# Towards Query-by-Speech Handwritten Keyword Spotting

Marçal Rusiñol, David Aldavert, Ricardo Toledo and Josep Lladós
Computer Vision Center, Dept. Ciències de la Computació
Edifici O, Univ. Autònoma de Barcelona
08193 Bellaterra (Barcelona), Spain.

*Abstract*—**In this paper, we present a new querying paradigm for handwritten keyword spotting. We propose to represent handwritten word images both by visual and audio representations, enabling a query-by-speech keyword spotting system. The two representations are merged together and projected to a common sub-space in the training phase. This transform allows to, given a spoken query, retrieve word instances that were only represented by the visual modality. In addition, the same method can be used backwards at no additional cost to produce a handwritten text-to-speech system. We present our first results on this new querying mechanism using synthetic voices over the George Washington dataset.**

## I. Introduction

Handwritten keyword spotting can be defined as the problem of locating within a collection of documents, the zones of interest where a particular queried word is likely to appear, without the explicit transcription of all the contents of the collection. Being a mature enough research topic, many different approaches have been presented in the literature aiming at the keyword spotting problem. In particular, one of the most prominent approaches is to follow a ***query-by-example*** paradigm. That is, the user provides a snippet image of the sought handwritten word that serves as example and the system retrieves a ranked list of words from the collection that are similar to the query. The simplicity of such approaches is quite attractive since by only using an adequate handwritten word description and similarity measure, the system can already deliver good retrieval performances. In addition, such spotting approaches can usually be pipelined to any indexing mechanism yielding efficient response times in large-scale scenarios. However, such simplicity hides an important usability flaw. In order to spot a word, the user has to browse the collection looking for an instance of the sought word. But forcing the user to manually extract a template word in large collections might be a really tedious task. Such approaches are thus unusable in real scenarios for plain users that might not be willing to make such an effort to just cast a query.

To overcome such limitations, more complex learning-based techniques have been proposed to bypass this burden. These techniques allow to query the system by just typing the query string, which is known as the ***query-by-string*** paradigm. Although this kind of techniques need a portion of the dataset to be transcribed beforehand in order to be used as training set, they present the advantage that they can be used in multi-writer scenarios. The latest trend in query-by-string learning-based techniques is to learn models for individual characters without needing a word segmentation technique. Machine learning approaches such as hidden Markov models or neural networks have been used for such purpose [1], [2]. Although such proposals provide a more user-friendly querying experience, they might not be really usable for large collections. At query time the whole collection needs to be processed to estimate the model output scores, and such step might be too computationally expensive. Therefore, recent query-by-string methods aimed at representing words in a numerical $n$-dimensional space, which can be efficiently indexed, have been proposed [3], [4]. Such approaches find a common subspace between textual and visual descriptions of the words, allowing the user to cast a textual query and retrieve words that were just described visually.

In this paper we propose a new querying paradigm to go one step beyond query-by-string. We propose to adapt our previous query-by-string publication [3] so that it can tackle audio signals instead of textual strings, thus allowing the user to cast spoken queries. To our best knowledge, this is the first work that proposes a ***query-by-speech*** paradigm for handwritten keyword spotting. Such paradigm provides several benefits. First, it produces a more user-friendly query experience than classical query-by-example methods. Second, since the final representation is a numeric feature vector, the solution is scalable to large collections, providing sub-linear query times when used in combination with off-the-shelf indexing strategies. Finally, we believe that the query-by-speech paradigm is even more ergonomic than query-by-string approaches, since we get rid of the keyboard, making it more easily integrable in some specific scenarios like museum exhibitions or keyword search engines integrated in smartphones.

Additionally, the proposed framework presents the advantage that without any additional cost, it can be used backwards. Once we have the trained model, the same system can be used either to cast spoken queries to retrieve handwritten words that were just represented visually, but also given an image of the word, it can retrieve the most similar utterance that we have indexed. Our proposed system can thus be seen either as a query-by-speech handwritten keyword spotting, or as a basic ***handwritten text-to-speech*** (TTS) system.

The remainder of the paper is organized as follows. Both audio and visual word representations are described in Section II. Subsequently, in Section III we detail how the combination of both representations is obtained through the use of latent semantic analysis. In Section IV, we overview the retrieval step and how from a query represented with one modality we are able to retrieve word instances that are only represented with the other modality. The experimental results are presented in Section V. We finally draw our conclusion remarks in Section VI.

## II. Word Representations

In this work, handwritten word images are represented by two different cues, one relying on an audio signal and the other relying on visual information. In this section, we will present the work-flow used to generate both representations.

### A. Audio Representation

Let us first detail how we create the audio signals of each document word and how we describe these audio signals.

*1) Synthetic Speech:* In order to carry out our experiments, we need recordings of the words appearing in a collection of handwritten documents. To produce a large enough set of audio records both for the train and the test phases of the algorithm, we decided to use synthetic voices rather than actually recording audio clips. We have used three different TTS engines, namely the Festival Speech Synthesis System [5], the Google TTS API[1] and the AT&T's Natural Voices $^{\text{TM}}$ software[2]. The Google TTS engine yields an utterance with a single voice for each word while Festival and AT&T TTS can create utterances with five different voices for the same word. The Google and AT&T software create audio records of a higher quality than Festival software obtaining in general audio recordings closer to natural voices. Finally, the AT&T software is only used to synthesize a selected set of words that we utilize as queries to evaluate the algorithm's robustness against unheard speakers in the training phase.

*2) Bag-of-Audio-Words:* In order to extract a feature vector from audio signals, we have chosen the Perceptual Linear Prediction (PLP) framework [6]. After applying an overlapped hamming window to the speech signal, PLP features are extracted on the short-term spectrum of speech. PLP uses several psychophysically based transformations to extract a set of cepstral coefficients. A subsequent filtering step named RASTA, proposed by Hermansky et al. in [7] is applied in order to make PLP analysis more robust to spectral distortions.

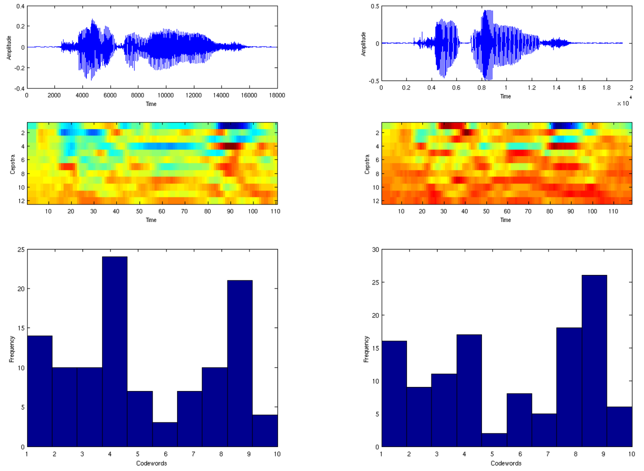


Fig. 1. Example of utterances of the word ***companies***, their RASTA-PLP cepstral coefficients and their BoAW histograms for two different voices.

However, the RASTA-PLP method still outputs a time signal that depends on the length of the pronounced utterance. In order to have a fixed-length feature vector, we have applied the Bag-of-Audio-Words (BoAW) framework over these time series features. From a set of utterances represented by their RASTA-PLP cepstral coefficients, we create a codebook by clustering the cepstral coefficients using the $k$-means algorithm. In this paper, we use a codebook of 8,192 audio words. A spoken word is then represented by an histogram which accumulates the frequencies of each audio word. We can see an example of a couple of utterances, their RASTA-PLP cepstral features and a simplified BoAW representation (just 10 codewords) in Figure 1. In order to encode some sequential information, we divide the audio information into different temporal bins and the histogram of audio words are independently accumulated for each of them. These temporal bins are created using a pyramid structure similar to the Spatial Pyramid Matching method [8] typically used for visual information. We employ a two level pyramid which halves the temporal bins at each new level, resulting in seven different temporal bins. These temporal bins are concatenated resulting in a 57,344-dimensional audio descriptor which is $L_2$-normalized to obtain the final audio representation $\mathbf{f}^a$.

### B. Visual Representation

Word image snippets are represented by a descriptor obtained using the Bag-of-Visual-Words (BoVW) framework. We used the same visual representation that was proposed in our previous work in [3], [9]. Let us briefly detail the followed steps.

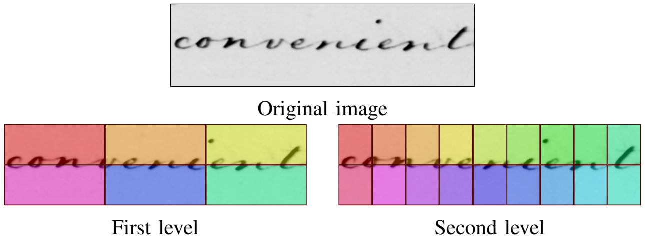


Fig. 2. Distribution of the spatial bins in the two levels of the spatial pyramid.

We first densely sample local regions over the image at different scales having sizes of 20, 30 and 40 pixels which are sampled at a constant step of 5 pixels. Then, the local regions are characterized by the SIFT local descriptor using its standard configuration. Once descriptors have been sampled from the image, we convert them into visual words by means of a codebook created by the $k$-means algorithm. In this paper, we use a codebook of 4,096 visual words. Local descriptors are then encoded into visual words using the Locality-constrained Linear Coding (LLC) algorithm [10] considering the three nearest neighbors. We then added spatial information by means of the SPM method [8]. We used a first level with 3 partitions in the $X$ axis and 2 partitions in the $Y$ axis and a second level with triple number of divisions in the $X$ axis while keeping the same amount of partitions in the $Y$ axis. We can see an example of the proposed configuration in Fig. 2.

[1] http://www.translate.google.com/translate_tts?tl=en\&q=Hello

[2] http://www2.research.att.com/~ttsweb/tts/demo.php

The visual descriptor is obtained concatenating the histograms of visual words of each pyramid level. This configuration results in a 98,304-dimensional visual descriptor. We then use a power normalization proposed by Perronnin et al. in [11] in order to increase the response of the visual words with low contribution resulting in a less sparse representation. Finally, an $L_2$-normalization is applied to the whole descriptor in order to obtain the final visual representation $\mathbf{f}^v$.

## III. Audio and Visual Fusion

Until now, word snippets are represented by two information modalities: an audio descriptor extracted from different utterances of the word and a visual representation generated from the visual words extracted from the document images. Our proposal is to bring together the audio and visual descriptors into a common representation space, so that we are able to use queries from one modality to retrieve words described using the other modality. We want to be able to use spoken queries to retrieve word snippets which are described solely by visual descriptors and vice-versa, to retrieve word utterances given an image of a handwritten word. This is achieved by searching a transform which projects both the visual and audio information into a common feature space, so that ideally the audio and visual descriptors of the same word will be similar in the transformed space.

This transform is obtained by assuming that both visual and audio features will co-occur for two different instances of the same word. Therefore, we can find a set of abstract topics that represent distributions of visual and audio features associated to some underlying semantics of the indexed words. We calculate these abstract topics with the Latent Semantic Analysis (LSA) algorithm [12] which uses the singular value decomposition (SVD) step to find a linear transform which projects both audio and visual information into a set of abstract topics in an unsupervised way.

In order to calculate the linear projection matrix, we first arrange the word descriptors of the training set in a descriptor-by-word matrix $\mathbf{A} = [\mathbf{f}_1 \dots \mathbf{f}_i \dots \mathbf{f}_M]$, where $M$ is the number of train samples and $\mathbf{f}_i = [\mathbf{f}_i^a, \mathbf{f}_i^v]$ is obtained by concatenating the audio $\mathbf{f}_i^a$ and the visual $\mathbf{f}_i^v$ descriptors of the $i$-th training sample. Since we can generate multiple utterances for each word, the number of train samples $M$ corresponds to the number of word snippet images multiplied by the number of voices used to create the audio descriptors. The LSA algorithm obtains the linear projection by decomposing this descriptor-by-word matrix in three matrices using a truncated SVD:

$$\mathbf{A} \simeq \hat{\mathbf{A}} = \mathbf{U}_T \mathbf{S}_T \left(\mathbf{V}_T\right)^\top ,$$

where $T$ is the number of abstract topics (i.e., the dimensionality of the transformed space) and $\mathbf{U}_T \in \mathbb{R}^{D \times T}$, $\mathbf{S}_T \in \mathbb{R}^{T \times T}$ and $\mathbf{V}_T \in \mathbb{R}^{M \times T}$. Finally, the transformation matrix $\mathbf{X}_T$ is calculated as

$$\mathbf{X}_T = \mathbf{U}_T \left(\mathbf{S}_T\right)^{-1} ,$$

so that the descriptor of $i$-th word snippet $\mathbf{f}_i$ is projected into the new feature space by simply $\hat{\mathbf{f}}_i = \mathbf{f}_i^\top \mathbf{X}_T$.

## IV. Retrieval Phase

The transformation matrix $\mathbf{X}_T$ has been calculated using words which has both visual and audio signatures available. However, in the retrieval phase just a single modality is present. The projection of either a visual or an audio signature into the topic space is computed as

$$\hat{\mathbf{f}}_i^v = \left[\mathbf{0}_a^\top, \mathbf{f}_i^{v\top}\right] \mathbf{X}_T, \qquad \hat{\mathbf{f}}_i^a = \left[\mathbf{f}_i^{a\top}, \mathbf{0}_v^\top\right] \mathbf{X}_T,$$

where $\mathbf{0}_a$ and $\mathbf{0}_v$ are zeros vectors with the same dimensionality as the audio or visual codebooks respectively.

### A. Query-by-speech keyword spotting

In this scenario, we have indexed the word image snippets with the projected signatures $\hat{\mathbf{f}}_i^v$, and an utterance is expected as the query. The queries are described with the projected audio signature $\hat{\mathbf{f}}_q^a$ and the cosine distance between them is used as a similarity measure to generate the ranked list.

This procedure allows to retrieve word instances using spoken queries from databases which have been described using only visual information. This is possible since the LSA algorithm has found relationships between the visual words and audio words in the training phase. Then, even if a source of information is not present in one of the descriptors, we are still able to rank and find relevant instances in the indexed documents.

### B. Audio Retrieval

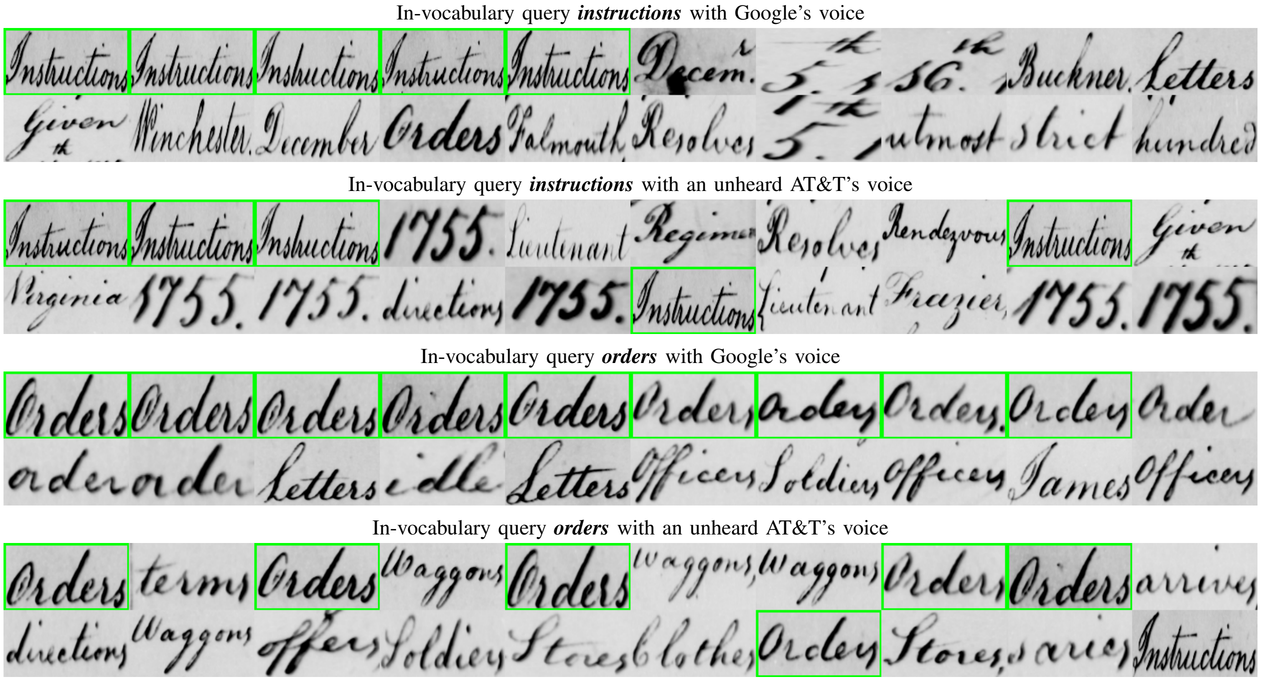
In addition, without any additional cost, we can use the proposed method backwards to reverse the retrieval task. That is to index audio information and use handwritten word images as queries. Such audio retrieval can be seen as a handwritten text-to-speech task which does not have an implicit recognition step. Furthermore, we generate multiple ranked lists for a single visual query by indexing the audios of the different available voices separately. These output lists are re-ranked into a single result using the Borda count algorithm to obtain a better retrieval performance. Finally, when only returning the first element of the ranked list to the user, such audio retrieval method can be seen as a handwritten text-to-speech system.

## V. Experimental Results

The proposed method has been evaluated in the George Washington (GW) database [13] consisting of 4864 segmented words. In order to train the LSA model that brings together the audio and visual information, we need a portion of the database to be transcribed. Therefore, the database is divided into four different folds. The system is trained using three of these folds and evaluated in the remaining one. At query time, all the words in the test set can be used. However, not all the words appearing in the test fold might be present in the train set. We will therefore report in our experiments two different measures, the performances reached when considering all words as queries and the performances reached when just considering the in-vocabulary words, i.e. just casting the queries that are present in both train and test sets.

### A. Query-by-speech Qualitative Results

We present in Fig. 3 some qualitative results of the system when casting queries pronounced by either Google's voice (used as well in training) or an AT&T voice, which has not


In-vocabulary query ***instructions*** with Google's voice

In-vocabulary query ***instructions*** with an unheard AT&T's voice

In-vocabulary query ***orders*** with Google's voice

In-vocabulary query ***orders*** with an unheard AT&T's voice

Fig. 3. Examples of the 20 most similar word images returned by the system for different spoken queries.

TABLE I. COMPARISON WITH OUR PREVIOUS WORK USING THE SAME VISUAL DESCRIPTION.

| Method | Evaluation | Queries | mAP |
|---|---|---|---|
| Query-by-speech (proposed) | 4 folds | All words | 51.24% |
| Query-by-speech (proposed) | 4 folds | In-vocabulary words | 78.38% |
| Query-by-string [3] | 4 folds | All words | 56.54% |
| Query-by-string [3] | 4 folds | In-vocabulary words | 76.2% |
| Query-by-example [9] | 1 fold | All words | 72.98% |

been heard by the system in the training phase. Framed in green appear the words that are considered relevant in our ground-truth. Note that in our experiments we have not filtered stop-words nor removed words with few appearances. We neither applied any stemming process, so for instance when querying the word ***orders***, results as ***order*** are accounted as negatives.

### B. Query-by-speech Quantitative Results

We report the obtained mean average previsions (mAP) results of our system in Figure 4 depending on the amount of topics $T$ when using Festival's and Google's voices as queries. The resulting mAP shows that the larger the dimensionality of the topic space is, the better the retrieval performance of the system. For the largest topic space size, we reached a 78.38% mAP when using only in-vocabulary queries. When considering all the queries in the test set, the performance drops to a 51.24% mAP.

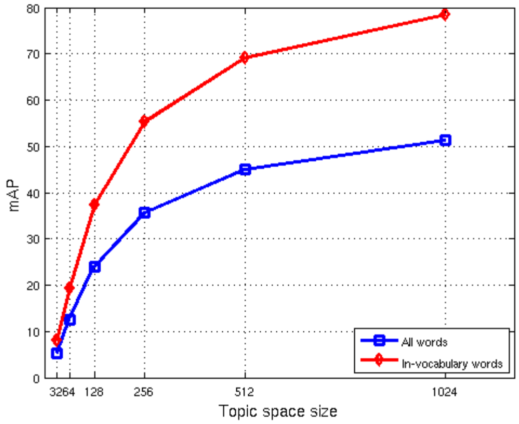


Fig. 4. mAP attained by the system using different number of dimensions in the topic space.

We report in Table I a performance comparison of the proposed query-by-speech method against our previous query-by-string proposal [3] and a query-by-example set up [9]. All three methods use exactly the same visual descriptors and both query-by-speech and query-by-string evaluation protocols are exactly the same. Obviously, the query-by-example experiment do not require any fold partition and each word snippet is used as query in a leave-one-out fashion. It is interesting to see that even if the audio queries are noisy while our previous string queries were not, the reached performances are quite comparable, even delivering better results in the in-vocabulary query setup. Obviously the reached performances in the case of multi-modal representations do not reach the performance levels of just using a visual descriptor in a query-by-example fashion, but we have to take into account that both query-by-string and query-by-speech do not require a manual search of the query template.

### C. Query-by-speech with Unheard Voices

In the previous experiment, we used as queries word utterances pronounced by the same voices that were using in

TABLE II. RETRIEVAL PERFORMANCE USING UNHEARD VOICES AS QUERIES.

| Voice | mAP all words from [14] | mAP for in-vocabulary words from [14] |
|---|---|---|
| AT&T's voice *Claire* | 15.46% | 16.12% |
| AT&T's voice *Crystal* | 14.72% | 15.27% |
| AT&T's voice *Lauren* | 10.69% | 11.13% |
| AT&T's voice *Mike* | 15.98% | 16.74% |
| AT&T's voice *Rich* | 13.01% | 13.64% |
| Average | 13.97% | 14.58% |

TABLE III. ACCURACIES FOR THE HANDWRITTEN TEXT-TO-SPEECH.

| Voice | Accuracy all words | Accuracy in-vocabulary words |
|---|---|---|
| Google voice | 31.13% | 37.86% |
| Festival voice 1 | 55.68% | 67.71% |
| Festival voice 2 | 58.40% | 71.06% |
| Festival voice 3 | 55.90% | 67.75% |
| Festival voice 4 | 61.04% | 74.18% |
| Festival voice 5 | 56.84% | 68.91% |
| Borda Count | 70.00% | 85.24% |

the training phase. We report in Table II the obtained results by the proposed system with $T = 1024$ when the casted queries are pronounced by a different voice than the ones used to train the LSA model. Here we have used the five different AT&T's voices to pronounce a small set of 38 queries, which are the ones used by Liang et al. in [14]. Here we can observe an important performance drop when compared with our previous experiment. However, such low mAP values might still be acceptable for plain users in certain scenarios, since despite the low mAP scores, usually some positive word instances are well ranked in the top positions as we have seen in Figure 3.

### D. Handwritten Text-to-speech via Audio Retrieval

Finally, we account the recognition accuracies of the handwritten text-to-speech task in Table III. The table reports the percentage of queries in which we retrieved the correct utterance at the first rank. As expected, we obtain a significant improvement when just querying in-vocabulary words in comparison to using the whole corpus. We also observe that the performance reached among different voices is quite disperse. Since our approach is not a TTS engine per se, but an audio retrieval system, given an image query we can combine several retrieval outputs from different voices to overcome such diversity and obtain better performances. In our case, the Borda count combination of the ranks obtained with each voice leads to an important improvement over the indexation of individual voices. The method yields promising results despite its simplicity and the fact that it does not entail an explicit recognition of the handwritten words.

## VI. CONCLUSION AND FURTHER STEPS

In this paper, we have proposed a method that enables two novel applications. On the one hand a query-by-speech handwritten word spotting system, and on the other hand a handwritten text-to-speech mechanism. We reach state of the art spotting performances when the same voice is used both for training and querying. However there is still room for improvement in the case that the system is queried with an unheard voice. Some adaptation step would be beneficial in that scenario. The proposed approach has been tested by using synthetic voices generated by TTS engines, we plan in the future to collect a large enough data sample of audio recordings of human voices to further test the generality of the approach. Concerning the handwritten text-to-speech applications, we reach acceptable recognition accuracies when combining different voices. It would be interesting to conduct an experiment indexing larger audio dictionaries such as WordNet.

## ACKNOWLEDGMENTS

This work has been partially supported by the Spanish Ministry of Education and Science under projects TIN2014-52072-P, and TIN2012-37475-C02-02, by the People Programme (Marie Curie Actions) of the Seventh Framework Programme of the European Union (FP7/2007-2013) under REA grant agreement no. 600388, and by the Agency of Competitiveness for Companies of the Government of Catalonia, ACCIÓ.

## REFERENCES

[1] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "Lexicon-free handwritten word spotting using character HMMs," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 934–942, May 2012.

[2] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke, "A novel word spotting method based on recurrent neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 211–224, February 2012.

[3] D. Aldavert, M. Rusiñol, R. Toledo, and J. Lladós, "Integrating visual and textual cues for query-by-string word spotting," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2013, pp. 511–515.

[4] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 12, pp. 2552–2566, July 2014.

[5] A. Black and P. Taylor, "The Festival speech synthesis system: System documentation," Human Communciation Research Centre, University of Edinburgh, Scotland, UK, Tech. Rep. HCRC/TR-83, 1997.

[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, April 1990.

[7] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, October 1994.

[8] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2169–2178.

[9] D. Aldavert, M. Rusiñol, R. Toledo, and J. Lladós, "A study of bag-of-visual-words representations for handwritten keyword spotting," *International Journal on Document Analysis and Recognition*, 2015.

[10] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proceedings of Computer Vision and Pattern Recognition*, 2010, pp. 3360–3367.

[11] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proceedings of the European Conference on Computer Vision*, 2010, pp. 143–156.

[12] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science and Technology*, vol. 41, no. 6, pp. 391–407, September 1990.

[13] V. Lavrenko, T. Rath, and R. Manmatha, "Holistic word recognition for handwritten historical documents," in *Proceedings of IEEE International Workshop on Document Image Analysis for Libraries*, 2004, pp. 278–287.

[14] Y. Liang, M. Fairhurst, and R. Guest, "A synthesised word approach to word retrieval in handwritten documents," *Pattern Recognition*, vol. 45, no. 12, pp. 4225–4236, December 2012.