# CLASSIFICATION OF COMPOUND IMAGES BASED ON TRANSFORM COEFFICIENT LIKELIHOOD

*Isaac Keslassy*, Mark Kalman, Daniel Wang, and Bernd Girod*

Information Systems Laboratory,
Department of Electrical Engineering
Stanford University, Stanford, CA 94305

## ABSTRACT

Applications like distance learning and teleconferencing often require compression of images that contain both text and graphics. Because text and graphics have different properties, a compression scheme can benefit by treating the textual and graphical portions of such compound images separately. In this paper, we propose new methods, called Transform Coefficient Likelihood (TCL) schemes, for separating the textual and graphical portions of a compound image. TCL schemes examine the DCT coefficient values of an 8 × 8 block. For each coefficient, they refer to stored histograms that give the likelihood that a certain value occurs in a text block, or in a graphics block. They then examine the differences in these two likelihoods over all the coefficients in the block to decide whether it contains text or graphics. Experimental results show that the best TCL methods significantly outperform previously proposed techniques.

## 1. INTRODUCTION

Compression concerns itself with the reduction of redundant, or less perceivable information. Consider common DCT-based compression schemes. These achieve compression in continuous tone images by exploiting the correlation among neighboring pixels. Text has different properties, however, and when it is compressed using a scheme developed for images, the compressed text will generally be at a lower quality for a given rate than surrounding graphics. It is beneficial, therefore, to identify regions in an image that contain text and treat them differently.

In this paper we present methods to identify text in an image and then compare the effectiveness of these methods. To facilitate this comparison, we implement each scheme as a function that takes as an input an 8 × 8 block of pixels or corresponding transform coefficients, and returns a real number we call the activity, $a$. If the activity is above some threshold, the scheme has decided that the 8 × 8 block is text.

To be more precise, let $\mathbf{X} \in \Re^{64}$ be the vector of intensity values (or DCT coefficient values) for a block. Then, let each scheme be $s = \{a_s, t_s\}$, where $a_s : \Re^{64} \rightarrow \Re$ is a function over the vector $\mathbf{X}$ and $t_s \in \Re$ is the threshold value for the scheme. Each scheme makes a decision $D_s$ as follows:

$$D_s(a_s(\mathbf{X})) = \begin{cases} \mathit{text} & \text{if } a_s(\mathbf{X}) > t_s \\ \mathit{graphics} & \text{otherwise} \end{cases}$$

The following pseudo-code illustrates:

```
foreach block in the image:
   load block values to $X$
   compute block activity $a(\mathbf{X})$
   return block type:  $D(a(X))$
end of loop
```

We organize this paper as follows. We begin in Section 2, by reviewing text identification methods previously introduced in the literature. Then, in Section 3, we introduce our new TCL schemes. In Section 4 we present the methodology that we use to assess and compare the performances of the various schemes. Finally, in Section 5 we show how well the schemes performed relative to one another.

## 2. STATE OF THE ART

Previously proposed techniques for text location in images include both methods that examine pixel values in the spatial domain and methods that examine DCT coefficient values. In our work, we implement these schemes to provide a basis for assessing the new schemes that we present in Section 3.

### 2.1. Spatial-Domain Schemes

The algorithms, *Range, Variance, Absolute Deviation,* and *Sobel Filter* examine the actual pixel values in an 8 × 8 block [1], [2].

*Isaac Keslassy is supported by a Wakerly Stanford Graduate Fellowship

0-7803-6725-1/01/$10.00 ©2001 IEEE

*Range* is based on the observation that text blocks are likely to have a higher dynamic range than non-text blocks. For *Range*, the activity value for a block is simply the range of its pixel values.

*Variance* computes the activity as the pixel variance in a block. Text blocks are likely to have higher pixel variances than graphics blocks.

*Absolute Deviation*, a similar scheme, computes the activity for a block as the mean absolute deviation from their average intensity of its pixel values.

The *Sobel Filter* scheme is based on the observation that text blocks are likely to have more edges than graphics blocks [1]. Based on the Sobel edge-detection filter, it computes the activity for a block as the sum of its pixels' Sobel Gradients.

## 2.2. DCT-Based Algorithms

The following algorithms are functions over the DCT coefficients of an $8 \times 8$ block. They exploit the differences between text and graphics blocks in the distribution of energy among their DCT coefficients. Figure 1 shows the mean absolute values of the 64 DCT coefficients, arranged in JPEG zig-zag scan order, for a typical compound image [3]. Plotted on a logarithmic scale, the top curve is for $8 \times 8$ blocks that contain text and the bottom curve is for blocks that contain graphics. Notice that the mean absolute value of DCT coefficients is much greater for text than for graphics, especially towards the higher frequencies.

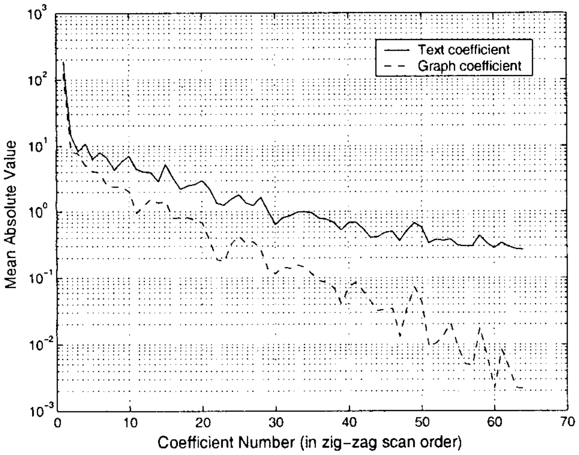

**Fig. 1**. Mean absolute values of DCT coefficients.

One DCT-based scheme that exploits the differences apparent in Figure 1, *DCT Energy*, takes as the activity value the squared sum of DCT coefficients 2 through 64. Let $\{c_n\}$ be the DCT coefficients of an $8 \times 8$ block arranged in zigzag scan order. The activity for the scheme is then:

$$a = \sum_{n=2}^{64} c_n^2$$

*DCT Absolute Sum*, a similar scheme, replaces the square in the energy calculation with an absolute value, which is easier to compute [1].

*DCT 18 Absolute Sum* uses only 18 of the 64 coefficients to compute the sum. These were determined to be the most reliable coefficients for separating text and graphics in [2]. For this scheme, then, the activity is given by:

$$a = \sum_{n \in S} |c_n|$$

Not all DCT-based schemes simply sum coefficient energies. *DCT Bitrate*, the scheme proposed by Konstantinides and Tretter, estimates how many bits it would cost to run-length encode a block [4]. It reasons that since text is more costly to encode than graphics, a high cost in bits indicates text. The scheme estimates that each non-zero JPEG-quantized transform coefficient, $\hat{c}_n$, in a block requires $\log_2 |\hat{c}_n| + 4$ bits. The activity value for this scheme is then:

$$a = \sum_{\hat{c}_n \neq 0} [\log_2(|\hat{c}_n|) + 4]$$

## 3. TRANSFORM COEFFICIENT LIKELIHOOD (TCL) SCHEMES

Having reviewed the schemes from the literature, we now introduce our TCL schemes. These rely on the relative frequencies of occurrence of coefficient values. While the algorithms in Section 2.2 treat each considered transform coefficient identically, TCL schemes make use of histograms specific to each coefficient. They employ two histogram tables for each of the 64 coefficients. One table stores the relative frequencies of occurrence of the coefficient's values given that a block contains text. The other does the same for blocks that contains graphics. TCL schemes look to these tables to determine the likelihood of each coefficient value given that a block contains text or given that a block contains graphics. They produce an activity value based on these likelihoods.

In order to reduce the size of the tables we quantize the coefficients according to the JPEG quantization matrix and limit their values to $[-255, 255]$ [3].

We populate the text tables using a set of 5 images that contain only text. A histogram is compiled for each quantized coefficient by counting the number of times each of the coefficient's 511 possible values occur in the training set. Then we normalize the histograms to sum to one. In a similar manner we generate a corresponding set of 64 tables of coefficient frequencies for graphics blocks.

In TCL schemes, we use these tables to estimate the probability that the nth quantized DCT coefficient takes on the value $\hat{c}_n$, given that the block contains text or graphics. We denote these conditional probabilities:

$$p(\hat{c}_n|text) \quad \text{or} \quad p(\hat{c}_n|graphics)$$

Let $\hat{\mathbf{C}}$ be the vector of quantized DCT coefficients. We can decide a block contains text according to the Maximum A Posteriori probability (MAP) rule. According to the MAP rule we decide a block is text if:

$$p(text|\hat{\mathbf{C}}) \geq p(graphics|\hat{\mathbf{C}})$$

Using Bayes' Rule the expression can be rewritten as:

$$\frac{p(\hat{\mathbf{C}}|text)}{p(\hat{\mathbf{C}}|graphics)} \geq \frac{p(graphics)}{p(text)}$$

Now we make two simplifying assumptions:

1. $p(text)$ and $p(graphics)$ are known.

2. $\{\hat{c}_n|text\}_{1\leq n\leq 64}$ and $\{\hat{c}_n|graphics\}_{1\leq n\leq 64}$ are sets of independent random variables.

Using these assumptions we can arrive at the scheme we call the *MAP Rule*:

$$a = \sum_{1}^{64}[\log p(\hat{c}_n|text) - \log p(\hat{c}_n|graphics)]$$

However, these assumptions don't generally hold. Moreover, this scheme is overly sensitive to the difference between the exact pmf of the coefficients and the approximation stored in the tables. Therefore we explore other schemes that operate on the differences in the marginal conditional distributions of the 64 coefficients.

$\Delta P$ is the simplest of these schemes. Its activity score is given as:

$$a = \sum_{n=1}^{64} [p(\hat{c}_n|text) - p(\hat{c}_n|graphics)]$$

*Delta Probabilities - High Probability*, a variation, biases the score towards the coefficients with higher probabilities by using the difference in the squares of probabilities:

$$a = \sum_{n=1}^{64} [p^2(x_n|text) - p^2(x_n|graphics)]$$

Another variant, $\Delta P$ - *High Difference* biases the score towards coefficients with the biggest differences:

$$a = \sum_{n=1}^{64} [p(\hat{c}_n|text) - p(\hat{c}_n|graphics)]^3$$

| **Scheme** | **activity** |
|---|---|
| *Range* | range of intensities |
| *Variance* | variance of intensities |
| *Abs. Deviation* | mean abs. dev. of intensities |
| *Sobel Filter* | Sum of Sobel Gradients |
| *DCT Energy* | $\sum_{n=2}^{64} c_n^2$ |
| *DCT Abs. Sum* | $\sum_{n=2}^{64} \|c_n\|$ |
| *DCT 18 Abs. Sum* | $\sum_{n\in S} \|c_n\|$ |
| *DCT Bitrate* | $\sum_{\hat{c}_n\neq 0}[\log_2(\|\hat{c}_n\|) + 4]$ |
| *MAP Rule* | $\sum_{n=1}^{64}[\log p(\hat{c}_n\|text) - \log p(\hat{c}_n\|graphics)]$ |
| $\Delta P$ | $\sum_{n=1^{64}} \Delta P_n$ |
| $\Delta P$ *High Prob.* | $\sum_n [p^2(\hat{c}_n\|text) - p^2(\hat{c}_n\|graphics)]$ |
| $\Delta P$ *High Diff.* | $\sum_{n=1}^{64} \Delta P_n^3$ |
| $\Delta P$ *Horizontal* | $\sum_{i,j} j\Delta P_{i,j}$ |
| $\Delta P$ *HF* | $\sum_{i,j} ij\Delta P_{i,j}$ |

**Table 1**. Summary of schemes.

Our final set of variants favors higher frequency DCT coefficients. High frequencies indicate edges in the block, a characteristic of text. This set of schemes consists of $\Delta P$ *Horizontal*, and $\Delta P$ *HF*.

Table 1 gives the activity values for these schemes using the notation:

$$\Delta P_{i,j} = [p(\hat{c}_{i,j}|text) - p(\hat{c}_{i,j}|graphics)]$$

where $i$ and $j$ are the quantized coefficient's row and column positions, respectively, in the matrix of DCT coefficients.

## 4. EVALUATION METHODOLOGY

Before evaluating the performance of the text identification schemes presented in the previous sections, we need to find an appropriate threshold value for each scheme. Recall that each scheme is a function over a 8 × 8 sized block that returns a real number we call the activity, $a$. If the activity is above a threshold, we label the block text.

To find appropriate thresholds, we use a training set of images whose text-blocks have already been identified manually. For each scheme we then vary the threshold over a suitable range and choose the one that gives the best match with the hand-generated answers. By "best-match" we mean the threshold value that gives the lowest error score:

$$\text{Error Score} = \frac{(\%\text{False Negative}) + (\%\text{False Positive})}{2}$$

False Negatives are text blocks mistakenly labeled graphics. False Positives are graphics blocks wrongly labeled text [5].

Using percentages prevents the biasing of the threshold in the case that there are more of one type of block than the other in the training set.

Thus, we train each algorithm by choosing the threshold value that minimizes the error score over a set of training images. Using these experimentally optimal thresholds, we then produce error scores for each algorithm over a different set of images. Having disjoint training and testing sets forces the algorithms to exhibit robust behavior with respect to threshold to perform well in our evaluation.

Please note that the histograms used by the TCL schemes are compiled over a set of ten images that is independent of the training and testing sets.

## 5. EXPERIMENTAL RESULTS

Error score results for the various schemes are tabulated in Figures 2 and 3. The results in figure 2 were generated using a set of 10 images that contained graphics and black, typed text. The results were averaged over ten trials where in each trial one image was held out as the test image and the remaining nine were used as training. We permuted the set of images in each successive trial so that in ten trials each image served as the test image exactly once. Figure 3 tabulates results when ten, more heterogenous, images were added to the set. These images included handwritten text and varying text hues. Here we averaged over twenty trials. The graphs show that our new algorithms consistently perform better than previous methods. In particular, our $\Delta P$ - $HF$ scored 22.8% better than the best previous method for the ten image set and 10.0% better for the more heterogenous 20 image set.

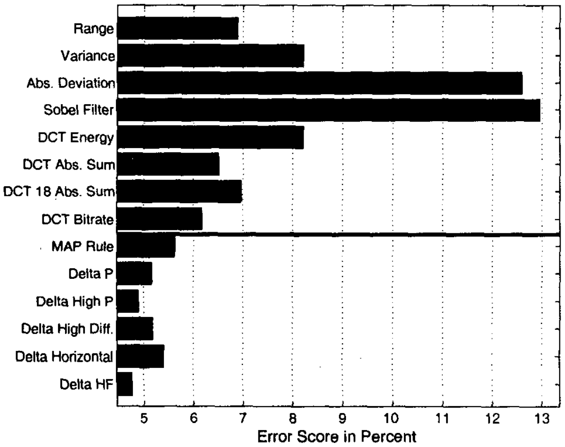

**Fig. 2.** Comparison of Error Score results for a set of 10 testing images that contain graphics and black, typed text. (TCL schemes shown below the dividing line).

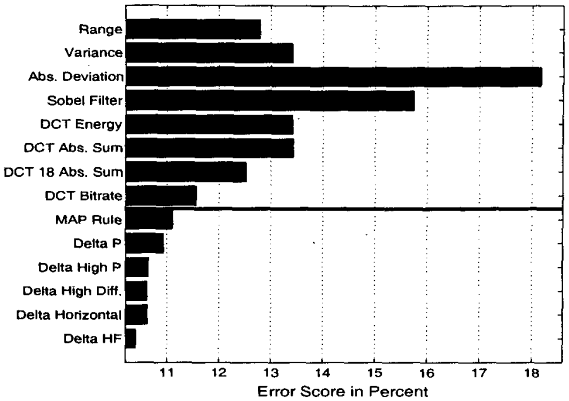

**Fig. 3.** Comparison of Error Score results for a set of 20 testing images that contain heterogenous text.

## 6. CONCLUSION

In this paper, we have proposed a new set of schemes for separating the textual and graphical portions of an image. These methods, which we call Transform Coefficient Likelihood (TCL) schemes, differ from previously proposed methods in that they are based on the relative frequency of occurrence of transform coefficient values. As shown, the most effective scheme in our TCL family, $\Delta P$ $HF$, significantly outperforms previously proposed techniques.

## 7. REFERENCES

[1] C.T. Chen, "Transform coding of digital images using variable block size DCT with adaptive thresholding and quantization", *SPIE* vol. 1349, pp.43-54, 1990.

[2] N. Chaddha, R. Sharma, A. Agrawal, and A. Gupta, "Text segmentation in mixed-mode images," in Proceedings of the 28th Asilomar Conference on Signals, Systems and Computers, vol. 2, pp. 1356–1361, 1995.

[3] ISO/IEC 10918-1 - ITU-T Recommendation T.81, *Information technology - Digital compression and coding of continuous-tone still images: Requirements and guidelines*, Jan. 1992.

[4] K. Konstantinides and D. Tretter, " A JPEG Variable Quantization Method for Compound Documents ", *IEEE Transactions on Image Processing,* Vol.9, No.7, p.1282, July 2000.

[5] N. Chaddha, "Segmentation-Assisted Compression of Multimedia Documents" in Proceedings of the 29th Asilomar Conference on Signals, Systems and Computers, vol. 2, pp.1452–1456, 1996.