Digital Image Processing: Term Project Proposal

Members:
R99222030 林昇慶
D01944015 蔡格昇
A01922201 唐杰

Purpose:
We would like to implement an Optical Character Recognition (OCR) technique for Captchas.

Description of problem:
A Captcha is an image containing characters distorted such that they are difficult to recognize by a machine. A common application is for websites to confirm that a user is human. An example can be seen in Figure 1. The key issue is to separate characters and then correctly identify them.

Figure 1: Examples of several Captchas (Source: http://www.bugtreat.com)

Final Goal:
We want to implement a practical OCR application to recognize text in images containing them. An additional goal is to process Captchas and correctly solve them with our OCR method.

Methodology:
Our method can be divided into two parts, first a *preprocessing* method to generate a recognizable input for the OCR algorithm, second the actual *character recognition*.

Character Recognition. We confine ourselves to English characters. The first step is to isolate single characters from a sentence. We then define a number of features, which we choose by analyzing and experimenting on popular font styles - we want to find common features in the character set. Next, we use a decision tree with these features to classify them and choose the top n most likely letters. In a post-processing step we do a final

decision based on dictionary lookup (assuming our input are valid words).

Preprocessing. To generate a clean input for our OCR method we first process the Captcha. We remove noise in the picture and do other processing steps according to the specific characteristics of the Captcha (which we will decide on later).

Implementation:

Our program will be coded in C++. We will also use OpenGL to display input and output.

Input constraints:

We focus on a few popular fonts, e.g., Arial. An input image consists of a single, existing word part of our dictionary. Furthermore, the alphabet is limited to alphanumeric characters.