

Determining text readability over textured backgrounds in augmented reality systems

Alex Leykin*

Department of Computer Science, Indiana University

Mihran Tuceryan†

Department of Computer Science, Indiana University - Purdue University Indianapolis

Abstract

This paper concerns the application of pattern classification techniques to the domain of augmented reality. In many augmented reality applications, one of the ways in which information is presented to the user is to place a text label over the area of interest. However, if this information is placed over very busy and textured backgrounds, this can affect the readability of the text. The goal of this work was to identify methods of quantitatively describing conditions under which such text would be readable or unreadable. We used texture properties and other visual features to predict if a text placed on a particular background would be readable or not. Based on these features, a supervised classifier was built that was trained using data collected from human subjects judgement of text readability. Using a rather small training set of about 400 human evaluations over 50 heterogeneous textures the system is able to achieve a correct classification rate of over 85%.

CR Categories: I.5.4 [Computing Methodologies]: Pattern Recognition—Applications

Keywords: text readability, texture analysis, augmented reality, pattern recognition, Gabor filter

1 Introduction

In this paper, we apply pattern classification techniques to determine whether the placement of textual information over a textured background in augmented reality systems will render the text unreadable. Augmented reality (AR) is a technology in which a user's view of the real world is enhanced or augmented with additional information generated from a computer model. The simplest such systems provide information in the form of a text placed on top of the existing real scene. If the placement is not carefully done, the visual properties of the background may interfere with the readability of the text so placed.

Some aspects of the placement in AR such as the layout and overlapping of the text have been studied and evaluation criteria and algorithms have been developed to generate the most readable layout of text labels [Azuma and Furmanski 2003; Rose et al. 1995].

*e-mail: oleykin@indiana.edu

†e-mail: tuceryan@iupui.edu

There are, however, considerations other than the text layout that affect the readability of the text labels. One such factor is the interference from the background texture.

Prior research has studied the visual properties of the background that affect the readability of text [Scharff et al. 1999; Scharff et al. 2000; Hill and Scharff 1999]. Scharff et al. have identified the contrast and spatial frequency content of the background texture as factors affecting the readability of text. These studies have been conducted in the context of augmented reality and web design applications, but problem was never treated as one of automated readability classification. The goal of their experiment was to study the factors affecting the readability of text in the human perceptual system and from this derive certain rules, according to which the appropriate background texture and text properties could be selected. This goal, of course, is practical in web design applications, but less so in augmented reality applications. One cannot always control the texture characteristics of the background images in the real world, other than possibly change the position of the text and place it on a more suitable background. In order to even decide that a particular background is going to adversely affect readability, one needs to have an automatic way of analyzing the background image and how it will interact with a specific text to be overlaid on it.

In this paper, we have tried to address precisely this issue. We develop a supervised classification method that, for a particular textured background, predicts if a text with certain features (e.g., font size and weight) superimposed on it will result in a readable or unreadable text. With text overlaid on textured backgrounds, our expectation was that this texture would interfere with the readability of the text at the particular frequencies and orientations. To capture this phenomenon, we used contrast features and Gabor filter based texture features at various frequencies and orientations to be used in our supervised classifier.

It is also worth mentioning that unlike previous human readability studies [Scharff et al. 1999; Scharff et al. 2000] where the reliance of the authors was on the speed with which subjects read paragraphs of text placed over textured backgrounds, we utilize the concept of readability of a single word or “a label”. This approach, we believe, has a number of advantages. Primarily it was chosen for the fact that label readability is what is mostly used in our field of application: automated object labeling in augmented reality. More-over readability of a single word is less prone to the influence of social factors in human subjects (such as education), which can influence the speed of reading of the connected paragraphs, but in much smaller degree the readability of a single short word.

In Section 2, we present our approach to classifying readability of text over textured background. In Section 2.1 we describe the experimental setup for obtaining the training data for the supervised classifier. In Section 2.2, we explain the visual features we use for the classification. Section 3 outlines the experimental results. Finally, in 4, we present our conclusions and outline potential research on this problem.

2 Architecture of the system and collection of training data

In order to automate the prediction of whether a text overlaid on a textured background will be readable, we built a supervised classifier that was trained on data collected from human subjects through a set of experiments. We implemented a number of different classifiers and used feature selection methods to identify the most important features in deciding readability. The training data was collected through a set of experiments with human subjects. The following subsections describe briefly each one of these components.

2.1 Collection of training data

The readability of text is intimately tied to the perception and cognition of the humans who will be looking at such texts [Legge et al. 1985; Solomon and Pelli 1994]. Unlike many other computer vision or pattern recognition tasks, in which the class labels are well defined (e.g., what a handwritten letter is), the degradation of the readability of a text must be determined by the performance of the human who is looking at the text. In order to collect the training data needed by our supervised classifier, we set up a series of experiments in which human subjects assigned class labels “readable” vs. “unreadable” to the sample texts displayed to them.

Six independent participants were presented on a computer screen with a series of 100 images each, in which some sample text was overlaid on a textured background (see Figure 1 for some examples). Three of the subjects were native speakers of English and three of them were non-native speakers. All participants had self-reported 20/20 or corrected to 20/20 vision. The subjects were approximately 50 cm away from the screen. As the overlaid text images were presented on the screen, the subjects had to assign a number between 1 to 8 for the level of readability of the text: 1 being least and 8 being most readable. The subjects were informed that numbers from 1 to 4 represent an unreadable text/texture combination, similarly numbers 5 to 8 represent a readable one. The textures were randomly chosen from a set of 50 texture images representing various materials, texture scales and objects. The textures were preselected to cover uniformly a range of frequencies and orientations. The texts presented to the subject were single words with approximately the same number of characters and the total set of characters was chosen so as to cover an entire Latin alphabet.

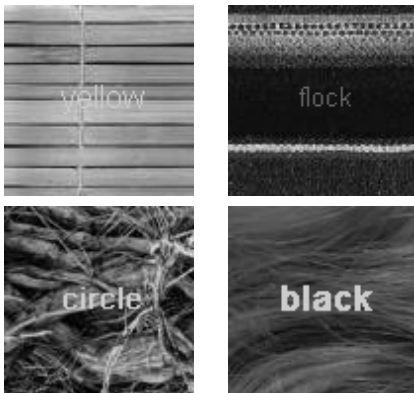


Figure 1: Samples presented to human subjects and the results of classification by our system. The first row shows samples classified as unreadable by the human subjects and the second row shows samples classified as readable by the human subject. The samples in the left column are classified by our supervised classifier as unreadable and the right column is classified as readable.

Several parameters were set as variable and further restrictions were imposed on them during the experiments. All the runs were performed on the same, calibrated monitor, under same lighting conditions (daylight). The viewing distance was set to be approximately an arm length (500mm) and viewing angle was fixed for all the subjects to position the eye-level at the same level as the center of the CRT screen. The dimensions of the display were 1024x768 with a pixel size of approximately 0.5mm.

To single out the luminance information all experiments were conducted in gray-scale. Five font intensity levels from 0 to 1 with a step of 0.2 were used. We have fixed font sizes (10, 12 and 14 point Times New Roman) so that the height of a character is located around the same angle, which can be estimated as:

$$\alpha = \frac{360 * h}{2 * \pi * r}, \quad (1)$$

where h is the height of the text and r is the distance from the screen. For the given font sizes (approx. 6 pixels high) and the viewing distance (approx. 500 mm) and measuring the size of the pixel as 0.5mm on the 1024x768 17 inch CRT display, we estimate the viewing angle height of the text to be approximately 0.35grad

Note that the ranges used for intensity level, font size and weight were selected as a result of a preliminary experiment to eliminate outlying configurations and approximately equalize the volume of two classes. A priori probability for each class was set to 0.5.

Overall each subject was presented with 100 random configurations of the following: 50 textures, 13 input words, 6 font intensities, 4 font weights and 3 font sizes. This constitutes to a space of 46800 possible configurations, a space hardly covered by the total of 600 exemplars taken by us.

As one can see from Figure 1, the judgement about readability by the human subjects may be questionable in borderline cases (upper right and lower left images in the figure). For such borderline cases, the variability among human subjects was also high.

In order to eliminate inconsistent judgements across different subjects, we removed the outliers in the following way: all the responses for the same texture image and text intensity with rankings differing more than 2 points (e.g., 5 and 8) were removed. See section 3 for the results of the simulation.

We had two sets of experiments in order to take into account the attentive vs. pre-attentive nature of the processing of the visual inputs. The first set of experiments were conducted with no time limit on the visual input presented to the subject. The subject could take as long as he wished until a decision was made on the readability of the input. It is interesting to note that in the first set of experiments, response times were recorded for each subjects’ choice. However they didn’t demonstrate any significant correlation with the text readability, as defined by subjects’ responses. This, of course, is a very deliberative process and we also wanted to see how much of the decision is affected by the pre-attentive processing by the visual system. Therefore, in the second set of experiments, all the original parameters were kept unchanged except that each image was shown for a brief period of time (150 msec or 300 msec). These intervals were chosen to capture pre-attentive and attentive visual system response correspondingly [Julesz 1981]. The subject had this amount of time to pick the word he or she was presented with from a list of 8 alternatives. These two timings were randomly mixed with equal distribution throughout the experiment.

2.2 Texture and Contrast Descriptors

As our visual features to be used in the classifier, we used font size, font weight, intensity contrast, and responses to a bank of Gabor filters at four spatial frequencies and four orientations convolved with the background texture. As shown by [Chubb et al. 1989] both

luminance contrast and texture of the text and background determine a so-called “perceived contrast.” In other words, the clarity of the text and therefore the ease of reading is stipulated by these two factors.

We computed the contrast feature as the absolute difference between the mean intensities of the text and the surrounding region. Given the mean intensity of the text, I_t , and the mean intensity of the background, I_b , the contrast measure is given by $C = |I_t - I_b|$. I_b was computed using pixels in the immediate surrounding vicinity of the text and not the entire background image.

Gabor filters have been used as successful texture descriptors for classification and segmentation tasks in the past [Jain and Farrokhnia 1991; Tuceryan and Jain 1998; Clark and Bovik 1987]. To extract both spatial and frequency descriptors of the background locally we used real-valued even-symmetric Gabor filters as given by Malik and Perona [Malik and Perona 1990]:

$$h(x, y) = \exp \left\{ -\frac{1}{2} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right) \right\} \cos(2\pi u_0 x), 3 \quad (2)$$

where σ_x and σ_y are the parameters controlling the width of the Gaussian envelope and u_0 is the frequency of a sine wave. For Gabor filters at different orientations, this basic function is rotated around the origin.

We have used filters at four orientations ($0, \pi/4, \pi/2$, and $3\pi/4$), and four spatial frequencies ($1/2, 1/4, 1/8$, and $1/16$), producing a bank of 16 Gabor filters.

The resulting bank was convolved with the original background image before text was superimposed on it, generating 16 filtered background images. The mean value over the entire image for each feature was computed to obtain the texture features for the image. This assumes that the image contains a single texture.

2.3 Supervised Classifier

We implemented a number of supervised classifiers using the training data obtained from our human experiments. The classifiers were trained using the features described in Section 2.2 and the class labels “readable” or “unreadable” obtained from our data collection. We assigned the class label “unreadable” for samples ranked 1 to 4 and the class label “readable” for samples ranked 5 to 8. We implemented and tested the classifiers in Table 1, a total of seven classifiers: Normal densities based linear classifier (LIN); Normal densities based quadratic (multi-class) classifier (SQR); Nearest mean classifier (MEAN); Scaled nearest mean classifier (MEANSC); Uncorrelated normal densities based quadratic classifier (BAYES); Fisher’s discriminant (minimum least square linear classifier) (FISHER); and Support vector machine classifier (SVM).

3 Experimental Results

We used 70% of the points as a training set to build a classifier and the remaining 30% to test its performance. Several classifiers mentioned above were used to classify the test data. The results are summarized in Table 1.

As we can see from this table, the best classification results are given by the Support Vector Machine classifier with $> 87\%$ classification accuracy. We looked at the reasons why we are unable to reach better classification rate, and our conclusion was the training data collected from the human experiments seems to be limiting the performance of the classifier. As discussed in Section 2.1, we did remove some outliers in the human classifications. The outliers were approximately 20% of the training from the training sample. Column 1a in Table 1 gives the error rates that include the outliers in the training data and column 1b gives them after taking the outliers out. As can be seen, the results improve substantially (2 – 3%).

Classifier	1a	1b	2a	2b
LIN	0.160	0.131	0.250	0.197
SQR	0.207	0.165	0.283	0.255
MEAN	0.157	0.138	0.270	0.284
MEANSC	0.155	0.128	0.266	0.269
BAYES	0.170	0.141	0.303	0.281
FISHER	0.160	0.132	0.249	0.198
SVM	0.161	0.122	0.250	0.206

Table 1: Classification error rates for: 1a - results of first experiment used as training data; 1b - same as 1a with outliers removed; 2a - results of second experiment (with 150 msec delay) used as training data; 2b - results of second experiment (with 300 msec delay) used as training data.

As mentioned in Section 2.1, the selection of ground truth is a highly subjective process, because it relies on the judgment of a few subjects. We believe that further improvement can be achieved with increased size of human subject set. In addition, if the number of samples in the training set is sufficiently high, the data can be subjected to more strict outlier removal process. For example removing identical samples with rank differing at least by 1 point (as opposed to 3 point in the current setup) could lead to a more consistent definition of “readability”.

To make a preliminary assessment of the class separation we projected the data from the multidimensional feature space down to two dimensions (Figure 2). Looking at this figure, one can see that the two classes have a certain amount of overlap and that explains to some extent the classification error rates.

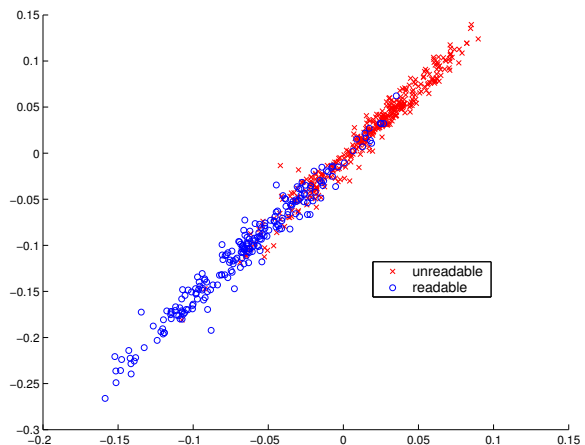


Figure 2: The features projected to two dimensions.

The results of the timed experiments, where the same samples were presented to subject for different amounts of duration, showed some correlation with the untimed data collected (Table 1; columns 2a and 2b).

4 Conclusions and Discussion

The results of the human subject experiments confirm that background variations only affect readability when the text contrast is low. The dominating frequency range that affects the readability: from 1.5 to 3.0 cycles per letter, also agrees with the results of [Scharff et al. 2000; Solomon and Pelli 1994].

Figure 3 shows the frequency most correlated to the readability ranking to be 8 cycles per filter (32 pixels), which is equal to 4 pix-

els. Given letter size of approximately 6 pixels we get 1.5 cycles per character. Dominant orientation was $\pi/4$ and the least correlated orientation was $\pi/2$.

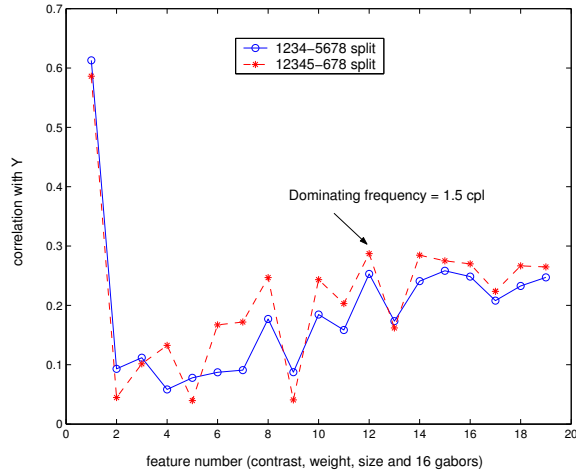


Figure 3: Correlation of features with classification

Another interesting observation is concerned with the alternative class separation 12345-678 which demonstrated improved classification performance. This indicates that these two classes are better separated than originally intended 1234-5678 split and suggests that subjects optimistically attribute barely readable text to the class of readable samples. In other words, what people deem almost readable in reality turns out to be unreadable by the very definition the same people have given for “readability”.

There has been some work in optics showing the influence of chrominance changes on text readability, hence it would be an improvement to try to incorporate chromatic contrast and color-texture features in the classification process.

5 Acknowledgments

This work has been supported by Information in Place, Inc.

References

- AZUMA, R., AND FURMAN, C. 2003. Evaluating label placement for augmented reality view management. In *Proceedings of the 2nd IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2003)*, 66–75.
- CHUBB, C., SPERLING, G., AND SOLOMON, J. A. 1989. Texture interactions determine perceived contrast. *Proc Natl Acad Sci U S A*, 86, 23 (Dec), 9631–9635.
- CLARK, M., AND BOVIK, A. 1987. Texture segmentation using gabor modulation/demodulation. *Pattern Recognition Letters* 6, 261–267.
- HILL, A., AND SCHARFF, L. 1999. Readability of computer displays as a function of colour, saturation, and background texture. In *Engineering psychology and cognitive ergonomics*, D. Harris, Ed., vol. 4. Ashgate, Aldershot, UK.
- JAIN, A. K., AND FARROKHNI, F. 1991. Unsupervised texture segmentation using gabor filters. *Pattern Recognition* 23, 12 (December), 1167–1186.
- JULESZ, B. 1981. Textons, the elements of texture perception, and their interactions. *Nature*, 290, 91–97.
- LEGGE, G., PELLI, D., RUBIN, G., AND SCHLESKE, M. 1985. Psychophysics of reading i. normal vision. *Vision Research* 25, 239–252.

- LEGGE, G., RUBIN, G., AND LUEBKER, A. 1987. Psychophysics of reading. v. the role of contrast in normal vision. *Vision Research* 27, 1165–1177.
- MALIK, J., AND PERONA, P. 1990. Preattentive texture discrimination with early vision mechanisms. *J. Opt. Soc. Am. A* 7 (May), 923–932.
- ROSE, E., BREEN, D., AHLERS, K., CRAMPTON, C., TUCERYAN, M., WHITAKER, R. T., AND GREER, D. 1995. Annotating real-world objects using augmented vision. In *Proceedings of Computer Graphics International '95*, 357–370.
- SCHARFF, L., AHUMADA, A., AND HILL, A. 1999. Discriminability measures for predicting readability. In *Human Vision and Electronic Imaging, SPIE Proc. 3644*, 270–277.
- SCHARFF, L., HILL, A., AND AHUMADA, A. 2000. Discriminability measures for predicting readability of text on textured backgrounds. *Optics Express* 6, 4, 81–91.
- SOLOMON, J. A., AND PELLI, D. G. 1994. The visual filter mediating letter identification. *Nature* 369, 395–397.
- TUCERYAN, M., AND JAIN, A. K. 1998. Texture analysis. In *The Handbook of Pattern Recognition and Computer Vision (2nd Edition)*, C. H. Chen, L. F. Pau, and P. S. P. Wang, Eds. World Scientific Publishing Co., 207–248.