SEMICCA: A NEW SEMI-SUPERVISED PROBABILISTIC CCA MODEL FOR KEYWORD SPOTTING

Giorgos Sfikas, Basilis Gatos

Computational Intelligence Laboratory / IIT
NCSR "Demokritos"
15310 Athens, Greece

Christophoros Nikou

Dpt. of Computer Science and Engineering
University of Ioannina
45110 Ioannina, Greece

ABSTRACT

In this paper we present a semi-supervised, attribute-based model suitable for keyword spotting (KWS) in document images. Our model can take advantage of available non-annotated segmented word images, as well as string annotations without a matching word image. We build our model by extending on the probabilistic interpretation of Canonical Correlation Analysis (CCA), solved using Expectation-Maximization (EM). On test-time, we back-project the query and database images to the embedded space by calculating the embedding space posterior density given the observations. Keyword spotting is then efficiently performed by computing query nearest neighbours in the embedded Euclidean space. We validate that our model offers superior performance given the presence of partially-labelled data, with keyword spotting trials on the *Bentham* and *George Washington* datasets.

1. INTRODUCTION

Keyword spotting has been established as an important application in the field of document image processing in the recent years. In cases where optical character recognition (OCR) of a scanned document is deemed to be very difficult and expected to give less than satisfactory results, or simple indexing is sufficient, keyword spotting has been proposed as an alternative to full OCR. In keyword spotting, the user queries the document database for a given word and the system is expected to return to the user a number of possible locations of the query in the original document. The query can either be a text string or an example word image. Various techniques have been proposed in the literature, covering the two scenarios, known as Query by Example (QbE) and Query by String (QbS) respectively [1, 2, 3, 4]. The taxonomy of word spotting systems further includes the distinction into segmentation-based and segmentation-free systems. In the former case, the datasets to be indexed are assumed to be segmented beforehand into line or word images [4].

[1]The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation Programme (H2020-EINFRA-2014-2015) under grant agreements n 674943 project READ.

Machine learning methods, and in particular, supervised learning methods [3, 5, 6], have been employed in document image processing with much success. Supervised learning methods require the existence of a set of manually transcripted documents, where word or line images are related to corresponding text strings. Learning using data of which only a (typically small) part is labelled is known as semi-supervised learning in the machine learning literature [7]. In [8], word-level segmented and annotated images have been used to learn attribute vectors for each input, and to embed them to a latent common subspace of image data and transcriptions. In that work, the embedding is performed using Canonical Correlation Analysis (CCA) [9]. In the current work we use a hierarchical, probabilistic model to learn projections of inputs instead of CCA. It has been shown [10] that CCA can be formulated as a hierarchical probabilistic model, in a manner that resembles the formulation of probabilistic PCA versus standard PCA. The proposed model, called "SemiCCA", uses the probabilistic formulation of CCA, further extending it to a semi-supervised model that can handle partially labelled data. Query-by-Example keyword spotting can then be performed by estimating the latent image y of the query in the common latent space, and comparing with the database using the Euclidean metric. Partially labelled data are defined in the current context as either (a) word images that have an unknown transcription or (b) word strings with no known word image match. Under the hypothesis that partially labelled data are more readily available than fully-transcribed ground truth data, we show that the proposed model is at a clear advantage compared to standard keyword spotting methods that cannot use unannotated data. Partially labelled data are used to better estimate the structure of the input spaces, and in turn learn better projections relating the input spaces with the common latent subspace. We confirm the validity of our assumption with numerical experiments on well-known collections of documents, such as "Bentham", which was used in the ICFHR'14 keyword spotting competition [11], and "George Washington" [1] databases.

The remainder of this paper is structured as follows. In Section 2, we present the proposed probabilistic SemiCCA

model and its solution with EM. In Section 3, we present numerical experiments that show that the proposed model can benefit from partially labelled data to boost keyword spotting performance. Finally, in Section 4 we discuss the paper's contribution and future work.

2. PROPOSED MODEL

In the current work we follow the (word level) segmentation-based paradigm, meaning that the document image content is assumed to be segmented into a set of word images. We further assume that we have training set, containing data that can be categorized either as *fully-labelled* or *partially-labelled*. With each datum, we associate (at most) two pieces of information: image content-based information and annotation information. Data are considered to be fully labelled when both word image content and annotation is available. If only one piece of information is available, the related datum is considered to be partially labelled.

For each type of information associated with each word token, i.e. image content or annotation, we compute a separate fixed-length descriptor. In order to describe word image content, we use an attribute-based representation [3, 8]. This representation is computed in two major steps: First, dense SIFT descriptors are extracted from all training data and are used to train a Gaussian Mixture Model (GMM). This is used as a means of describing the variability of input in terms of image content. Fisher Vectors (FVs) for an unseen word image are in turn computed [8] as a function of the SIFT information and the GMM. For each one of the defined attributes, a SVM is trained using FVs of all training images as input. In this work we shall denote an attribute vector as $x_\alpha \in \Re^D$. We use the Pyramidal Histogram Of Characters (PHOC) descriptor [8] to represent word annotation information. PHOC is constructed as a set of binary histograms, describing letter appearance in a hierarchy of different spatial levels of the word string. Based on an analogous hierarchical scheme, binary ground-truth responses for each attribute are concatenated into a single D-sized vector. Note that the PHOC vector is of the same size as the attribute vector. We shall denote a PHOC vector as x_ϕ, with $x_\phi \in \{0,1\}^D$. D stands for the number of attributes, and is the same for both the word image and the annotation (PHOC) representation. Dimensionality D depends on the characteristics of the target language and the number of histogram levels.

We proceed by using both the available fully labelled and partially labelled data to learn an embedding/relation of attributes and PHOCs onto a common, latent, low-dimensional subspace. We formally define labelled data as $|N_{\alpha\phi}|$ pairs of attribute vectors and matching transcription vectors, i.e. $\{x_\alpha^n, x_\phi^n\}_{n\in N_{\alpha\phi}}$ and partially labelled data as $|N_\alpha|$ attribute vectors and $|N_\phi|$ PHOC transcription vectors, i.e. $\{x_\alpha^n\}_{n\in N_\alpha}$ and $\{x_\phi^n\}_{n\in N_\phi}$ respectively. N_α, N_ϕ are sets of indices of partially labelled data and $N_{\alpha\phi}$ is a set of indices of (fully) labelled data. The aforementioned sets are disjoint. Let us stress that attribute and transcription vectors in the partially labelled set *do not* form matching pairs, that is, in this set, attribute vector x_α^n does not have a transcription that is described by PHOC vector x_ϕ^n, and in general $|N_\alpha| \neq |N_\phi|$. All attribute and PHOC vectors are D-dimensional. The proposed graphical model can be examined in Fig.1. The fully/partially labelled data are the model observations, while $\{y^n\}_{n\in N}$ where $N = N_\alpha \cup N_\phi \cup N_{\alpha\phi}$ are latent d-dimensional random variables, with $d \leq D$. The variables y are independent and identically distributed as:

$$y^n \sim N(0, I), \forall n \in N_{\alpha\phi} \tag{1}$$

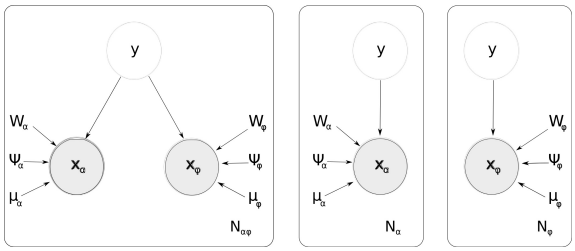


Fig. 1. The graphical model for SemiCCA, proposed in this work. See text for details.

We shall dub the d-dimensional space where the y latent variables y reside at y-space or common latent subspace. For each pair of attribute and PHOC vector a common variable y is assumed to exist. Latent variables and observations are related through the assumptions

$$x_\alpha^n \sim N(W_\alpha^T y^n + \mu_\alpha, \Psi_\alpha), \forall n \in N_{\alpha\phi} \tag{2}$$

$$x_\phi^n \sim N(W_\phi^T y^n + \mu_\phi, \Psi_\phi), \forall n \in N_{\alpha\phi} \tag{3}$$

where W_α and W_ϕ are $d \times D$ projection matrices, μ_α and μ_ϕ are D-dimensional vectors and Ψ_α and Ψ_ϕ are $D \times D$ covariance matrices. Expressions (1), (2), (3) hold for all fully-labelled data and together correspond to the leftmost plate shown in the graphical model of Fig.1. It has been shown in [10], that the Maximum Likelihood (ML) solution for the model where only labelled data are available, is identical to the solution for a corresponding CCA model where x_α and x_ϕ are the two observed views for each pair/labelled datum. This model is extended here to handle single-view observations with the two additional plates of fig.1. We shall refer to this model as SemiCCA in this paper. In a manner analogous to what has been assumed for labelled data, we further assume for partially labelled data $y^n \sim N(0, I), \forall n \in N_\alpha \cup N_\phi$; $x_\alpha^n \sim N(W_\alpha^T y^n + \mu_\alpha, \Psi_\alpha), \forall n \in N_\alpha$; $x_\phi^n \sim N(W_\phi^T y^n + \mu_\phi, \Psi_\phi), \forall n \in N_\phi$. These equations correspond to the two right-most plates of the graphical model in Fig.1. Note that the model parameters $\Theta = \{\{W_\alpha, W_\phi\}, \{\Psi_\alpha, \Psi_\phi\}, \{\mu_\alpha, \mu_\phi\}\}$ are the same for either fully or partially labelled data.

The proposed SemiCCA model can be solved, that is compute $\Theta^\star = \arg\max_\Theta \ln p(x_\alpha, x_\phi; \Theta)$, using the EM algorithm [12]. In EM, after selecting initial values for the model parameters, updates for latent variable moments (E-step) and parameters (M-step) are applied and reiterated until convergence.

The E-step updates are computed as follows:

$$\begin{aligned}
\text{cov}\{y^n\}^{(t+1)}_{n\in N_{\alpha\phi}} &= \{W_\alpha^{T(t)}\Psi_\alpha^{-1(t)}W_\alpha^{(t)} \\
&+ W_\phi^{T(t)}\Psi_\phi^{-1(t)}W_\phi^{(t)} + I\}^{-1}, \quad (4)\\
\text{cov}\{y^n\}^{(t+1)}_{n\in N_\alpha} &= \{W_\alpha^{T(t)}\Psi_\alpha^{-1(t)}W_\alpha^{(t)} + I\}^{-1}, \quad (5)\\
\text{cov}\{y^n\}^{(t+1)}_{n\in N_\phi} &= \{W_\phi^{T(t)}\Psi_\phi^{-1(t)}W_\phi^{(t)} + I\}^{-1} \quad (6)\\
<y>^{n(t+1)}_{n\in N_{\alpha\phi}} &= \text{cov}\{y^n\}^{(t+1)}[W_\alpha^{T(t)}\Psi_\alpha^{-1(t)}(x_\alpha^n - \mu_\alpha^{(t)}) \\
&+W_\phi^{T(t)}\Psi_\phi^{-1(t)}(x_\phi^n - \mu_\phi^{(t)})], \quad (7)\\
<y>^{n(t+1)}_{n\in N_\alpha} &= \text{cov}\{y^n\}^{(t+1)}[W_\alpha^{T(t)}\Psi_\alpha^{-1(t)}(x_\alpha^n - \mu_\alpha^{(t)})] \quad (8)\\
<y>^{n(t+1)}_{n\in N_\phi} &= \text{cov}\{y^n\}^{(t+1)}[W_\phi^{T(t)}\Psi_\phi^{-1(t)}(x_\phi^n - \mu_\phi^{(t)})] \quad (9)\\
<yy^T>^{n(t+1)}_{n\in N} &= \text{cov}\{y^n\}^{(t+1)} + <y><y^T>^{n(t+1)}. \quad (10)
\end{aligned}$$

The M-step updates are computed as follows:

$$\mu_\alpha^{n(t+1)} = \sum_{n\in N_\alpha\cup N_{\alpha\phi}} (x_\alpha^n - W_\alpha^{(t)} <y>^{n(t+1)}) \big/ |N_\alpha \cup N_{\alpha\phi}|, \quad (11)$$

$$\mu_\phi^{n(t+1)} = \sum_{n\in N_\phi\cup N_{\alpha\phi}} (x_\phi^n - W_\phi^{(t)} <y>^{n(t+1)}) \big/ |N_\phi \cup N_{\alpha\phi}|, \quad (12)$$

$$W_\alpha^{n(t+1)} = \frac{\sum_{n\in N_\alpha\cup N_{\alpha\phi}} \hat{x}_\alpha^{n(t+1)} <y^T>^{n(t+1)}}{\sum_{n\in N_\alpha\cup N_{\alpha\phi}} <yy^T>^{n(t+1)}} \quad (13)$$

$$W_\phi^{n(t+1)} = \frac{\sum_{n\in N_\phi\cup N_{\alpha\phi}} \hat{x}_\phi^{n(t+1)} <y^T>^{n(t+1)}}{\sum_{n\in N_\phi\cup N_{\alpha\phi}} <yy^T>^{n(t+1)}} \quad (14)$$

$$\Psi_\alpha^{n(t+1)} = \sum_{n\in N_\alpha\cup N_{\alpha\phi}} \{\hat{W}_\alpha^{n(t+1)} + \hat{A}_\alpha^{n(t+1)} - 2\hat{B}_\alpha^{n(t+1)}\} \big/ |N_\alpha \cup N_{\alpha\phi}|, \quad (15)$$

$$\Psi_\phi^{n(t+1)} = \sum_{n\in N_\phi\cup N_{\alpha\phi}} \{\hat{W}_\phi^{n(t+1)} + \hat{A}_\phi^{n(t+1)} - 2\hat{B}_\phi^{n(t+1)}\} \big/ |N_\phi \cup N_{\alpha\phi}|, \quad (16)$$

where we have defined $\hat{x}_\alpha^{n(t+1)} = (x_\alpha^n - \mu_\alpha^{n(t+1)})$, $\hat{A}_\alpha^{n(t+1)} = \hat{x}_\alpha^{n(t+1)}\hat{x}_\alpha^{n(t+1)T}$, $\hat{B}_\alpha^{n(t+1)} = \hat{x}_\alpha^{n(t+1)} <y^T>^{n(t+1)} W_\alpha^{n(t+1)}$, $\hat{W}_\alpha^{n(t+1)} = W_\alpha^{(t+1)} < yy^T >^{n(t+1)} W_\alpha^{T(t+1)}$. Quantities $\hat{x}_\phi, \hat{A}_\phi, \hat{B}_\phi, \hat{W}_\phi$ are defined in an analogous manner, simply by substituting ϕ in place of α.

Let us note that in the equations of the M-step, the quantities indexed by α are computed $\forall n \in N_\alpha \cup N_{\alpha\phi}$, while every quantity indexed by ϕ is computed $\forall n \in N_\phi \cup N_{\alpha\phi}$.

After the ML parameters of the model are estimated using the EM algorithm we can perform QbE word spotting, i.e. we assume that the word image of the query is available.

The required equations to compute expected values of y are already available as part of the EM algorithm (E-step). By combining equations (5) and (8) we can compute the required expectation as

$$<y>^{new} = [W_\alpha^T\Psi_\alpha^{-1}W_\alpha + I]^{-1}[W_\alpha^T\Psi_\alpha^{-1}(x_\alpha^{new} - \mu_\alpha)] \quad (17)$$

where we use the EM-optimized values for the model parameters. After comparing y^{new} with common subspace images of words in the queried database, the nearest neighbors of y are returned as the query result. A summary of the full algorithm to perform keyword spotting using the proposed SemiCCA model can be examined below (Algorithm 1).

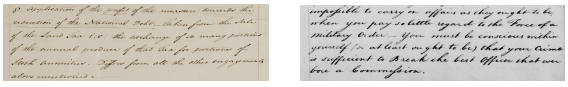
Algorithm 1 Keyword spotting using SemiCCA

Process database
- Compute attribute vectors for fully and partially-labelled data
- Compute PHOC vectors for fully and partially-labelled data

Train SemiCCA model
- Compute E-step (eq. 4-10)
- Compute M-step (eq. 11-16)
- Reiterate E-step and M-step until convergence
- Store model parameters $\Theta^\star$

Perform query
- Compute attribute vector for query word image
- Use eq. 17 and $\Theta^\star$ to compute latent subspace image for query
- Return nearest neighbours of query image

3. EXPERIMENTAL RESULTS



(a) (b)

Fig. 2. Samples from the (a) *Bentham* [11] and (b) *George Washington* [1] datasets, used in our experiments.

We have run experiments on two different datasets (Fig. 2): *Bentham*, and *George Washington* [1]. Dataset *Bentham* has been used in the ICFHR'14 Keyword Spotting competition [11]. We have used a number of different partitions of these sets into fully-labelled training/validation sets, partially labelled training sets and test sets. We have used the naming convention *database-number* to identify each partition, where *number* corresponds to the total number of fully-labelled data used (fully-labelled training + validation). For example *Bentham100* refers to the Bentham dataset partitioned so that 100 fully-labelled data, that is 100 pairs of word images and corresponding transcriptions, are available. We kept the size of the test set fixed, and vary the size of the fully-labelled training set versus the size of the partially-labelled training set. [1]

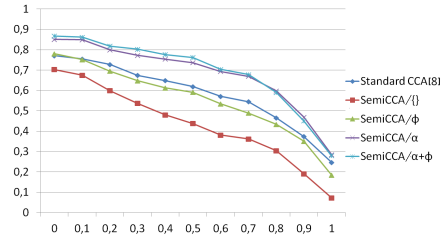
We used both the fully-labelled training and partially-labelled training sets to train the SIFT-based GMM. Fisher vectors and PHOC vectors were then computed over the whole database. SVM-based attribute models are computed over only fully-labelled data (since both image and annotation content is necessary). Attribute vectors are then computed

[1] The exact indices, corresponding to the index of each segmented word image in the collection in reading order, are as follows: Washington50: $1-40$, $41-3000$, $3001-4849$, $4850-4859$, Bentham50: $1-40$, $41-5300$, $5301-10638$, $10639-10648$, Bentham100: $1-80$, $81-5300$, $5301-10628$, $10629-10648$, Bentham500: $1-400$, $401-5300$, $5301-10548$, $10549-10648$.

for the whole set ($\{x_\alpha^n\}$) and fed into the SemiCCA model along with PHOC vectors ($\{x_\phi^n\}$). Following [8], we set the dimensionality of the common latent subspace to 80.

Concerning the experiment and benchmarking layout we used, we first selected a number of word classes as queries. We chose queries following [11] for *Bentham*, and for *George Washington* we used the query classes suggested in [13]. We have first compared the performance of the proposed model given differing types of partially labelled data. The standard CCA-based model of Almazán et al. [8] was used as a baseline. In Fig. 3, MAP figures and precision-recall (PR) curves are shown for results over the *Washington50* dataset. SemiCCA/$\{\varnothing\}$ corresponds to the proposed model with *no* partially labelled data available. In the sense of the type of data that can be used, it is equivalent to the standard CCA model. However, results are markedly inferior to standard CCA. The difference in performance is explained by the way the two models are solved. Standard CCA offers an eigenanalysis-based solution which is (globally) optimal in the sense that computed projections are indeed maximally correlated by construction of the solution. On the other hand, the solution of SemiCCA is EM-based, which is known to offer only a locally optimal, hence in general suboptimal, solution. SemiCCA/ϕ corresponds to the proposed model with PHOC partially labelled data available (x_ϕ), that is extra transcriptions and PHOC vectors for these transcriptions are available, which are not linked to a specific word image. SemiCCA/α corresponds to the proposed model with attribute partially labelled data available (x_α), that is, data coming from extra word images with no known transcription. This result suggests that string transcription information is nowhere near as useful as word image information as partially labelled data. This result is confirmed by comparing SemiCCA/$\alpha+\phi$, which uses both untranscripted word images and strings (with no matching word image) with SemiCCA/α. The improvement is only minimal.

We proceed with numerical tests on partitions of datasets where the factor that varies is the ratio of fully-labelled data to partially-labelled. The type of partially-labelled data that we chose to vary in this experiment is the number of unannotated word images, as this has the greatest impact on performance (according to the previous experiment, cf. Fig.3). For the same reason, the variant of the proposed model used here is SemiCCA/α. In semi-supervised learning, a usual underlying hypothesis is that fully-labelled data are only a very small portion of the training set [14]. The results show that the proposed framework exhibits its best performance when this is indeed the case. Conversely, it does not perform as well when there are enough fully-labelled data available. This can be seen in table 1, where we show MAP figures for word spotting trials that we have run on three different partitions of *Bentham*, where the number of fully-labelled data is 50, 100 and 500 data respectively.



Method	MAP(%)
Standard CCA [8]	58.2
SemiCCA/$\{\varnothing\}$	43.1
SemiCCA/ϕ	55.2
SemiCCA/α	**68.0**
SemiCCA/$\alpha+\phi$	**69.0**

Fig. 3. Performance comparison between standard CCA (used in [8]) and different variants of SemiCCA (proposed model). Trials were run on *Washington50* dataset. The MAP table below the graph summarizes the results of the PR curve. Variants differ in what type of partially-labelled data is available. The proposed model corresponds to SemiCCA/α and SemiCCA/$\alpha+\phi$.

Table 1. Performance comparison between standard CCA and SemiCCA (proposed model) on dataset partitions with different ratios of fully-labelled to partially-labelled data (Ratio F to P). Our model outperforms standard CCA when few fully-labelled data are available but there is an abundance of partially-labelled data.

Dataset	Ratio F to P	CCA [8]	SemiCCA
Bentham50	$\sim$ 1 to 100	31.9	**42.3**
Bentham100	$\sim$ 1 to 50	47.2	**50.4**
Bentham500	$\sim$ 1 to 10	**57.2**	53.1

4. CONCLUSION

We have presented a new method for keyword spotting, formulated as a semi-supervised probabilistic keyword spotting model. When unannotated word images and lexicon text strings are available, our model can take advantage of it to improve performance. Experiments have shown that the proposed model outperforms the state-of-the-art supervised learning model of [8] when sufficient partially labelled data are available. Compared with other well-known learning-based algorithms, like Hidden Markov Model or Neural Network-based models, our model is also at an advantage as (a) neither they can exploit partially labelled data and (b) they require an amount of annotated/fully-supervised data that is significantly larger than the corresponding amounts we have used (for example, as few as 50 annotated words). As future work, possible directions could include extending the model to a kernel-CCA like version, or integrating with Deep Learning-based features [2].

5. REFERENCES

[1] Toni Rath and Raghavan Manmatha, "Features for word spotting in historical manuscripts," in *Proceedings of the International Conference on Document Analysis and Recognition*. IEEE, 2003, pp. 218–222.

[2] Giorgos Sfikas, George Retsinas, and Basilis Gatos, "Zoning aggregated hypercolumns for keyword spotting," in *Proceedings of the* 15^{th} *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2016, pp. 283–288.

[3] Giorgos Sfikas, Angelos P. Giotis, Georgios Louloudis, and Basilis Gatos, "Using Attributes for Word Spotting and Recognition in polytonic Greek documents," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2015, pp. 686–690.

[4] Angelos P. Giotis, Giorgos Sfikas, Basilis Gatos, and Christophoros Nikou, "A survey of document image word spotting techniques," *Pattern Recognition*, vol. 68, pp. 310 – 332, 2017.

[5] Gundram Leifert, Tobias Strauß, Tobias Grüning, Welf Wustlich, and Roger Labahn, "Cells in multidimensional recurrent neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3313–3349, 2016.

[6] Alejandro Héctor Toselli and Enrique Vidal, "Fast HMM-filler approach for keyword spotting in handwritten documents," in *Proceedings of the International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 501–505.

[7] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, *Semi-Supervised Learning*, MIT Press, 2006.

[8] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny, "Word spotting and recognition with embedded attributes," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 36, no. 12, pp. 2552–2566, 2014.

[9] David Hardoon, Sandor Szedmak, and John Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[10] Francis R. Bach and Michael I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Tech. Rep., 2005.

[11] Ioannis Pratikakis, Konstantinos Zagoris, Basilis Gatos, Georgios Louloudis, and Nikolaos Stamatopoulos, "Competition on handwritten keyword spotting (H-KWS 2014)," in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, 2014.

[12] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer New York, 2006.

[13] Yann Leydier, Frank Lebourgeois, and Hubert Emptoz, "Text search for medieval manuscript images," *Pattern Recognition*, vol. 40, no. 12, pp. 3552–3567, 2007.

[14] Marcin Olof Szummer, *Learning from partially labeled data*, Ph.D. thesis, Massachusetts Institute of Technology, 2002.