Evaluation Issues in ImageRefiner

Kristen Summers, Luis Hernandez

November 4, 2005

Introduction

- **Goal:** automatically determine how to transform document images, to improve OCR quality
- **Motivation**
  - OCR accuracy depends on image quality
  - Image transforms may improve or degrade images
  - Automatic selection of methods desirable for:
    - Accuracy: improvement of OCR suitability vs. improvement of appearance to human
    - Processing speed: computer-based classification is faster than human-based
- Joint work with Ilya Zavorin, Eugene Borovikov, Yaguang Yang, Mark Turner
- Sponsored by Army Research Lab (ARL) for the Language and Speech Exploitation Resources Advanced Concept Technology Demonstration (ACTD)

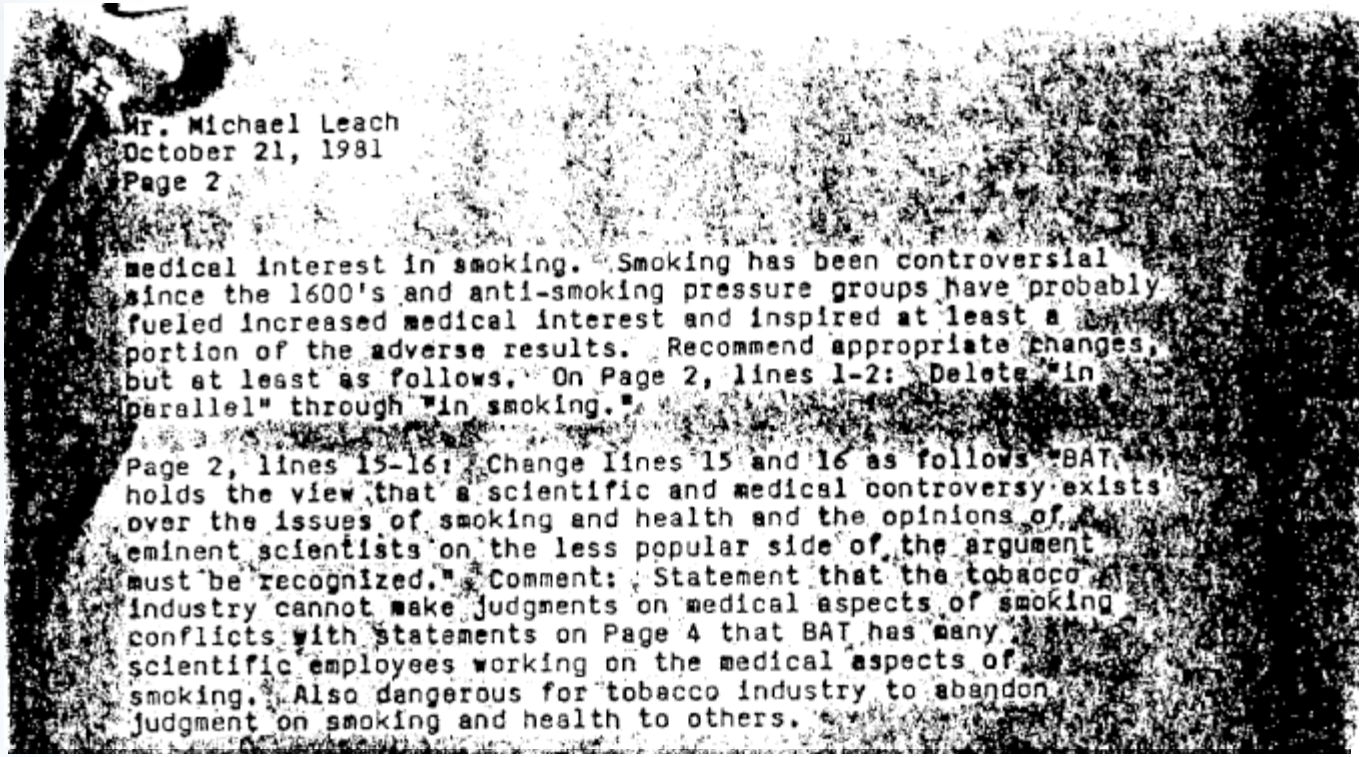
ImageRefiner

- **Flexible framework for document image enhancement:**
  - Learns what image transformation to apply, given the characteristics of a document image
  - Machine-learning (ML) based
  - Handles bitonal (b/w) or grayscale images
  - Able to incorporate any new or existing image characteristic measures, image transformations, ML methods
- **Includes:**
  - 22 image characterization methods
  - 15 image transformations
  - 5 machine learning methods
- **Strategies:**
  - Transformation based features
  - Adaptive image transformations (e.g. via image segmentation)
  - Multi-step image processing (e.g. iterative processing)
- **Tested with Latin, Arabic and Thai scripts**

Document Image Enhancement

- **Many methods (transformations) are available to clean up document images before OCR.**
- **Applying the wrong transformation(s) can result in *lowered* OCR accuracy.**
- **Which transformation(s) should be applied to a specific image?**

Image Transformation:
improvement

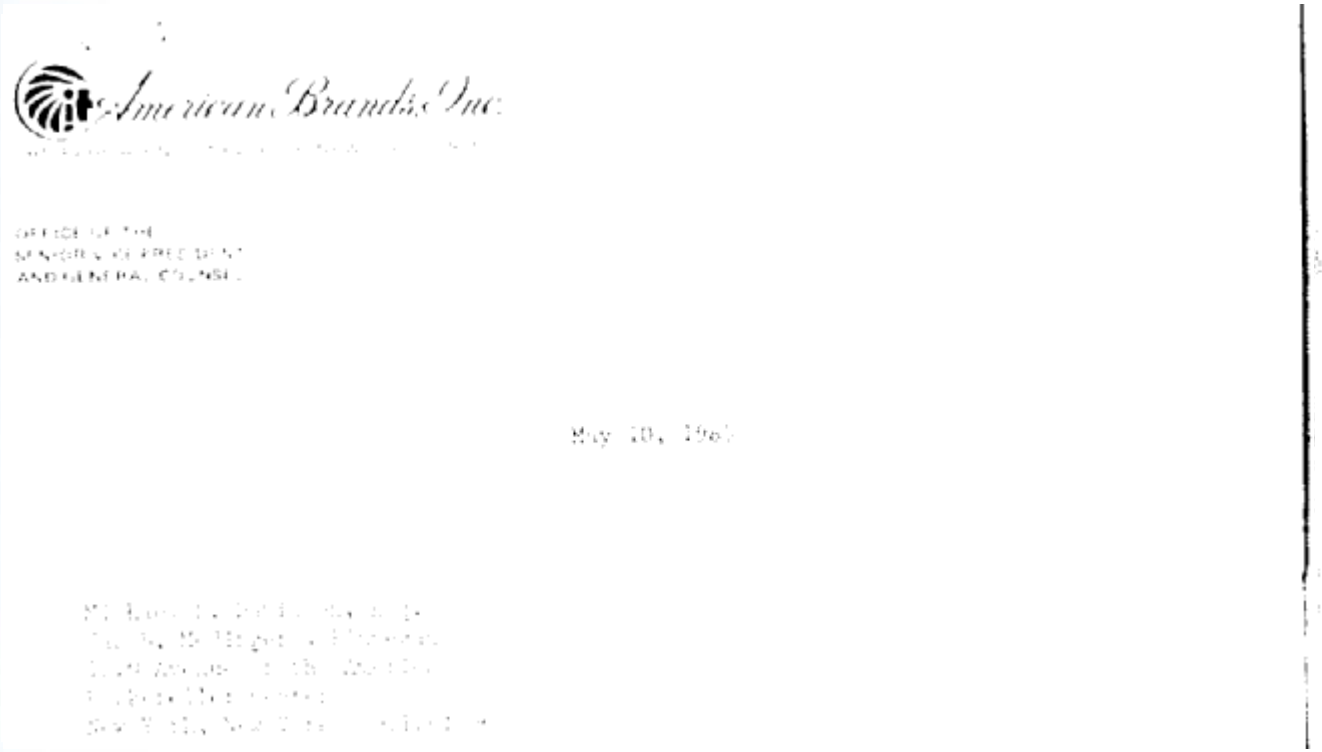


Mr. Michael Leach
October 21, 1981
Page 2

medical interest in smoking. Smoking has been controversial since the 1600's and anti-smoking pressure groups have probably fueled increased medical interest and inspired at least a portion of the adverse results. Recommend appropriate changes, but at least as follows. On Page 2, lines 1-2: Delete "in parallel" through "in smoking."

Page 2, lines 15-16: Change lines 15 and 16 as follows "BAT holds the view that a scientific and medical controversy exists over the issues of smoking and health and the opinions of eminent scientists on the less popular side of the argument must be recognized." Comment: Statement that the tobacco industry cannot make judgments on medical aspects of smoking conflicts with statements on Page 4 that BAT has many scientific employees working on the medical aspects of smoking. Also dangerous for tobacco industry to abandon judgment on smoking and health to others.

Image Transformation: *degradation*



ImageRefiner Approach

- **Treat selection of improvement methods as machine learning problem:** classify images according to preferred transformations for improving OCR output
  - Measure characteristics of images
  - Consider a set of candidate image transformations
  - Use a training set with text ground truth provided
    - Evaluate OCR output of each transformation on each image, to determine their effects
    - Classify each image according to its best transform
  - Use Machine Learning to generalize
  - Optionally segment the image and apply above process to individual segments

Previous Work

- **McNamara, Casey, Smith, and Bradburn (1993)**
  - Characteristics based on statistics of runs of black pixels, connected components
  - Select from 2 transformations (thinning, modified thickening) or keep original
- **Cannon, Hochberg, and Kelly (1999): QUARC**
  - Designed for very noisy, old, typewritten documents (fixed width) in English
  - Measure characteristics designed to reflect types of additive noise in fixed-width English documents
  - Select from 13 transformations, or keep original. 4 transformations use “typewriter grid”

Script and Language

- **Script-Specificity**
  - Expect to need script-specific choices of image features and possible transformations, as well as learning.
  - New script requires basis for understanding its marks and OCR issues, in addition to training corpus.
    Is there a kind of ground truth that could help with this?
- **Language-Specificity**
  - Expect that training may be language-specific, but features and transformations should be applicable across languages in a script.
  - New language requires new training corpus.
- **Currently operates in English, Arabic, Thai**

Areas of Evaluation

- **OCR Accuracy:** string edit distance for training and testing
- **Transformation Selection**
  - Goal: Select the best improvement
  - Effects of selected transformation (improve/degrade, magnitude, comparison with effects of best choice)
- **Segmentation**
  - Evaluate choice of regions
  - Effect of transforming a region on OCR of neighboring regions
- **Generality:** how are results affected by inconsistency between training and application data? By heterogeneity of training data?

OCR Accuracy Evaluation

- **Based on string edit distance (Levenshtein)**
- **Alternative measures**
  - Line-Based: Calculate at line level for alignment, then character level for character matching (as in Esakov, Lopresti & Sandberg 1994; Chen 1993)
  - Character-Based: Apply string edit distance directly to sequence of characters in document
  - Both rely on reading order matching up correctly

OCR Accuracy Scoring

- **Precision, Recall, Combination**
- **Desirable characteristics in a score measure:**
  - Falls in a defined range
  - Single value
  - Reflects effort to correct and/or errors occurring in recognition
  - More script- or language-specific
- **Could be much more precise with ground truth that gives location on page**
  - Expensive to produce on real data
  - Issue: trade-offs with synthetic data

Classification Evaluation

- **Core question:** did we select the right operation for the image?
  - Binary evaluations of transformation choice:
    - Is this the best?
    - Is the change an improvement?
  - Continuous value:
    - Distribution of OCR improvement over transformation set
    - How much an improvement does the change present?

Classification Evaluation, ctnd.

- **Issues**
  - Target is the best transformation; other transformations may help
  - Magnitude matters
    - Impact on OCR
    - Difference between impact on OCR and that of ideal choice
  - Expanding to sequences of transformations
- **Derived from OCR evaluation**
- **Indirect measure:** Improvement for other purposes (e.g., human readability) would require additional ground truth.

Segmentation Evaluation

- **Currently applies in testing only:** segmentation approach is not trained.
- ***Not* conventional segmentation:** seek regions that are consistent in their OCR challenges (noise sources and possibly also font size, style, etc.).
- **Core question:** how good is this segmentation?
  - Infinite number of possible segmentations
  - Quality of segmentation should reflect “purity” of regions and also preference for fewer regions
  - How to account for effects of transforming one region on OCR of neighboring region?

Segmentation Examples: Arabic

مشكلة X سطرين

• زوجة ابني تعاملني معاملة قاسية وتكيل لي أقذع الكلمات والأوصاف امام ابني الوحيد الذي وهبت حياتي له بعد وفاة والده وليس لي مكان آخر اذهب اليه مع العلم انهما يسكنان معي في منزلي؟!

المعذبة - ر. ف البحرين

- هذه المشكلة ليست قاصرة عليك وحدك يا سيدتي العزيزة، فهناك الكثير من الأمهات اللاتي يعانين من تسلط زوجات الأبناء والسنتهم الطويلة التي تحتاج الى القطع حتى يتعلمن كيف يتعاملن مع أمهات ازواجهن، ومن رأى مصائب غيره هانت عليه مصيبته.

وانا لا اقول ذلك للتخفيف عنك ولكن الحقيقة تستوجب توبيخ ابنك بأقذع الأوصاف التي يتسحي قلمي ان يكتبها فهذا الولد ولا اقول الرجل ليس سوى العوبة في يد زوجته لأنه لو كان رجلا بحق ما سمح لها بالتطاول عليك سواء في حضوره او في غيابه ولو أنني اعلم ان لديك القدرة على طردهم من منزلك لطلبت منك ذلك على الفور ولكن للاسف الشديد فقد تأخر الوقت بعد ان اصبحت زوجة ابنك هي المسيطرة على كل الامور وليت ابنك هذا امامي الآن او امام احد من قرائنا حتى نعلمه كيف يحترم امه ويقدرها خاصة بعد ان أضاعت شبابها من اجله، ومع ذلك يا سيدتي العزيزة، اطلب منك الصبر والايمان والتوجه الى الله بالدعاء لهما بالهداية، وعليك ايضا تفادي المواقف التي تعرضك للتعامل المباشر مع هذه المرأة الملعونة ذات اللسان السليط التي ستقع حتما في يوم من الايام فريسة لتقدم السن وستكون ايضا ضعيفة امام زوجة ابنها التي ستأخذ حقك من هذه المرأة المتسلطة وكما تدين تدان.

رياضة الفكر

فلتشحذ العقل ونروّح عن النفس في آن معا

● هل تعلم؟

• ان عقار الإحباط يفيد في المناعة والحصانة من الجلطة الدماغية؟

• ان الرضع الانجليز يعانون من اجازات الأمومة القصيرة فيصابون بالاكتئاب؟

• ان هيلين والتون البريطانية أغنى امرأة بالوراثة في العالم؟

• انه تم اكتشاف بعض حجارة القمر التي يعود تاريخها الى حوالي ٣٩٠٠ مليون سنة في الرياض؟

• ان العلماء اخترعوا قمرا صناعيا يساعد الكلاب التي تجر المكفوفين؟

● ابتسامات

جلس طبيب القرية يراقب برامج التلفزيون بعد يوم من الارهاق والتعب، فسمع قرعا على الباب.

- نعم، من هناك؟

• نعم يا دكتور انا، لقد عضني احد الكلاب.

- الا تعرف ان دوامي ينتهي الساعة السابعة يا رجل.

فأجاب الرجل بسرعة:

حسنا، عالجني هذه المرة يا دكتور وفي المرة القادمة سأخبر الكلب بمواعيدك.

● حكمة العدد

دار السفيـــــــة ولا تمار تكرمـــا
يرجع بأنف راغم مهـــــشـــوم
وكـوامن الحســـاد لا تخـفى وكم
زند يبـــوح بســـره المكتـــوم

● كلمات ودلالات

العالم الساكت:

الخائم الذي يأمر بالخير وينهى عن المنكر في زيادة من الله.

والشاجب الناطق بالخفاء والمعين على الظلم.

أكتب أحسن ما تسمع واحفظ احسن ما تكتب وحدث بأحسن ما تحفظ.

● الغار

١ - للاذكياء فقط:

مرت عصفورة على مجموعة من العصافير على غصن الشجرة فقالت لهم: السلام عليكم يا مائة، فردت إحداهن وقالت: نحن لسنا بالمائة ولكن مثلنا ونصفنا وربعنا وانت تكون مائة، فكم كان عدد العصافير الموجودة فعلا على غصن الشجرة؟

٢ - فزورة: ما هو الجرح الذي لا ينزف دما؟

تراجع الحانظين.. الا... هنا!

يتوارى المحافظون في الدول التي تنشد التقدم إلى الأمام والنظر إلى المستقبل بعيون مفتوحة عدا منطقتنا العربية!

في إيران، حرب الشعب الإيراني حكم المحافظين أكثر من (١٧) عاما، فماذا جنوا؟!

في بريطانيا العظمى جربوا تشدد المحافظين ومنجهيتهم، فماذا جنوا؟

الفرق بدا واضحا مع وصول خاتمي في إيران، علاقات خارجية جيدة مع دول مجلس التعاون، علاقات أفضل مع من كان يسمى الشيطان الأكبر، انفتاح في الشارع الإيراني أثر على كل مؤيدي إيران في العالم العربي وفي... الكويت!

في بريطانيا، لا اضرابات، لا تفجيرات داخلية، لا عداءات، سعي نحو الوحدة الأوروبية بخطى حثيثة، ولعل أنجح دليل، فوز حزب العمال فوزاً تاريخياً، فللأول مرة يفوز في دورتين متتاليتين وبالنسبة نفسها تقريبا.

ونلتفت الى أحوالنا، فماذا نرى؟

انفلاقاً بصوة رهيبة وغير منطقية، انفلاقاً في الفكر وفي الرؤية وفي الخطط، هذا إن كانت هناك حقا خطط!!

يتقدمون نحو المستقبل بخطوات واثقة، ونتراجع الى الوراء بخطوات واثقة؟!

لا اقول ذلك تشاؤما، ولكن انظروا الى ايران وبريطانيا قبل (٢٠) سنة وقارنوهما بحالتهما الآن واحكموا بأنفسكم!

هل لديهما شعب يفكر ويخطط أفضل منا؟ام انهم ذبذوا تفكير القبيلة، وعدنا اليه راكضين لاهثين وكأننا افتقدناه بعنف!!

يا أصحاب العقول المفكرة؟ يا من تعشقون أوطانكم بصدق، فكروا، قليلا وقارنوا واعملوا، فلم نعد نتحمل المزيد من التراجع!

سوالف

عهود الجريد - السعودية

رامية السهام

كانت تهوى رمي السهام
تصيب او تخيب ولا تعرف ماذا تريد
تصوب اجمل النظرات فيغرق في بحرها العشرات
يطلبون الود ويريدون إسداء الخدمات
كانت لا تطلب منهم شيئا
فما تريده هو ان يصيب السهم ويصل للمكان الصحيح
تريد ان تراهم يتساقطون كأوراق الخريف
واحدا تلو الآخر يستجدون العطف ويطلبون المزيد
لكنها لا تعرف ماذا تريد
لا تلبي طلبهم ولا تكون كما يريدون
فما تريده هو ان تكون الجميلة المحاطة بكل هؤلاء المعجبين
يسقط الواحد منهم صريعا لجمالها ولا يلبث ان يتركها
تتساءل بعد ان تهمل لماذا يتركونني؟؟
وفي احدى الامسيات بينما كانت تصوب السهام عشوائيا
امتدت يد وأمسكت بأحد السهام... لم تشاهدها
وسمعت خطوات تقترب منها
وصوت قوي يقول لها يبدو ان هذا السهم يخصك سقط منك
ازاحت بوجهها عنه ولم ترد فقال لها ما كان ينبغي ان تفعلي ذلك فتجاهلته
فقال سيدتي ليس هكذا تصوب السهام
السهم اذا لم يكن مصوباً بعناية فسوف يسقط في منتصف الطريق
وهو قد سقط في يدي قبل ان يصل لهدفه سيدتي
انصحك بترك رمي السهام وتعلم رمي الرماح
فهي اقوى واسرع وتصيب شخصاً وليس عدة اشخاص
طأطأت رأسها ولم ترد فكانت لا تعرف ماذا تريد.

Evaluating Generality

- **Question:** How are results affected by inconsistency between training and application data?
  Requires measuring similarity/difference between these data sets
  - Distribution of measured features? What if the relevant similarity/difference is in what the system does not measure?
  - Human judgments of type of noise?
  - Known sources of noise?
  - Source (type of document, age, paper quality ...)
  - Language/script
- **Question:** How are results affected by heterogeneity of training data?
  Requires same kind of measurement
- **Wish list:** ground truth corpus annotated with noise sources, document sources, human judgments of image "quality."

Acknowledgements

The research reported in this document/presentation was performed in connection with Contract No. DAAD19-03-C-0059 with the U.S. Army Research Laboratory. The views and conclusions contained in this document/presentation are those of the authors and should not be interpreted as presenting the official policies or position, whether expressed or implied, of the U.S. Army Research Laboratory or the U.S. Government unless so designated by other authorized documents. Citation of manufacturers or trade names does not constitute an official endorsement or approval of the use thereof. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.