High Performance Query-by-Example Keyword Spotting Using Query-by-String Techniques

Enrique Vidal, Alejandro H. Toselli and Joan Puigcerver - [ahector , evidal , joapuipe]@prhlt.upv.es

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

1. Introduction

Key Word Spotting (KWS): find probable occurrences of a keyword in a collection of document images.

Query-by-Example (QbE): The keyword is a text image snippet.
Do not generally need training, but KWS performance is *low*

Query-by-String (QbS): The query is the text itself.
Training from transcribed text images, KWS performace can be *high*

Can QbE be approached using QbS technology?

2. Training-based Query-by-String KWS

Two models are trained from training transcribed text images: *Optical Model* (OM) and *Language Model* (LM). Let $\mathcal{M}$ denote these models.

For a given word v and (line shaped) image region x from a test image, $\mathcal{M}$ can be used to compute the *probability that v is present in x, at horizontal position i:* $P(v \mid x, i), 1 \leq i \leq n = |x|$.

The Relevance of x for keyword v is a binary random variable, $\mathcal{R}$. Then:

$$P(\mathcal{R} = 1 \mid x, v) \;\equiv\; P(\mathcal{R} \mid x, v) \;\approx\; \max_{1 \leq i \leq n} P(v \mid x, i) \quad (1)$$

3. Query-by-Example through QbS KWS

For QbE, $P(\mathcal{R} \mid x, q)$ is needed, where q is the query *text image snippet*.

• The true transcript of q is *unknown* (or *"hidden"*), but using $\mathcal{M}$ we can compute $P(v \mid q)$ for all possible transcripts, v. Then:

$$\begin{aligned} P(\mathcal{R} \mid x, q) &= \sum_v P(\mathcal{R}, v \mid x, q) \\ &= \sum_v P(\mathcal{R} \mid x, q, v)\, P(v \mid x, q) \\ &= \sum_v P(\mathcal{R} \mid x, v)\, P(v \mid q) \quad (2) \\ &\approx \max_v P(\mathcal{R} \mid x, v)\, P(v \mid q) \quad (3) \end{aligned}$$

• A further approximation (which explains a simple, intuitive idea): use $\mathcal{M}$ to automatically transcribe the query snippet q; then:

$$\begin{aligned} v^\star &= \arg\max_v P(v \mid q) \quad \text{// query image recognition} \\ P(\mathcal{R} \mid x, q) &\approx P(v^\star \mid q)\, P(\mathcal{R} \mid x, v^\star) \quad (4) \end{aligned}$$

• A final, totally naïve *baseline* approach:
use $\mathcal{M}$ to automatically transcribe also the text image region x ; then do just text-based KWS using the (noisy) recognized word $v^\star$ and text $w^\star$:

$$\begin{aligned} w^\star &= \arg\max_w P(w \mid x) \quad \text{// image region recognition} \\ P(\mathcal{R} \mid x, q) &\approx \begin{cases} 1 & \textit{if } v^\star \in w^\star \\ 0 & \textit{otherwise} \end{cases} \quad (5) \end{aligned}$$

4. Datasets and Keywords

Test images & keywords: exactly as in the ICFHR-2014 KWS competition.

OM training data: Transcribed images from ICFHR-2014 HTRtS contest; additional Bentham texts used to train the LM:

		Training	Validation
Image Data (Bentham page images)	Pages	300	50
	Lines	8 019	1 291
	Running chars.	373 604	61 859
	Character set	93	84
Text Data (Bentham + other texts)	Running words	10 855 571	12 221
	Lexicon size (words)	78 311	2 602

5. Line Image Regions

To allow catching *linguistic context*, line-shaped image regions are needed. Two empirical conditions explored:

1. Line regions are given, but the system determines horizontal positions of the spotted words (conventional line-level QbS KWS).
2. Automatic line segmentation and horizontal word position determination (the most standard segmentation-free QbE KWS setting, as used in the ICFHR-2014 KWS competition).

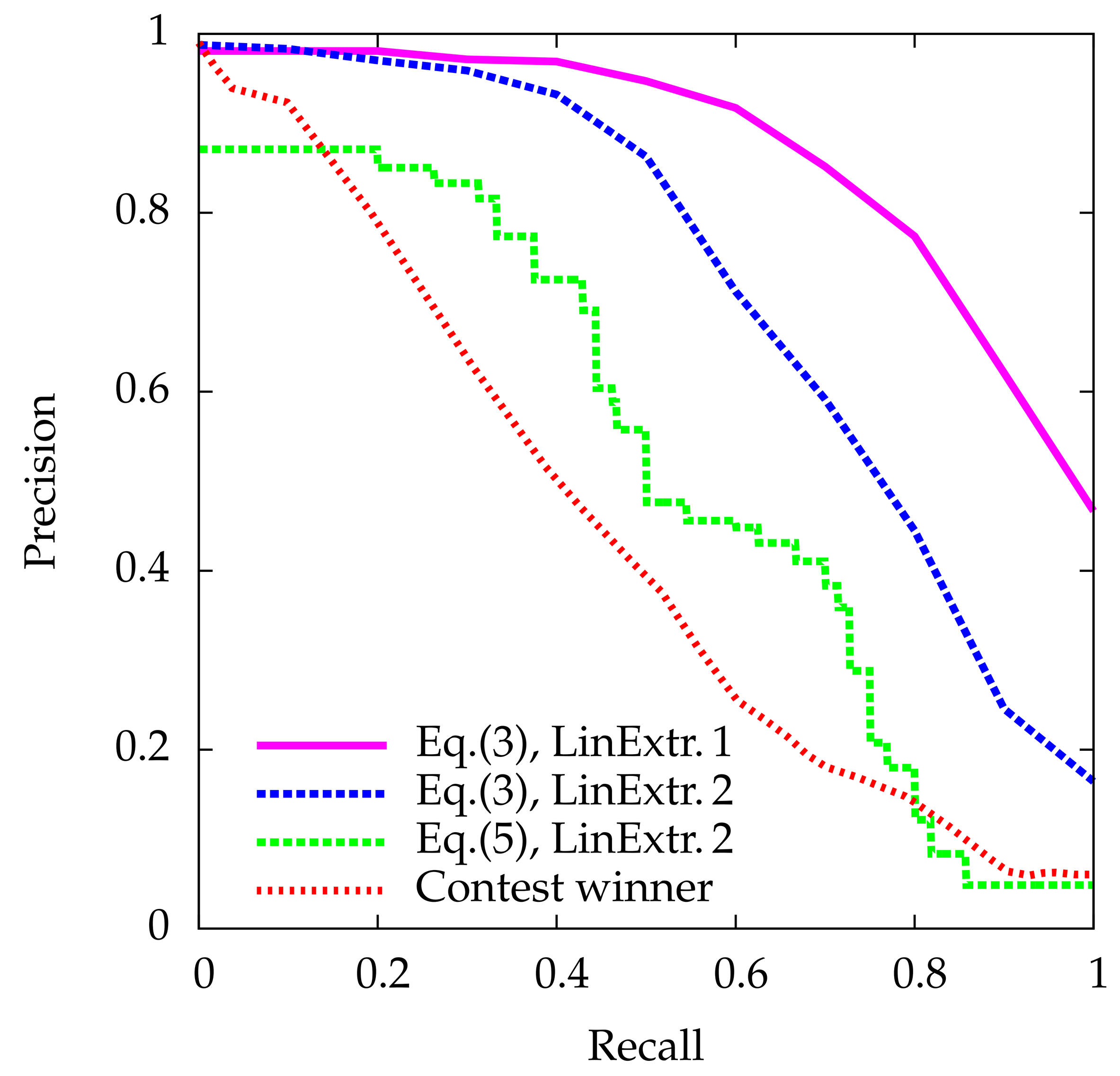
6. Experimental Results

• KWS results for different equations and line extraction assumptions:

Equation Line Extraction	(1) 1	(3) 1	(3) 2	(5) 2
mAP	0.863	0.865	0.715	0.547

Results for Eq.(2-4) are very similar to Eq.(3); Eq.(1) result corresponds to QbS; Result in red: identiacal test conditions as in ICFHR-2014 KWS contest.

• Mean Recall-Precision curves achieved in this work, along with that of the winner of the ICFHR-2014 KWS competition:



• Comparison between the official scoreboard of the ICFHR-2014 KWS contest and this work:

	P@5	NDCG (bin.)	NDCG	mAP
Team 1	0.611	0.640	0.657	0.419
Team 3	0.568	0.518	0.536	0.372
Team 4	0.341	0.363	0.376	0.209
Team 5	0.550	0.513	0.531	0.347
This work	0.879	0.822	0.823	**0.715**

7. Discussion and Future Work

• Yes: QbE can be properly approached from the QbS perspective.

• QbS leverages the use of training data, leading to superb KWS results.

• Even higher performance would be possible by avoiding or reducing line segmentation errors $\Longrightarrow$ *Future work*

• Results of this work were achieved using large image and text training data sets $\Longrightarrow$ *Future work*

• ***Future work***: Determine how much training data is really enough, and whether optical and language models trained for a different, large collection can be used to obtain competitive results.