Data Quality In The Cultural Heritage Sector: From An Image Processing Perspective

Tan Lu

Department of Mathematics and Data Science
Vrije Universiteit Brussel (VUB)

Department of Digitization
Royal Library of Belgium (KBR)
The ADOCHS Project

KBR
images

VUB
VRIJE UNIVERSITEIT BRUSSEL

ADOCHS

arch
CEGESOMA

metadata

ULB
UNIVERSITÉ LIBRE DE BRUXELLES
Digitization in the Cultural Heritage Sector

- Google Books: over 40 million books
- Europeana Newspapers: aggregating 18 million historic newspaper pages and converting 10 million newspaper pages to full text
- Royal Library of Belgium (KBR): 4500 medieval codices and about one million prints and drawings

Challenges and Opportunities

- Image Quality Assessment (IQA)
- Image Understanding
  - Document Image Segmentation (DIS)
  - Damage Recognition (DR)
A Unified Approach to Image Quality Assessment

- A unified model to process natural and document images simultaneously
- Content-aware such that different types of quality information is provided according to different types of input images

DMOS

OCR accuracy
Document Image Quality Assessment based on Transfer Learning

- **AlexNet**
  - Conv: 11x11, 96, /4
  - Max pool: 3x3, /2
  - 5x5, 256
  - 3x3, /2
  - 3x3, 384
  - 3x3, 256
  - 3x3, /2
  - FC 4096

- **Feature Extraction**
  - Document images
  - Knowledge base
  - Image features
  - OCR accuracy regression
  - Fine-tuning

- **Document Image Quality Assessment (DIQA)**
  - Document images
  - Task-specific segment
  - Regressor
  - Quality score
The knowledge learned on natural image processing can be effectively exploited for the OCR accuracy prediction of document images.

### Cross-Domain Homogeneity between Natural and Document Images

The knowledge learned on natural image processing can be effectively exploited for the OCR accuracy prediction of document images.
Unified Image Quality Assessment

- **Cross-Domain Homogeneity between Natural and Document Images**
  - Possible to process natural and document images simultaneously within one quality assessment model.
  - Balanced performance on these two types of images can be obtained with the UIQA model.
  - The process of learning a common representation is mixed with that of regressing the common representation towards respective quality scores – difficult to investigate and develop.
Unified Image Quality Assessment based on Contractive GAN

- **Cross-Domain Homogeneity between Natural and Document Images**
  - Learning a common representation (i.e., a generalization) of natural and document images in a latent domain
  - The process of generalization is separated from that of regression
  - The quality assessor operates as if it is processing a single type of images
Unified Image Quality Assessment based on Contractive GAN

- **Main Objective:**

\[
\min_{R,f} \max_D \left\{ \mathbb{E}_{x \sim p_B} \{ \log D(f(x)) \} + \mathbb{E}_{x \sim p_A} \{ \log [1 - D(f(x))] \} \right\} \\
+ \mathbb{E}_{x \sim p_A} \{|R(f(x)) - t_A|\} + \mathbb{E}_{x \sim p_B} \{|R(f(x)) - t_B|\}
\]

- **Quality Discriminator:**

\[
\begin{align*}
\max_{D_A} & \left\{ \mathbb{E}_{x \sim p_A} \{ \log [1 - D_A(f(x)) - \sigma(x)] \} \right\} \\
\max_{D_B} & \left\{ \mathbb{E}_{x \sim p_B} \{ \log [1 - D_B(f(x)) - \sigma(x)] \} \right\}
\end{align*}
\]

where:

\[
\sigma(x) = \begin{cases} 
1, & \text{if } |R(x) - t| \leq \epsilon \\
0, & \text{otherwise}
\end{cases}
\]
Unified Image Quality Assessment based on Contractive GAN

- Qualitative evaluation: visualization of the operation of the C-GAN model
Unified Image Quality Assessment based on Transfer Learning

LIVE + SOC

CSIQ + SOC
Unified Image Quality Assessment based on Contractive GAN

- Comparing to content-specific IQA and DIQA models

<table>
<thead>
<tr>
<th>IQA Models</th>
<th>CSIQ</th>
<th>SOC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PLCC</td>
<td>SRCC</td>
</tr>
<tr>
<td>BLIINDS2</td>
<td>-</td>
<td>0.880</td>
</tr>
<tr>
<td>DIQA</td>
<td>-</td>
<td>0.870</td>
</tr>
<tr>
<td>CORNIA</td>
<td>-</td>
<td>0.854</td>
</tr>
<tr>
<td>NRSL</td>
<td>-</td>
<td>0.896</td>
</tr>
<tr>
<td>CNN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNN</td>
<td>N.A.</td>
<td></td>
</tr>
<tr>
<td>LDA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sparse Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed method</td>
<td>0.92</td>
<td>0.89</td>
</tr>
</tbody>
</table>

- Cross-dataset evaluation of the proposed UIQA model on the natural scene images

<table>
<thead>
<tr>
<th>IQA Models</th>
<th>LIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PLCC</td>
</tr>
<tr>
<td>BLIINDS2</td>
<td>-</td>
</tr>
<tr>
<td>DIQA</td>
<td>-</td>
</tr>
<tr>
<td>CORNIA</td>
<td>-</td>
</tr>
<tr>
<td>NRSL</td>
<td>-</td>
</tr>
<tr>
<td>Proposed method</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Gestalt Principles and Text Homogeneity

- Proximity
- Similarity
- Symmetry
- Conceptualization

**Text homogeneity** is the homogeneous pattern displayed in text regions, which consists of **proximately** and **symmetrically** arranged units with **similar morphological** and **texture** features.
Probabilistic Local Text Homogeneity – A Neighborhood Graph

- **Description of local text homogeneity on** $G(V,E)$
  
  If we take a one-step walk from a Gestalt $CC_i$ by following an arbitrary (symmetry) direction, and arrives at another Gestalt, say $CC_j$, the probability that $CC_j$ is located within a short (proximity) distance and resembles (similarity) $CC_i$ is higher when $CC_i$ is a text component (e.g. a letter from a paragraph).

  • probabilistic weighting $w_{ij} = P(S_{ij} = s^+_lij)$

  $S_{ij} = \begin{cases} 
  s^+_{ij}, & \text{if } CC_i \text{ and } CC_j \text{ are homogeneous,} \\
  s^-_{ij}, & \text{if } CC_i \text{ and } CC_j \text{ are heterogeneous;}
  \end{cases}$
Text Homogeneity Revisit
- Text Homogeneity pattern
- Neighborhood transition
Propagation of Wavelet Approximation

- **Wavelet Propagation**

  propagation of wavelet approximation (PWA) and propagation of cone-of-influence wavelet approximation (PCWA).

  - PWA
    \[ \alpha_{n \rightarrow k, l} \triangleq \log_2 \left( \frac{1}{k-n} \sum_{j=n}^{k-1} \frac{|w_{j,l}^i|}{|w_{j,l}^i|} \right) \]
  
  - PCWA
    \[ \beta_{n \rightarrow k, l} \triangleq \log_2 \left( \frac{1}{k-n} \sum_{j=n}^{k-1} \frac{|I_{j+1,l}|}{|I_{j,l}|} \right), \]
    \[ I_{j,l} \triangleq \sum_{m \in C(j,l)} |w_{j,l}^m| \]
Bayesian Distortion Recognition
Bayesian Distortion Recognition
Conclusion

• Qualitative evaluation: visualization of the operation of the C-GAN model
Thank you for your attention!

Q & A