

Non-Writer Specific Handwriting Generation for the CAPTCHA Application

Achint Oommen Thomas, Amalia Rusu, Smruthi Mukund and Venu Govindaraju

Abstract

Automated recognition of unconstrained handwriting continues to be a challenging research task. Unavailability of large training datasets adds to these challenges and therefore there is a special interest in generating synthetic “human-like” handwritten samples. In this paper we present a novel synthetic handwriting generator where the generated text lines need to be close to human handwriting. However, the text lines need not be writer-specific. Not only are we using the handwriting generator for generating a large number of samples to improve recognizers' accuracy but also for Cyber security applications such as CAPTCHAs. Our paper proposes a new application of handwriting recognition in design of CAPTCHAs (Completely Automatic Public Turing test to tell Computers and Humans Apart), which can exploit the differential in the reading proficiency between humans and computers when dealing with handwritten text images, so they can be used for human verification for online services.

To be suitable for online applications, we are automatically generating infinitely-many distinct artificial handwritten samples. Various models of human-like writing generation are available in the literature. Most of the existing approaches are on-line based since it is more convenient to change the trajectory and shape of the letters based on the on-line information such as pen-down, pen-up, and velocity profiles. However, the on-line information is not always available and as an alternative, researchers are applying various perturbations directly on real characters or templates. We describe a method for generation of cursive English handwriting samples that uses character templates.

The generation algorithm consists of several steps: i) character auto-scaling, ii) automatic baseline determination, iii) ligature endpoint detection, iv) ligature parameterization, v) ligature joining, vi) skeleton perturbation, and vii) skeleton thickening. We first construct what is known as a preliminary image, which is a concatenation of individual character templates. Character templates are one-pixel wide representations (skeletons) of the original character image. The preliminary image contains individual characters templates strung together to form a word. Since the text lines have to be close to human handwriting style, important aspects like character baseline alignment, scaling of character sizes and ligature joins have to be handled properly. In this paper we present techniques for automatic baseline detection and ligature generation. We also present a character auto-scaling technique. Once we have the preliminary image, we apply a set of geometric perturbations that randomly distort the preliminary images. The perturbations can be parameterized so that the values are picked at random over a range. Finally, the distorted image is thickened. Thickening can be controlled so that different parts of the image are thickened by different amounts. Several examples of final images are shown in Figure 1.

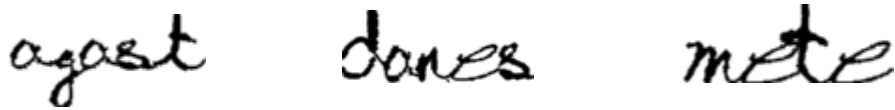


Figure 1. Synthetic handwritten word images (non-distorted)

We have further designed synthetic handwritten word images that exploit the knowledge of the common source of errors in automated handwriting recognition systems and also take advantage of the cognitive aspects of human reading. We have generated synthetic handwriting samples and then applied various transformations to make them unreadable by automatic computer programs, such as adding lines, grids, arcs, circles, background noise, and occlusions, etc, so that they can be used as CAPTCHA challenges over the Internet. We have compared humans to handwriting recognizers' abilities in recognition of a set of synthetically generated word images and have identified the gap that could be used to distinguish between them in online services.

Based on our tests we claim that, using handwritten word images for the CAPTCHA application is an effective approach in differentiating between humans and automated programs on the web. Our ability to generate infinitely many handwritten word images means that we are not limited by a finite size database of CAPTCHA challenge images. Since our technique does not generate writer-specific handwritten word samples, writer-specific training will not work well against our method. The generation approach can be expanded to allow saving of the generated templates in any format suitable for training of character recognition programs based on the recorded pen strokes instead of the scanned bitmaps (e.g. for use in PDAs or Tablet PCs). Many cyber security applications will benefit from our proposed method.