

Progress Report - Analysis on Relationship between Document Difficulty Score and Visual Features

10/31/2019

Mike Pack

1. Introduction

As the 3rd iteration of Project 4 (Document Image Quality Assessment), in this experiment, we aim to reveal a relationship between a *difficulty score* and *visual features*. One of the expected beneficial outcomes from this experiment is to build a *difficulty score prediction model* that could give the Library of Congress a capability of controlling and managing challenging document images, especially for tasks that involve human perception such as transcription.

2. Dataset

The dataset used in this experiment is a subset of document images (15,592 images) collected from the Library of Congress archive along with corresponding difficulty score.

The *difficulty score* of a document image—collected by Library of Congress—is the number of trials on transcription by human volunteers carried out on the document image. This is based on the intuition behind that poorly readable document images—due to various visual artifacts (e.g., noise or ugly handwriting)—would have a higher number of resubmissions by multiple transcription volunteers. Note here that the scores have not been verified by the human experts.

3. Experiment 1: Visual Inspection

Before directly diving into the numerical correlation analysis between visual features and difficulty scores, we first visually inspect a handful of images to investigate to what extent the difficulty scores reflect the human perception of difficulty, particularly for transcription-like tasks. Particularly, we focus on finding any notable visual cues that makes distinctive differences between different difficulty scores. To this end, we sampled two images (i.e., acceptable and not acceptable for human perception of difficulty to the difficulty score) from two different types (i.e., *handwritten* and *typed* document images) with six different difficulty scores as shown in Table 1. From the inspection, we have the following two observations:





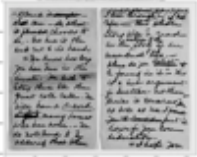



Observation 1. The same *visual feature* deemed relevant to the *difficulty score* in *typed* documents is not deemed relevant to the difficulty score in *handwritten* documents.

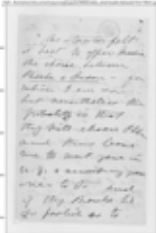



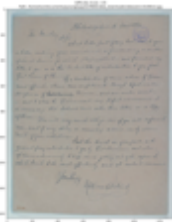



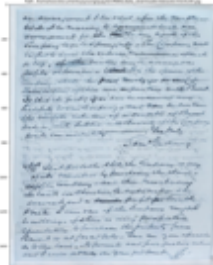



Observation 2. It would be challenging to identify correlation between a simple standalone *visual feature* (e.g., number of characters or low contrast) and a *difficulty score*.

For a better comprehensive understanding of the above observations, consider the following examples. About the first observation, for the *typed* documents, note that the visually perceived amount of contents/characters in an image (i.e., density) only slightly corresponds to the difficulty score (see the rightmost column in Table 1), meanwhile it is not the case for the *handwritten* documents (see the third column in Table 1.)

About the second observation, note that it is hard to find notable visual similarity between *unacceptable* images and *acceptable* images within the same difficulty score. For example, document images with a difficulty score of 9, *unacceptable handwritten* image show poor image quality with lower contrast and higher density compared to the *acceptable handwritten* image, as shown in Table 1. This is also the case in the *typed* document images.

Table 1. Document samples for different difficulty scores. The empty cells meaning no more images exist for the corresponding difficulty score.

| Difficulty Score | Not Acceptable | | Acceptable | |
|------------------|---|---|--|---|
| Type | Handwritten | Typed | Handwritten | Typed |
| 9 |  <p>Low contrast + comparably large amounts of contents, but ONLY 9</p> |  <p>Complicated layout + decent amounts of contents, but ONLY 9</p> |  |  |
| 20 |  <p>Range-effect, but comparably small amounts of contents, but 20</p> |  <p>Amounts of characters looks similar to 9</p> |  |  |

| | | | | |
|---|---|---|--|---|
| 70 |  |  |  |  |
| | Small amounts of contents, but ONLY 70 | Amounts of characters is way small | | |
| 135 |  |  |  |  |
| | Low contrast, but relatively pretty writing, but 135 | Looks quite similar to the difficulty score of 9 or 20 images | | |
| 350 | - | - |  |  |
| | | | Bleed-through + Ugly writing | |
| 748 (handwritten) 3064 (typed) | - | - |  |  |
| | | | Ugly writing + Medium contents | |

From the above observations, we can set the following two assumptions:

Assumption 1. A feature indicating whether an image is *handwritten* or *typed* seems promising to be somewhat related with the *difficulty score*.

Assumption 2. It is necessary to find more high-level visual features (e.g., expert knowledge-based engineered features or deep-features learned by a deep-learning model) hard to expect to find a correlation between a simple standalone *visual feature* (e.g., number of characters or low contrast) and the *difficulty score*.

4. Experiment 2: Pearson's Correlation

In this experiment, we aim to find a set of visual features showing a meaningful correlation to the difficulty score using the Pearson's Correlation. The set of features that we consider consists of low-level visual features obtained by relatively simple image processing techniques, such as contrast measure or counting the number of connected components (i.e., letters or characters.) Along with these low-level visual features, based on the above two assumptions, we added four additional high-level visual features: (1) *prediction*, (2) *density*, (3) *number of zones*, and (4) *zone size abnormality*.

- First, the *prediction* feature is a categorical value indicating whether the type of document image is *handwritten*, *typed*, or *mixed*. This feature is obtained by our deep-learning-based document type prediction model developed in our Project 2, which showed promising classification performance with 0.9 of f1-score (best value at 1 and worst value at 0.)
- Second, the *density* feature measures how dense the document is by considering the area of non-background regions. This feature is obtained by dividing the area of non-background regions by the resolution of the image.
- Third, the *number of zones* feature represents how many zones (i.e., visually homogeneous regions) are presented in the image. This feature is obtained by segmenting the image by our deep-learning-based document segmentation model developed in our Project 1, which showed promising segmentation performance with 0.7 of mIoU (best value at 1 and worst value at 0.)
- Fourth, the *zone size abnormality* feature measures the size of zones and calculates the degree of outliers. This feature is obtained by counting the number of outliers in terms of zone size and divide it by the resolution of the image for the normalization purpose. The intuition behind this feature is that the output of our segmentation algorithm tends to generate the relatively regular and uniform size of zones for straight forward and clear document images whereas it tends to generate a number of abnormal size of zones (i.e., extremely small zones and big zones simultaneously) for noisy document images.

After obtaining the above set of visual features, before conducting Pearson's correlation, we carry out a histogram analysis to visually inspect how images are distributed for each visual feature, as shown in Figure 1 below. From this analysis, we can observe that some visual features follow a normal distribution at a certain level, such as *density*, *contrast*, and the *number of letters*. Note that one assumption behind Pearson's correlation is that variables (i.e., visual features) should be normally distributed. In this regard, we can expect that those three features are likely to show relatively high correlation coefficient values. We can observe that this expectation does actually match with the result of Pearson's correlation, as shown in Table 2.

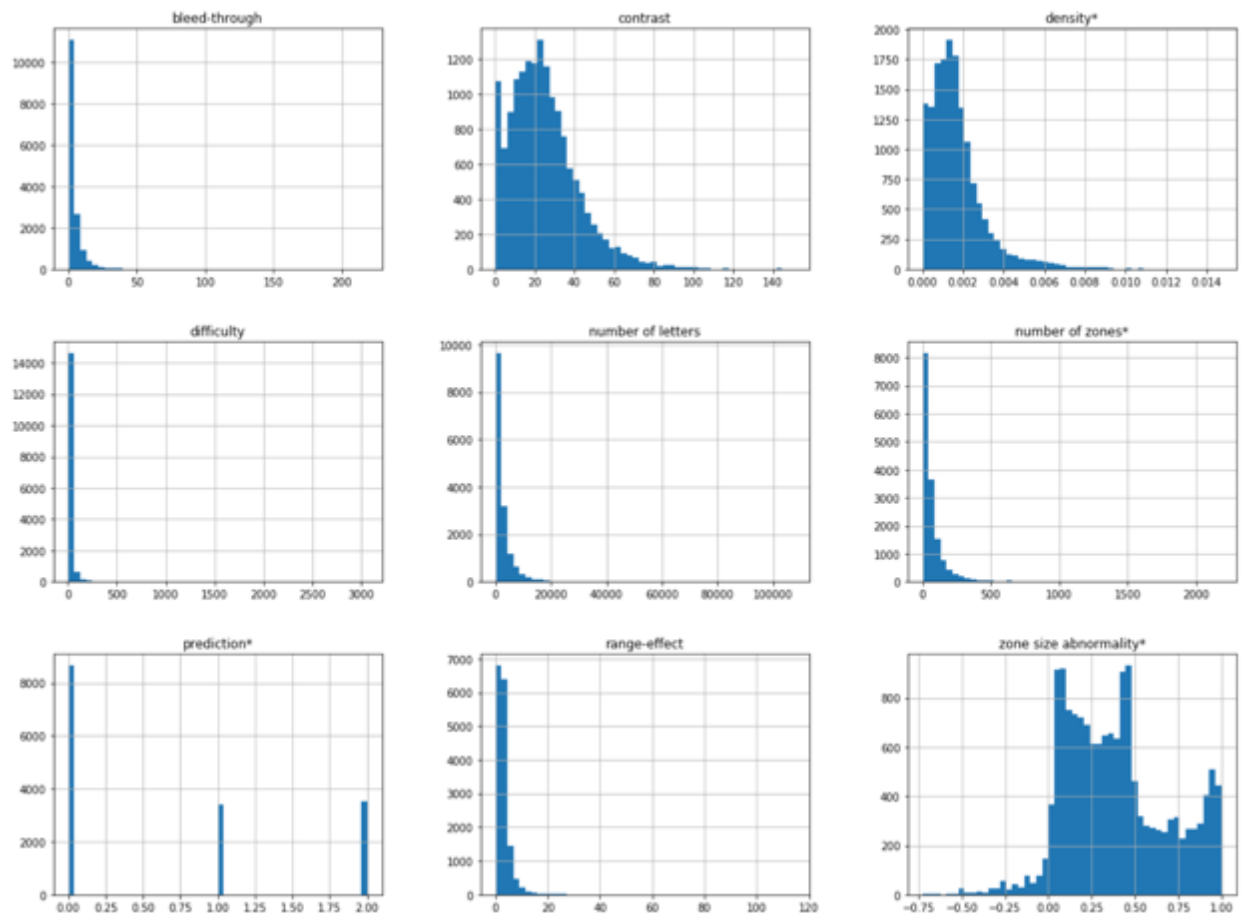


Figure 1. Distribution of visual features over 15,592 images. Note that one of the assumptions behind the Pearson's correlation coefficient is that variables (i.e., visual features) should be normally distributed.

Table 2. The size of correlation for various visual features. Note that a visual feature with asterisks (*) is high-level engineered features using low-level visual features.

| Visual features | Size of correlation |
|------------------------|---------------------|
| Density* | 0.17 |
| Contrast | 0.15 |
| Number of Letters | 0.15 |
| Number of Zones* | 0.10 |
| Zone Size Abnormality* | 0.07 |
| Bleed-through | 0.03 |
| Range-effect | 0.02 |
| Prediction* | 0.01 |

It is worth noting that there is no rule for determining what correlation value is considered strong, moderate, or weak. The interpretation of the value depends, in part, on the topic and context of the study. When we are conducting research that is difficult to measure, in our case, the difficulty of the document image in the context of human perception, we should expect the correlation value to be lower. With this in mind, we can interpret this result as follows.

First, we can observe that one of our high-level engineered visual features, *density*, shows the highest correlation with the difficulty score.

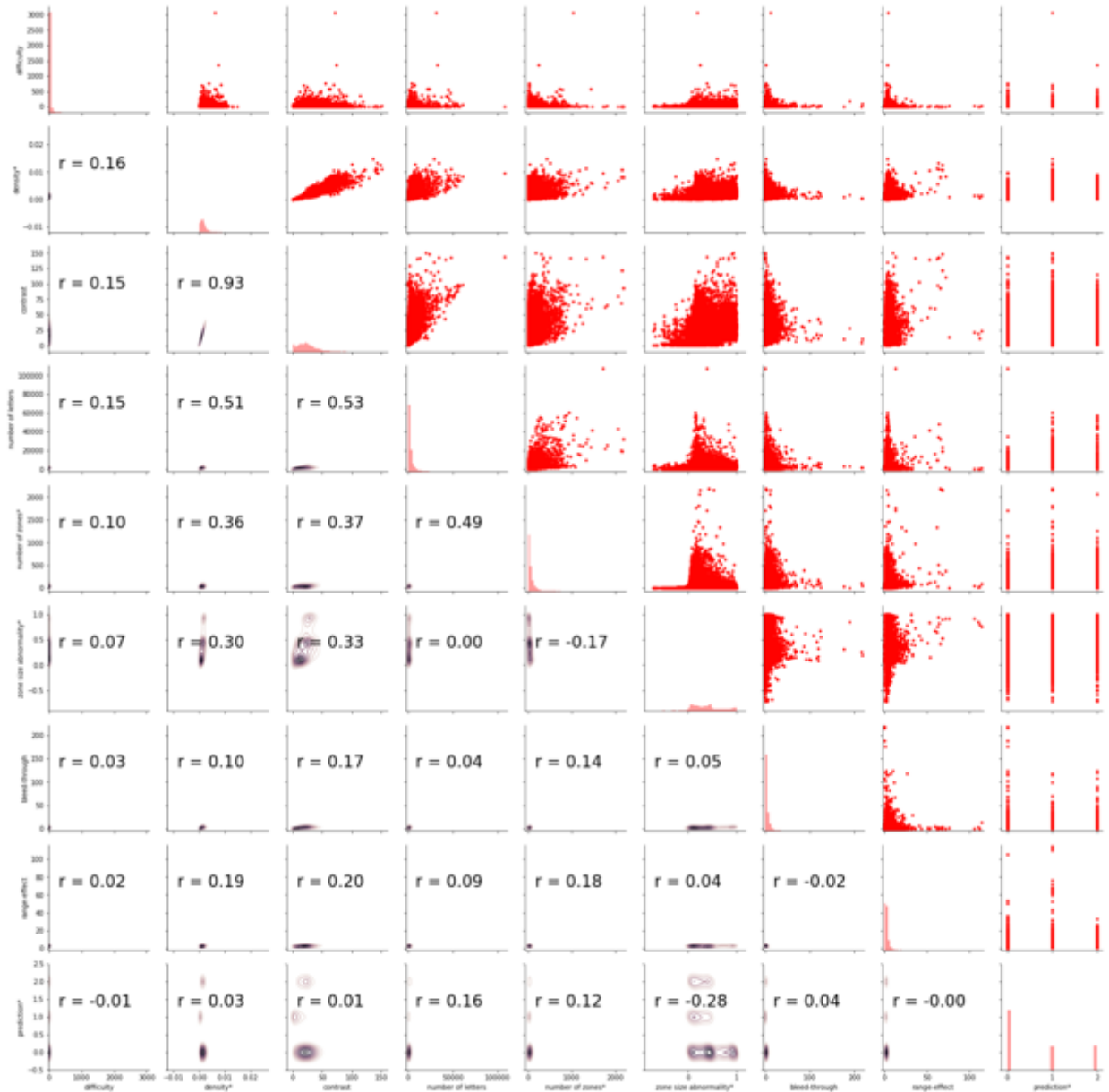


Figure 2. Pairs plot to visualize scattered distribution and relationship between two visual features. The r value represents the size of correlation (range from -1 to 1, best value at -1 or 1, worst value at 0.) It is worth noting that the relationship between most of the *visual features* with the *difficulty score* is not linear (first row). Especially, the *prediction* and the *difficulty score* does not show any linear relationship (top right cell.)

Second, a standalone *prediction* feature shows negligible correlation with the difficulty score. This result is anticipated since its distribution (see Figure 1) does not follow the normal distribution. Also, since the Pearson’s correlation is limited to reveal a “linear relationship” between variables, if there is a non-linear relationship between variables (see Figure 2),

correlation between the visual feature and the difficulty score can be very low. In this regard, based on our Assumptions 1 and 2, we expect that this *prediction* feature should be combined with other variables in a non-linear way, for example, by using the polynomial regression or support vector machine, to reveal, if any, correlation with the *difficulty score*.

5. Conclusion

In this experiment, we show that both low-level and high-level engineered visual features that we used were *not* able to capture any correlation with the difficulty score. As shown in the pairs plot, most of the visual features rarely show any linear relationship with the *difficulty score* (see the first row in Figure 2). From this outcome, we can think of two future directions to reveal a more comprehensive understanding of the relationship between visual features and the difficulty score.

- First, instead of low-level or engineered visual features, we can explore deep features, which is learned by a neural network model. Because of non-linearity property inherent in the neural network model, the features extracted by the model are known to be significantly high-level non-linear property.
- Second, we can explore models that are capable of dealing with non-linear data, such as polynomial regression, support vector machine, or neural network. These models are mapping the low-level features into high dimensional space, which has an effect of embedding non-linearity property and the interaction between different low-level visual features, and we can expect a better understanding of the relationship between *visual features* and the *difficulty score*.