit distance

This section introduces the concept of *Graph Edit Distance* (GED) jointly with relevant cubic and quadratic time approximations.

The *Graph Edit Distance* (GED) [14,15,43] evaluates the similarity of two graphs in terms of edit operations. The GED is inspired by the *String Edit Distance* (SED), also known as the *Levenshtein* distance [44,45]. In this specific case, the order of the characters allows the efficient use of dynamic programming to find the string-to-string correspondence. However, for general graphs, the correspondence cannot rely on a specific ordering of nodes and edges. The idea of GED is to compute the minimum cost transformation from the source graph g_1 to the target one g_2 in terms of a sequence of edit operations $e_1, \ldots, e_k$. This sequence is named *edit path* between g_1 and g_2. Usually *insertion, deletion* and *substitutions* for nodes and edges are considered. GED is formally defined as

Definition 3.1 (Graph Edit Distance)**.** Let $g_1 = (V_1, E_1, \mu_1, \nu_1)$ and $g_2 = (V_2, E_2, \mu_2, \nu_2)$ be the source and the target graphs respectively. The *graph edit distance* between g_1 and g_2 is defined by

$$d(g_1, g_2) = \min_{(e_1,\ldots,e_2)\in\Upsilon(g_1,g_2)} \sum_{i=1}^{k} c(e_i),$$

where $\Upsilon(g_1, g_2)$ denotes the set of edit paths transforming g_1 into g_2, and $c(e_i)$ denotes the cost function measuring the strength of the edit operation e_i.

The GED is a known NP-complete problem (see [46] for a detailed proof), exponential with respect to the number of nodes. Thus, in addition to exact GED algorithms, some efficient approximations have been proposed [47,48]. In the following, we review in detail two relevant techniques

The *Assignment Edit Distance* (AED), also known as *bipartite graph matching*, proposed by Riesen et al. [16], is a cubic time approximation of GED with respect to the number of nodes of the involved graphs. It provides an upper bound of order $\mathcal{O}\big((n_1+n_2)^3\big)$ where $n_1 = |V_1|$ and $n_2 = |V_2|$.

The main idea is to transform the GED computation to an assignment problem between nodes and their local structure. This method defines a matrix of edit costs between the nodes of both graphs. Afterwards, the best correspondence between nodes is found by a linear assignment method [49]. The matrix definition for the AED algorithm takes into consideration both the local structure of the vertices and their attributes. The cost matrix C is defined as

$$C = \begin{bmatrix} c_{1,1} & \ldots & c_{1,m} & c_{1,\varepsilon} & \ldots & \infty \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ c_{n,1} & \ldots & c_{n,m} & \infty & \ldots & c_{n,\varepsilon} \\ c_{\varepsilon,1} & \ldots & \infty & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \infty & \ldots & c_{\varepsilon,n} & 0 & \ldots & 0 \end{bmatrix}$$

where $c_{i,j}$ denotes the cost of a node substitution $c(u_i \to v_j)$; $c_{i,\varepsilon}$ denotes the cost of a node deletion $c(u_i \to \varepsilon)$; and $c_{\varepsilon,j}$ denotes the costs of a node insertion $c(\varepsilon \to v_j)$ where $v_i \in V_1$ and $v_j \in V_2$.

Other works have focused on speeding up this method in some particular settings. For example, Serratosa et al. [50] define an algorithm able to reduce the computation time with the only constrain that the edit costs should define the GED as a real distance function.

Despite obtaining a good approximation, the time complexity of the AED algorithm is still a problem for some applications where the time is an important constrain. To alleviate this issue, Fischer et al. [17] proposed the *Hausdorff Edit Distance* (HED) which is a lower bound approximation of the real GED with a quadratic time complexity of $\mathcal{O}(n_1 \cdot n_2)$ where $n_1 = |V_1|$ and $n_2 = |V_2|$. HED is based on the *Hausdorff distance* which is formally defined as

Definition 3.2 (Hausdorff distance)**.** Let A and B be two non-empty subsets of a metric space (M, d). The Hausdorff distance $d_H(A, B)$ is defined as

$$d_H(A,B) = \max\left(\sup_{a\in A}\inf_{b\in B} d(a,b), \sup_{b\in B}\inf_{a\in A} d(a,b)\right).$$

For finite sets A,B the Hausdorff distance is reformulated as

$$d_H(A,B) = \max\left(\max_{a\in A}\inf_{b\in B} d(a,b), \max_{b\in B}\inf_{a\in A} d(a,b)\right).$$

By definition, the Hausdorff distance is very sensitive to outliers. Hence, they propose to replace the maximum with the summation operation, which forces the distance to take into account all nearest neighbour distances and becomes more robust to noise than the original one. Thus, the new distance is defined as

$$d_{\hat{H}}(A,B) = \sum_{a\in A}\min_{b\in B} d(a,b) + \sum_{b\in B}\min_{a\in A} d(a,b). \tag{3}$$

Until now, this distance $d_{\hat{H}}(\cdot)$ does not consider node insertions and deletions. Therefore, from Eq. (3) they define a new distance on graphs. Given two graphs $g_1 = (V_1, E_1, \mu_1, \nu_1)$ and $g_2 = (V_2, E_2, \mu_2, \nu_2)$ and a matching cost defined as c, the HED is defined as

$$\mathrm{HED}(g_1,g_2,c) = \sum_{u\in V_1}\min_{v\in V_2\cup\{\varepsilon\}} c_n^*(u,v) + \sum_{v\in V2}\min_{u\in V_1\cup\{\varepsilon\}} c_n^*(v,u), \tag{4}$$

where $c_n^*(u,v)$ is a modified node matching cost defined as,

$$c_n^* = \begin{cases} \frac{c_n(u,v)}{2}, & \text{if } (u\to c) \text{ is a substitution} \\ c_n(u,v), & \text{otherwise.} \end{cases}$$

This redefinition of the node matching cost is needed because HED does not enforce bidirectional substitutions. Eq. (4) is composed by a summation, hence, only if both directions are considered, the full cost will be taken into account. The same matching algorithm is considered, if needed, for the edge matching. In this setting, the HED finds an optimal assignment per node instead of a global optimal assignment as pretended by the real GED.

A typical drawback of GED approximation algorithm is that it only relies on local edge structures rather than global information. Some efforts have been made to improve the performance by increasing the node context at matching time [51]. However, obtaining a better knowledge on the relation of each node within the graph is still an open issue that we aim to address by the new advances on graph neural networks.

4. The learned graph distance framework

This section presents our proposed learning framework for graph distance inspired by the Hausdorff edit distance [17]. As a learning setting, the proposed architecture can be trained either with pairs or triplets of graphs. Hence, as ground-truth, only the information on whether or not two graphs belong to the same class is required. Note, that in our proposed approach, we do not require the node correspondence information nor the real graph edit distance. Instead, the node assignment, i.e.the edit costs for insertions, deletions and substitution, is implicitly learned. Therefore, we do not require to manually tune these parameters following the traditional pipeline. Although our framework can be trained using pairs, in this work we will focus on the triplet setting. Hence, we use three GNN with shared weights.

Fig. 1 shows a graphical outline of the proposed approach. Our pipeline can be divided in two stages. First, a graph neural network $\phi(\cdot)$ is used to obtain a node-level embedding which codifies the local context information, in terms of structure, for each node. Second, a novel graph similarity algorithm based on the Hausdorff edit distance is proposed as a technique to compare two graphs $d_\theta(\cdot)$. Observe that $d_\theta(\cdot)$ can be replaced by any differentiable graph distance approach. Each stage is carefully described next. First, the GNN architecture is explained in detail. Afterwards, the proposed graph similarity is developed.

4.1. Learning node embeddings

The first stage of our framework is a graph neural network architecture $\phi(\cdot)$ able to learn a new graph representation in terms of node embeddings. Our architecture consists of K propagation layers that map the input graph to an enriched representation. Thus, each propagation layer takes a set of node representations at layer k, $\{h_i^{(k)}\}_{i\in V}$ and maps it to a new node representation $\{h_i^{(k+1)}\}_{i\in V}$ at layer $k+1$. We evaluate the following two different architectures according to two different message passing strategies.

GAT-based model This model uses graph attention networks (GAT), introduced by Velikovi et al. [31]. A GAT layer is formally defined as

$$h_v^{(k+1)} = \sum_{u\in\mathcal{N}(v)} \alpha_{vu} W^{(k)} h_u^{(k)}, \tag{5}$$

where α_{vu} is the attention score between node v and node u, $W^{(k)}$ are the learned weights and $h_v^{(k)}$ is the hidden state of node v, both, at layer k. The attention score α_{vu} is learned by

$$\begin{aligned} \alpha_{vu}^{(k)} &= \mathrm{softmax}_v(e_{vu}^l) \\ e_{vu}^{(k)} &= \mathrm{LeakyReLU}\big(\vec{a}^t[W^{(k)}h_v^{(k)} \| W^{(k)}h_u^{(k)}]\big), \end{aligned} \tag{6}$$

where $\vec{a}$ and $W^{(k)}$ are a vector and a matrix of learned weights. Moreover, a multi-head attention can be used to enrich the model capacity and to stabilize the learning process.

In our setting, we use these layers with residual connections, four attention heads and BatchNorm layers [52] with the exception of the last layer. The attention heads are concatenated at the intermediary layers and averaged for the final layer.

GRU-based model This architecture is based on the gated graph neural networks (GG-NN) proposed by Li et al. [32]. Originally, the message function is formulated as $M(h_v^{(k)}, h_u^{(k)}, e_{vu}) = A_{e_{vu}} h_u^{(k)}$, where $A_{e_{vu}}$ is a learned matrix for each possible edge label. Note that we are restricted to a discrete set of labels. In order to overcome this constrain, Gilmer et al. [2] proposed to use a modified message function defined as $M(h_v^{(k)}, h_u^{(k)}, e_{vu}) = A(e_{vu})h_u^{(k)}$, where $A(e_{vu})$ is a neural network which maps the edge vector to a matrix $d \times d$. This modification allows the use of non-discrete information as edge attributes. Finally, the update function is defined as $U(h_v^{(k)}, m_v^{(k)}) = \mathrm{GRU}(h_v^{(k)}, m_v^{(k)})$, where GRU is the Gated Recurrent Unit [53]. In its original formulation, the first node hidden state is padded with zeros to meet the size defined by the GRU, however, we propose to use a fully-connected layer as a first node embedding. Moreover, we propose to incorporate edge features according to the source and destination node according to, $e_{vu} = \mathrm{MLP}(|h_v^{(1)} - h_u^{(1)}|)$ as proposed in Garcia and Bruna [54]. There, MLP stands for multi-layer perceptron and the absolute value is used to preserve the simmetry of the edge direction.

4.2. Graph distance or similarity

Following the idea of HED defined in Eq. (4), we propose to dynamically adapt the insertions and deletion costs according to

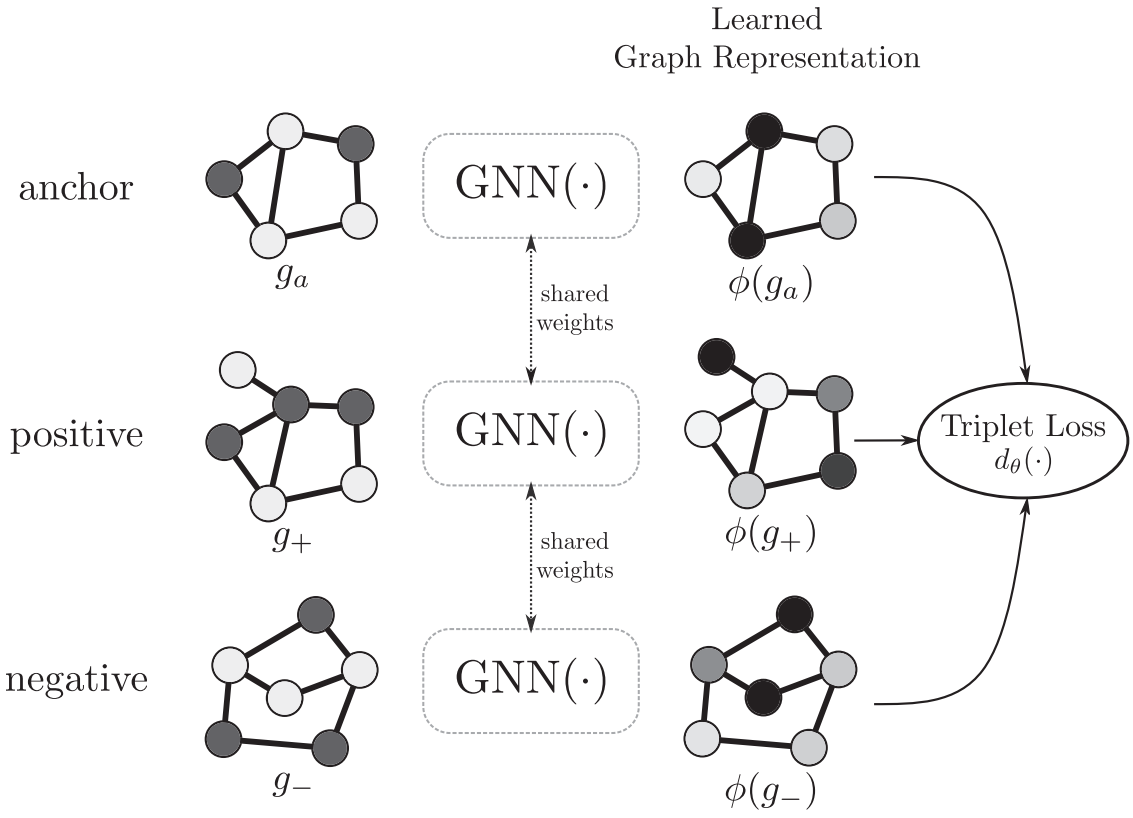


Fig. 1. Overview of our learning framework. Given a triplet of graphs (g_a, g_+, g_-) as the anchor, positive and negative samples respectively, the GNN $\phi(\cdot)$ learns a graph representation per each one $(\phi(g_a), \phi(g_+), \phi(g_-))$, which can be matched by means of a learned distance $d_\theta(\cdot)$.

the application domain. With this aim, we introduce two learnable costs $c(\varepsilon \to v)$ and $c(u \to \varepsilon)$ for the insertion and deletion of nodes. Thus, taking advantage of the computed node embeddings, our nodes are enriched with information aggregated from their local context and, therefore, its importance within the graph. Thus, we propose to take advantage of this in order to define two neural networks $\varphi_i(\cdot)$ and $\varphi_d(\cdot)$, defined as $\varphi_* : \mathbb{R}^n \to \mathbb{R}^+$, able to decide the corresponding cost of this operation. In our experiments, $\varphi_i(\cdot)$ and $\varphi_d(\cdot)$ are the same network with shared weights as we consider the insertion and deletion operations to be symmetric. Moreover, we take the absolute value as the insertion and deletion costs must be positive.

Therefore, we define the distance between two graphs $g_1 = (V_1, E_1, \mu_1, \nu_1)$ and $g_2 = (V_2, E_2, \mu_2, \nu_2)$ as

$$d_\theta(g_1, g_2) = \frac{1}{|V_1|+|V_2|} \begin{array}{l} \Big(\sum_{u \in V_1 \cup \{\varepsilon\}} \min_{v \in V_2 \cup \{\varepsilon\}} c_\theta(u, v) \\ \quad + \sum_{v \in V_2 \cup \{\varepsilon\}} \min_{u \in V_1 \cup \{\varepsilon\}} c_\theta(u, v)\Big), \end{array} \tag{7}$$

where θ are learnable parameters and $c_\theta(\cdot, \cdot)$ is the corresponding learnable cost function defined as

$$c_\theta(u, v) = \begin{cases} \varphi_d(u; \theta) & \text{if } (u \to \varepsilon) \text{ is a deletion,} \\ \varphi_i(v; \theta) & \text{if } (\varepsilon \to v) \text{ is an insertion,} \\ \frac{d(u,v)}{2} & \text{otherwise.} \end{cases} \tag{8}$$

In our scenario, the edges are not taken into account as we consider the local structures to be already encoded during the message passing phase. However, edges can be incorporated to Eq. (8), considering the adjacent edges as nodes and applying the same distance $d_\theta(\cdot)$ with different learned weights.

Observe that an important aspect of the proposed distance is the fact that the node correspondence might not be symmetric. Fig. 2 illustrates this issue, moreover, we also show the effect of considering insertions and deletions as a ε node in Fig. 2(b). Note that not considering ε nodes as proposed in Riba et al. [24] and illustrated in Fig. 2(a), can limit the learning capabilities of the framework.

A limitation of our approach is that in some scenarios it might lose some node feature information in favor of encoding the local node structure. As a solution, we optionally combine the original graph information into Eq. (8), that can be redefined as,

$$c_\theta(u, v) = \begin{cases} \tau_d + \varphi_d(u; \theta) & \text{if } (u \to \varepsilon) \text{ is a deletion,} \\ \tau_i + \varphi_i(v; \theta) & \text{if } (\varepsilon \to v) \text{ is an insertion,} \\ \frac{d(u,v)+d'(u,v)}{2} & \text{otherwise,} \end{cases} \tag{9}$$

where $\tau_d, \tau_i > 0$ are user-defined parameters to fix the minimum cost for node deletion and insertion respectively; $d'(\cdot, \cdot)$ corresponds to the distance computed on the original node attributes. We find this setting helpful to avoid incorrect matchings between nodes that share structurally similar neighborhoods.

5. Training setting and learning objective

In this work, we follow the idea of triplet networks to exploit the ranking properties of the desired metric. Thus, we use three GNN models with shared weights following the architecture illustrated in Fig. 1. Our model is trained in a supervised manner, so we know which pairs of graphs belong to the same class. Compared to other approaches, we do not require node assignments nor a pre-computed similarity score.

The objective function is the triplet loss, also known as margin ranking loss. This learning objective receives three samples in which we already know its ranking, i.e. which pair should have a higher similarity score or distance. Let $\{g_a, g_+, g_-\}$ be a triplet training sample where, g_a is the anchor graph, g_+ is a positive graph sample i.e. a sample different from g_a but belonging to the same class and g_- is a negative graph example i.e. a sample be-

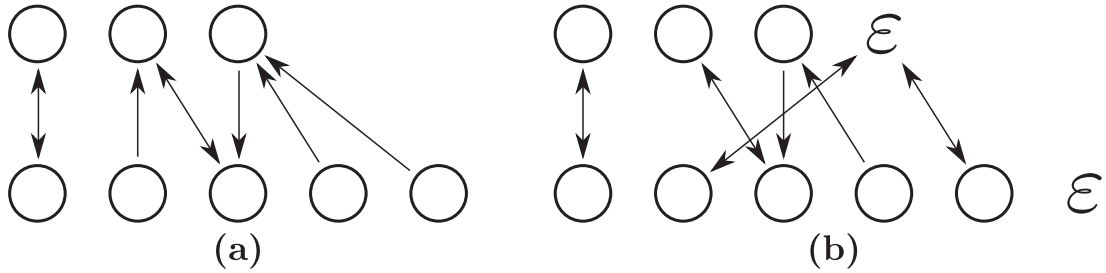


Fig. 2. Assignment problem according to the proposed distance. (a) only substitutions are considered, (b) insertions and deletions are included as an extra epsilon node.

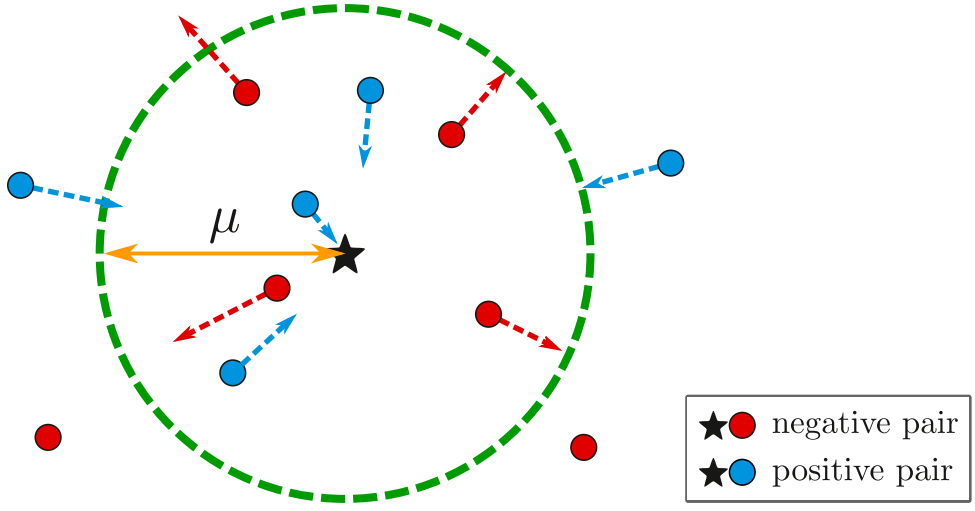


Fig. 3. Illustration of the triplet learning objective. The anchor graph illustrated as a star should be close to its positive pair, in blue, and farther than a margin μ to its negative counterpart. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

longing to a different class. Then, the triplet loss is defined as

$$\mathcal{L}(\delta_+, \delta_-) = \max(0, \mu + \delta_+ - \delta_-), \tag{10}$$

where μ is a fixed margin parameter, $\delta_+ = d_\theta(\phi(g_a), \phi(g_+))$ is the distance with respect to the positive sample and $\delta_- = d_\theta(\phi(g_a), \phi(g_-))$ is the distance with respect to the negative sample. Fig. 3 illustrates how this loss performs. Note that positive pairs are pushed to be close each other whereas negative samples are separated at least by the predefined margin μ.

Moreover, following the idea introduced in Balntas et al. [55], we apply an in-triplet hard negative mining which means that the anchor and positive samples can be swapped in case the positive sample is harder than the anchor one. Hence, we define $\delta'_- = d_\theta(\phi(a), \phi(n))$ and $\delta_* = \min(\delta_-, \delta'_-)$. Finally, the new loss with the anchor swap is defined as

$$\mathcal{L}(\delta_+, \delta_*) = \max(0, \mu + \delta_+ - \delta_*). \tag{11}$$

Algorithm 1 presents our training strategy. $\Gamma(\cdot)$ denotes the optimizer function. All models are trained using the *Adam* optimizer [56] with weight decay i.e. L_2 regularization. The learning rate of 0.001 is multiplied by 0.95 every 5 epochs to decrease its value, and we applied early stopping to finish our training process. Regarding the edit cost operation described in Eq. (9), we experimentally set the parameters τ_i and τ_d to 0.5. If not specified, all the models used $K = 3$ GNN layers. In addition, both the final node embedding and the hidden state of each layer has a size of 32. Moreover, in the case of the GAT-based model, 4 attention heads have been used whereas in the case of the GRU-based model, (i) the node embedding is a simple linear layer that maps our initial node features to a tensor of size 32; (ii) the edge features e_{vu} uses two linear layers with ReLU activation functions; (iii) the neural network $A(e_{eu})$ is defined as two linear layers with a ReLU activation in between that generates a matrix of the hidden state dimension (32×32).

Algorithm 1 Training algorithm for our proposed model.

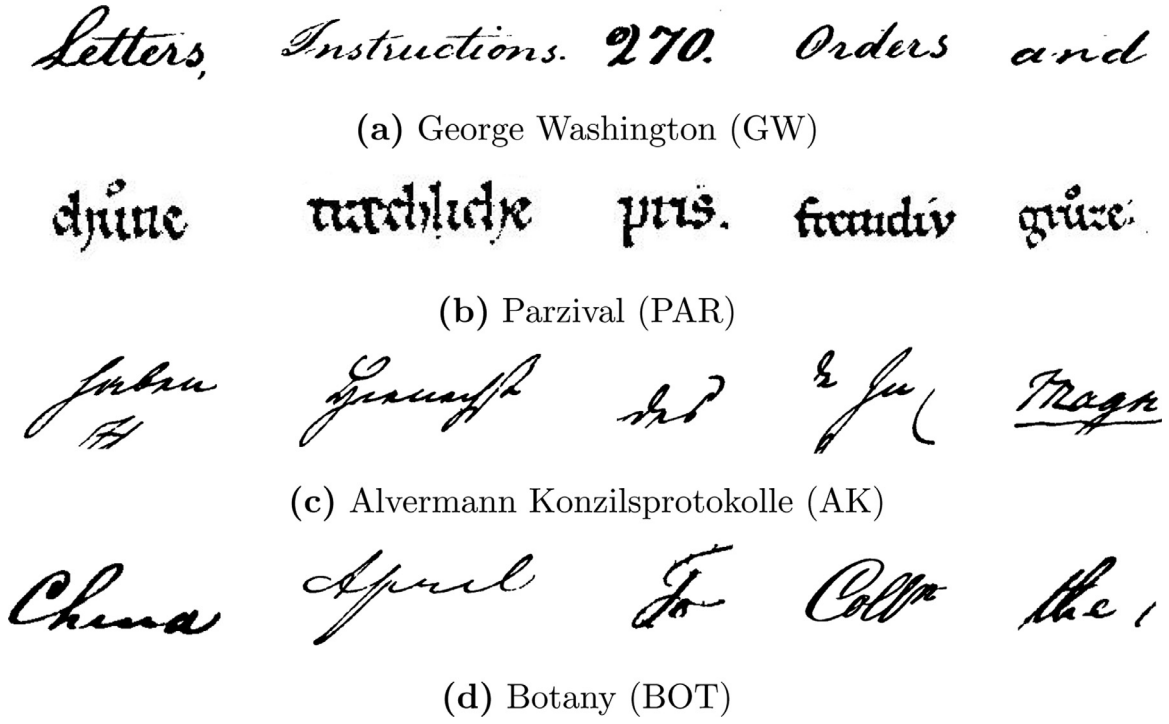
Require: Input data $\mathcal{G}$; max training iterations T
Ensure: Networks parameters $\Theta = \{\Theta_\phi, \theta\}$.
1: **repeat**
2: Sample triplet mini-batches $\{g_a, g_+, g_-\}_{i=1}^{N_B}$
3: $\mathcal{L} \leftarrow$ Eq. 11
4: $\Theta \leftarrow \Theta - \Gamma(\nabla_\Theta \mathcal{L})$
5: **until** Convergence or max training iterations T

6. Experimental validation

For validating our approach, a keyword spotting application in historical manuscripts has been considered as our main application scenario. Moreover, a final experiment on a classical mesh graph dataset is conducted. Our empirical evaluation demonstrate that the proposed approach provides competitive results when compared to the state-of-the-art. All the code is available at github.com/priba/graph_metric.pytorch.

6.1. Historical keyword spotting

In Document Image Analysis and recognition, Keyword Spotting (KWS), also known as word spotting, has emerged as an alternative to handwritten text recognition for documents in which the transcription performance is not satisfactory. Therefore, KWS is formulated as a content-based image retrieval strategy which relies upon



(a) George Washington (GW)

(b) Parzival (PAR)

(c) Alvermann Konzilsprotokolle (AK)

(d) Botany (BOT)

Fig. 4. Pre-processed word examples of the four datasets.

obtaining a robust word image representation and a subsequent retrieval scheme.

6.1.1. Dataset description

The HistoGraph dataset [57,58] is a graph database for historical keyword spotting evaluation.[2] It consists of different well known manuscripts.

George Washington (GW) [59]: This database is based on handwritten letters written in English by George Washington and his associates during the American Revolutionary War in 1755.[3] It consists of 20 pages with a total of 4894 handwritten words. Even tough several writers were involved, it presents small variations in style and only minor signs of degradation.

Parzival (PAR) [59]: This collection consists of 45 handwritten pages written by the German poet Wolfgang Von Eschenbach in the 13th century. The manuscript is written in Middle High German with a total of 23,478 handwritten words. Similarly to GW, the variations caused by the writing style are low, however, there are remarkable variations caused by degradation.

Alvermann Konzilsprotokolle (AK) [60]: It consists of German handwritten minutes of formal meetings held by the central administration of the University of Greifswald in the period of 1794 to 1797. In total 18,000 pages were used with small variations in style and only minor signs of degradation.

Botany (BOT) [60]: It consists of more than 100 different botanical records made by the government in British India during the period of 1800 to 1850. The records are written in English and contain certain signs of degradation and especially fading. The variations in the writing style are noticeable especially with respect to scaling and intra-word variations.

Fig. 4 provides some examples of pre-processed word images from which the graphs are created. Observe that the word segmentation of AK and BOT datasets is imperfect [60]. Moreover, these two datasets do not provide a validation set. So, some images from the training set have been used for validation. Table 1 provides an overview of the dataset in terms of number of words.

Table 1
Dataset overview in terms of number of keywords and word images for training, validating and testing respectively.

Dataset	Keywords	Train	Validation	Test
GW	105	2447	1224	1224
PAR	1217	11,468	4621	6869
BOT	150	1684	–	3380
AK	200	1849	–	3734

To obtain a graph for each word in these datasets, the two most promising graph constructions introduced in Stauffer et al. [58] have been used:

- **Keypoint**: Characteristic points are extracted from the skeletonized word image. Moreover between the connected characteristic points, equidistant nodes are inserted on top of the word skeleton.
- **Projection**: An adaptative grid is generated according to the vertical and horizontal projection profiles. Then, nodes are inserted in the corresponding center of mass of each grid cell. Moreover, undirected edges are inserted if nodes are directly connected by a stroke.

All the datasets presented in this work only contain spatial and structural information. This means that nodes are labelled with its normalized (x, y)-position on the image and that edges are unlabelled.

6.1.2. Experimental protocol

Following the evaluation schemes of previous word spotting methodologies, we performed our experiments on two evaluation protocols.

[2] Available at http://www.histograph.ch/

[3] George Washington Papers at the Library of Congress from 1741 to 1799, Series 2, Letterbook 1, pages 270–279 and 300–309, https://www.loc.gov/collections/george-washington-papers/about-this-collection/

Table 2
Study on the GNN model and margin parameter of the proposed model. Mean average precision (mAP) and standard deviation (average on four runs) for graph-based KWS system on George Washington (GW), Parzival (PAR) Alvermann Konzilsprotokolle (AK) and Botany (BOT) datasets.

	Model	μ	GW		PAR		AK		BOT	
			mAP	±	mAP	±	mAP	±	mAP	±
Keypoint	GAT	1	72.49	1.169	66.46	3.162	62.90	1.325	39.86	0.396
		10	**76.92**	2.309	**73.14**	0.973	62.72	1.783	**41.52**	0.782
	GRU	1	72.86	3.331	67.27	1.281	**64.42**	1.003	39.69	0.532
		10	68.45	2.715	48.59	9.571	60.84	1.127	38.22	0.778
Projection	GAT	1	67.86	2.379	70.77	1.906	63.44	1.233	39.12	2.037
		10	**70.25**	3.431	**75.19**	0.755	62.72	1.518	38.83	2.801
	GRU	1	68.09	1.234	71.07	1.933	**65.04**	1.226	**42.83**	0.568
		10	63.39	4.222	52.32	1.298	60.51	1.451	37.59	0.778

- **Individual**: Each query image follows the traditional retrieval pipeline. Thus, queries are matched against the elements in the gallery and each ranking is evaluated independently.
- **Combined**: A query can consist of a set of graphs $Q = \{q_1, \dots, q_t\}$ where all the instances $q \in Q$ represent the same keyword. In this case, we consider the minimal distance achieved on all t query graphs. This second evaluation protocol was adopted in some previous graph-based word spotting works [58,61]. The motivation of this setting is to mitigate the structural bias provided by the query instance, i.e. different handwriting styles can provide extremely different graphs, so, in terms of graph distances, it is unrealistic to consider them from the same class.

For the evaluation, we use the mean Average Precision (mAP), a classic information retrieval metric [62]. First, let us define Average Precision (AP) as

$$\mathrm{AP} = \frac{\sum_{n=1}^{|\mathrm{ret}|} P@n \times r(n)}{|\mathrm{rel}|}, \tag{12}$$

where $P@n$ is the precision at n and $r(n)$ is a binary function on the relevance of the nth item in the returned ranked list. Then, the mAP is defined as:

$$\mathrm{mAP} = \frac{\sum_{q=1}^{Q} \mathrm{AP}(q)}{Q}, \tag{13}$$

where Q is the number of queries.

6.1.3. Ablation study

We first empirically investigate the influence of the margin parameter μ to each model, as well as the importance of the GNN layers choice. Table 2 presents a comparison of the different settings, providing the averaged mAP of 4 runs and its corresponding standard deviation. This evaluation has been done for all the datasets and for both graph representations *viz.* Keypoint and Projection. The evaluation protocol in this experiment is Individual as we believe that it is the natural experimental setting for this problems.

From these results, we observe that GRU-based models are slightly better, allowing a higher degree of deformations between words from the same class. Note that both datasets AK and BOT do contain imperfect word segmentations. In addition, these datasets contain samples written with a more artistic calligraphic style as shown in Fig. 4. Thus, the artistic strokes are drivers of a higher degree of complexity. Observe that the performance drop in BOT dataset can be also explained by the artistic nature of the dataset.

Additionally, the Keypoint representation performs the best on the GW dataset, since the strokes are simpler and its structure in terms of the binary image skeleton is more relevant for a proper retrieval. Finally, we observe that a higher margin μ is more adequate for the GAT-based models whereas it's harmful for the GRU-based one.

Table 3
State-of-the-art on graph-based KWS techniques. Mean average precision (mAP) for graph-based KWS system on GW, PAR, AK and BOT datasets.

Distance	Representation		GW	PAR	AK	BOT
AED [16]	Keypoint [61]		68.42	55.03	77.24	50.94
	Projection [61]		60.83	63.35	76.02	50.49
	Ensemble [58]	min	70.56	67.90	82.75	65.19
		max	62.58	67.57	82.09	67.57
		mean	69.16	79.38	84.25	**68.88**
		sum_{α}	68.44	74.51	**84.77**	68.77
		$\mathrm{sum}_{\mathrm{map}}$	70.20	76.80	84.25	**68.88**
HED [17]	Keypoint [61]		69.28	69.23	79.72	51.74
	Projection [61]		66.71	72.82	**81.06**	51.69
Ours	Keypoint		**78.48**	79.29	78.64	51.90
	Projection		73.03	**79.95**	79.55	**52.83**

Table 4
Comparison against non-graph learning based systems. Mean average precision (mAP) for graph-based KWS system on AK and BOT datasets.

Method	Representation	AK	BOT
CVCDAG [64]	–	77.91	75.77
PHOCNet [63]	–	96.05	89.69
QTOB [65]	–	82.15	54.95
Ours	Keypoint	64.42	41.52
	Projection	65.04	42.83

6.1.4. Results and discussion

Table 3 compares with graph-based methodologies. In this setting we follow the Combined evaluation protocol reported by Stauffer et al. [58], Ameri et al. [61]. For each dataset and graph representation, we use the best model reported in the previous section. Observe that, for a fair comparison, that is, using the same graph representation, we outperform both AED [16] and HED [17] on all the datasets but AK, where we obtain very similar results. However, the ensemble methods reported in Stauffer et al. [58] are able to obtain a better performance on the datasets with more variability. Note that these ensembles combine, in different ways, the graph distances computed on different graph representations of the same images. Therefore, we do not consider it a fair comparison but a remarkable fact that the performance of our system is able to outperform them in two of the datasets while obtaining competitive results on the other two.

Table 4 shows a comparison with non-graph based approaches. In particular, we compare against three state-of-the-art learning-based reference systems of the ICFHR2016 competition [60]. In this case, the evaluation of these learning-based frameworks follows the protocol described in the competition. Thus, queries of the same query keyword are considered to be independent. Note that, due to this query protocol, the learning-based frameworks are not directly comparable to the state-of-the-art graph-based KWS

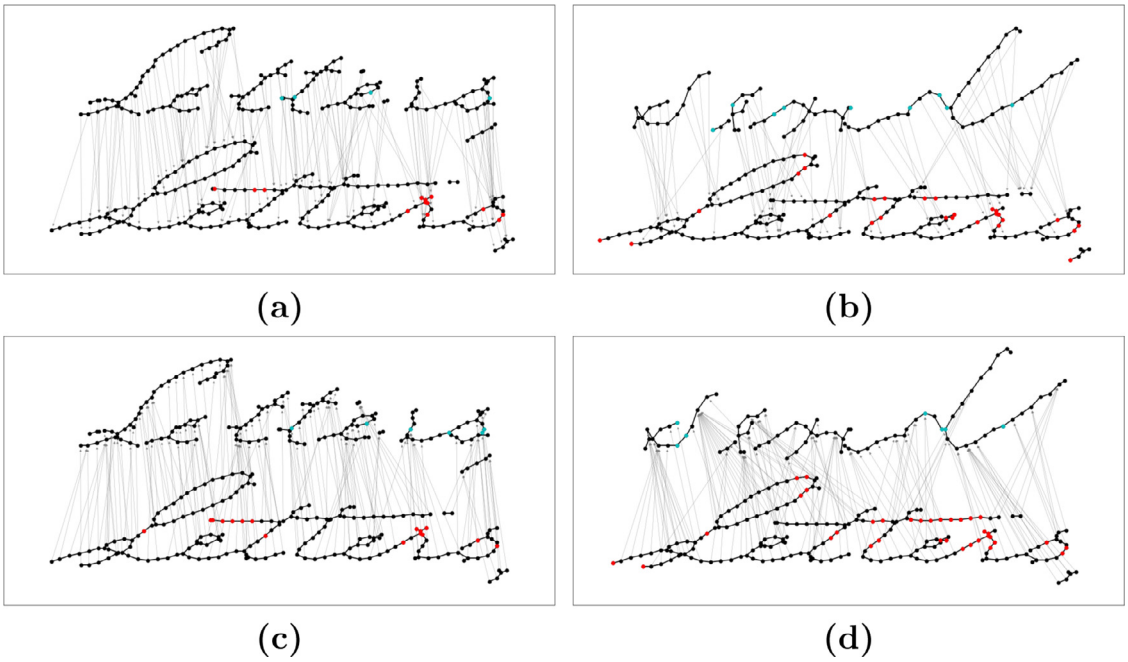


Fig. 5. Visualization of the learned node correspondence. First row, shows the node matching from top to bottom; the second row of the figure shows de opposite. (a)–(d) "Letters"–"Letters"; (b)–(d) "send"–"Letters".

Table 5
Median and maximum number of nodes, mean runtime per graph pair in milliseconds for AED, HED and the two proposed architectures on the George Washington (GW) dataset.

Representation	Statistics		GED		Ours	
	$\|V\|_{\text{med}}$	$\|V\|_{\max}$	T_{AED}	T_{HED}	T_{GAT}	T_{GRU}
Keypoint	81	366	303.0	3.2	17.33	7.77
Projection	80	391	344.1	3.9	17.28	7.81

results that we have reported above. In this table, we can observe the superiority of learning methods working directly on the image domain. In particular, PHOCNet [63] leads to stunning accuracies for this task. However, the proposed graph-based approach is able to provide a new step towards closing the gap between structural and statistical methodologies on this kind of tasks. In addition, the proposed approach is able to deal with the noise introduced by the graph construction.

Table 5 shows the inference time of both models, GAT and GRU in the GW dataset. Even though we have to take into account the different environment conditions such as the programming language, or GPU vs CPU; we observe that despite being significantly slower than the plain HED we are able to obtain 39x speed up in comparison to the AED. From the theoretical perspective, our distance formulation has a quadratic time complexity $\mathcal{O}(n_1 \cdot n_2)$ where $n_1 = |V_1|$ and $n_2 = |V_2|$, following the HED previously discussed. In addition, the GNN used to enhance our graph representation has a huge impact in temrs of time. As reported in Gilmer et al. [2], the complexity of each message-passing phase for a dense graph is $\mathcal{O}\left(|V|^2 d^2\right)$ where d are the dimension of the node embeddings. Note that the GAT-based model has an over-head due to its attention heads. In their work, when computing d' features they report a time complexity of $\mathcal{O}\left(|V|dd' + |E|d'\right)$.

Finally, Fig. 5 provides qualitative examples of our matching framework for a positive and negative sample. The first row provides the top to bottom matching whereas the second row shows the opposite, from the bottom to the top. Notice that in the positive sample case, both directions are much more consistent than in the negative sample case.

6.2. Experimental comparison to GMN

Among the graph metric learning approaches in the literature, the graph matching networks (GMN) work [38] is the most prominent one. In this section, we propose an extra experiment to compare with their work.

6.2.1. Dataset description

The IAM Graph Database Repository [66] provides several graph datasets covering a wide spectrum of different applications. In particular, we focused on the COIL-DEL dataset. The COIL-100 [67] consists of 100 object images at different poses. In order to construct the COIL-DEL dataset, these images were converted into mesh graphs by means of the Harris corner detection algorithm followed by a Delaunay triangulation. COIL-DEL is divided in 2400, 500 and 1000 graphs for training, validation and test respectively. In average these graphs have 21.5 nodes and 54.2 edges. Thus, they are rather small graphs.

6.2.2. Experimental protocol

In these experiments, we followed the same experimental protocol introduced by Li et al. [38], so we evaluated our method on two different metrics:

- **pair AUC**: The area under the ROC curve for classifying pairs of graphs as similar or not on a fixed set of 1000 pairs.
- **triplet accuracy**: The accuracy of correctly assigning a higher similarity to the positive pair than a negative pair on a fixed set of 1000 triplets.

Note that the fixed pairs and triplets are not the same from the original paper. We performed a random selection of these pairs and triplets while trying to balance the number of examples per class.

Table 6
Performance comparison on the COIL-DEL dataset against the methodologies introduced in Li et al. [38]. We studied the effect of the proposed GAT and GRU models, as well as, the number of layers and margin parameter μ.

Model		# Layers (K)	μ	Pair AUC	Triplet Acc
GCN [38]		–	–	94.80	94.95
Siamese-GCN [38]		–	–	95.90	96.10
GNN [38]		–	–	98.58	98.70
Siamese-GNN [38]		–	–	98.76	98.55
GMN [38]		–	–	98.97	98.80
Our	GAT	3	1	97.82	96.74
			10	96.22	96.94
		5	1	97.85	96.94
			10	97.70	97.54
	GRU	3	1	98.08	97.50
			10	96.25	95.69
		5	1	97.56	97.60
			10	95.36	96.07

6.2.3. Results and discussion

Table 6 presents a comparison with the state-of-the-art in graph metric learning. In particular, we compare with the different architectures proposed in Li et al. [38].

It is not surprising that their GMN technique outperforms our proposed model as they do incorporate cross-graph connections following an attention paradigm. Therefore, the correspondence is learned end-to-end in a much robust way. However, this is only feasible in rather small datasets as it incorporates a huge computational overhead. Notice also, that their GNN and Siamese-GNN models just obtain a slightly better performance than our proposed approach. However, when dealing with such small graphs it is hard to compare against embedding based approaches as they are able to encode the graph characteristic features without a huge loss of information.

In this extra experiment, we have also evaluated the effect on the choice of GNN models, the number of layers and the margin parameter μ. From this table, we conclude that the GRU-based models are drivers of a better performance on these experiments. Moreover, we find important to set our margin parameter to 1. The number of layers has not proven to bring a boost on performance on this particular dataset. Overall, our best model is able to obtain comparable results against GMN in this small dataset.

7. Conclusions and future work

In this paper, we have proposed a triplet GNN architecture for learning graph distances. Our architecture is able to learn node embeddings based on structural information of nodes local contexts. These learned features lead to an enriched graph representation which is later used in the distance computation. Moreover, we extended the graph edit distance approximation *viz.* Hausdorff edit distance, to the new learning framework in order to learn its operation costs within an end-to-end fashion. We have validated our proposed architecture on a graph retrieval scenario, in particular, we faced a keyword spotting task for handwritten words. Finally, we demonstrated competitive results against state-of-the-art, learning-based methods for graph distance learning.

Several future research lines emerge taking our proposed framework as starting point. For instance, our framework does not exploit the edge structure at matching time as we considered it implicitly encoded as node features. However, leveraging the edges information at this stage might lead to better results. Another promising line of research relates to the use of different graph pooling layers for reducing the size of large graphs before computing the learned Hausdorff edit distance.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been partially supported by the Spanish project RTI2018-095645-B-C21, the FPU fellowship FPU15 / 06264, the Ramon y Cajal Fellowship RYC-2014-16831, and the CERCA Program / Generalitat de Catalunya. We thank NVIDIA for the donation of a Titan Xp GPU.

References

[1] J. Yang, J. Lu, S. Lee, D. Batra, D. Parikh, Graph R-CNN for scene graph generation, in: Proceedings of the European Conference on Computer Vision, 2018.
[2] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural message passing for quantum chemistry, in: Proceedings of the International Conference on Machine Learning, 2017, pp. 1263–1272.
[3] G. Gkioxari, J. Malik, J. Johnson, Mesh R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9785–9795.
[4] L. Zhou, X. Bai, X. Liu, J. Zhou, E.R. Hancock, Learning binary code for fast nearest subspace search, Pattern Recognit. 98 (2020) 107040.
[5] B. Xiao, E.R. Hancock, R.C. Wilson, Graph characteristics from the heat kernel trace, Pattern Recognit. 42 (11) (2009) 2589–2606.
[6] D. Conte, P. Foggia, C. Sansone, M. Vento, Thirty years of graph matching in pattern recognition, Int. J. Pattern Recognit. Artif. Intell. 18 (3) (2004) 265–298.
[7] P. Foggia, G. Percannella, M. Vento, Graph matching and learning in pattern recognition in the last 10 years, Int. J. Pattern Recognit. Artif. Intell. 28 (1) (2014) 1–40.
[8] K.M. Borgwardt, Graph Kernels, Ludwig-Maximilians-Universität München, 2007 Ph.D. thesis.
[9] F. Zhou, F. de la Torre, Factorized graph matching, IEEE Trans. Pattern Anal. Mach. Intell. 99 (2015) 1.
[10] J. Gibert, E. Valveny, H. Bunke, Graph embedding in vector spaces by node attribute statistics, Pattern Recognit. 45 (9) (2012) 3072–3083.
[11] A. Dutta, P. Riba, J. Lladós, A. Fornés, Hierarchical stochastic graphlet embedding for graph-based pattern recognition, Neural Comput. Appl. 32 (2020) 11579–11596.
[12] R. Kondor, H. Pan, The multiscale Laplacian graph kernel, in: Advances in Neural Information Processing Systems, 2016, pp. 2982–2990.
[13] A. Sanfeliu, K.-S. Fu, A distance measure between attributed relational graphs for pattern recognition, IEEE Trans. Syst. Man Cybern. (3) (1983) 353–362.
[14] H. Bunke, G. Allermann, Inexact graph matching for structural pattern recognition, Pattern Recognit. Lett. 1 (4) (1983) 245–253.
[15] X. Gao, B. Xiao, D. Tao, X. Li, A survey of graph edit distance, Pattern Anal. Appl. 13 (1) (2010) 113–129.
[16] K. Riesen, H. Bunke, Approximate graph edit distance computation by means of bipartite graph matching, Image Vis. Comput. 27 (7) (2009) 950–959.
[17] A. Fischer, C.Y. Suen, V. Frinken, K. Riesen, H. Bunke, Approximation of graph edit distance based on Hausdorff matching, Pattern Recognit. 48 (2) (2015) 331–343.
[18] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[19] M.M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, P. Vandergheynst, Geometric deep learning: going beyond euclidean data, IEEE Signal Process. Mag. 34 (4) (2017) 18–42.
[20] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, R. Shah, Signature verification using a "siamese" time delay neural network, in: Advances in Neural Information Processing Systems, 1994, pp. 737–744.
[21] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (2009) 207–244.
[22] I. Elezi, S. Vascon, A. Torcinovich, M. Pelillo, L. Leal-Taixe, The group loss for deep metric learning, arXiv preprint arXiv:1912.00385(2019).
[23] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, arXiv preprint arXiv:2002.05709(2020).
[24] P. Riba, A. Fischer, J. Lladós, A. Fornés, Learning graph distances with message passing neural networks, in: Proceedings of the International Conference on Pattern Recognition, 2018, pp. 2239–2244.
[25] M. Gori, G. Monfardini, F. Scarselli, A new model for learning in graph domains, in: IEEE International Joint Conference on Neural Networks, 2, 2005, pp. 729–734.
[26] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Trans. Neural Netw. 20 (1) (2009) 61–80.
[27] J. Bruna, W. Zaremba, A. Szlam, Y. LeCun, Spectral networks and locally connected networks on graphs, arXiv preprint arXiv:1312.6203(2013).
[28] M. Henaff, J. Bruna, Y. LeCun, Deep convolutional networks on graph-structured data, arXiv preprint arXiv:1506.05163(2015).

[29] M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering, in: Advances in Neural Information Processing Systems, 2016, pp. 3844–3852.
[30] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: Proceedings of the International Conference on Learning Representations, 2017.
[31] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, in: Proceedings of the International Conference on Learning Representations, 2017.
[32] Y. Li, D. Tarlow, M. Brockschmidt, R. Zemel, Gated graph sequence neural networks, in: Proceedings of the International Conference on Learning Representations, 2016, pp. 1–20.
[33] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: a review of methods and applications, arXiv preprint arXiv:1812.08434(2018).
[34] P.W. Battaglia, J.B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, et al., Relational inductive biases, deep learning, and graph networks, arXiv preprint arXiv:1806.01261(2018).
[35] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S.Y. Philip, A comprehensive survey on graph neural networks, IEEE Trans. Neural Netw. Learn. Syst. 32 (1) (2020) 4–24.
[36] V.P. Dwivedi, C.K. Joshi, T. Laurent, Y. Bengio, X. Bresson, Benchmarking graph neural networks, arXiv preprint arXiv:2003.00982(2020).
[37] P. Baldi, Y. Chauvin, Neural networks for fingerprint recognition, Neural Comput. 5 (3) (1993) 402–418.
[38] Y. Li, C. Gu, T. Dullien, O. Vinyals, P. Kohli, Graph matching networks for learning the similarity of graph structured objects, in: Proceedings of the International Conference on Machine Learning, 2019, pp. 3835–3845.
[39] U. Chaudhuri, B. Banerjee, A. Bhattacharya, Siamese graph convolutional network for content based remote sensing image retrieval, Comput. Vis. Image Underst. 184 (2019) 22–30.
[40] J. Zhang, Graph neural distance metric learning with graph-bert, arXiv preprint arXiv:2002.03427(2020).
[41] Y. Bai, H. Ding, S. Bian, T. Chen, Y. Sun, W. Wang, SimGNN: a neural network approach to fast graph similarity computation, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019, pp. 384–392.
[42] Y. Bai, H. Ding, K. Gu, Y. Sun, W. Wang, Learning-based efficient graph similarity computation via multi-scale convolutional set matching, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020.
[43] A. Sanfeliu, K. Fu, A distance measure between attributed relational graphs for pattern recognition, IEEE Trans. Syst. Man Cybern. SMC-13 (3) (1983) 353–362.
[44] V. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Sov. Phys. Dokl. 10 (8) (1966) 707–710.
[45] R.A. Wagner, M.J. Fischer, The string-to-string correction problem, J. ACM 21 (1) (1974) 168–173.
[46] Z. Zeng, A.K. Tung, J. Wang, J. Feng, L. Zhou, Comparing stars: on approximating graph edit distance, Proc. VLDB Endow. 2 (1) (2009) 25–36.
[47] D. Justice, A. Hero, A binary linear programming formulation of the graph edit distance, IEEE Trans. Pattern Anal. Mach. Intell. 28 (8) (2006) 1200–1214.
[48] M. Neuhaus, K. Riesen, H. Bunke, Fast suboptimal algorithms for the computation of graph edit distance, in: Structural, Syntactic, and Statistical Pattern Recognition, 2006, pp. 163–172.
[49] J. Munkres, Algorithms for the assignment and transportation problems, J. Soc. Ind. Appl. Math. 5 (1) (1957) 32–38.
[50] F. Serratosa, Fast computation of bipartite graph matching, Pattern Recognit. Lett. 45 (2014) 244–250.
[51] A. Fischer, S. Uchida, V. Frinken, K. Riesen, H. Bunke, Improving Hausdorff edit distance using structural node context, in: Proceedings of the Workshop on Graph-Based Representation in Pattern Recognition, 2015, pp. 148–157.
[52] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: Proceedings of the International Conference on Machine Learning, 2015, pp. 448–456.
[53] K. Cho, B. Van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: encoder-decoder approaches, Workshop on Syntax, Semantics and Structure in Statistical Translation, 2014.
[54] V. Garcia, J. Bruna, Few-shot learning with graph neural networks, in: Proceedings of the International Conference on Learning Representations, 2018.
[55] V. Balntas, E. Riba, D. Ponsa, K. Mikolajczyk, Learning local feature descriptors with triplets and shallow convolutional neural networks, in: Proceedings of the British Machine Vision Conference, 2016, pp. 119.1–119.11.
[56] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980(2014).
[57] M. Stauffer, A. Fischer, K. Riesen, A novel graph database for handwritten word images, in: Joint International Workshops on Statistical techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition, 2016, pp. 553–563.
[58] M. Stauffer, A. Fischer, K. Riesen, Keyword spotting in historical handwritten documents based on graph matching, Pattern Recognit. 81 (2018) 240–253.
[59] A. Fischer, A. Keller, V. Frinken, H. Bunke, Lexicon-free handwritten word spotting using character HMMs, Pattern Recognit. Lett. 33 (7) (2012) 934–942.
[60] I. Pratikakis, K. Zagoris, B. Gatos, J. Puigcerver, A.H. Toselli, E. Vidal, ICFHR2016 handwritten keyword spotting competition (H-KWS 2016), in: Proceedings of the International Conference on Frontiers in Handwriting Recognition, 2016, pp. 613–618.
[61] M.R. Ameri, M. Stauffer, K. Riesen, T.D. Bui, A. Fischer, Graph-based keyword spotting in historical manuscripts using Hausdorff edit distance, Pattern Recognit. Lett. 121 (2019) 61–67.
[62] M. Rusiñol, J. Lladós, A performance evaluation protocol for symbol spotting systems in terms of recognition and location indices, Int. J. Doc. Anal. Recognit. 12 (2) (2009) 83–96.
[63] S. Sudholt, G.A. Fink, PHOCNet: a deep convolutional neural network for word spotting in handwritten documents, in: Proceedings of the International Conference on Frontiers in Handwriting Recognition, 2016, pp. 277–282.
[64] J. Almazán, A. Gordo, A. Fornés, E. Valveny, Word spotting and recognition with embedded attributes, IEEE Trans. Pattern Anal. Mach. Intell. 36 (12) (2014) 2552–2566.
[65] T. Wilkinson, A. Brun, Semantic and verbatim word spotting using deep neural networks, in: Proceedings of the International Conference on Frontiers in Handwriting Recognition, 2016, pp. 307–312.
[66] K. Riesen, H. Bunke, IAM graph database repository for graph based pattern recognition and machine learning, in: Joint International Workshops on Statistical techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition, 2008, pp. 287–297.
[67] S. Nayar, S. Nene, H. Murase, Columbia Object Image Library (COIL 100), Tech. Rep. CUCS-006-96, Dept. of Comp. Science, Columbia University, 1996.

Pau Riba received his B.Sc. in Mathematics and Computer Science and his M.Sc. in Computer Vision from the Universitat Autnoma de Barcelona. Currently, he is pursuing his Ph.D. under the supervision of Dr. Josep Llads and Dr. Alicia Forns. His main research interests include Graph Representations and Deep Learning.

Andreas Fischer received his M.Sc. and Ph.D. degrees in Computer Science from the University of Bern in 2008 and 2012, respectively. He then conducted postdoctoral research at CENPARMI and the Ecole Polytechnique de Montreal. Currently, Andreas Fischer is an Associate Professor at the University of Applied Sciences Western Switzerland and a Lecturer at the University of Fribourg. He has more than 90 publications related to historical document analysis, handwriting recognition, keyword spotting, signature verification, hidden Markov models, recurrent neural networks, and graph matching for structural pattern recognition.

Josep Lladós received the degree in Computer Sciences in 1991 from the Universitat Politecnica de Catalunya and the Ph.D. degree in Computer Sciences in 1997 from the Universitat Autonoma de Barcelona (Spain) and the UniversitL Paris 8 (France). Currently he is an Associate Professor at the Computer Sciences Department of the Universitat Autonoma de Barcelona and a staff researcher of the Computer Vision Center, where he is also the director since January 2009. He is associate researcher of the IDAKS Lab of the Osaka Prefecture University (Japan). He is chair holder of Knowledge Transfer of the UAB Research Park and Santander Bank. He is the coordinator of the Pattern Recognition and Document Analysis Group (2014SGR-1436). His current research fields are document analysis, structural and syntactic pattern recognition and computer vision. He has been the head of a number of Computer Vision R+D projects and published more than 230 papers in national and international conferences and journals, and supervised 12 Ph.D. theses. J. Llads is an active member of the Image Analysis and Pattern Recognition Spanish Association (AERFAI), a member society of the IAPR. He is currently the chairman of the IAPR-EC (Education Committee). Formerly he served as chairman of the IAPR-ILC (Industrial Liaison Committee), the IAPR TC-10, the Technical Committee on Graphics Recognition, and also he is a member of the IAPR TC-2 (Structural Pattern Recognition), IAPR TC-11 (Reading Systems) and IAPR TC-15 (Graph based Representations). He is chief editor of the ELCVIA (Electronic Letters on Computer Vision and Image Analysis). He is co-Editor in Series in Machine Perception and Artificial Intelligence (SMPAI) of World Scientfic Publishing Company. He serves on the Editorial Board of the Pattern Recognition journal, in IJDAR (International Journal in Document Analysis and Recognition), the Frontiers in Digital Humanities journal, and also a PC member of a number of international conferences. He was the recipient of the IAPR-ICDAR Young Investigator Award in 2007. He was the general chair of the International Conference on Document Analysis and Recognition (ICDAR 2009) held in Barcelona in July 2009, and co-chair of the IAPR TC-10 Graphics Recognition Workshop of 2003 (Barcelona), 2005 (Hong Kong), 2007 (Curitiba) and 2009 (La Rochelle). Josep Llads has also experience in technological transfer and in 2002 he created the company ICAR Vision Systems, a spin-off of the Computer Vision Center working on Document Image Analysis, after winning the entrepreneur award from the Catalonia Government on business projects on Information Society Technologies in 2000.

Alicia Fornés is a senior research fellow at the Universitat AutLonoma (UAB) de Barcelona and the Computer Vision Center. She obtained the Ph.D. degree in Computer Science from the UAB in 2009. She was the recipient of the AERFAI (Spanish brand of the IAPR, International Association for Pattern Recognition) best thesis award 20092010, and the IAPR/ICDAR Young Investigator Award in 2017. Her research interests include document image analysis, handwriting recognition, optical music recognition, writer identification and digital humanities.