デジタル化画像の活用事例

深層学習手法を用いた古書籍画像の可読性向上

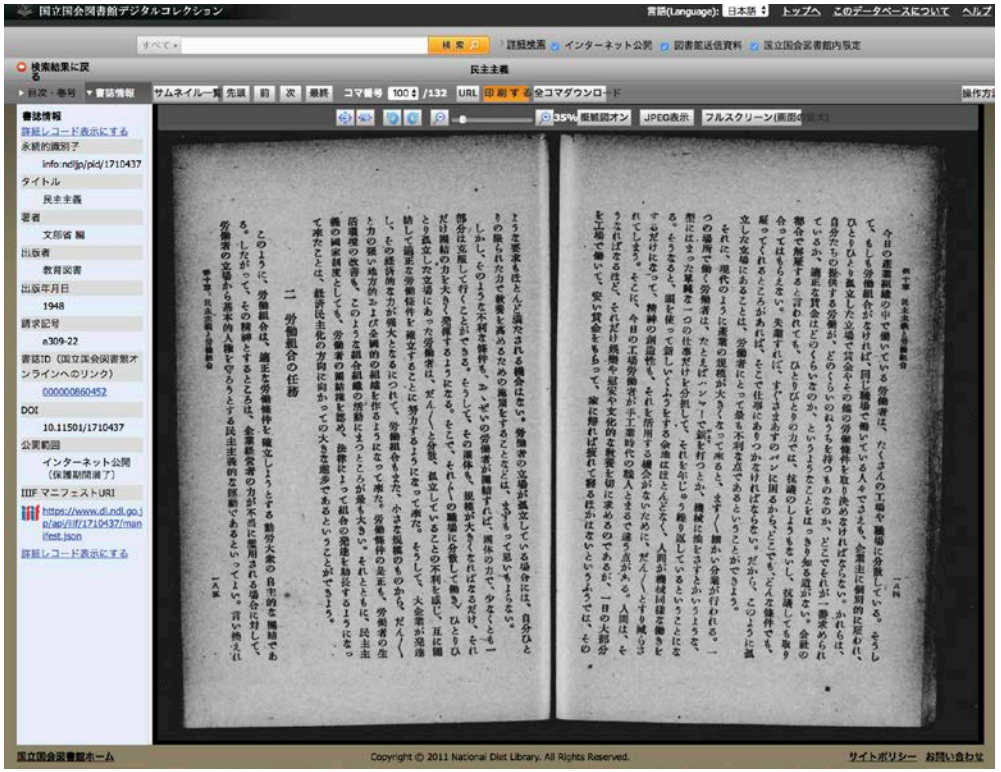
国立情報学研究所 特任准教授
国立国会図書館電子情報部次世代システム開発研究室 委嘱研究員
特定非営利活動法人連想出版 理事

阿辺川武

2018年10月30日

この研究のきっかけ (1/3)

- NDLデジタルコレクション
  - 著作権保護期間が満了したが切れた古書籍が多数公開

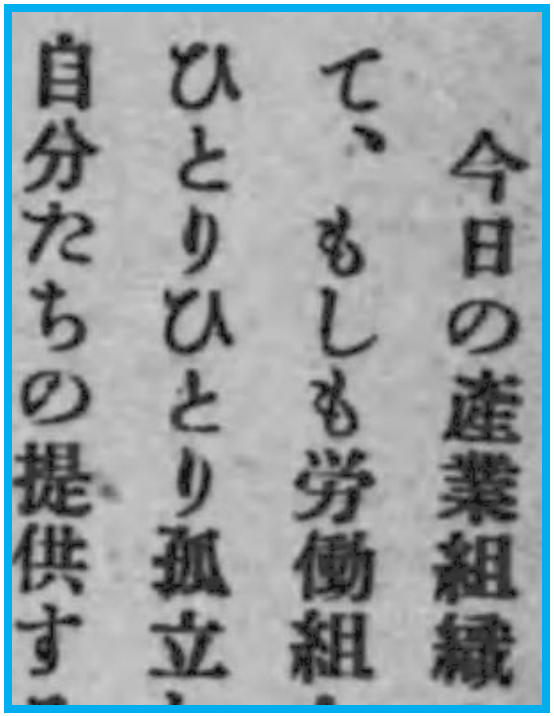


http://dl.ndl.go.jp/info:ndljp/pid/1710437

写真撮影の後、
マイクロフィルムで
保存されているため
画像が美しくなく
可読性が悪い

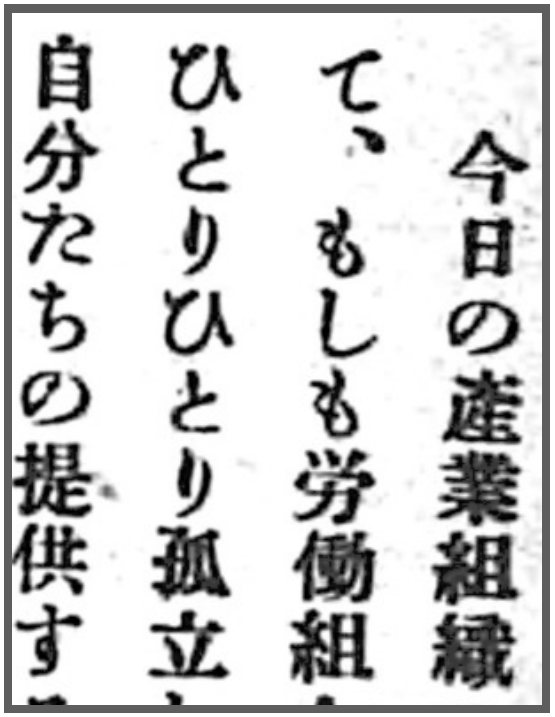
この研究のきっかけ (2/3)

- きれいな画像で書籍を読みたい
  - コントラストや明るさを調整して背景を白色化



元文書画像

各種レベル補正



補正後画像

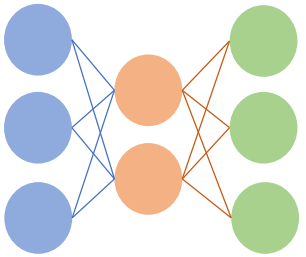
背景にゴミが残ったり、文字のカスレがそのまま

この研究のきっかけ (3/3)

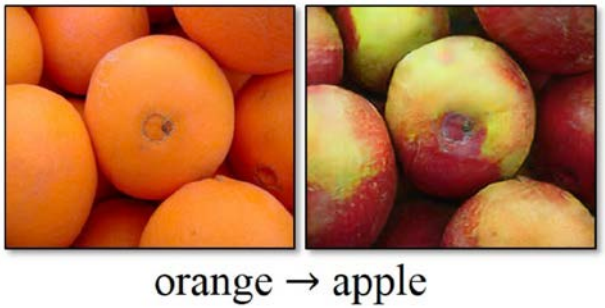
NDL所蔵古書POD (プリントオンデマンド)

https://www.amazon.co.jp/dp/4802005970/

CycleGAN

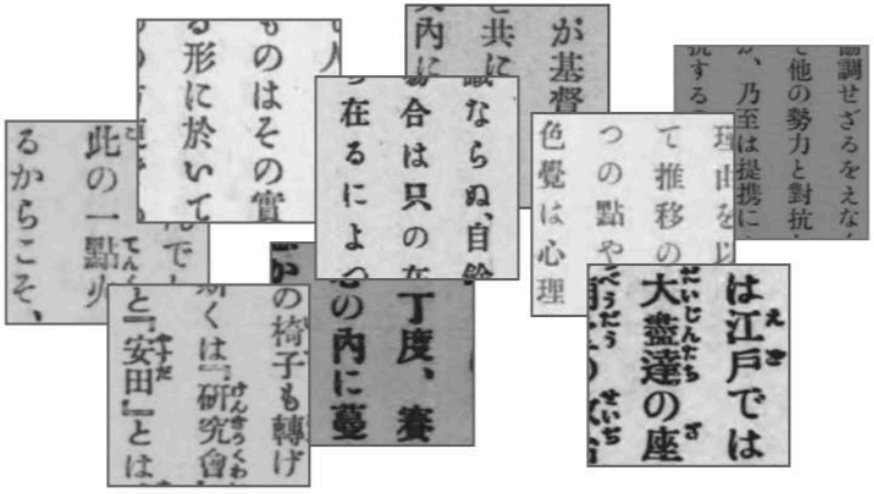


- ディープラーニングを用いた画像変換手法の１つ (Yan, 2017) https://github.com/junyanz/CycleGAN
  - 馬をシマウマに
  - 夏景色を冬景色に
  - 写真をモネ風、セザンヌ風、ルノワール風に
  - オレンジをりんごに

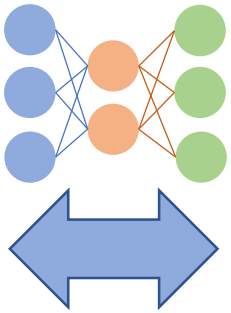


フォントスタイル変換

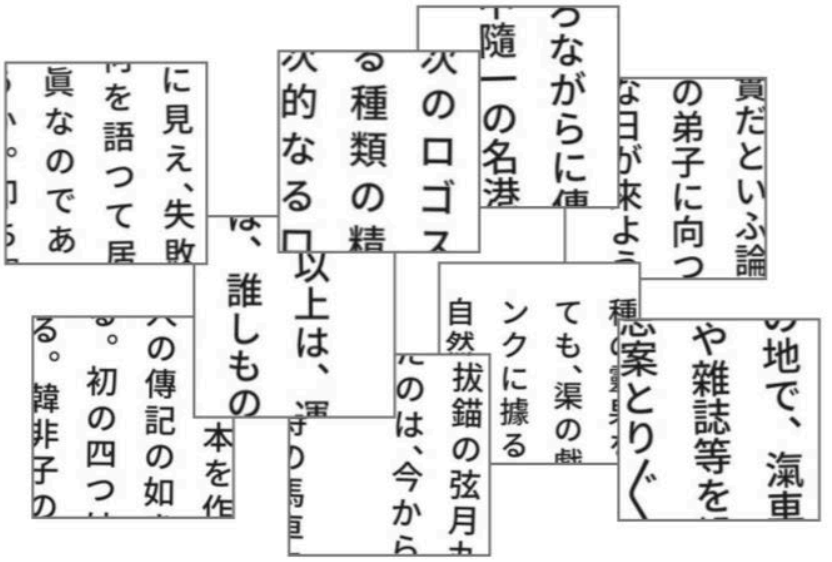
- CycleGANを古書籍の文書画像に応用できないか？
- ２つの画像集合を用意し、対応関係を機械学習させる
  - 古書籍の文書画像集合：画像から切り出す
  - きれいな文字の画像集合(参照画像集合)：組版ソフトで作成



古文書画像集合

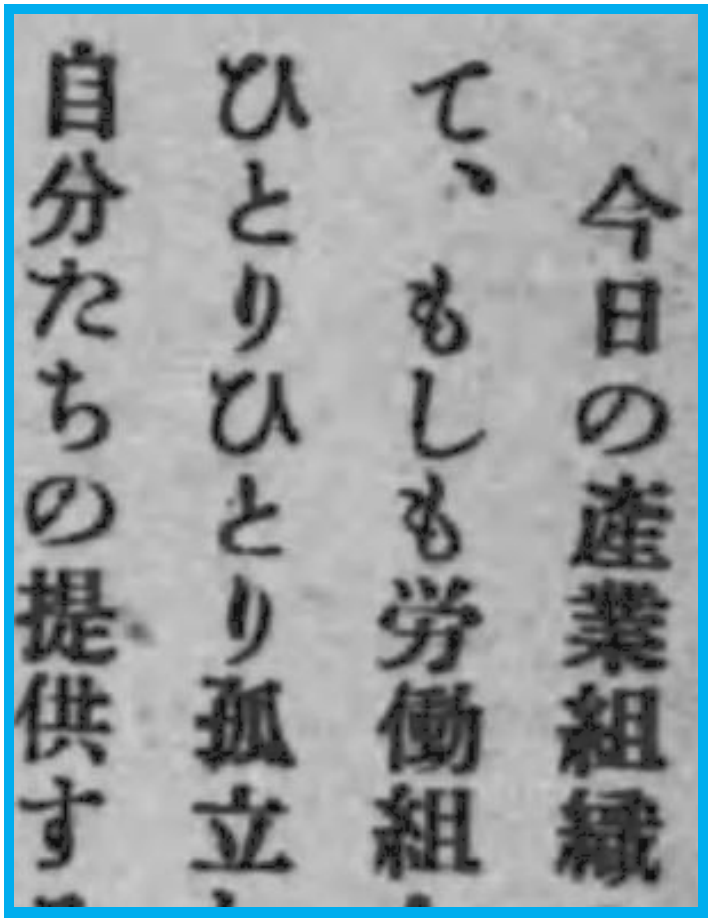


CycleGANで
対応関係を
学習する



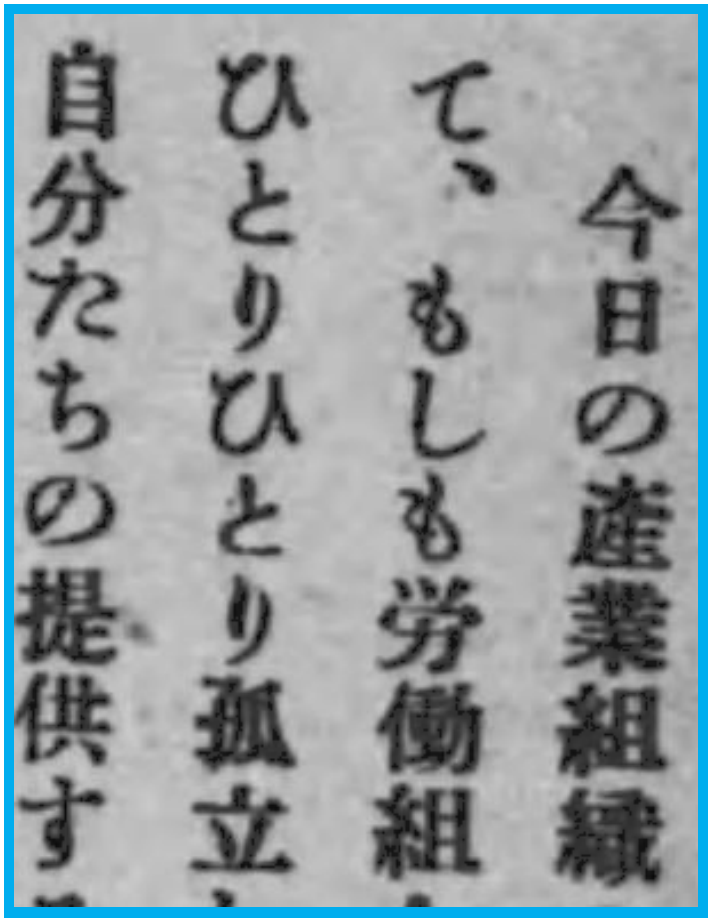
参照画像集合

結果（参照画像に明朝体）



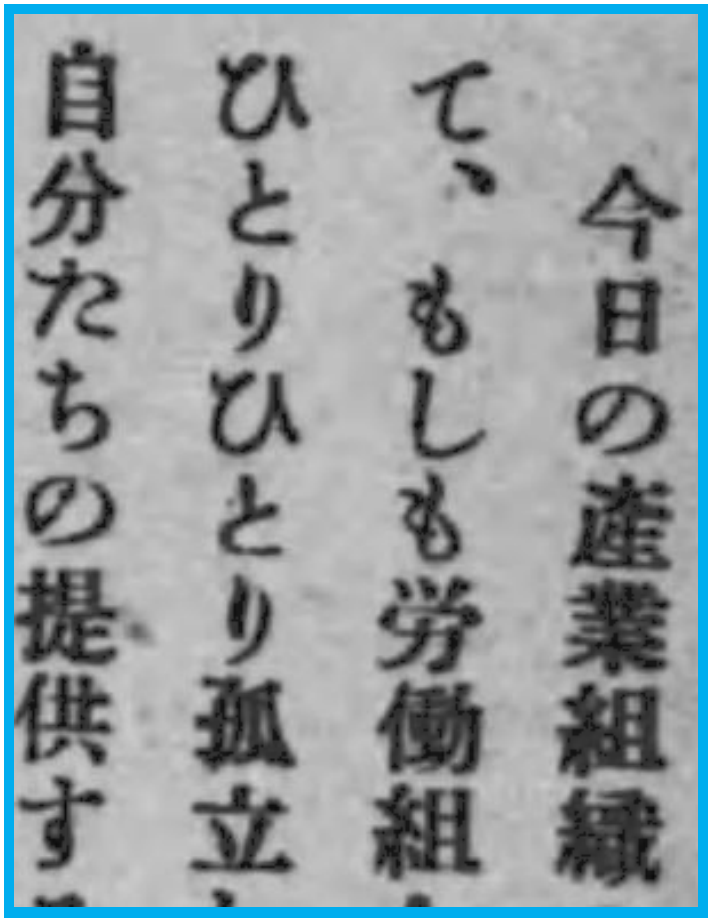
参照画像集合：源ノ明朝Heavy

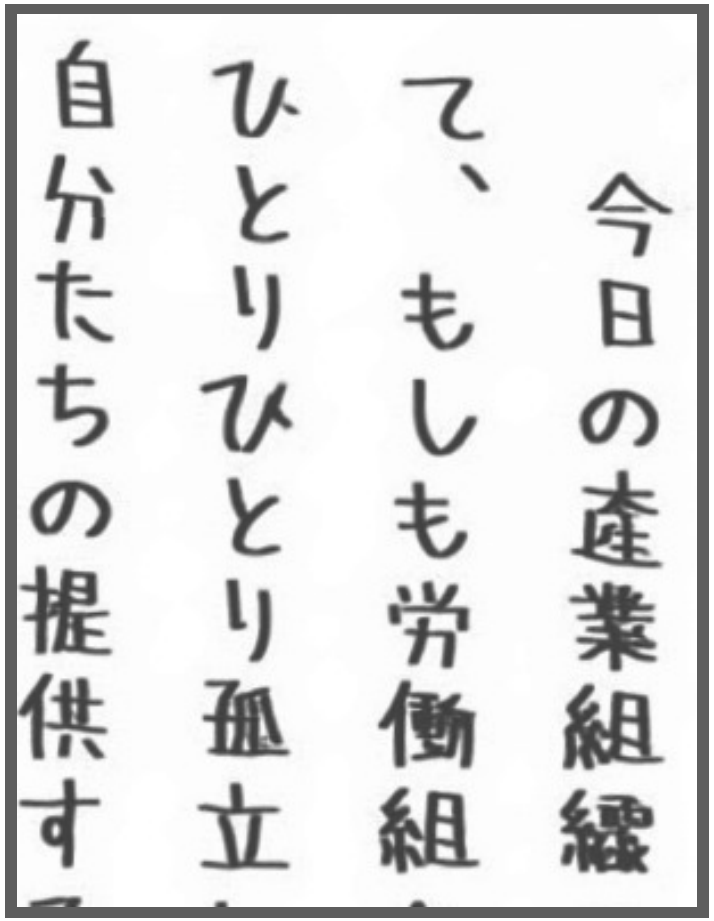
結果（参照画像にゴシック体）



参照画像集合：源ノ角ゴシックMedium

結果（参照画像に手書きフォント）





参照画像集合：手書きフォント「無心」http://modi.jpn.org/font_mushin.php

今日の産業組織の中で働いている労働者は、たくさんの工場や職場に分散している。そうして、もしも労働組合がなければ、同じ職場で働いている人々でさえも、企業主に個別的に雇われ、ひとりひとり孤立した立場で賃金やその他の労働條件を取り決めなければならない。かれらは、自分たちの提供する労働が、どのくらいのねうちを持つものなのか、どこでそれが一番求められているか、適正な賃金はどのくらいなのか、というようなことをはっきり知る道がない。会社の都合で解雇すると言われても、ひとりひとりの力では、抗議のしようもないし、抗議しても取り合ってはもらえない。失業すれば、すぐさまあすのパンに困るから、どこでも、どんな條件でも、雇ってくれるところがあれば、そこで仕事にありつかなければならない。だから、このように孤立した立場にあることは、労働者にとって最も不利な点であるということができよう。

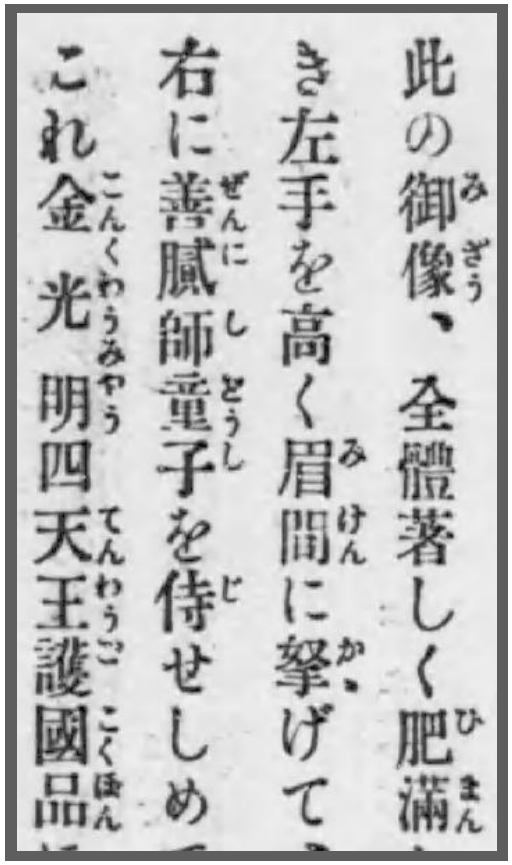
それに、現代のように産業の規模が大きくなって來ると、ます〳〵細かい分業が行われる。一つの場所で働く労働者は、たとえばハンマーで鋲を打つとか、機械に油をさすとかいうような、型にはまった單純な一つの仕事だけを分担して、それを年じゅう繰り返しているということになる。そうなると、頭を使って新しいくふうをする余地はほとんどなく、人間が機械同様な働きをするだけになって、精神の創造性も、それを活用する機会がないために、だん〳〵とすり減らされてしまう。そこに、今日の工場労働者が手工業時代の職人とまるで違う点がある。人間は、そうなればなるほど、それだけ娯樂や慰安や文化的な教養を切に求めるのであるが、一日の大部分を工場で働いて、安い賃金をもらって、家に帰れば疲れて寝るほかはないというふうでは、その

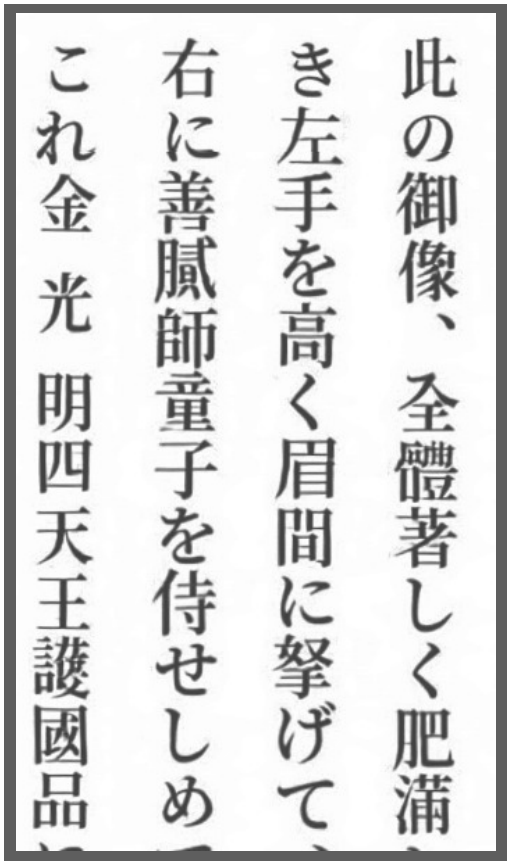
今日の産業組織の中で働いている労働者は、たくさんの工場や職場に分散している。そうして、もしも労働組合がなければ、同じ職場で働いている人々でさえも、企業主に個別的に雇われ、ひとりひとり孤立した立場で賃金やその他の労働條件を取り決めなければならない。かれらは、自分たちの提供する労働が、どのくらいのねうちを持つものなのか、どこでそれが一番求められているか、適正な賃金はどのくらいなのか、というようなことをはっきり知る道がない。会社の都合で解雇すると言われても、ひとりひとりの力では、抗議のしようもないし、抗議しても取り合ってはもらえない。失業すれば、すぐさまあすのパンに困るから、どこでも、どんな條件でも、雇ってくれるところがあれば、そこで仕事にありつかなければならない。だから、このように孤立した立場にあることは、労働者にとって最も不利な点であるということができよう。

それに、現代のように産業の規模が大きくなって來ると、ます〳〵細かい分業が行われる。一つの場所で働く労働者は、たとえばハンマーで鋲を打つとか、機械に油をさすとかいうような、型にはまった單純な一つの仕事だけを分担して、それを年じゅう繰り返しているということになる。そうなると、頭を使って新しいくふうをする余地はほとんどなく、人間が機械同様な働きをするだけになって、精神の創造性も、それを活用する機会がないために、だん〳〵とすり減らされてしまう。そこに、今日の工場労働者が手工業時代の職人とまるで違う点がある。人間は、そうなればなるほど、それだけ娯樂や慰安や文化的な教養を切に求めるのであるが、一日の大部分を工場で働いて、安い賃金をもらって、家に帰れば疲れて寝るほかはないというふうでは、その

ルビの削除

- ルビが全く存在しない参照画像集合を用いるとルビが消える





OCRの精度向上に寄与する

親と頼んで隠れつゝ、一心讀誦、毘沙門經、咽も裂けよと叫ぶ折から、ア、ラ不思議や、大木自らにどうと仆れて、三つの妖鬼は其の下敷、板の如くに潰されて、これは又見るも無慚、敢へなき最期を遂げてしまつた——こゝから最急行。其の翌朝伊勢人登山して、峯延と師檀を約す、鞍馬山寺の隆運、こゝに源を發するといふ。

閑話休題、鞍馬安置の毘沙門天は、伊勢人の手に勸請した靈像。——尤も藤原前期との説もある——都南門樓の兜跋毘沙門に對して、鞍馬は正しく帝京の坤位、王城鎮守の靈山だ。こゝに北方天の靈體を安置して、千代萬代の末永く、皇基を守護し、都城を衛護せしめむとは、忠臣伊勢人の、夢寐なほ忘れ得ぬ宿題であつた。さればこそ此の御像、全體著しく肥滿して強剛に、分けても顔面の緊張物々しく、寶塔を捧ぐべき左手を高く眉間に拏げて、四六時中皇都俯瞰の相丰生けるが如く、且又左に吉祥天、右に善膩師童子を侍せしめて、天王の威光更に一段を加ふ、而して兩脇侍立の形式は、これ金光明四天王護國品に説けるところ、伊勢人の意を用ふる、まことに厚く深きものがある。

親と頼んで隠れつゝ、一心讀誦、毘沙門經、咽も裂けよと叫ぶ折から、ア、ラ不思議や、大木自らにどうと仆れて、三つの妖鬼は其の下敷、板の如くに潰されて、これは又見るも無慚、敢へなき最期を遂げてしまつた——こゝから最急行。其の翌朝伊勢人登山して、峯延と師檀を約す、鞍馬山寺の隆運、こゝに源を發するといふ。

閑話休題、鞍馬安置の毘沙門天は、伊勢人の手に勸請した靈像。——尤も藤原前期との説もある——都南門樓の兜跋毘沙門に對して、鞍馬は正しく帝京の坤位、王城鎮守の靈山だ。こゝに北方天の靈體を安置して、千代萬代の末永く、皇基を守護し、都城を衛護せしめむとは、忠臣伊勢人の、夢寐なほ忘れ得ぬ宿題であつた。さればこそ此の御像、全體著しく肥滿して強剛に、分けても顔面の緊張物々しく、寶塔を捧ぐべき左手を高く眉間に拏げて、四六時中皇都俯瞰の相丰生けるが如く、且又左に吉祥天、右に善膩師童子を侍せしめて、天王の威光更に一段を加ふ、而して兩脇侍立の形式は、これ金光明四天王護國品に説けるところ、伊勢人の意を用ふる、まことに厚く深きものがある。

OCRの精度

フォントスタイル変換前と変換後でOCR精度を比較

	1710437-100		983386-156（ルビあり）	
	無修正	変換後	無修正	変換後
OCR A	0.949	0.969	0.465	0.922
OCR B	0.898	0.961	0.860	**0.938**
OCR C	**0.982**	0.981	0.903	0.925

今後の予定

- 古書籍に特化したOCRの作成
  - フォントスタイル変換手法を使って、
    全ての文字の古文書画像を自動生成する