

ICDAR–2009 Tutorial:

Interactive Multimodal Transcription of Text Images

I – Introduction

Alejandro H. Toselli & Enrique Vidal

atoselli@irisa.fr (on leave from PRHLT)

evidal@iti.upv.es



Pattern Recognition and Human Language Technology Group

Instituto Tecnológico de Informática – Universidad Politécnica de Valencia



Spain

July 2009

ICDAR09: Interactive Multimodal Transcription

A B L A N K P A G E

Tutorial Contents and Schedule

- I **Introduction**
 - Multimodal Interaction in Pattern Recognition
 - Interactive-Predictive Pattern Recognition and Document Image Analysis
 - Quick Survey of Handwritten Text Recognition (HTR) concepts and techniques
- I-p **Off-line HTR in practice**
 - HTR Preprocessing
 - Training HMMs using the "Hidden Markov Model Toolkit" (HTK)
 - Training Language Models and Dictionaries for HTR
 - HTR Experiments
- II **Computer-Assisted Transcription of Text Images (CATTI)**
 - Human interaction in HTR
 - A CATTI formal framework
 - Feedback-derived dynamic language modelling and search
 - Performance measures and results achieved in typical applications
- *** **COFFEE BREAK**
- II-p **CATTI in practice**
 - Adapting Language Models and Search for CATTI
 - CATTI Experiments
 - Analyzing quantitatively the CATTI performance
- III **Multimodality in CATTI (MM-CATTI)**
 - Touchscreen based multimodal user correction
 - A MM-CATTI formal framework
 - Multimodal language modelling and search
 - Performance measures and results achieved in typical applications
- III-p **Demonstration of a complete MM-CATTI System in a real HTR task**

Index

- 1 Multimodal Interaction in Pattern Recognition (PR) ▷ 3
- 2 Handwritten Text Recognition (HTR) concepts and techniques ▷ 13
- 3 HTR and Interactive-Predictive PR ▷ 19
- 4 Bibliography ▷ 21

Computer Assisted Pattern Recognition: Motivation

- In most PR problems and applications, development purportedly aims at fully automated systems
- But full automation often proves elusive or unnatural in many applications where technology is expected to *assist*, rather than replace the human agents
- In these and many other cases, *practical* PR developments typically end up just in “semiautomatic systems” or systems for “*computer assisted*” operation, where it is a human expert who makes the final decisions
- These facts are very seldom acknowledged: typically, full automation is pretended and the “eventual” need of human intervention is ignored in the mathematical formulation
- *Human interaction* issues are often seen as “uninteresting final implementation details”.

However...

Computer assistance and/or human interaction do entail very interesting opportunities rather than bothersome problems.

Computer Assisted Pattern Recognition: Opportunities

Human interaction entails three types of opportunities:

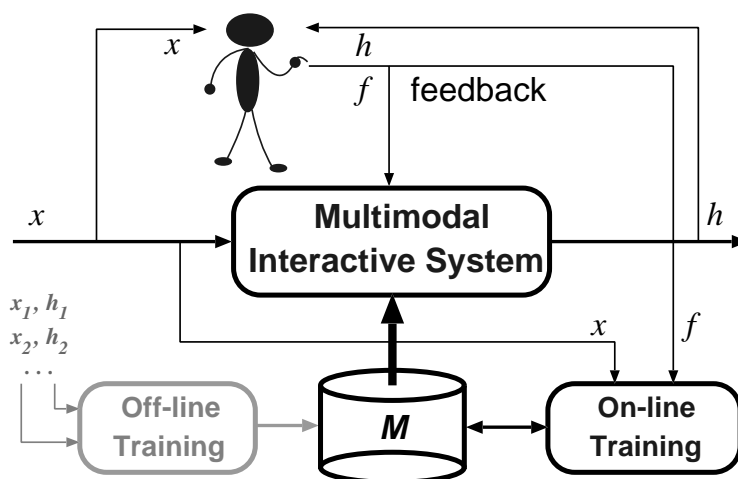
1. *Feedback information* directly derived from the interaction process can be used to significantly *improve system performance*
2. Interaction feedback signals or data are generally of a different nature from those of the original PR problem. This naturally leads to *multimodal operation*, where synergy among different input modalities may help *improving overall system behavior and usability*.
3. Each interaction step generally yields ground-truth data, which can be advantageously used as very valuable *adaptive training data to tune system performance* for the specific task and/or user mode of operation

These challenges and opportunities can be approached through the “*Interactive-Predictive Pattern Recognition*” (IPPR) framework

Interactive-Predictive Pattern Recognition (IPPR): Notation

- \mathcal{X} is the system's input domain; i.e., the domain where input stimuli, observations, signals or data come from
- \mathcal{H} is a possibly infinite set of possible system outputs, actions or hypotheses. $h \in \mathcal{H}$ (also $h(x)$) is a hypothesis which the system derives from a certain input $x \in \mathcal{X}$
- \mathcal{F} is the domain where feedback signals come from. $f(h, x)$, or just $f \in \mathcal{F}$ is a specific feedback signal which the user provides as a response to the system's hypothesis $h(x)$
- \mathcal{M} is any model which the system uses to derive its hypotheses
- True probabilities will be written as $\text{Pr}()$, while $P_{\mathcal{M}}()$ or just $P()$, will denote probabilities computed with some model \mathcal{M}

Diagram of an Interactive-Predictive PR system



In general, \mathcal{M} is initially obtained through a (“batch”) training process from a certain sequence of *training data*, consisting of pairs (x_i, h_i) from the task being considered.

IPPR Opportunity 1: Directly using feedback information

Without varying \mathcal{M} , *human interaction* offers a unique opportunity to improve the quality of system hypotheses, h , using information directly derived from the interaction process.

In traditional PR, for fixed \mathcal{M} a best hypothesis is one which maximizes the posterior probability:

$$\hat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmax}} \operatorname{Pr}(h | x) \approx \underset{h \in \mathcal{H}}{\operatorname{argmax}} P_{\mathcal{M}}(h | x)$$

Now interaction allows adding more *conditions*:

$$\hat{h} \approx \underset{h \in \mathcal{H}}{\operatorname{argmax}} P_{\mathcal{M}}(h | x, f) \quad (1)$$

where f stand for the feedback, interaction-derived informations; e.g., in the form of *partial hypotheses* or *restrictions on \mathcal{H}* .

The richer the feedback signals included in f the greater the opportunity to obtain better \hat{h} . But, also, solving (2) may be increasingly more difficult than solving our familiar (1).

IPPR Opportunity 2: Multimodal Interaction

- How the interaction feedback informations, f , can be obtained?
- Generally, these informations do not naturally belong to the original domain from which the main data, x , come from; i.e., $\mathcal{F} \neq \mathcal{X}$.

For instance, in a vehicle plate recognition system, it is quite unlikely that user's feedback comes in the form of images captured by the plate image capturing camera. Instead, it will arrive in form of keystrokes, mouse gestures, or perhaps spoken utterances.

- Therefore, interaction naturally entails some form of *multimodal operation*, for which interesting mathematical formulations can be developed.
- Using only traditional keyboard & mouse, and/or any other *deterministic* feedback modality, is *not* considered *multimodal*:

Multimodality arises when the *additional* feedback signals are *non-deterministic* and need to be “decoded”, also using PR techniques

The challenge is how to achieve an adequate *modality synergy* which finally allows taking maximum advantage of all the modalities involved

IPPR: Multimodal Interaction

Assume for simplicity that both the input x and the feedback f are unimodal. Interaction leads to the following *modality fusion* problem:

$$\hat{h} = \underset{h}{\operatorname{argmax}} \Pr(h | x, f) = \underset{h}{\operatorname{argmax}} \Pr(x, f | h) \cdot \Pr(h)$$

In many applications it can be assumed independence of x and f given h (e.g., x is an image and f the acoustic signal of a feedback spoken command). This allows for a *naive Bayes* decomposition:

$$\hat{h} \approx \underset{h}{\operatorname{argmax}} P_{\mathcal{M}_X}(x | h) \cdot P_{\mathcal{M}_F}(f | h) \cdot P_{\mathcal{M}_H}(h)$$

- Independent models, \mathcal{M}_X , \mathcal{M}_F and \mathcal{M}_H , can now be estimated separately for the image and speech components and for the prior hypotheses distribution (e.g., the command language), respectively.
- The resulting search problem amounts to the joint optimization of the conditional probability product

IPPR Opportunity 3: Adaptive Training

So far \mathcal{M} has been kept fixed. But now *human interaction* offers another unique opportunity to improve system's behavior by tuning \mathcal{M} .

The f_i obtained through the successive steps of the interaction process can generally be converted into new, fresh training data, useful for *adapting* the system to changing environment.

For many years, *adaptive learning* has been the focus of thorough studies. One outstanding framework is *Bayesian learning*, where *priors* are used to statistically model how to modify \mathcal{M} when new training data is available.

However, practical applications of these theoretical results are generally scarce, mainly because only the *interactive* paradigm offers a natural framework where *adaptive learning* can be used advantageously.

On-line, adaptive training will not be considered here; therefore \mathcal{M} is assumed to be fixed throughout this talk.

Performance evaluation in IPPR

- Perhaps one of the most influential factors for the rapid development of PR in the last few decades, is the *assessment paradigm* based on *labeled training and testing corpora*,

It allows to easily, objectively and automatically test different approaches or algorithms without requiring human intervention.

- In the IPPR framework, a human being is embedded “in the loop”, and system performance has to be gauged mainly in terms of *human effort*.
- Although evaluating system performance in this new scenario apparently requires human work and judgment, by carefully specifying precise goals and ground-truth, the corpus-based assessment paradigm is still applicable in most IPPR tasks.

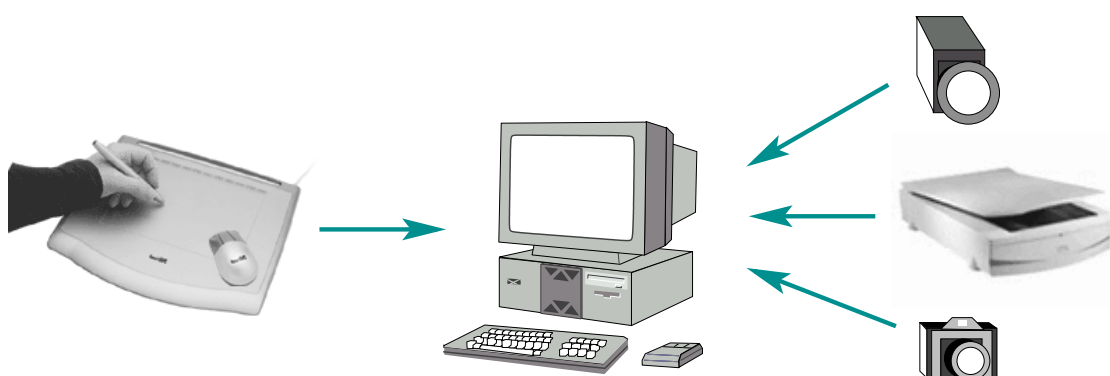
IPPR: Case Studies

- Computer Assisted Transcription of Text Images (CATTI)
- Multimodal Interaction in Text Image Transcription
- Multimodal Interaction in other Document Processing tasks
- Computer Assisted Speech Transcription (CAST)
- Interactive Machine Translation (IMT)
- Multimodal Interaction in MT: Speech-enabled CAT
- Relevance-based Information Retrieval
- Many other possible applications; see:
<http://miprcv.iti.upv.es/>

Handwritten Text Recognition (HTR)

- ▷ Handwritten text and computers:
 - Text images: “off-line” HTR
 - E-pen input (tablet, touchscreen, etc.): “on-line” HTR
 - Block letter handwriting (“OCR”) vs. cursive text “HTR”
- ▷ Interest of off-line HTR:
 - Huge collections of legacy and historical manuscripts
- ▷ Architectures for HTR:
 - Preprocessing, feature extraction and recognition
 - Segmentation-free, HMM approaches to HTR
 - Document layout analysis

Handwritten Text and computers



ON-LINE

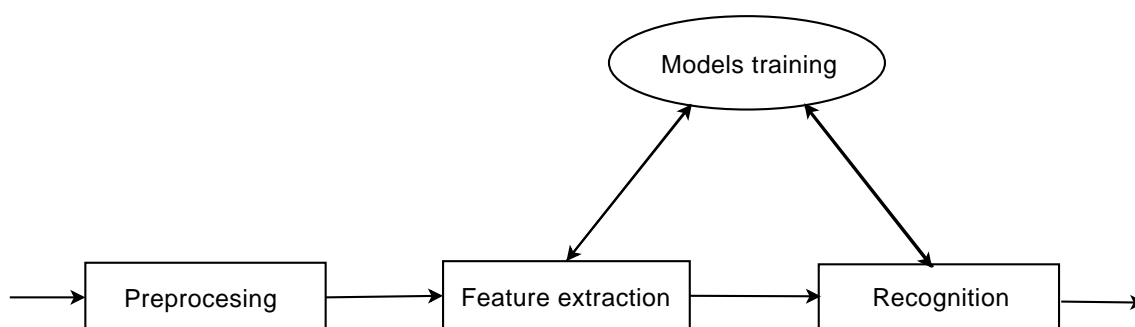
Point sequence representation
(digital pen, tablet, etc.)

OFF-LINE

Bitmap (image) representation
(camera, scanner, video, etc.)

HTR Segmentation-Free Architecture

Classical ASR-like architecture, composed of three main modules:



- *Preprocessing*: noise removal, line detection and geometric normalizations
- *Feature Extraction*: grey-levels or point coordinates and their derivatives
- *Modeling*: morphological Hidden Markov Models + N-gram Language model
- *Recognition*: Viterbi search
- *Modeling & Recognition* are identical for both *on-line* and *off-line* HTR.

Preprocessing and feature extraction for On-Line HTR

- *Repeated points elimination*
- *Noise reduction*: removing high-frequency components by low-pass filtering
- *Size normalization*: scaling the points sequence to a fixed range, preserving the original aspect-ratio of the character sample
- *Writing speed normalization*: redistribution of data points to enforce even spacing between them
- The preprocessed points sequence is transformed into a sequence of (6 or) 7-dimensional real-valued feature vectors: *normalized (x and) y coordinates, normalized x and y first and second time derivatives and curvature*

Preprocessing and Feature Extraction for Off-Line HTR

- **Page or text-block preprocessing:** *background removal, noise reduction, skew correction and text line extraction.*
- **Line preprocessing:** *Slope/slant corrections and non-linear size normalization.*
- **Feature extraction:** A grid is applied to divide the image into $N \times M$ squared cells. Three smoothed features are calculated in each cell:
 - Normalized gray level
 - Horizontal gray level derivative
 - Vertical gray level derivative
- Each text line image is represented as a sequence of $(3 \times M)$ -dimensional feature vectors. This sequence can itself be displayed as a grey-level image:



Statistical framework for HTR

Handwritten Text Recognition: Given a stream of feature vectors representing text (line) image, x , and a set of morphological character, lexicon and language models, \mathcal{M} , obtain a sequence of words (transcription) from which x can be produced with maximum likelihood; that is:

$$\hat{w} = \underset{w}{\operatorname{argmax}} P_{\mathcal{M}}(w | x)$$

Using the Bayes theorem (and dropping \mathcal{M} to simplify notation):

$$\hat{w} = \underset{w}{\operatorname{argmax}} P(x | w) \cdot P(w) \quad (2)$$

Popular models:

- $P(x | w)$: *morphological HMMs*
- $P(w)$: *N-Gram Language Model*

Segmentation-free training: Embedded Baum-Welch *HMM* training from text-image / transcription pairs and standard *N-Gram* training from transcriptions

Segmentation-free integrated decoding: Based on the *Viterbi* algorithm

Interactive-Predictive PR and HTR

- HTR typical state-of-the-art results
 - Practically useful results for small vocabulary and/or syntax restricted text (e.g., recognition of bank check legal amounts, form-constrained text, etc.)
 - Poor results with large vocabularies and/of unrestricted text (e.g., comments in survey questionnaires, historic books, etc.): word accuracy ranges from 80% to 50%, or even worse for difficult texts, noisy images and/or insufficient training data.
- Such accuracy could be enough for tasks such as document *indexing and searching* in some (or many?) applications.
- Too low for *high quality transcription* of most handwritten text images of interest.
 - Human *post-editing can be very expensive* and hardly acceptable by profesional transcribers (paleographers, e.g.).
 - *Computer Assisted, Interactive-Predictive processing* offers promise for *improvements in practical performance and user acceptance*.

Statistical framework for Computer Assisted HTR (CATTI)

Given a feature vector stream, x , a set of morphological, lexicon and language models, \mathcal{M} and a *transcription prefix*, p , validated by the user in the previous step, obtain a proper completion (*suffix*) of p from which x can be produced with maximum likelihood; that is:

$$\hat{s} = \underset{s}{\operatorname{argmax}} P_{\mathcal{M}}(s \mid x, p)$$

With respect to the general IPPR formulation (1), here s and p constitute the *system hypothesis* h and the *human feedback* f .

Using the Bayes theorem (and dropping \mathcal{M} to simplify notation):

$$\hat{s} = \underset{s}{\operatorname{argmax}} P(x \mid p, s) \cdot P(s \mid p)$$

The concatenation of p and s is a full sentence (w), then this is very similar to the main HTR equation (2) if $P(s \mid p)$ is understood as a special “prefix-conditioned language model”.

Bibliography

- I. Bazzi, R. Schwartz, J. Makhoul. "An Omnifont Open-Vocabulary OCR System for English and Arabic". IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) Vol.21 pp.495-504, 1999.
- A. Vinciarelli, S. Bengio, H. Bunke. "Offline Recognition of Unconstrained Handwritten Texts Using HMMs and Statistical Language Models". IEEE Trans. on PAMI, Vol.26, pp.709-720, 2004.
- A. H. Toselli, A. Juan, D. Keysers, J. Gonzalez, I. Salvador, H. Ney, E. Vidal and F. Casacuberta. "Integrated Handwriting Recognition and Interpretation using Finite-State Models". Int. Journal of Pattern Recognition and Artificial Intell., 18(4):519-539, June 2004.
- M. Zimmermann, J.C. Chappelier and H. Bunke. "Off-line Grammar-Based Recognition of handwritten sentences". IEEE Trans. on Pattern Analysis and Machine Intelligence, 28(5):818-821, May 2006.
- E. Vidal, L. Rodriguez, F. Casacuberta and I. García-Varea: "Interactive Pattern Recognition". 4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI-07), Volume 4892 of LNCS, pp.60-71. Brno, Czech Republic, June 2007.