Martín-Albo Simón, D.; Romero Gómez, V.; Vidal Ruiz, E. (2015). Escritoire: A Multi-touch Desk with e-Pen Input for Capture, Management and Multimodal Interactive Transcription of Handwritten Documents. En Pattern Recognition and Image Analysis. Springer. 471-478. doi:10.1007/978-3-319-19390-8_53.

Escritoire: a Multi-Touch Desk with e-Pen Input for Capture, Management and Multimodal Interactive Transcription of Handwritten Documents

Daniel Martín-Albo, Verónica Romero, and Enrique Vidal

{damarsi1,vromero,evidal}@prhlt.upv.es
Pattern Recognition and Human Language Technology Research Center
Camí de vera, 46022, Valencia, Spain

Abstract. A large quantity of documents used every day are still handwritten. However, it is interesting to transform each of these documents into its digital version for managing, archiving and sharing. Here we present Escritoire, a multi-touch desk that allows the user to capture, transcribe and work with handwritten documents. The desktop is continuously monitored using two cameras. Whenever the user makes a specific hand gesture over a paper, Escritoire proceeds to take an image. Then, the capture is automatically preprocesses, obtaining as a result an improved representation. Finally, the text image is transcribed using automatic techniques and finally the transcription is displayed on Escritoire.

Keywords: Multimodal Interaction, Handwritten Text Recognition, Multi-Touch Desk, User Gestures

1 Introduction

Although digital data are increasingly more widely used, many documents are still on paper. Ideally, those documents should become accessible as a machine-readable text for searching, browsing, and editing. To bridge the gap between the *analog* and the digital, paper handwritten documents need to be captured [9] and transcribed.

This paper presents Escritoire, a system that takes a step into this direction. The user works in a digital desk environment that combines the advantages of paper documents and the digital world, allowing an intuitive, natural and comfortable way to annotate, modify and work with both paper and digital documents in a seamless manner.

This desk is continuously monitored using two cameras. The first one allows Escritoire to perform high-resolution scans. The second one is used by the built-in gesture recognizer. Escritoire automatically preprocesses the captured images, obtaining an adequate representation for the subsequent steps. Finally, if the document is handwritten, a multimodal interactive transcription is carried out using a tablet where the user interacts with the comfort provided by a high-resolution e-pen.

The major challenges in designing and implementing such system are: real-time performance, accurate detection of documents, reliable detection and interpretation of the user gestures, preprocessing and layout analysis of camera-based captured handwritten documents, interactive transcription and customized interfaces design.

Below we comment on prior works that bear direct relevance with the different research areas involved in the digital desk presented here.

2 Related Work

There is a huge body of research on capture of paper documents. In [9] a survey regarding the state-of-the-art on the document capture and detection is presented. In [8] a prototype where a capture is triggered by some pointing gestures and performed using a consumer camera is presented. Unfortunately, this system can only deal with printed documents, for which transcriptions can be carried out using OCR.

User interface design is not an easy task, mainly because designers do not tend to follow any strict rule-based procedure. Several studies on user-friendly interfaces, have been carried out. In [15] a set of shortcomings in current user interfaces along with some guidelines are presented. Given that design involves personal stylistic preferences, a system that applies an adaptive algorithm to interface design is described in [3].

Some important work has been carried out [10] on detecting and interpreting user gestures. The applications of gesture recognition are manifold: sign language, medical rehabilitation, virtual reality, etc. In [14] a survey on gesture recognition is provided with particular emphasis in hand gestures and facial expressions. Applications involving Hidden Markov models, particle filtering and condensation, finite-state machines, optical flow and skin color are discussed in detail. Existing challenges and future research possibilities are also highlighted. In [19], the requirements of hand-gesture interfaces and the challenges in meeting the needs of various application types are described. Moreover, in [20] ideas for extracting user behavior and intentions from objects and gestures are presented.

With respect to handwritten text recognition, the results provided by automatic text recognition technologies have improved dramatically in recent years, although results are still far from being perfect. In [17], a multimodal interactive scenario where the user and the system collaborate to generate a better solution was presented for handwritten text recognition (HTR). The situation in layout analysis and text line segmentation is rather similar. The segmentation quality does not reach the acceptable levels needed by end-user applications [11,6].

3 Interacting with Escritoire

Here the system is presented by following Alice while she uses it. Alice wants to transcribe a handwritten document. So she places the sheet in the capture zone (Figure 1a) and points with her index finger the *capture* button (Figure 1b). Escritoire captures and preprocesses the document, generating a better representation. Escritoire automatically detects that the document is handwritten and proceeds to transcribe it. After this, Alice decides to save the document for now. She aims with her index finger at the digital document and moves it into one of the available folders (Figure 1c).

Sometime later, Alice realizes that the system proposed transcription was not entirely correct. So she aims with her index finger at the folder to open it and extracts the saved transcribed document. Alice points at the Tablet button (Figure 1b) to send

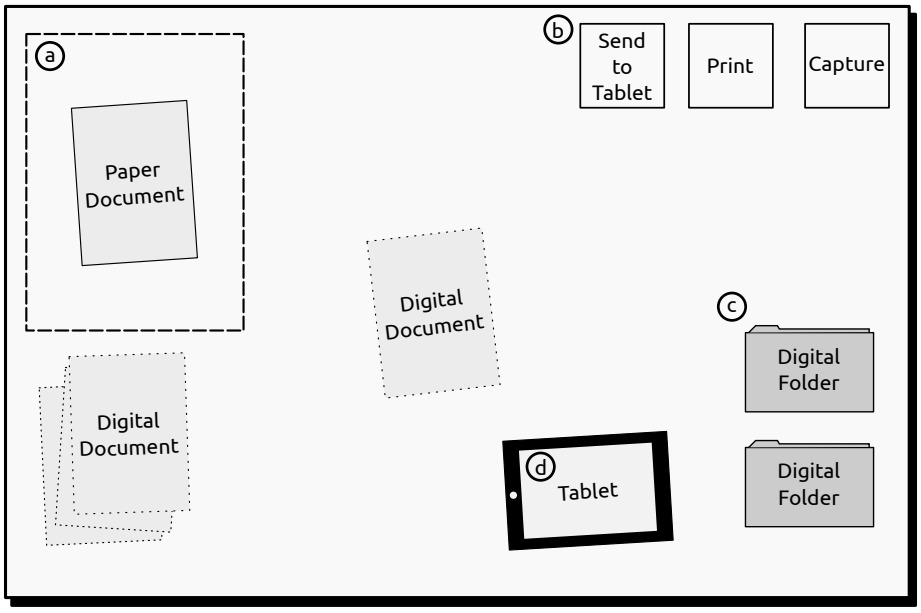


Fig. 1. Escritoire user interface mock-up. a) Capture zone where a *real* document is placed to be digitized. b) Action buttons: The first button, labeled on the real interface with a tablet icon, allows the user to transfer a transcribed document to the tablet to improve the transcription. The second one, tagged with a printer icon, allows the user to print a copy of a digitized document. Finally, by pressing the button tagged with a camera and after a five-seconds countdown, any document located in the capture zone will be digitized. c) Similarly to traditional desktops, documents can be stacked or arranged into folders in Escritoire. d) The tablet used for interactive transcription.

the document to the (physical) interactive transcription tablet (Figure 1d). When she is happy with the result, she clicks the *Return to Escritoire* button at the tablet interface and the document returns modified to the digital desktop.

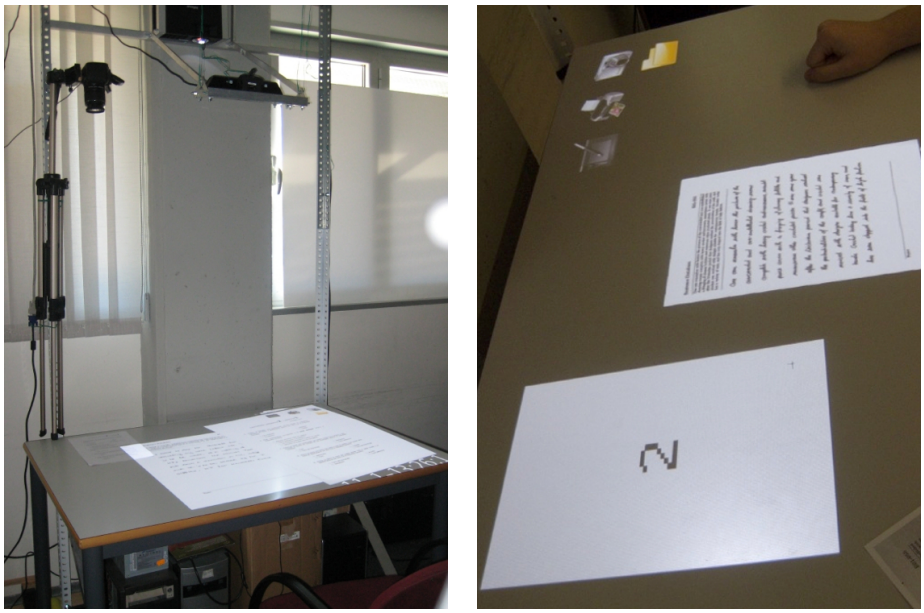
4 System Implementation

A system comprising a digital desktop can be designed in many different ways: it can be composed by a real physical desk, with a conventional screen and an *e-pen* or digital tablet; projecting the display onto the desktop and interacting using a digital pen or touch screen; or directly work in a large desktop-sized multi-touch screen.

Figure 2a shows the prototype that we have built. We decided to use a physical desk where the screen was projected. In our opinion, this was the simplest and cheapest way to create a prototype. Due to the distance between the desktop and the projector, a short throw projector (*InFocus IN1503*) was chosen, allowing us to display the proper image size.

Two cameras were used: a fixed-location high-resolution camera (*Canon EOS 1100D*) is responsible for capturing documents. This camera is zoomed and focused on the *capture zone* (Figure 1a). This type of set-up provides us with a more robust configuration (same light conditions, less geometric distortions due to perspective, etc.), therefore simplifying the subsequent steps. A camera with a depth sensor (*Microsoft Kinect*) was used to detect the finger gestures.

Finally, the interactive transcription is carried out using a *Lenovo Thinkpad Tablet 2* instead of directly on the desktop. This decision is based on previous experience. We realized that, due to the minimum font size and the projector resolution, trying to write directly on the desktop is cumbersome and inaccurate.



(a) (b)

Fig. 2. (a) Escritoire prototype showing two digitized documents. (b) A sheet of paper near the capture zone, showing the five-seconds countdown.

4.1 Gesture monitoring and detection

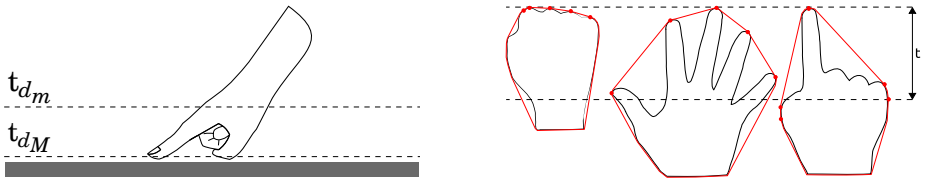
As we said before, Escritoire is a desktop on top of which a Kinect device has been mounted. This monitors the work area searching for possible user gestures. To date, we use a simple set of gestures that are performed using one or the two index fingers. The current set of gestures includes:

Select: the system will select the item shown on Escritoire under the finger.

Move: after selecting an item, the user can translate it by just moving the hand around.
Rotate: pointing both index fingers to a document and rotating hands.
Zoom: pointing both index fingers to a document and pinch open or close.

Since the set of gestures is performed with the index fingers, our system must be able to detect them. The first step in order to achieve this is to be able to differentiate the hands from the background. As the virtual desktop is displayed over the desk surface, we could not use any color-based technique (e.g., skin detection) to segment the hands.

Here we used the depth map provided by the Kinect, which captures depth information under any ambient light condition. From this depth image we need to isolate the pixels of interest (Figure 3a). We will calculate for every pixel contained in the depth



(a) (b)

Fig. 3. (a) Maximum and minimum threshold for defining a pixel of interest. t_{d_m} and t_{d_M} were empirically tunned. (b) Example (with $n = 4$) of convex hull for different hand positions.

image if they are a pixel of interest. Equation (1) provides the formal definition of pixel of interest ($\mathcal{S}_{ij}$). We compute the depth median value for every pixel ($\eta_{d_{ij}}$) during a period of 2 seconds. This way we can obtain a more stable value, minimizing the influence of out-layers derived from the sensor. Then, we calculate the difference between $\eta_{d_{ij}}$ with respect to the current depth value ($c_{d_{ij}}$). If this difference is within a minimum (t_{d_m}) and a maximum (t_{d_M}) threshold we can say that the current pixel is a pixel of interest.

$$\mathcal{S}_{ij} = \begin{cases} \text{True} & \text{if } t_{d_m} \leq \eta_{d_{ij}} - c_{d_{ij}} \leq t_{d_M} \\ \text{False} & \text{otherwise} \end{cases} \quad (1)$$

After segmenting the pixels of interest from the depth image, we want to know whether a group of pixels of interest are a hand. Previous to this, we apply a *closing* to the image, to remove any possible internal small hole and an *opening*, to remove any small noise object. As a simplification, we will assume that the biggest volumes, with a maximum of two, exceeding a certain area threshold, will be considered hands. This area threshold was empirically tunned to distinguish between noise and actually hands.

Once the hand (or hands) has been segmented from the background, the location of the index finger(s) must be found. We assumed that the user is always interacting with the system in front of the desk, thus the hands will always point *forward*. We also assume that the user hand has only 3 different states: pointing with the index finger, hand with the fingers clenched or flat.

Therefore, to distinguish between these cases we perform the following process. First, we compute the convex hull [1] of the contours that we consider hands. Then we apply the following technique: if the distance between the higher y-value vertex of the convex hull contour and the next n y-value vertex is for any case greater than a threshold d_t, we will say that the highest y-value point of this contour is the tip of an index finger (where n and d_t are parameters to be optimized). Otherwise, we assume that the user has the hand flat or with the fingers clenched. Figure 3b illustrates this process. Once we have found the tip of the index finger(s), a simple Nearest Neighbor tracking algorithm was applied to track their consecutive positions. After this, a Kalman filter [5] is applied to the tracked path(s) in order to reduce the noise.

Finally, depending on the number of fingers that the system has recognized and their position with respect to the desktop (located over a document, a button, etc), we can clearly identify the user gesture and react with the corresponding action.

4.2 Document capture and management

The document capture is carried out using a *Canon EF-S 18-55mm f/3.5-5.6 IS II* objective. The capture is performed when Alice selects the camera button. The system automatically will show the capture area, where the sheet must be placed in order to be captured. Figure 2b shows a captured document and the countdown.

Alice can work with the document on Escritoire once it has been captured. As previously explained, there are several options: the document can be moved, rotated and zoomed. In addition the document can be archived in a folder.

4.3 Handwritten Text Recognition

Once the document has been captured, it is automatically transcribed. The handwritten text recognition (HTR) system employed here follows the classical architecture composed of: preprocessing, feature extraction and recognition. Each captured image is preprocessed in order to remove margins, noise and geometric distortions [16]. After these steps, each line is detected and extracted [2]. Each preprocessed line image is represented as a sequence of feature vectors [7]. The recognition module accepts a sequence of feature vectors and is based on Hidden Markov Models (HMMs), n-gram language models. The search (or decoding) is optimally carried out by using the Viterbi algorithm [4]. After this, the transcript will be shown on Escritoire and Alice will be visually notified.

4.4 Multimodal Interactive Handwritten Text Transcription

Finally, as we commented before, the edition of the transcribed document is by means of a tablet (see Figure 4), giving the process greater comfort. When Alice wants to edit the transcription, she moves the document to the tablet icon on the desktop and the document is transfered.

Here we will follow a multimodal interactive approach [18,12,13], in which an automatic HTR system and the user cooperate to generate a better transcript. At each

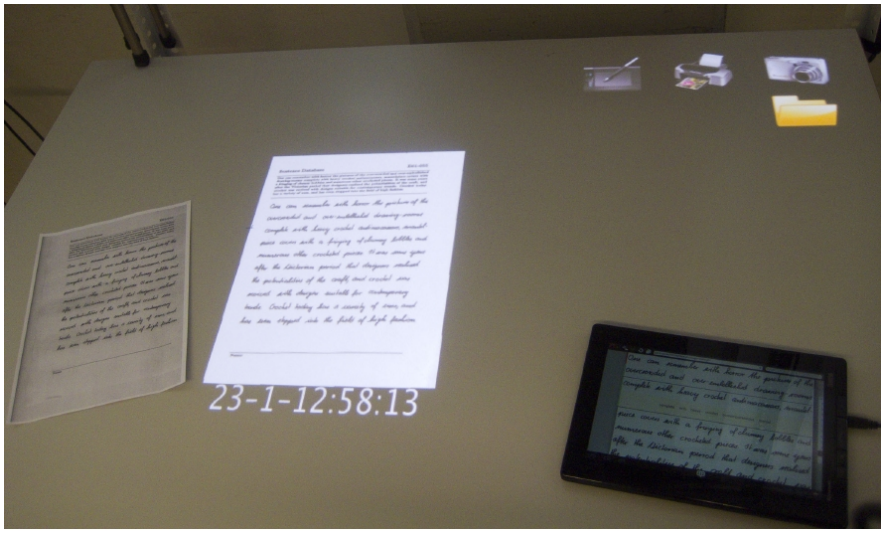


Fig. 4. From left to right: a paper document, its digital version shown on Escritoire and tablet.

interaction step, the system proposes its best transcript. If Alice finds the proposal correct, she can accept it and the process goes on. Otherwise, Alice can correct the first erroneous element and the system reacts with a revised output where this error is fixed and other related errors are potentially fixed.

5 Summary and Future Work

We have presented Escritoire, a document management system in which digital and paper documents coexist. We have integrated well-known computer vision and handwriting recognition techniques. We plan to carry out experiments with real users in order to test the gesture recognizer performance or the computer assisted transcription. Additionally, we intend to include more input modalities, for example, allowing the user to perform certain actions by voice.

Acknowledgment

This work was partially supported by the Spanish MEC under FPU scholarship (AP2010-0575), STraDA research project (TIN2012-37475-C02-01) and MITTRAL research project (TIN2009-14633-C03-01); the EU's 7th Framework Programme under tranScriptorium grant agreement (FP7/2007-2013/600707).

References

1. Andrew, A.: Another efficient algorithm for convex hulls in two dimensions. Information Processing Letters 9(5) (1979)
2. Bosch, V., Toselli, A.H., Vidal, E.: Statistical text line analysis in handwritten documents. In: Proc. ICFHR (2012)
3. Eisenstein, J., Puerta, A.: Adaptation in automated user-interface design. In: Proc. Int. Conf. on Intelligent User Interfaces (2000)
4. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press (1998)
5. Kalman, R.E.: A new approach to linear filtering and prediction problems. Transactions of the ASME–Journal of Basic Engineering (1960)
6. Keysers, D., Shafait, F., Breuel, T.M.: Document image zone classification - a simple high-performance approach. In: Proc. Int. Conf. on Computer Vision Theory (2007)
7. Kozielski, M., Forster, J., Ney, H.: Moment-based image normalization for handwritten text recognition. In: Proc. ICFHR (2012)
8. Lampert, C.H., Braun, T., Ulges, A., Keysers, D., Breuel, T.M.: Oblivious document capture and real-time retrieval. In: Int. Workshop on Camera Based Document Analysis and Recognition (2005)
9. Liang, J., Doermann, D., Li, H.: Camera-based analysis of text and documents: a survey. Int. J. Document Analysis and Recognition (2005)
10. Liwicki, M., Rostanin, O., El-Neklawy, S.M., Dengel, A.: Touch & write: a multi-touch table with pen-input. In: Proc. Int. Workshop on Document Analysis Systems (2010)
11. Marti, U.V., Bunke, H.: Text Line Segmentation and Word Recognition in a System for General Writer Independent Handwriting Recognition. In: Proc. ICDAR (2001)
12. Martín-Albo, D., Romero, V., Toselli, A.H., Vidal, E.: Multimodal computer-assisted transcription of text images at character-level interaction. Int. J. Pattern Recognition and Artificial Intelligence (2012)
13. Martín-Albo, D., Romero, V., Vidal, E.: Interactive off-line handwritten text transcription using on-line handwritten text as feedback. In: Proc. ICDAR (2013)
14. Mitra, S., Acharya, T.: Gesture recognition: A survey. IEEE T. Systems, Man and Cybernetics (2007)
15. Terry, M., Mynatt, E.D.: Recognizing creative needs in user interface design. In: Proc. C&C (2002)
16. Toselli, A.H., Juan, A., Keysers, D., González, J., Salvador, I., H. Ney, Vidal, E., Casacuberta, F.: Integrated Handwriting Recognition and Interpretation using Finite-State Models. Int. J. Pattern Recognition and Artificial Intelligence (2004)
17. Toselli, A.H., Romero, V., Pastor, M., Vidal, E.: Multimodal interactive transcription of text images. Pattern Recognition (2010)
18. Toselli, A.H., Romero, V., Vidal, E.: Computer assisted transcription of text images and multimodal interaction. In: Proc. Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (2008)
19. Wachs, J.P., Kolsch, M., Stern, H., Edan, Y.: Vision-based hand-gesture applications. Communications of the ACM (2011)
20. Wobbrock, J.O., Morris, M.R., Wilson, A.D.: User-defined gestures for surface computing. In: Proc. CHI (2009)