

# Fisher kernels for handwritten word-spotting

**Florent Perronnin**

Xerox Research Centre Europe  
Meylan, France

[Florent.Perronnin@xrce.xerox.com](mailto:Florent.Perronnin@xrce.xerox.com)

**Jose A. Rodriguez-Serrano**

Computer Vision Centre (UAB)  
Barcelona, Spain  
*(now at Loughborough University, UK)*

[cojar@lboro.ac.uk](mailto:cojar@lboro.ac.uk)

**xerox**



**Centre de Visió  
per Computador**

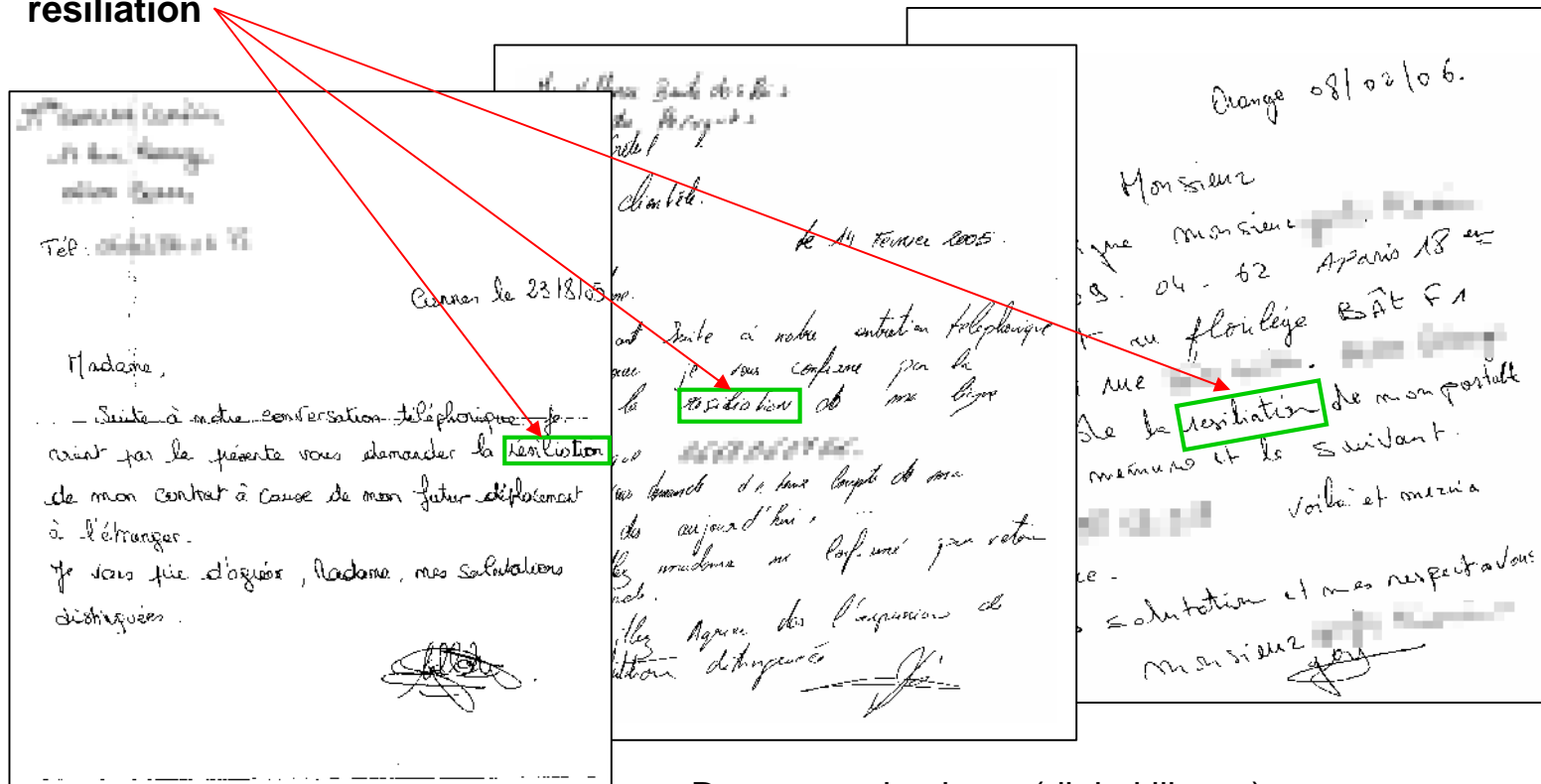
**UAB**

Universitat Autònoma de Barcelona

# Handwritten word-spotting

Handwritten word-spotting: task of finding keywords in handwritten document images (Manmatha et al., 1996)

Query:  
**résiliation**

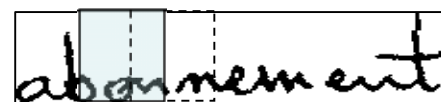


Document database (digital library)

# Word-spotting with hidden Markov models

We assume standard segmentation and pre-processing operations

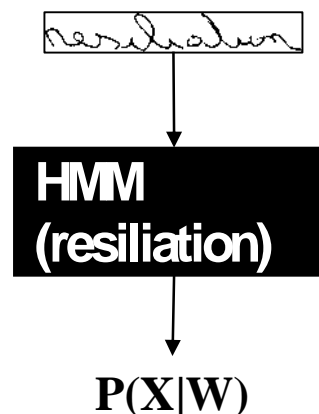
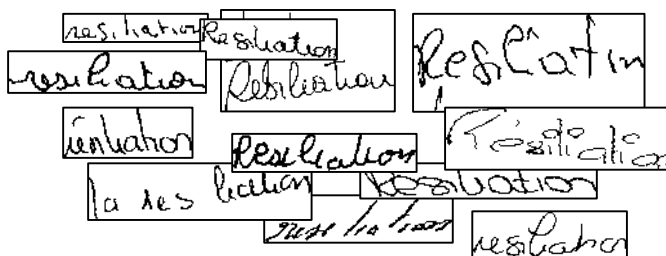
Words images described  
as sequences of feature vectors



$X_1$   $X_t$   $X_{t+1}$  ...  $X_T$  } Sequence  
of feature  
vectors

Local gradient histogram features (Rodriguez and Perronnin, 2008)

“resiliation”



Baum-Welch algorithm  
(Baum et al., 1970)

Forward-backward algorithm  
(Rabiner, 1989)

Since the number of keywords is small (and keywords known in advance), we use whole-word models

# Generative vs. discriminative classifiers

HMMs are generative classifiers

## Generative classifiers

Attempt to model the class-conditional probabilities, i.e.  $p(X|W)$

Classification performed e.g. by Bayes' rule  
 $p(W|X) \propto p(X|W)p(W)$

Pro: Able to handle variable length sequences or missing data

Cons: Not accurate if the underlying sample generation does not follow the chosen model  
Optimal only with infinite training data

## Discriminative classifiers

Attempt to model the class boundaries

Posteriors  $P(W|X)$  modelled directly

Pros: Focus directly on the problem of interest  
Theoretical and often practical superiority

Con: Kernels on strings are costly

**Fisher kernels:** framework for combining the advantages of generative and discriminative classifiers

# Fisher kernel principle

- Idea [JH99]: given a generative model  $p$  with parameters  $\lambda$ , compute a fixed-length representation of the vector set  $X = \{x_t, t = 1 \dots T\}$  using the following gradient vector:

$$\nabla_{\lambda} \log p(X|\lambda)$$

- Gradient vector = in which direction should the parameters of the model be modified to best fit the data

- Use the Fisher information matrix to measure the similarity between gradients:

$$F_{\lambda} = E_X [\nabla_{\lambda} \log p(X|\lambda) \nabla_{\lambda} \log p(X|\lambda)']$$

- Equivalent to normalizing the gradient vectors:

$$F_{\lambda}^{-1/2} \nabla_{\lambda} \log p(X|\lambda)$$

# Fisher kernel on hidden Markov models

$$G_\lambda(X) = \nabla_\lambda \mathcal{L}(X|\lambda)$$

We focus on the derivatives with respect to the means and covariances

$$\left. \begin{aligned} \frac{\partial \mathcal{L}(X|\lambda)}{\partial \mu_{i,j}^d} &= \sum_{t=1}^T \gamma_t(i,j) \left[ \frac{x_t^d - \mu_{i,j}^d}{(\sigma_{i,j}^d)^2} \right], \\ \frac{\partial \mathcal{L}(X|\lambda)}{\partial \sigma_{i,j}^d} &= \sum_{t=1}^T \gamma_t(i,j) \left[ \frac{(x_t^d - \mu_{i,j}^d)^2}{(\sigma_{i,j}^d)^3} - \frac{1}{\sigma_{i,j}^d} \right] \end{aligned} \right\}$$

Size of the gradient vector is 2DM, with  
*D = dimensionality of the feature vectors*  
*M = number of Gaussians in the HMM*

D = 128, M ≈ 100  
 Dimensionality ≈ 10,000

High dimensionality  $\stackrel{?}{=}$  more discriminative

These derivatives are computed as a combination of the following statistics

$$\sum_{t=1}^T \gamma_t(i,j)$$

$$\sum_{t=1}^T \gamma_t(i,j) x_t^d$$

$$\sum_{t=1}^T \gamma_t(i,j) (x_t^d)^2$$

But these quantities are already accumulated during MLE. Since MLE is a building block in any HMM system, the FK implementation represents a small add-on to the already existing code.

# Classification

We work in the framework of kernel classifiers

$$s(z) = f \left( \sum_{i=1}^N \alpha_i y_i K(z, z_i) + b \right)$$

We use sparse logistic regression

f: sigmoid

Laplacian prior on  $\alpha$  that induces a sparse solution

**Linear kernel**

$$z^T \left( \sum_{i=1}^N \alpha_i y_i z_i \right) + b$$

(fast)

**Non-linear kernel**

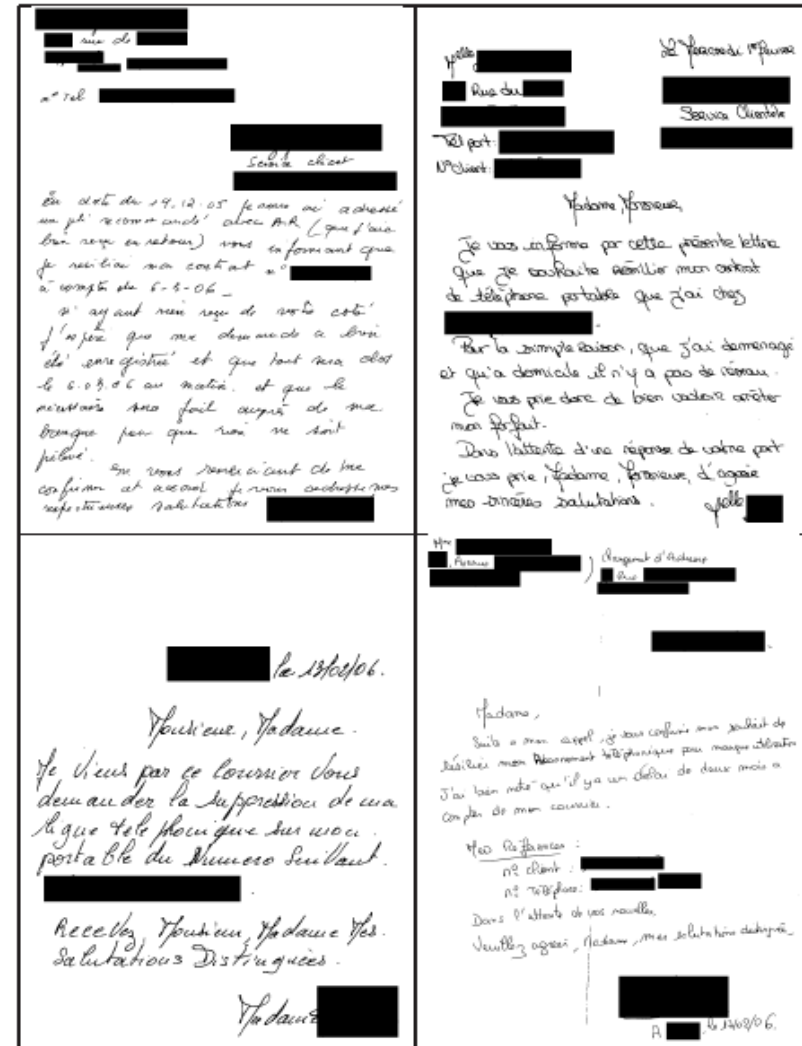
$$D(z, z_i) = \left\| \frac{z}{\|z\|_1} - \frac{z_i}{\|z_i\|_1} \right\|_1$$

$$K(z, z_i) = \exp(-\beta D(z, z_i))$$

(compares directions)  
(vectors distributed  
according to Laplacian)

# Experimental benchmark

- Dataset: real mailroom data (1<sup>st</sup> page of 630 letters)
  - ~630 different writers
- Word segmentation: produces 180,000 word hypotheses
- The 10 most frequent keywords considered in the experiments
- 5-fold cross-validation
- Baseline: HMM with 16 Gaussians per state and score normalisation (Rodriguez and Perronnin 2008b)
- We use an HMM with 1 Gaussian per state



# Experimental benchmark

- Dataset: real mailroom data (1<sup>st</sup> page of 630 letters)
  - ~630 different writers
- Word segmentation: produces 180,000 word hypotheses
- The 10 most frequent keywords considered in the experiments
- 5-fold cross-validation
- Baseline: HMM with 16 Gaussians per state and score normalisation (Rodriguez and Perronnin 2008b)
- We use an HMM with 1 Gaussian per state

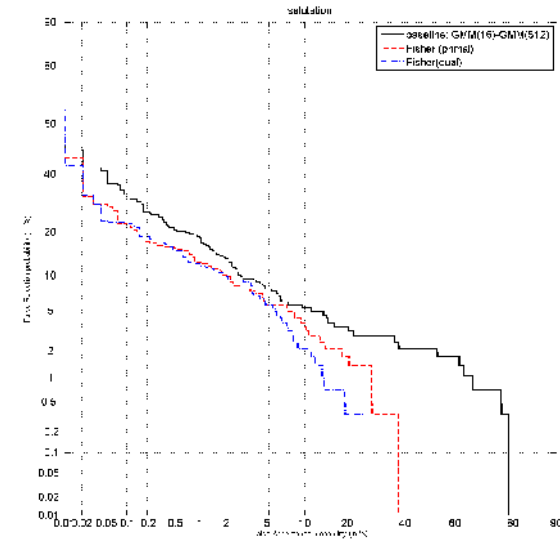
Keyword	Translated as	# labeled examples
<i>Monsieur</i>	Sir	750
<i>Madame</i>	Madam	543
<i>contrat</i>	contract	514
<i>résiliation</i>	cancellation	446
<i>salutation</i>	greeting	382
<i>résilier</i>	to cancel	349
<i>demande</i>	request	337
<i>abonnement</i>	subscription	299
<i>(company name)</i>		308
<i>veuillez</i>	if you please	303



# Experimental results

Accuracy  
(Average precision)

word ID	[11]	FK-L	FK-NL
1	90.6	<b>93.1</b>	92.6
2	94.8	95.5	<b>95.6</b>
3	90.4	91.8	<b>92.3</b>
4	82.7	83.8	<b>85.8</b>
5	90.1	91.8	<b>92.8</b>
6	86.1	80.3	<b>88.9</b>
7	85.9	85.9	<b>86.1</b>
8	71.6	70.4	<b>74.6</b>
9	86.9	90.7	<b>91.0</b>
10	<b>90.1</b>	88.8	90.0
Average	86.9	87.2	<b>89.0</b>



DET plot for "salutation"

Speed  
(time per keyword image  
per word model)

35 ms    2.5 ms    0.5s

(but can be further speeded up with  
(Maji et. al 2008)

# Speeding up the non-linear kernel

Non-linear kernel (intersection kernel)

$$D(z, z_i) = \left\| \left\| \frac{z}{\|z\|_1} - \frac{z_i}{\|z_i\|_1} \right\|_1 \right\|_1 \quad K(z, z_i) = 1 - \frac{1}{2} D(z, z_i)$$

Assume kernels that can be expressed as a sum over the dimensions  
(such as the previous non-linear kernel)

$$y = \sum_{i=1}^N \alpha_i y_i \sum_{d=1}^D K(x^d, x_i^d)$$

Interchanging the summations leads to

$$y = \sum_{d=1}^D \sum_{i=1}^N \alpha_i y_i K(x^d, x_i^d) = \sum_{d=1}^D \underbrace{f_d(x_d)}$$

(Maji et al. 2008): piecewise linear function of  $N+1$  segments  
approximated by a reduced set of piecewise constant/linear segments  
**constant runtime**

# Conclusions and perspectives

## Conclusions

- We applied the Fisher kernel framework to the task of handwritten word spotting
- Generate a feature vector that describes how the parameters of the generative model must be modified in order to better fit the word image
- Classification with two types of kernels
  - Linear kernel: 15 times faster
  - Non-linear kernel: 15% relative reduction of error

## Perspectives

- Use of Fisher kernels for QBS approaches, where the number of words to spot can be of thousands
- Experiment with other types of kernels (e.g. similarity measure CVPR'2009)

# Bibliography

- Baum et al., “A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains”, *The Annals of Mathematical statistics*, 1970
- T. Jaakola and D. Haussler, *Exploiting generative models in discriminative classifiers*, NIPS, 1999.
- Maji, A. et al., "Classification using Intersection Kernel Support Vector Machines is Efficient", CVPR 2008.
- Manmatha et al., “Word spotting: a new approach to indexing handwriting”, CVPR, 1996
- Perronnin and Dance. Fisher kernels on visual vocabularies for image categorization, *CVPR*, 2007.
- Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition”, *Proceedings of the IEEE*, 1989
- Rodriguez and Perronnin, “Local gradient histogram features for word spotting in unconstrained handwritten documents”, *ICFHR*, 2008
- J. A. Rodriguez and F. Perronnin. Score normalization for HMM-based handwritten word spotting using a universal background model, *ICFHR*, 2008.