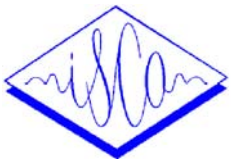
ISCA Archive
http://www.isca-speech.org/archive



SPECOM'2004: 9th Conference
Speech and Computer
St. Petersburg, Russia
September 20-22, 2004

Integration and Fusion Aspects of Speech and Handwriting Media

Sascha Schimke, Thomas Vogel, Claus Vielhauer, Jana Dittmann

Department of Computer Science, ITI Research Group on Multimedia and Security,
Universitätsplatz 2, 39106 Magdeburg, Germany
{sascha.schimke, thomas.vogel, claus.vielhauer, jana.dittmann}@iti.cs.uni-magdeburg.de

Abstract

In this paper we discuss synchronization approaches for fusion of speech and handwriting data on a signal representation level. There are many advantages in utilizing additional modalities to speech, for example bimodal signals have the potential of increasing accuracy of recognition systems. Further we intend to provide users more flexibility for human to computer communication by allowing them to choose their preferred modality. After discussion of goals, we analyze different ways for synchronization of media streams. Besides approaches based on synchronized time stamp protocols as additional metadata, we dwell on a concept for synchronization based on embedding the data stream of one modality into the other by using digital watermarking techniques. Here we introduce the general concept of direct embedding and analyze the necessary watermarking capacity (payload) for synchronization. Finally we have a look at aspects of information retrieval in multimodal documents.

1. Introduction

We use the term *modality* in this paper as human to computer interface *input channel*. Examples for modalities in this sense are human voice, handwritten inputs, gestures or keyboard typing. A multimodal interface is any combination of more than one input channel. We will concentrate on modality combinations consisting of spoken and handwritten (as well as hand sketched) inputs.

Giving a user the possibility to enter her inputs with more than one modality offers her a more seamless and flexible handing of the machine. That is the way, persons communicate with each other; most things are spoken but some pieces of information are expressed using script (in particular terms that are not easy to pronounce, like foreign names) or by drawing sketches. A possible application scenario, which regards user's desire for free selection of input modality, is the form filling. In form filling applications, the system presents the user a form (e.g. questionnaires, transaction orders or medical forms) with fields of different types and the user has to fill these fields. In section 3.1 we will describe the form filling scenario more detailed.

Besides the scenario of form filling applications, there are scenarios of multimodal document gathering. These scenarios are tangent to the research area of information retrieval. Examples for multimodal document gathering are *making handwritten notes while an oral presentation* (like a lecture or a conference talk) or *writing minutes in an interrogation or an interview*. In these examples, both of the modalities (handwriting and speech) are used by different subjects. Of course, in the presentation scenario, also the presenter could produce written information, e.g. by using an electronic whiteboard. In section 3.2 and 3.3 we give a detailed discussion of both example scenarios. Aspects of retrieval in multimodal documents are the topic of section 4.

2. Recording and Synchronization of Multimodal Signals

In this section we describe the different techniques for the recording of speech and handwriting input signals. A further part of this section covers the discussion of aspects of synchronization of these signals.

2.1. Speech Signals

As every audio signal, even speech is recorded by using audio sampling hardware like soundcards. "*Sound is produced through vibration of matter, which creates pressure variations in the surrounding matter (usually air). The vibrations generate a waveform […] In order to represent the sound waveform digitally it is sampled by an analog-to-digital converter.*" [1]

The sampled signals can be saved, among other ways, as a sequence of sampled values or as a sequence of coefficient of frequency domain representation of the audio signal. A lot of file formats for audio data is nowadays available, each one with its own advantages and disadvantages. Examples for widely spread audio formats are wav and mp3.

2.2. Handwriting Signals

While the audio domain is well investigated and there are some standard file formats for audio data, the domain of handwriting is nowadays not that popular. Although the research on handwriting signals started some decades ago [2][3], only for a short time the successes in this domain allow the adoption of handwriting techniques in real-user applications [4].

The gathering of handwritten inputs is mainly differentiated between so called off-line and on-line methods. The basis of off-line processing is a two dimensional image of the handwriting input. Therefore a sheet of paper with handwritten words is scanned. In contrast to

this, the basis of on-line processing is a sequence of successional pen tip positions. So the on-line handwriting data is something that can be called digital ink [5]. There are some different types of devices for recording on-line handwriting data.

- The most popular devices consist of an electronic writing surface and a special active pen. The writing surface can sense the position of the pen tip and in some cases the pen has a sensor for the pressure onto the surface. Examples for this kind of devices are graphic tablets or TabletPCs. In a TabletPC the writing surface is embedded into the screen, so the user can directly see her input displayed on the screen.

- Another wide spread type of handwriting device consists of a pressure sensitive surface, so any kind of pen can be used for input. The major advantage of these devices is, that a sheet of paper can be placed on the surface and a traditional pen can be used to write. Examples for this device type are handheld computers (PDA – *personal digital assistant*) or the CrossPad [6].

- A special case of handwriting devices are active whiteboards. These are large boards as known from presentations, which have the ability to collect the pen movement data. These electronic whiteboards often use special pens, but some devices consist of a sensitive surface, so any pen can be used [7][8].

- Besides these two device types, there are handwriting devices, which consists only of an active pen, that is able to decide its position by sensing its environment. For example, the io Pen [9] has a scanner to read puniest markers on a special sheet of paper to decide its location.

All these handwriting devices have the common ability to acquire pen tip position data (x, y). Most of them additionally have sensors for the pen pressure and some also can sense the pen inclination. Except of a few kinds, all devices give a time index for each set of position, pressure and inclination data. Dependant on the device, these data is acquired with a more or less high frequency.

To save a handwriting signal, the acquired tuples of position, pressure, inclination data and time index are written into special formats. There are different file formats (e.g. InkML [10], Jot [11] or Unipen [12], as well as some proprietary formats [13][14][15]) but until now, none of them became widely accepted.

2.3.Synchronization

One aspect of multimodal signals is, that the single signals can be produced (by one or more user) at the same time (synchronously) or in alternating manner. If, for example, speech and handwriting signals, whose contents are equivalent, are recorded at the same time, then the speech recognizer could use the results of the handwriting recognizer to disambiguate its results and vice versa.

A method to preserve this simultaneity of different recorded signals has to be found, to exploit it for further processing steps like recognition or information retrieval. Having such synchronization, it is possible to determine the order of certain *events* (spoken or written words, or other kinds of input) in the signals. The most intuitive idea is to save a time index along with the speech and the handwriting signals, so for every sampling point in each of the signals the time of occurrence can be determined. Since all audio file formats have an implicit time index and since even most digital ink file formats ([10][11][12]) save the time index of samples (if this time index information is available from the input device), this time index synchronization method is a possible solution. The disadvantage of this method is the necessity for handling different files; one file for each input signal.

To overcome this disadvantage, a solution is to store the data of all signals into one single file. This could be done by defining a special multimodal file format, which is able to handle several signals of different characteristics – in most cases, e.g. the sampling rates of speech and handwriting signals is complete different. Another idea for storing audio signals along with handwriting signals could be to use a second channel in the audio file format (stereo) and dump the handwriting signal into this channel. If there are numerous handwriting signals (see sections 3.2 and 3.3), this approach of *misusing* an audio channel is not applicable.

Another idea for synchronous storage of audio and handwriting signals is to embed one of the signals into the other by using watermarking techniques. The easiest watermarking technique is the embedding into the LSB (*least significant bit*) of sampled audio data [16]. Table 1 shows, how many bytes of information can be embedded into audio (.wav) files with different sampling rates by using LSB watermarking.

Table 1: LSB watermarking capacity of mono channel audio files in wav-format.

Sampling rate	Capacity per seconds
48 kHz	48,000 bits = 6,000 bytes
44.1 kHz	44,100 bits ≈ 5,512 bytes
32 kHz	32,000 bits = 4,000 bytes
22.05 kHz	22,050 bits ≈ 2,756 bytes
16 kHz	16,000 bits = 2,000 bytes
11.025 kHz	11,025 bits ≈ 1,378 bytes
8 kHz	8,000 bits = 1,000 bytes
6 kHz	6,000 bits = 750 bytes

High quality handwriting sampling devices have a resolution of 1,000 “pixels” per cm (px/cm) in x- and y-direction, can differentiate 1,024 steps of pen tip pressure and have a sampling rate of 200 Hz [17]. So while recording of handwritten input on a tablet with the size of A4 (round about 21x30 cm), for each second 8,000 bits accumulate (8,000 = 200Hz × (10bit + 15bit +

15bit) — 1,024 pressure steps: 10bit, 21cm × 1,000px/cm: 15bit, 30cm × 1,000px/cm: 15bit). As we can see in table 1, these 8,000 bits fit well in nearly every common sampling rate, if LSB watermarking is used. For good quality speech recording, the sampling rate should be higher than 8 kHz (which is the sampling rate in telephone networks). Additionally, the amount of handwriting data can be reduced by coding only the position and pressure differences, instead of the absolute sampled values [14]. Moreover, in most cases a spatial resolution of handwriting signals less than 1,000 px/cm and sampling rate less than 200 Hz is acceptable, so the amount of data further can be reduced.

3. Multimodal Application Scenarios

In this section we describe different examples for multimodal applications. The first example application is a user interface, while the second and third ones are the basis for multimodal information retrieval.

3.1. Form Filling

The first scenario for multimodal applications, we describe, is the filling of forms by speaking or writing. The filling of forms is a usual action in the office. The classical way was to fill a sheet of paper using a pen. Since, on the one hand, nowadays computers are nearly everywhere available and many documents are produced electronically, even many forms can be filled in that way. But on the other hand, not every subject is familiar with computers. It seems to be a good idea, to change the human to computer interface, so that the user is able to interact with the machine in a more convenient way. Nearly everyone is familiar with handwriting, since this is one of the first lessons, persons learn in school. The same is true for the modality of speech; nearly everyone is able to speak. So handwriting and human voice could have the power to be the basis of a natural user interface. To be helpful, these user inputs have to be processed by a recognizer engine.

Especially form filling applications are appropriate for speech and handwriting interfaces, since in most cases the user input is limited to a small set of possible inputs, since most forms consist of fields of definite types. Common types of fields are numbers, names, dates, amounts of items or money, but there are many more types, of course. The recognizer algorithm can profit by information about the type of field [18][19]. Of course, in case of free text fields, this is not easy possible.

Having a multimodal system, the user can select her preferred input modality for each form field. For complicate names or words, most users will probably prefer to write these inputs into the respective fields, while for example with numbers they could decide to enter them in the oral way. Another reason for choosing the one or the other modality could be the environment; having no hand free for writing, the user would prefer to make inputs via voice while in a noisy environment she would use instead the pen [20].

3.2. Interrogation or Interview Scenario

In situations of interrogations or interviews, primary one person is speaking while another one writes notices, minutes or a word-by-word protocol [20]. At least in case of a word-by-word protocol, the recognizing algorithms for the speech and for the handwriting signals could exploit the semantic redundancy of the bi-modal input to enhance their recognition quality (see section 2.3). If the pen signal is available while writing, this signal could be immediately embedded into the audio signal, as described in section 2.3.

Besides enhancement of recognizer results, multimodal information, recorded in an interrogation or interview scenario, as well as in a presentation scenario as described in the next subsection, could be the basis for document retrieval applications (see section 4).

3.3. Presentation Scenario

While presentations, lectures or conference talks, the normal case is, that one person is speaking. Often some persons in the auditorium make their own notes in a written form. Besides this, also the presenter or lecturer sometimes produces writings e.g. on black- or whiteboards. If all these information streams (the speech of lecturer and the several handwriting streams of listeners and the lecturer) are recorded in digital form, they establish a multimodal scenario. Landay and Davis discuss in [21] a related scenario, but without the speech signal. They describe an experiment for shared note taking while meetings and conference talks. Their goal is to automatically assemble minutes based on notes of different persons.

In our presentation scenario, the lecturer should use a microphone (as usual in large lectures or talks, anyway) to record her voice. Instead of using a classical black- or whiteboard, comments, sketches and outlines shall be written or drawn on an electronic whiteboard. An alternative solution is to use a pen based computer like a TabletPC that is connected to a projector/beamer (as used for slide presentations). Equipped with these devices, the lecturer is able to write or draw on her display and the auditory can see her writing or drawing.

To record the handwriting signals of the persons in the auditory, they have to use any pen based input device, e.g. even a TabletPC or a PDA (see 2.2). Since in most cases there are more than just some handwriting signals of the different persons in the auditory, it would not be possible to use watermarking techniques for synchronous storage of writing into audio signals. So for this reason, another method has to be used for storage and preserving the time synchronization information (see 2.3).

Of course, persons in the auditory may have caveats against this publication of their personal notes, but on the one hand, everyone could have a benefit of the set of

notes of different persons and on the other hand, it could be enabled to set a flag for notes, which are for private use only, so these notes will not become part of the public multimodal document, as mentioned in [21].

4. Multimodal Information Retrieval

With the onward spread of pen enabled devices like TabletPCs, more and more handwritten documents will be produced [5], especially if scenarios, as those described in sections 3.2 and 3.3, increase in popularity. To make information in these documents accessible, the development of potent retrieval techniques becomes essential.

The goal of information retrieval, as we use this term, is to search for documents with special features. These features for example can be the occurrence of keywords or the fact that the respective document concentrates on a special topic.

The first idea for information retrieval in multimedia data like speech and handwriting, one could consider, is to use speech and handwriting recognizers to convert the contents of the different inputs to ASCII. These ASCII documents then would be used for classical text retrieval techniques. This approach indeed has been put into practice, but it has the problem of imperfect recognizers [5][22]. Until now no automatic system exists, which is able to transcribe human speech or handwriting to text without recognition errors. But there are approaches, nevertheless to use the results of these recognitions with errors as a basis information retrieval [21][23].

At least for the handwriting and handdrawing modality, a solution could be to use techniques like QBE (*query by example*), as mentioned in [5][24][25]. The main idea behind QBE for handwritten inputs is, to formulate the document retrieval query by using a pen. In this case, written words or drawn sketches are compared with those in documents to find matches.

An interesting feature of multimodal documents as described in sections 3.2 and 3.3 is that it is satisfactory to find matches of keywords only in one of the single modalities. So, the more single modality streams are available in a document, the higher is the chance to find this document. In the following, this will be explained with an example.

Assume there are a lot of lectures in digitized form, consisting of recorded speech stream of the lecturer as well as several channels of on-line recorded handwriting signals, originated by lecturer using an electronic whiteboard and by different persons from the auditory, using TabletPCs or PDAs. It's not suitable to listen to all speech audio files and read all handwritten notices in order to find special information from one lecture. If the modalities are synchronized and a search system for keywords and handwritten queries is available, then information retrieval in all modality streams should not only find the relevant lecture but furthermore the concrete position of occurrence in the audio stream and all notices, written at the same time, as well.

5. Conclusions and Future Work

We have discussed new approaches for multimodal applications and proposed techniques for synchronizing signals of speech and handwriting modality. Furthermore we addressed the issue of information retrieval in multimodal documents. The next steps have to be to implement prototypic these applications and test them in a real environment.

6. Acknowledgement

This work has been partly supported by the EU Network of Excellence SIMILAR (Proposal Reference Number: FP6-507609). The contents of this publication are the sole responsibility of the authors and can in no way be taken to reflect the views of the European Union.

7. References

[1] Mauthe, A., and Thomas, P., Professional Content Management Systems, John Whiley & Sons, Ltd, 2004.

[2] Earnest, L.D., "Machine Reading of Cursive Script", IFIP Congress, pp. 462-466, Amsterdam, 1963.

[3] Burr, D.J., "Designing a Handwriting Reader", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 5, pp. 554-559, 1983.

[4] Plamondon, R., "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 1, 2000

[5] Jain, A.K., and Namboodiri, A.M., "Indexing and Retrieval of On-Line Handwritten Documents", *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR)*, 2003.

[6] CrossPad, A.T. Cross Company, http://www.cross.com/

[7] Elrod, S., Pier, K., Tang, J., Welch, B., Bruce, R., Gold, R., Goldberg, D., Halasz, F., Janseen, W., Lee, D., McCall, K., and Pedersen, E., "Liveboard: a large interactive display supporting group meetings, presentations, and remote collaboration", *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 599-607, Monterey, 1992.

[8] Virtual Ink Corporation, http://www.mimio.com/

[9] Logitech Inc., http://www.logitech.com/

[10] W3C – Ink Markup Language (InkML), http://www.w3.org/TR/InkML/

[11] JOT – A Specification for an Ink Storage and Interchange Format, http://hwr.nici.kun.nl/unipen/jot.html

[12] Unipen Version 1.0 format, http://hwr.nici.kun.nl/unipen/unipen.def

[13] SVC 2004: First International Signature Verification Competition, http://www.cs.ust.hk/svc2004/
[14] Ferri, L.C., Mayerhöfer, A., Frank, M., Vielhauer, C., and Steinmetz, R., "Biometric authentication for ID cards with Hologram Watermarks", *Proceedings of SPIE*, pp. 629-640, San Jose, 2002
[15] Pen&Internet, http://www.ritemail.net/
[16] Dittmann, J., Digitale Wasserzeichen. Grundlagen, Verfahren, Anwendungsgebiete, Springer-Verlag, 2000.
[17] Wacom Intuos 2, WACOM Europe GmbH, http://www.wacom-europe.com/
[18] Seni, G., Rice, K., and Mayoraz, E., "Online Handwriting Recognition in a Form Filling Task – Evaluating the Impact of Context-Awareness", *Proceedings of SPIE-IS&T Electronic Imaging*, Vol. 5296, 2004.
[19] W3C – Speech Recognition Grammar Specification Version 1.0, http://www.w3.org/TR/2004/REC-speech-grammar-20040316/
[20] Vielhauer, C., Schimke, S., Thanassis, V., and Stylianou, Y., "Fusion Strategies for Speech and Handwriting Modalities in HCI", submitted to *SPIE EI 2005 – Conference on Multimedia on Mobile Devices*.
[21] Landay, J.A., and Davis, R.C., "Making sharing pervasive: Ubiquitous computing for shared note taking", *IBM Systems Journal*, Vol. 38, No. 4, 1999.
[22] Foote, J., "An Overview of Audio Information Retrieval", *Multimedia Systems*, Vol. 7, No. 1, pp. 2-10, 1999.
[23] Nielsen, J., Phillips, V.L., and Dumais, S.T., "Information Retrieval of Imperfectly Recognized Handwriting", http://www.useit.com/papers/handwriting_retrieval.html
[24] Aref, W.G., Kamel, I., and Lopresti, D.P., "On Handling Electrinc Ink", *ACM Computing Surveys*, Vol. 27, No. 4, pp. 564-567, 1995.
[25] Lopresti, D., and Tomkins, A., "On the Searchability of Electronic Ink", *Proceedings of the 4th International Workshop on Frontiers in Handwriting Recognition*, 1997.