Label embedding for text recognition

Jose A. Rodriguez Serrano and Florent Perronnin

Xerox Research Centre Europe

BMVC 2013

Table of Contents

Table of Contents

Problem: Text recognition in natural images

From: IIIT-5K set (Mishra et al. [6])

Problem: Text recognition in natural images

From: IIIT-5K set (Mishra et al. [6])

Goal of word recognition:

Output the sequence of characters present in an input **word image**.

E.g.:  $\rightarrow$ {W,E,L,C,O,M,E}

Problem: Text recognition in natural images

From: IIIT-5K set (Mishra et al. [6])

Goal of word recognition:

Output the sequence of characters present in an input **word image**.

E.g.:  $\rightarrow$ {W,E,L,C,O,M,E}

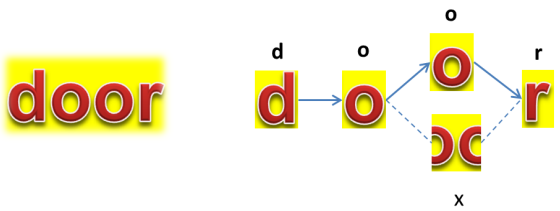
Output can be constrained by the **lexicon** (list of valid words).

Related work

xerox

Bottom-up approaches

1. Detect character hypotheses
2. Use a high-level model to reach consensus between the hypotheses.



✓ Scales to large lexicons

✓ Enables open vocabularies (zero-shot learning)

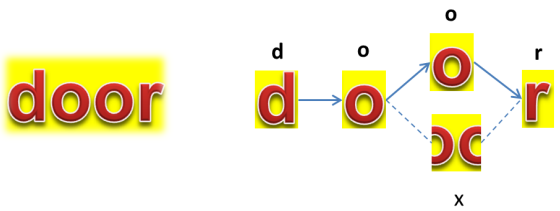
✗ Relies on the character detector

See: Wang, ECCV'10 [16]; Wang, ICCV'11 [15]; Neumann, CVPR'12 [8]; Mishra, CVPR'12 [7]; Mishra, BMVC'12 [6]; Novikova, ECCV'12 [9]

Related work

Bottom-up approaches

1. Detect character hypotheses
2. Use a high-level model to reach consensus between the hypotheses.



✓ Scales to large lexicons

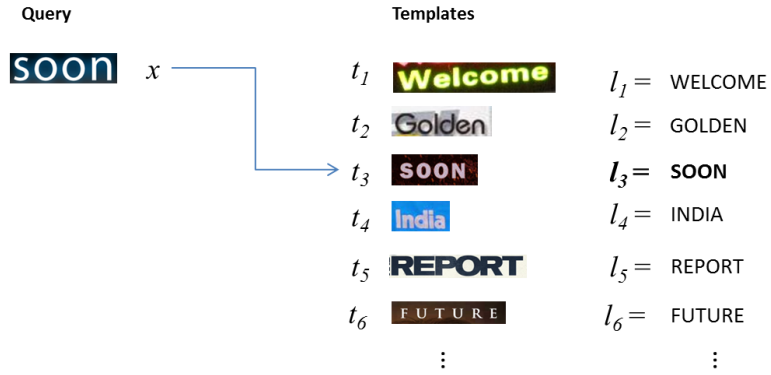
✓ Enables open vocabularies (zero-shot learning)

✗ Relies on the character detector

See: Wang, ECCV'10 [16]; Wang, ICCV'11 [15]; Neumann, CVPR'12 [8]; Mishra, CVPR'12 [7]; Mishra, BMVC'12 [6]; Novikova, ECCV'12 [9]

Holistic approaches

1. Extract 1 feature vector for the whole word image
2. Find nearest template



✓ Simplicity and efficiency at runtime (for small lexicons)

✓ Does not require pre-processing

✗ No zero-shot learning

See: Rath CVPR'03 [12]; Rodriguez, PAMI 2012 [13]; Rodriguez, CVVT 2012 [14]

Desired framework

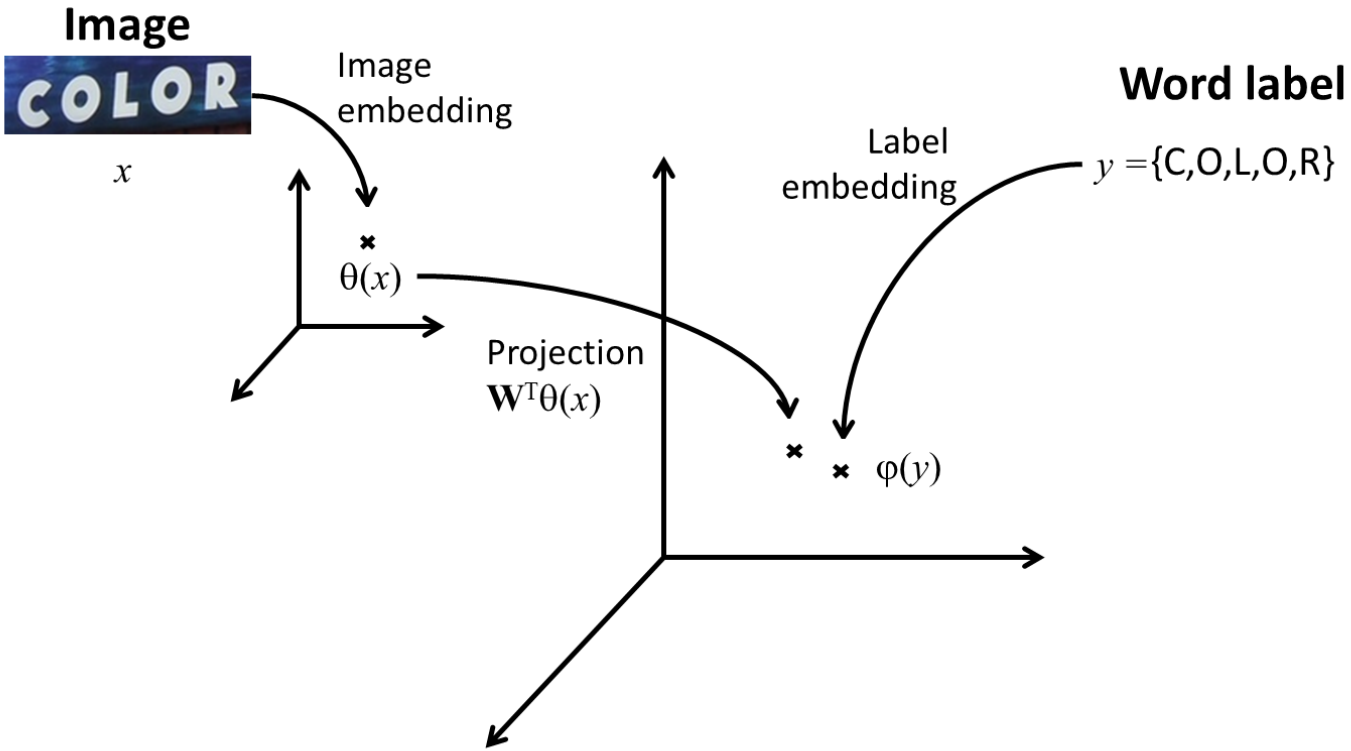
We want:

- Extract 1 single feature vector per image
- Compute match between images and labels using a simple function
- Preserve zero-shot learning

Table of Contents

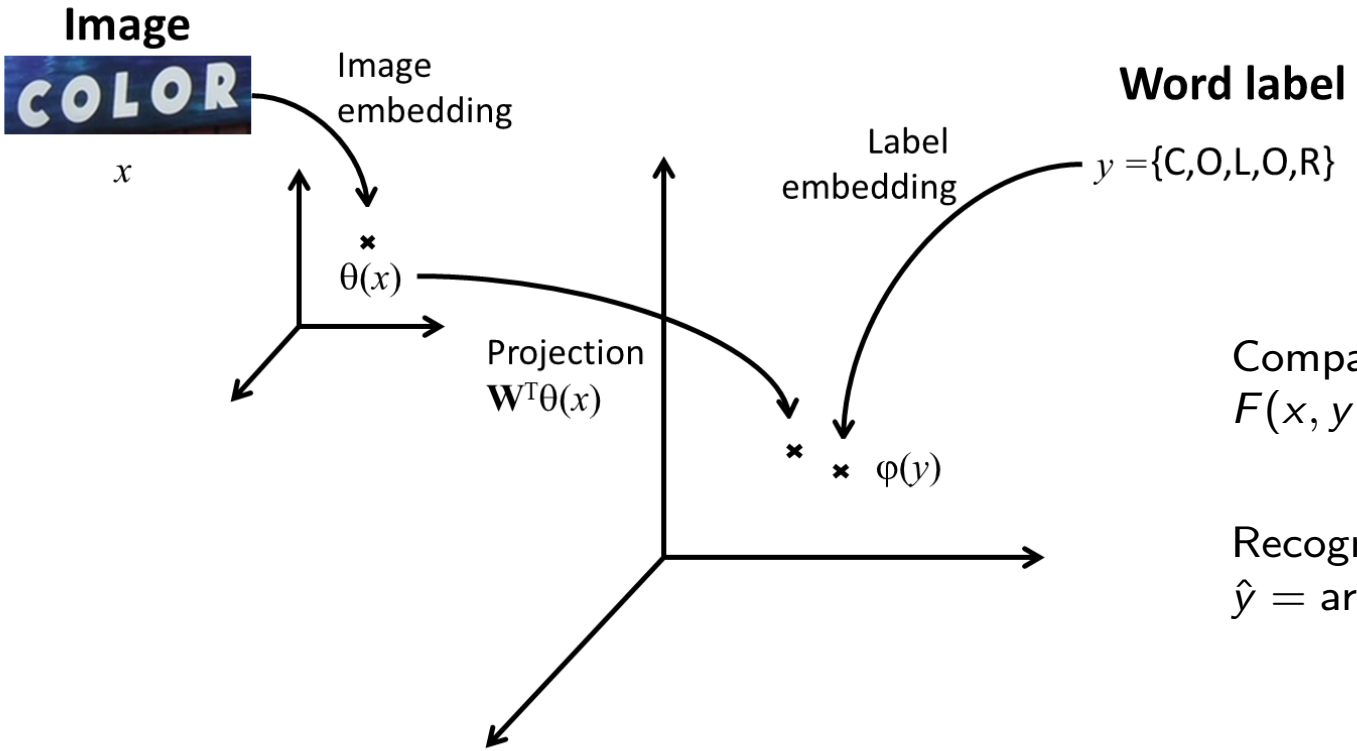
Model formulation

Embed both input **and output** space



Model formulation

Embed both input **and output** space



Compatibility

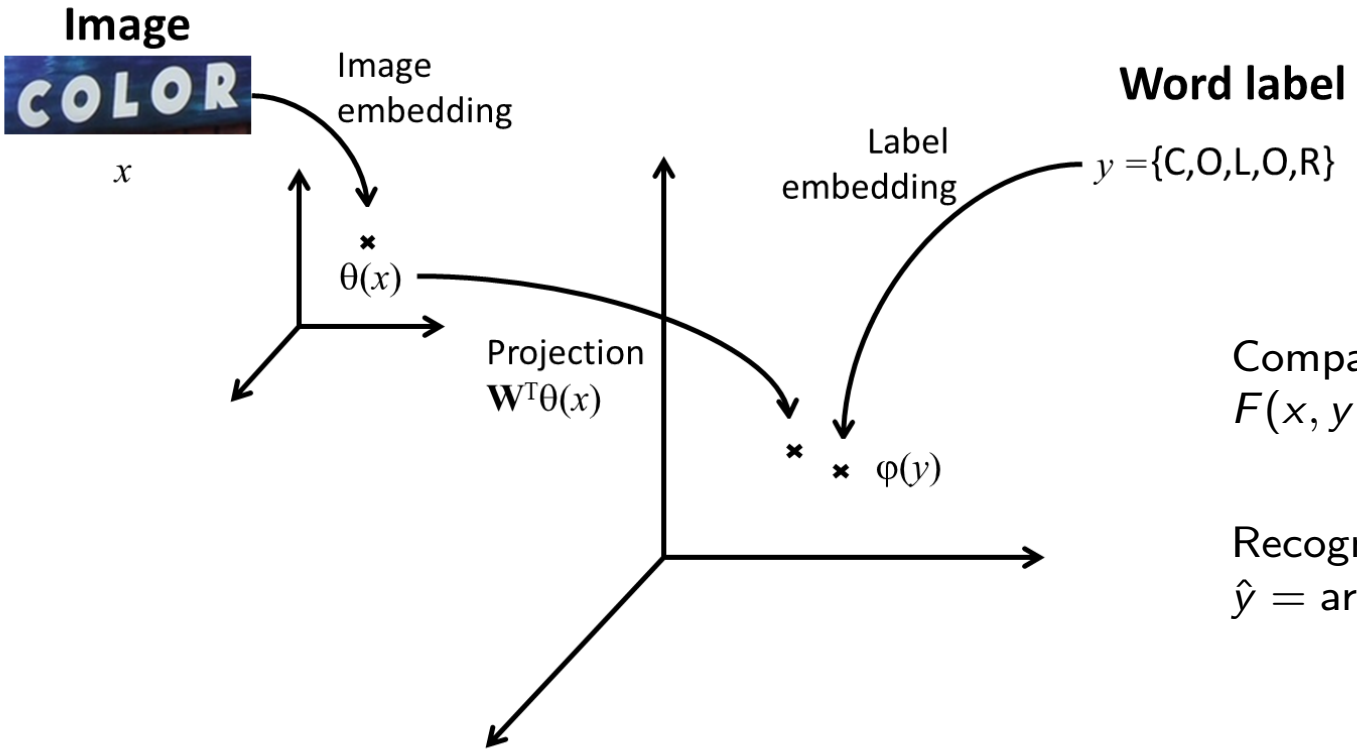
$F(x, y; W) = \theta^T(x) W \varphi(y)$

Recognition

$\hat{y} = \arg\max_{y \in \mathcal{Y}} F(x, y; W)$

Model formulation

Embed both input **and output** space



Compatibility

$F(x, y; W) = \theta^T(x) W \varphi(y)$

Recognition

$\hat{y} = \arg\max_{y \in \mathcal{Y}} F(x, y; W)$

Questions

- Q1 Define image embedding $\theta(x)$
- Q2 Define label embedding $\phi(y)$
- Q3 Learn projection matrix W

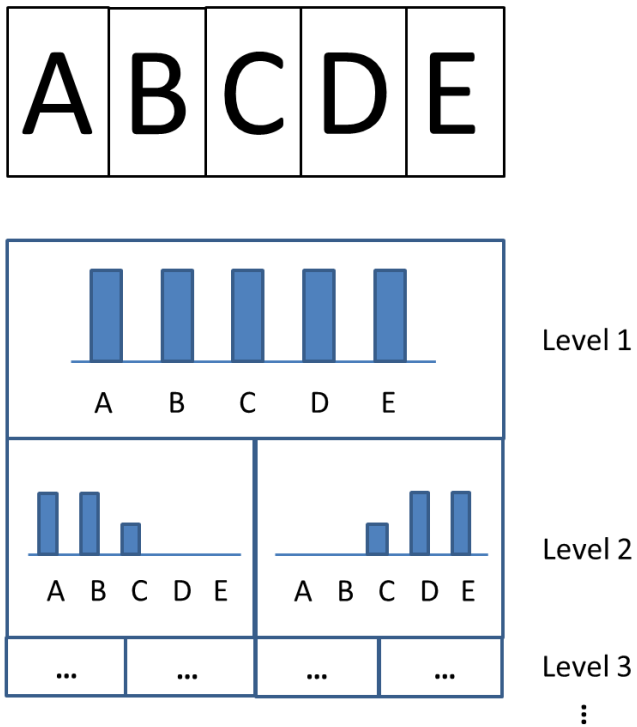
Q1. Image embedding $\theta(x)$

- The proposed framework is independent of the specific image coding
- We use Fisher vectors [11] as a test case (state-of-the-art results in image retrieval [4] and classification [3]).

Q2. Label embedding $\phi(y)$

- Goal: Embed sequence of characters into a vector space
- Spatial Pyramid of Characters (SPOC)
  = Bag of characters with linear spatial pyramid

Example with word {A,B,C,D,E}, alphabet=[A-E]



Q3. Learning the W in the compatibility function

Optimize a ranking criterion: Structured SVM [10]

Force $F(x_n, y_n; w) > F(x_n, y; w)$, $y \neq y_n$

Q3. Learning the W in the compatibility function

Optimize a ranking criterion: Structured SVM [10]

Force $F(x_n, y_n; w) > F(x_n, y; w),\ y \neq y_n$

$$R(\mathcal{S}; w) = \underbrace{\frac{1}{N}\sum_{n=1}^{N} \Delta_{0/1}(y_n, f(x_n))}_{L} + \frac{\lambda}{2}||w||^2. \quad (1)$$

$$B_2(y_n, f(x_n)) = \sum_{y \in \mathcal{Y}} \delta(y_n, y) - F(x_n, y_n; w) + F(x_n, y; w) \geq L \quad (2)$$

Optimized using Stochastic Gradient Descent [2]

Q3. Learning the W in the compatibility function

Optimize a ranking criterion: Structured SVM [10]

Force $F(x_n, y_n; w) > F(x_n, y; w),\ y \neq y_n$

$$R(\mathcal{S}; w) = \underbrace{\frac{1}{N}\sum_{n=1}^{N} \Delta_{0/1}(y_n, f(x_n))}_{L} + \frac{\lambda}{2}||w||^2. \quad (1)$$

$$B_2(y_n, f(x_n)) = \sum_{y \in \mathcal{Y}} \delta(y_n, y) - F(x_n, y_n; w) + F(x_n, y; w) \geq L \quad (2)$$

Optimized using Stochastic Gradient Descent [2]

Optimize a reconstruction criterion: Ridge regression

$$E(W) = \frac{1}{N}\sum_i \|\varphi(y_i) - W\theta(x_i)\|^2 + \eta_R \|W\|_F^2, \quad (3)$$

Closed-form solution

Table of Contents

Experiments on License Plate Recognition

- Database of 45K US license plate images [14] (train: 34K, test: 11K)
- Challenging because US plates contain graphical backgrounds, symbols, many plate types allowed, many concurrent numbering systems.

Table : Summary of parameters

Parameter	Value
Low-level descriptors	SIFT (128 dimensions)
PCA on descritors	32 dimensions
Visual vocabulary	64 Gaussians
Pyramid	4×1
SPOC levels	5
Alphabet	$\mathcal{L} = \{0 \dots 9\} \cup \{A \dots Z\}$
Lexicon	set of unique $\sim$5K plate numbers
Optimization	SSVM

Examples of US plates (for illustration)

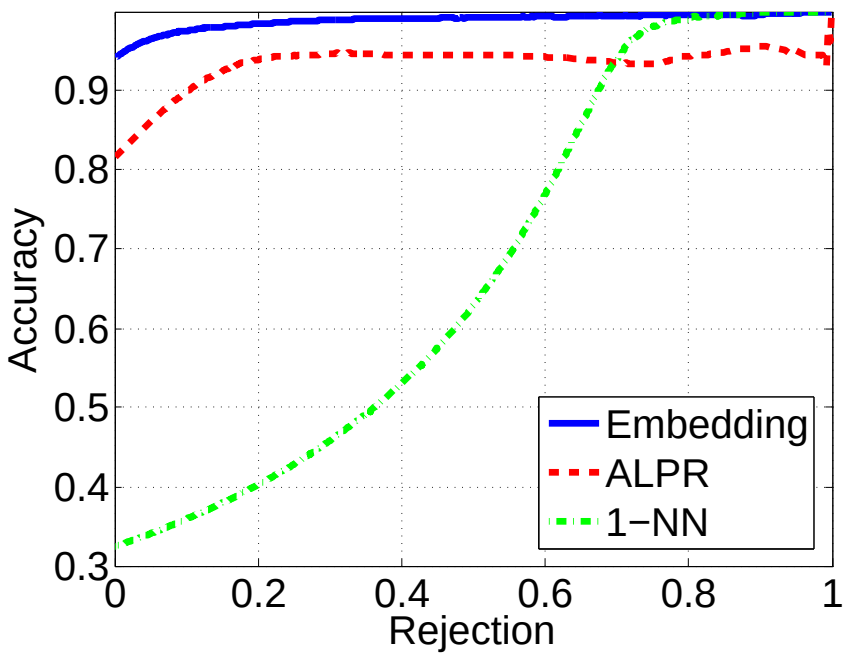
Source: Flickr Creative Commons (Attrib. License). By: timparkinson, stijlfoto, tobyotter. Images cropped manually.

Experiments on License Plate Recognition

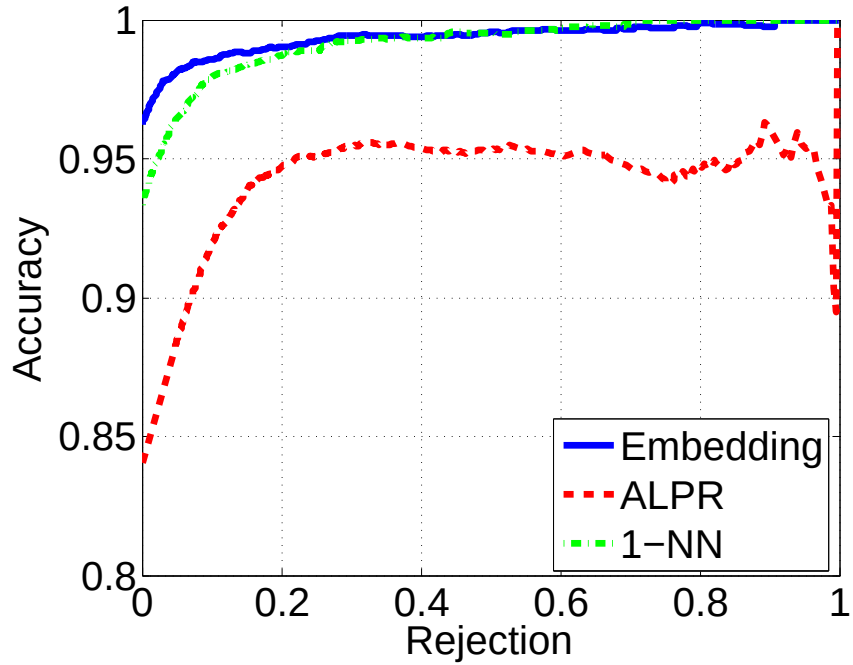
Baselines

- Commercial ALPR system (which looks for characters)
- Holistic approach using NN match [14]

Whole test set



Only labels "seen" during training



Experiments on Scene Text Recognition

- IIIT-5K set [6] (5000 word images from street scenes)

- Optimized with ridge regression

Table : Recognition results on the IIIT-5K dataset

Method	Accuracy $\lvert\mathcal{Y}\rvert = 50$	Accuracy $\lvert\mathcal{Y}\rvert = 1000$
Mishra et al. [6]	64.1%	57.5 %
Label embedding	76.1%	57.4 %

Correctly recognized words

ADVERTISING

AFTER

DROPPED

HOURS

REGENCY

RESEARCH

VILLA

WEBS

WISTERIA

11 AM

172

Failure cases

Confusions with similar words

CATE

GET

OOK

MAIL

BANK

Failure cases

Confusions with similar words

CATE

GET

OOK

MAIL

BANK

Words with common n-grams tend to score very high

COMMITMENT

SERVICES

STATE

THE

Failure cases

Confusions with similar words

CATE

GET

OOK

MAIL

BANK

Words with common n-grams tend to score very high

COMMITMENT

SERVICES

STATE

THE

Frequent errors in numbers...

12 PM

36

14

...and simply bad projections

DEVICE

Conclusions

xerox

Contribution

- Label embedding framework for text recognition

Benefits

- No need to localize/classify individual characters
- Recognition = linear search with cosine similarity
- Zero-shot learning (generalize to "unseen" words)

Future work

- Scale to large lexicons
  (leverage abundant literature on efficient large-scale image retrieval).

Thanks

Jose A. Rodriguez-Serrano and Florent Perronnin

Xerox Research Centre Europe (France)
BMVC, 2013

Related work I


B. Bai, J. Weston, D. Grangier, R. Collobert, O. Chapelle, and K. Weinberger.
Supervised semantic indexing.
In *CIKM*, 2009.


Léon Bottou.
SGD.
http://leon.bottou.org/projects/sgd.


K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman.
The devil is in the details: an evaluation of recent feature encoding methods.
In *BMVC*, 2011.

Related work II


Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid.
Aggregating local image descriptors into compact codes.
IEEE Trans. Pattern Anal. Mach. Intell., 34(9), 2012.


T. Joachims.
Optimizing search engines using clickthrough data.
In *SIGKDD*, 2002.


A. Mishra, K. Alahari, and C. V. Jawahar.
Scene text recognition using higher order language priors.
In *BMVC*, 2012.


A. Mishra, K. Alahari, and C. V. Jawahar.
Top-down and bottom-up cues for scene text recognition.
In *CVPR*, 2012.

Related work III


Lukas Neumann and Jiri Matas.
Real-time scene text localization and recognition.
In *CVPR*, 2012.


Tatiana Novikova, Olga Barinova, Pushmeet Kohli, and Victor Lempitsky.
Large-lexicon attribute-consistent text recognition in natural images.
In *ECCV*, 2012.


S. Nowozin and C. Lampert.
Structured learning and prediction in computer vision.
Foundations and Trends in Computer Graphics and Vision, 2011.


F. Perronnin, J. Sánchez, and Thomas Mensink.
Improving the Fisher kernel for large-scale image classification.
In *ECCV*, 2010.

Related work IV


Toni M Rath and Raghavan Manmatha.
Word image matching using dynamic time warping.
In *CVPR*, 2003.


José A. Rodríguez-Serrano and Florent Perronnin.
A model-based sequence similarity with application to handwritten word spotting.
IEEE Trans. PAMI, 34(11), 2012.


José A. Rodríguez-Serrano, Harsimrat Sandhawalia, Raja Bala, Florent Perronnin, and Craig Saunders.
Data-driven vehicle identification by image matching.
In *ECCV Workshop on Computer Vision for Vehicle Technology*, 2012.


Kai Wang, Boris Babenko, and Serge Belongie.
End-to-end scene text recognition.
In *ICCV*, 2011.

Related work V


Kai Wang and Serge Belongie.
Word spotting in the wild.
In *ECCV*, 2010.

Backup: Bottom-up approaches

xerox

Work	Char detection	Char classification	High-level model
Wang, ECCV'10 [16]	Sliding window	HOG + nearest-neighbor	Pictorial structure
Wang, ICCV'11 [15]	Sliding window	HOG + random ferns	Pictorial structure
Neumann, CVPR'12 [8]	Extremal regions	Shape + AdaBoost	Pairwise rules
Mishra, CVPR'12 [7]	Sliding window	HOG + SVM	Pairwise CRF
Mishra, BMVC'12 [6]	Sliding window	HOG + SVM	High-order CRF
Novikova, ECCV'12 [9]	MSER	HOG + nearest-neighbor	Weighted finite-state transducer

- Advantages: Learning and recognition scale to large vocabularies.
  Capability to model words not seen during training.
  [**Zero-shot learning**]
- Disadvantages: Very sensitive to character detector
  Requires pre- and post-processing
  (e.g. binarization, string edit correction)

Backup: Holistic approaches

Principle

1. Extract 1 feature vector for the whole word image
2. Find nearest template

- Small vocabulary $\rightarrow$ word spotting
  (Rath & Manmatha [12], Rodriguez and Perronnin, 2012 [13])
- Not-so-small vocabulary (5K) is not-so-infeasible
  (Rodriguez et al., 2012 [14])

Advantage: Recognition = 1 NN query (simple, nowadays efficient)
Less pre-processing

Disadvantages: No open-vocabulary recognition (zero-shot learning)

Backup: Experiments on Scene Text Retrieval

- Projection of images to space of labels: $\tilde{\theta}(x) = W^T \theta(x)$
- Induces a new similarity between images:

$$s(x, y) = \theta(x)^T W^T W \theta(y) \tag{4}$$

- Is it a good similarity between images?

Table : Retrieval results on IIIT-5K

Method	Top-1 acc
FV	38.0
FV+ SSI (Bai [1])	42.2
DTW (Rodriguez [13])	37.0
Label emb	43.7

Table : Sample query images and top-1 match using the image similarity.

Backup. Structured SVM (I)

Remember our "compatibility function":

$$F(x, y; W) = \tilde{\theta}^T(x)\varphi(y) = \theta^T(x)W\varphi(y) \quad (5)$$

Backup. Structured SVM (I)

Remember our "compatibility function":

$$F(x, y; W) = \tilde{\theta}^T(x)\varphi(y) = \theta^T(x)W\varphi(y) \tag{5}$$

If $vec(A)$ denotes the row-wise stacking operation of matrix A, and

- $w = vec(W)$
- $\psi(x, y) = \theta(x) \otimes \varphi(y) = vec([\theta_i(x)\varphi_j(y)])$

Backup. Structured SVM (I)

Remember our "compatibility function":

$$F(x, y; W) = \tilde{\theta}^T(x)\varphi(y) = \theta^T(x)W\varphi(y) \quad (5)$$

If $vec(A)$ denotes the row-wise stacking operation of matrix A, and

- $w = vec(W)$
- $\psi(x, y) = \theta(x) \otimes \varphi(y) = vec([\theta_i(x)\varphi_j(y)])$

Then we can re-write the problem as

SSVM formulation

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} F(x, y; W) \quad (6)$$

with

$$F(x, y; w) = w^T\psi(x, y) \quad (7)$$

Backup. Structured SVM (I)

Remember our "compatibility function":

$$F(x, y; W) = \tilde{\theta}^T(x)\varphi(y) = \theta^T(x)W\varphi(y) \tag{5}$$

If $vec(A)$ denotes the row-wise stacking operation of matrix A, and

- $w = vec(W)$
- $\psi(x, y) = \theta(x) \otimes \varphi(y) = vec([\theta_i(x)\varphi_j(y)])$

Then we can re-write the problem as

SSVM formulation

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} F(x, y; W) \tag{6}$$

with

$$F(x, y; w) = w^T\psi(x, y) \tag{7}$$

Which is the well-known structured SVM [10].

Backup. Structured SVM (II)

Minimize

$$R(\mathcal{S}; w) = \frac{1}{N} \sum_{n=1}^{N} \Delta(y_n, f(x_n)) + \frac{\lambda}{2} ||w||^2. \quad (8)$$

For Δ, we use the 0/1 loss, i.e. $\Delta(y_i, \hat{y}_i) = 0$ if $y_i = \hat{y}_i$ and 1 otherwise, since we are interested in top-1 recognition accuracy.
In the paper we bound it by

$$B_2(y_n, f(x_n)) = \sum_{y \in \mathcal{Y}} \delta(y_n, y) - F(x_n, y_n; w) + F(x_n, y; w) \geq B_1(y_n, f(x_n)). \quad (9)$$

This is similar in spirit to the ranking SVM of Joachims [5]. We call this formulation Ranking SSVM (RSSVM).

Backup. Optimization

xerox

SGD [2] optimization

1. Randomly sample (x_n, y_n)
2. Randomly sample $y \in \mathcal{Y} - y_n$
3. Compute $\xi = \Delta(y_n, y) - F(x_n, y_n; w) + F(x_n, y; w)$
4. If $\xi > 0$, update: $w \leftarrow (1 - \eta_t \lambda) w + \eta_t \left[\psi(x_n, y_n) - \psi(x_n, y)\right]$

Initialization choices:

- W random values from a Normal distribution with mean 0 and standard deviation $\sqrt{D}$
- Regularized ridge regression

$$E(W) = \|\varphi(y) - W\theta(x)\|^2 + \eta_R \|W\|_F^2, \tag{10}$$

$$W = AB^{-1}, \qquad A = \sum_i \varphi_i \theta_i^T, \qquad B = \sum_i \theta_i \theta_i^T + \eta_R I \tag{11}$$

Backup: Data-driven vehicle identification

Rodriguez-Serrano et al., 2012 [14]