2009 10th International Conference on Document Analysis and Recognition

Pen Acoustic Emissions for Text and Gesture Recognition

Andrew Seniuk and Dorothea Blostein
School of Computing
Queen's University
Kingston, Ontario, Canada
andrew.seniuk@gmail.com, blostein@cs.queensu.ca

Abstract

The sounds generated by a writing instrument provide a rich and under-utilized source of information for pattern recognition. We examine the feasibility of recognition of handwritten cursive text, exclusively through an analysis of acoustic emissions. Our recognizer uses a template matching approach, with templates and similarity measures derived variously from: raw power signal with fixed resolution, discrete sequence of magnitudes obtained from peaks in the power signal, and ordered tree obtained from a scale space signal representation. Test results are presented for isolated lowercase cursive characters and for whole words. Recognition rates of over 70% (alphabet) and 90% (26 words) are achieved, based solely on acoustic emissions, with samples provided by a single writer. We also present qualitative results for recognizing gestures such as circling, scratch-out, check-marks, and hatching. These preliminary results demonstrate that acoustic emissions are a rich source of information, usable – on their own or in conjunction with image-based featuresi – to solve pattern recognition problems. In future work, this approach can be used in applications such as writer identification, handwriting and gesture-based computer input technology, emotion recognition, and temporal analysis of sketches.

1. Introduction

The sounds produced during any activity carry information about what is occurring. In many pattern recognition situations, sound can be recorded conveniently, and can supplement or replace visual sources of information. In this paper, we investigate the acoustic emissions produced by a human writer, analyzing the sounds that are generated by the friction between the writing instrument and surface. These sounds are generally sibilant, giving the informal impression that the acoustic signals are very noisy. However, the experiments we report here demonstrate that these signals have the potential to be quite useful for pattern recognition. The tests are preliminary, based on a single writer, but they provide convincing evidence that pen acoustics provide a promising and so far underutilized source of information for pattern recognition.

The studies we present here focus on the use of pen sounds to address character, word and gesture recognition problems in online handwriting. With straightforward classification methods based only on acoustic information, we obtain respectable recognition rates for isolated characters (over 70% on 26 character classes) and whole words (over 90% on 26 word classes) written in a cursive style. These error rates are too large for a usable stand-alone handwriting recognizer, but they provide evidence that acoustic classifiers could be a strong addition to a multiple classifier system allied with image-based classifiers. Multiple-classifier systems are most effective in the situation when the errors that classifiers make are uncorrelated [6]; this is likely to be the case when acoustic classifiers are combined with image-based classifiers. Also, it is possible that further research could produce audio-based handwriting classifiers that achieve significantly higher recognition rates.

2. Related Work

Handwriting recognition and speech recognition both have a rich research history [1, 9]. To the best of our knowledge, no previous studies have been conducted on the use of pen sounds for handwriting recognition. The use of pen sounds for author detection and verification is discussed in a patent [2] and a publication [10].

Sound-based gesture recognition is discussed in a recent paper by Harrison *et al.* [5]. This work was carried out concurrently with our own, indicating that various researchers are appreciating the potential benefits of sound-based online recognizers. The classifier used by Harrison *et al.* is a decision tree, using features based on peak count and relative amplitude in the gated audio power signal. They use a stethoscope to improve the signal to noise ratio for a contact microphone; this is a technique which we could profitably adopt. They report recognition rates of about 90% for a set

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.251

of six gestures, consisting of single or double taps, and up to four swipes (lines). It is interesting that their recognizer achieves the highest recognition accuracy on the simpler gestures, while our recognizer achieves higher accuracy on more complicated gestures. (As we report in §7, our classification accuracy is higher for words than for isolated characters.) This may be due to differences in the task: it may be that the complicated gestures in the Harrison *et al.* test set are newly learned by the writer, showing greater variability than is observed between instances of the same written word.

A novel input device, the Stane, is described in recent work by Murray-Smith *et al.* [7]. This is an avocado-sized, textured, hand-held controller, which both senses scratching by the fingernails and affords haptic feedback through vibrations. The Stane is demonstrated by the inventors as a controller for an MP3 player. Part of the power of the Stane is that it detects location of the scratching using multiple microphones. Accurate positional information relaxes the need for recognition of a large number of texture varieties, since the same scratched texture can mean different things on different parts of the surface of the device. In this connection it is interesting that the TAI CHI project [8] has developed a method by which a *single* microphone can relay positional information, provided the impulse response of the scratched surface is a known and sufficiently varying function of position. For handwriting applications, we believe microphones would be unfeasible for detecting pen position with sufficient accuracy to support recognition based on position, except under the most exacting conditions.

3. Equipment for Recording Sound

Our apparatus consists of readily available materials. The writing surface is a single sheet of standard 60 lb. HP laserprinter paper attached to a Masonite clipboard. The writing instrument is a Bic Round Stic Grip (fine). The microphone is a Labtec PC Mic 333 with the plastic housing removed, taped to the midpoint of the pen shaft, with the business end of the microphone facing the writing tip. All the results reported in this paper are obtained on data collected using this setup. We have obtained similar performance using other microphone placements and other writing instruments, such as pencils and felt tip pens. Audio was captured in a quiet environment on an Asus Eee PC 701 (Intel HD audio) running Audacity under Ubuntu Linux, at a modest sample rate of 16 kHz, with 8 bits per sample, and saved to a file in WAV format.

4. Data Sets

4.1. Data sets of letters and words

We define two data sets of 26 classes each, to train and test our classification algorithms. The data set of letters, which we denote ABC, consists of isolated characters of the cursive lowercase alphabet. The data set of words, which we denote FOOD, consists of words drawn from a vocabulary of 26 food items. Audio recordings were made as the same writer (Seniuk) wrote each data set eight times. We refer to these recordings as ABC-1 to ABC-8, and FOOD-1 to FOOD-8, and to specific character or word segments as, for example, d-7 or banana-4. The visual trace of the FOOD-6 recording is shown in Figure 1. In both the ABC and FOOD data sets, the text is written without dotting the i's and j's, and without crossing t's and x's.

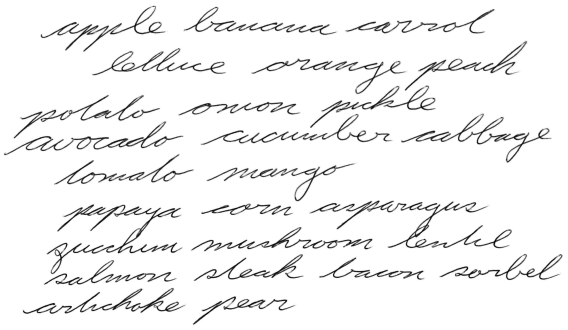

Figure 1. The visual trace of the writing from which the audio samples in the FOOD-6 data set were recorded.

4.2. Data sets of gestures

The set of gestures we study includes scratch-out, circle (single and double), check and ex, dots, and hatch. These gestures are chosen for their familiar use in handwritten documents, and are illustrated in Figure 5 along with their audio signals. We discuss these signals qualitatively in §7.3.

5. Preprocessing

A signal is preprocessed by smoothing, amplitude normalization, segmentation, and timescale normalization. Optionally, a nonlinear distortion step is used to create test data.

In the smoothing step, Gaussian smoothing is applied to the modulus (absolute value) of the audio signal. The resulting signal is non-negative and represents the power of the original audio signal local to the Gaussian. The smoothing occurs over a time interval of roughly 10ms. More precisely, the width of the Gaussian is chosen so that 95% of the area under the Gaussian falls within 10ms of the sample. Additional smoothing at coarser scales is used in the scale space algorithm described in Section 6.3.

Amplitude normalization rescales the power signal to achieve unit amplitude at its maximum.

In the segmentation step, the signal is semi-automatically segmented into constituent letters or words. Our segmentation algorithm detects the relatively

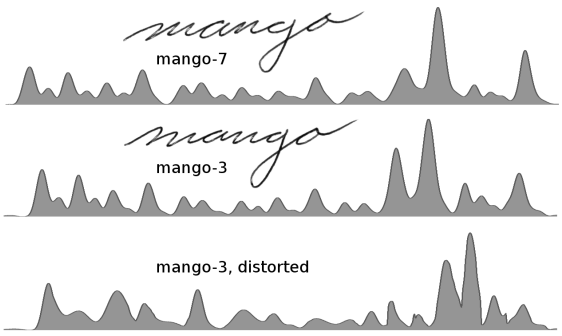

Figure 2. Typical pen acoustic emission power signals. The bottom one is corrupted by synthetic nonlinear distortion.

long silences that occur when the pen is lifted. Manual correction is applied in a few cases where undersegmentation occurs due to ambient noise. Development of a more sophisticated segmentation algorithm is planned as future work.

In timescale normalization, each segment is normalized to span one time unit.

The optional nonlinear distortion step distorts both the time and the amplitude of the signal. The distorted signals are used as test data for evaluating the robustness of the classification algorithms. The time axis is distorted by inserting and deleting samples, according to a probability density that varies smoothly with time. The amplitudes are distorted by applying a smooth perturbation function. An example is shown in Figure 2 for two renditions of the word 'mango'.

6. Classification Algorithms

We investigate three approaches for comparing the audio power signals: integrating the absolute difference between two signals (§6.1), comparing peaks in the two signals (§6.2), and comparing the structures obtained from scale space representations [11] of the signals (§6.3). All three classification algorithms are based on some form of template matching, with templates constructed from the training data. Each test sample is classified according to its top-matching template.

6.1. Subtracting the power signals

The first, and simplest, classification algorithm measures the similarity of two signals by integrating the absolute value of the difference of the signals. The training data (after preprocessing, as described in §5) is used directly as a set of templates. Similarity of a test signal to a template is then defined as the integral of the absolute value of the difference between the two timespan-normalized and amplitude-clipped signals. An integral of zero arises when the two signals are identical, and larger values indicate lesser similarity.

6.2. Comparing peaks in the power signals

As can be seen in Figure 2, the signals have distinctive peaks (maxima). Suppose $\{p_i\}_{i=1}^{M}$ and $\{q_j\}_{j=1}^{N}$ are the sequences of peaks for signals P and Q, where here a peak is encoded as a pair $p_i = (t_i, h_i)$ of offset in time t_i to the crest of the peak, and height h_i at the crest. Our similarity measure is defined as the edit distance between these sequences. Edit distance $d(P, Q)$ is the least cost among all sequences of edit operations transforming P into Q, with the cost of a sequence being the sum of the costs of its operations. We allow the edit operations usual in edit distance methods: substitution, insertion, and deletion. The cost of a substitution of an existing peak for a new peak is defined to be the product of the timeshift with the difference in height. The two factors are weighted equally, given that the power signal has been normalized in both time and amplitude dimensions in the course of preprocessing. The cost of deletion, as well as of insertion, is the height of the peak in question. (The use of peak integrals instead of heights is expected to improve the results, and is planned for future work.) The global cost minimization over all edit sequences has an efficient dynamic programming solution.

6.3. Scale space representation

The algorithms described so far operate on input that is smoothed at a fixed resolution (§5), but it is difficult to choose an effective global resolution, since the various features of the input will stand out best at different scales of smoothing. Here we describe an algorithm that avoids the dependence on a window size, instead using structure and quantities derived from a scale space representation [11] of the signal.

The finest resolution of the scale space is formed by the preprocessed audio signal; as described in §5, this uses Gaussian smoothing in which 95% of the Gaussian lies within 10ms of the sample. All maxima and minima are identified at the finest resolution. Then the scale is increased stepwise until the power signal is a unimodal swell, having only a single maximum and no minima. Power signals at intermediate scales are generated, so that the positions of the maxima and minima trace out coherent arch-like structures as seen in Figure 3.

We define a similarity measure based on edit distance between the sequences of ridges traced by the maxima (the dark halves of the arches in Figure 3). We also explored tree edit distance [3] based on trees derived from the nested structure of the arches, but we found that this hierarchical structure can vary significantly between samples of the same word, as can also be seen in the figure. Better results

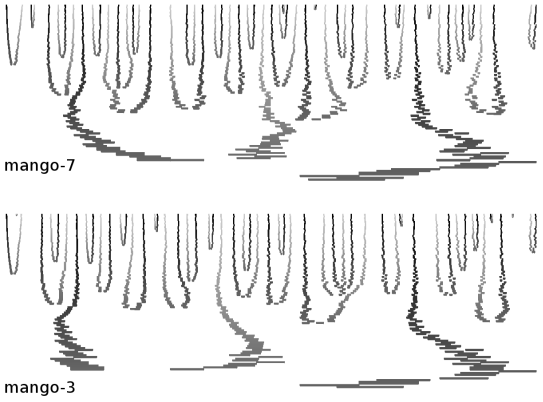

Figure 3. This figure depicts the scale space representations for the two audio traces of the handwritten word 'mango' shown in Figure 2. A nested structure emerges as pairs of maxima (dark) and minima (light) of the power signal are tracked through increasing scale.

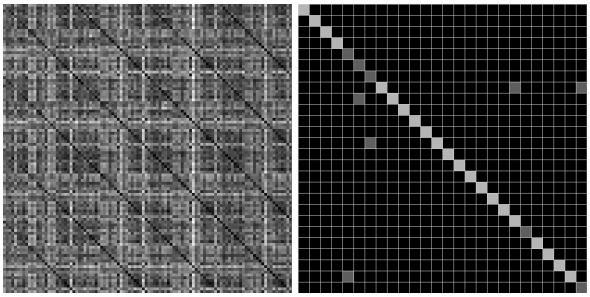

Figure 4. The similarity matrix on the left summarizes waveform comparisons, with test waveforms (FOOD-1,3,5,7) forming the horizontal axis and training data (FOOD-2,4,6,8) forming the vertical axis. Dark values indicate highly-similar waveforms; dark diagonal streaks arise when the recognition rate is high. The confusion matrix on the right shows correct classifications along its principal diagonal and errors in the off diagonal areas.

were obtained by ignoring the nested structure, basing the similarity measure on just the time positions and scale extents of the maxima ridges.

7. Results

We discuss the performance of our recognizers using summary statistics (percent correct, percent error, percent reject). This is followed by more detailed analyses of selected cases, using similarity and confusion matrices.

7.1. Classifier performance on cursive characters and words

The summary statistics in Tables 1 and 2 show that recognition rates are higher for the FOOD data set than for the ABC data set. This is not surprising, since the individual characters in the ABC data set are brief, minimal gestures. In contrast, the FOOD data set consists of whole words, which offer longer audio recordings with greater classification information. The best recognition rate for the FOOD data set was roughly 95%, achieved by both the function difference algorithm (§6.1) and the peak comparison algorithm (§6.2). For the ABC dataset the best recognition rate was about 70%, using the function difference method. The results in Table 1 summarize algorithm performance, under the fixed conditions as outlined in §5. and both in the absence and presence of the synthetic noise in the test inputs. We use test sets of size 104; according to Figure 9.10 in [4], this results in a performance estimate that is expected to be accurate to $\pm 8\%$. Use of rejection was tested, both by voting and by maxmium permissable dissimilarity. This resulted in a greater proportional reduction in error rate than in recognition rate. Possibly with larger datasets this approach will prove worthwhile. Our current results (Tables 1 and 2) suggest that the peak comparison algorithm outperforms the naive function difference algorithm when the test set is subjected to our nonlinear distortions, and that both outperform the scale space methods in every case. We will test this more rigorously in future work with larger datasets and with multiple writers.

7.2. Similarity and confusion matrices

Similarity and confusion matrices permit a more detailed analysis of classifier performance. Figure 4 shows an example. This data is for the function difference based classifier, using FOOD-2,4,6,8 for training, and FOOD-1,3,5,7 for testing. Analogous similarity and confusion matrices

Table 1. Performance with the FOOD-n Data Set

Algorithm		% Correct	
		Clean	Distorted
(§6.1) Signal Subtraction		93	27
(§6.2) Peak Comparison		95	46
(§6.3) Scale Space	Ridges	89	
	Trees	61	

Table 2. Performance with the ABC-n Data Set

Algorithm		% Correct	
		Clean	Distorted
(§6.1) Signal Subtraction		70	10
(§6.2) Peak Comparison		55	20
(§6.3) Scale Space	Ridges	30	
	Trees	35	

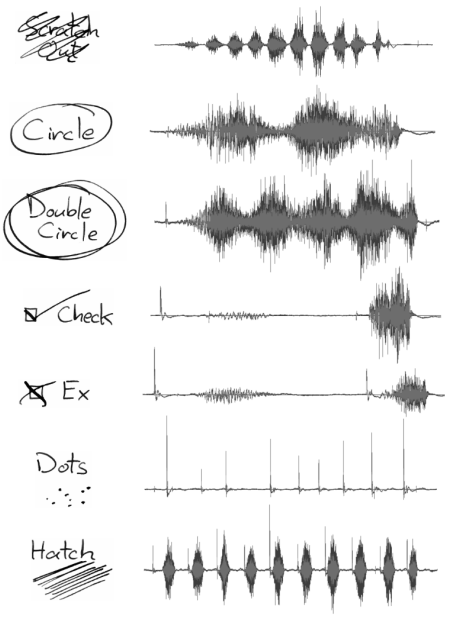

Figure 5. Simple gestures, with audio.

are available for all of the results reported in the tables. In the confusion matrix in Figure 4, the error nearest the diagonal (for example) shows that 'peach' was mistaken for 'pickle' once, of the four 'peach' samples tested. Analysis of confusion matrices for the ABC data sets shows that 'e', 'i', 'l', 't' and 'x' are frequently confused.

7.3. Some common gestures

Our observation of scratch-out gestures in our recorded samples gave strong experiential evidence of their ease of recognition: the scratching has a regular picket of peaks (Figure 5), and moreover the amplitude is typically double that of the highest peaks in normal writing. The actual amplitude of a scratch is proportional to its length, as can also be seen in Figure 5. The impulse-like character of dots, and generally of moments of first contact between pen and paper, are also evident: sharply defined, tall lines in the signal.

8. Acknowledgements

Financial support from Canada's Natural Sciences and Engineering Research Council in gratefully acknowledged.

9. Conclusions

Handwriting is traditionally analyzed based on image data. We propose that audio information provides a useful alternative source of information. We demonstrate that an inexpensive microphone can be used to record the sound created by friction between the writing stylus and the writing surface, and that this data can be effectively utilized for classification. We describe classification algorithms based on template matching, with templates obtained from peaks in the acoustic power signals, and present recognition results on an alphabetic and a word database. These promising, preliminary results point to rich future possibilities in applying sound analysis to applications such as writer identification, handwriting and gesture based computer input technology, emotion recognition, and temporal analysis of sketches.

References

[1] R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, and V. Zue, editors. *Survey of the State of the Art in Human Language Technology*. 1996.

[2] P. E. Cox. Personal handwriting verification. *British Patent No. GB2159998*, 1985.

[3] E. D. Demaine, S. Mozes, B. Rossman, and O. Weimann. An optimal decomposition algorithm for tree edit distance. In *In Proceedings of the 34th International Colloquium on Automata, Languages and Programming (ICALP*, pages 146–157, 2007.

[4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley and Sons, Inc, New-York, USA, 2001.

[5] C. Harrison and S. E. Hudson. Scratch input: creating large, inexpensive, unpowered and mobile finger input surfaces. In *UIST '08: Proceedings of the 21st annual ACM symposium on User interface software and technology*, pages 205–208, New York, NY, USA, 2008. ACM.

[6] L. I. Kuncheva. Diversity in multiple classifier systems. *Information Fusion*, 6(1):3–4, 2005.

[7] R. Murray-Smith, J. Williamson, S. Hughes, and T. Quaade. Stane: synthesized surfaces for tactile input. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1299–1302, New York, NY, USA, 2008. ACM.

[8] D. T. Pham, Z. Ji, M. Yang, Z. Wang, and M. Al-Kutubi. A novel human-computer interface based on passive acoustic localisation. In J. A. Jacko, editor, *HCI (2)*, volume 4551 of *Lecture Notes in Computer Science*, pages 901–909. Springer, 2007.

[9] R. Plamondon and S. N. Srihari. On-line and off-line handwriting recognition: A comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(1):63–84, 2000.

[10] Y. Sato and K. Kogure. Online signature verification based on shape, motion, and writing pressure. In *In Proceedings of the 6th Int. Conf. Pattern Recognition*, pages 823–826, 1982.

[11] A. P. Witkin. Scale space filtering: A new approach to multiscale description. In S. Ullman and W. Richards, editors, *Image Understanding*, chapter 3, pages 79–95. Ablex, Norwood, N.J., 1984.