Human-Document Interaction systems - a new frontier for document image analysis

Dimosthenis Karatzas*, Vincent Poulain d'Andecy†, Marçal Rusiñol*, Antonio Chica‡ and Pere-Pau Vazquez‡
*Computer Vision Centre, Universitat Autónoma de Barcelona, Spain, Email: {dimos, marcal}@cvc.uab.es
†Yooz, France, Email: Vincent.PoulaindAndecy@yooz.fr
‡ViRVIG Group, Universitat Politecnica de Barcelona, Spain, Email: {achica, pere.pau}@cs.upc.edu

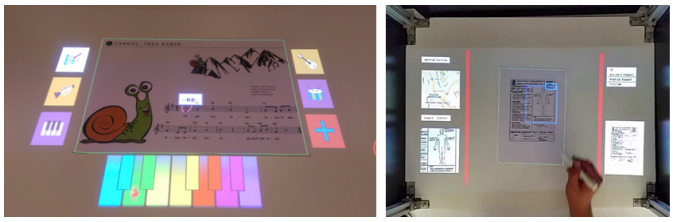

Fig. 1: Applications of our HDI system for (left) interpreting and interacting with a musical score in a public library and (right) automating a business work-flow (laboratory simulation).

***Abstract*—All indications show that paper documents will not cede in favour of their digital counterparts, but will instead be used increasingly in conjunction with digital information. An open challenge is how to seamlessly link the physical with the digital – how to continue taking advantage of the important affordances of paper, without missing out on digital functionality. This paper presents the authors' experience with developing systems for Human-Document Interaction based on augmented document interfaces and examines new challenges and opportunities arising for the document image analysis field in this area. The system presented combines state of the art camera-based document image analysis techniques with a range of complementary technologies to offer fluid Human-Document Interaction. Both fixed and nomadic setups are discussed that have gone through user testing in real-life environments, and use cases are presented that span the spectrum from business to educational applications.**

I. Introduction

It is widely accepted that the shift to paperless living is not going to materialize any time soon, as paper presents certain affordances that serve users in ways difficult to match with electronic devices and digital documents. A number of processes initially depend upon and require paper, while in many contexts people prefer to read on paper, while when it comes to performing certain kinds of cognitive tasks, paper has many advantages [1]. The seminal work of Sellen and Harper on this topic [2] reveals that the use of paper is only going up when people turn to digital, simply because they tend to print out more. The average office worker produces on average 10,000 sheets of paper per year, and more than half of what is printed is only read once and then discarded.[1]

A number of authors have pointed out the respective affordances of paper and digital documents [2][3][4]. Attempting a quick summary: paper is tangible, spatially flexible and tailorable – we can pick up a document, scribble on it, and arrange it on the desk; digital documents on the other hand are quick to edit, copy, transmit, share, can be easily searched, filed and retrieved. It is evident that a lot is to be gained by bridging the two domains.

The fluid interaction between the digital and the physical domain through augmented or mixed reality interfaces has been a focus of human computer interaction research for many decades, starting with the seminal work of Paul Wellner on the Digital Desk two decades ago [4]. Numerous approaches have been proposed for what has been variably called interactive paper [5], paper interface [6][7], digital desk [4], or paper augmented digital documents [8]. Surprisingly, the document image analysis community has not engaged in such research. Most concepts and prototypes in the literature make use of markers or rudimentary computer vision technology, instead of incorporating state of the art camera-based document image analysis techniques. To some extent this is an indication of the inadequacy of current DIAR methods for fluid, real-time *Human-Document Interaction* (HDI) and at the same time a call for action for the community.

Current technology permits constructing inexpensive and small-size projector/camera setups for real-life augmentation and interaction with physical documents. Lately, the Digital Desk concept has been revisited both by industry (Fujitsu's *Fingerlink*, LBO's *Light Touch*, HP's Sprout) and by academia (e.g. MIT's *LuminAR* [9] and *SixthSense* [10] projects).

In this paper we detail our own experience putting together a system for Human-Document Interaction. The novelty of this paper is in combining state of the art document image analysis algorithms with non-DIAR technologies (tracking, gesture recognition, visualisation) into a flexible, modular, extensible framework for Human-Document Interaction.

We give technical details (section III) and comment on useful lessons learnt (section V) from the application of our system in both fixed and nomadic setups. We put emphasis on the design decisions both at the software and hardware side and although we describe the whole framework we focus particularly on the document image analysis elements. Document image analysis is a sine-qua-non component of our system, enabling the link between the physical and the digital document, while it also provides application level functionality.

The paper presents two real-life implementations of the system (section IV) addressing different use cases: first the application to an administrative paper based work-flow and

[1]source: US Environmental Protection Agency; "Rethinking printing" report on printing habits in the UK, Kyocera, 2010

second an educational scenario. We extract useful intuition from the experience with real-users in the *Library Living Lab*, a real-life, open, participatory, experimentation space that the researchers have setup into a public library.[2]

II. Background

Considerable work has been done on the interface between physical and digital documents by the human computer interaction community. Augmented reality based interfaces are of particular interest for this work, especially the intersection with camera-based document image analysis techniques

Typical mobile Augmented Reality (AR) interfaces are screen bounded and thus require the use of a display as a mediator of the actual experience. This introduces various limiting factors, most importantly the fact that users have to switch context between the physical document and the screen while the hands of the user are not free to interact with the paper. As such, mobile AR systems for documents have been restricted to displaying predefined content over selected patterns and typically used for advertising. The authors have employed mobile AR in the past to create interactive experiences over musical scores[3]; these suffer from the above limitations.

Projected augmented reality interfaces on the other hand, like the one presented in this paper, allow the user to continue working in the physical space in an habitual way. One of the earliest attempts is Wellner's Digital Desk [4]. Such approaches did not evolve substantially until the last few years with projects like MIT's fixed desktop setup LuminAR [9] and wearable gestural interface SixthSense [10]. Other interesting prototypes include Disney's HideOut [11] that uses a handheld pocket projector to guide the interaction and StripTic [12] that explores the use of augmented paper strips for air traffic controllers.

Invariably, all above approaches require some kind of image based registration between the physical and the digital document for the spatial link to be established. Generally, this is achieved through the use of visible or infrared markers. In a few cases content based registration is used. Hull *et al.*[13] uses spatial layout features based on basic word arrangements to index predefined hotspot areas on documents. In subsequent work, Erol *et al.*[14] introduce the Brick Wall Coding that makes use of both layout and aspect ratios of word boundaries. These ideas are very similar to Nakai *et al.*[15] and Liu and Doermann [16]. Such approaches have been demonstrated to gracefully scale to large document databases. Nevertheless, they require a minimum amount of textual content and are best suited to printed full-text pages, while they tend to present problems when applied to more complex layouts such as forms, when text is sparse or does not follow Latin language text conventions. Alternatively, pixel level image features are used by a handful of methods, like FACT [17] which uses FIT features [18] (SIFT-like local visual features) for enabling fine-grained interaction. The system described here is using SIFT features.

[2]The Library Living Lab, Barcelona (http:\l3.cvc.uab.es), setup within the public library Miquel Batllori in Barcelona, is a member of the European Network of Living Labs and a unique infrastructure contextualised around a library theme, facilitating direct interaction with the end users in a real-life functional space.

[3]Augmented Songbook http://www.cvc.uab.es/songbook

An explicit mapping between the physical and the digital document is not always desired, feasible or necessary. Treating a paper document as a transient information source, without persistent linkage to its digital counterpart is for example useful when dealing with previously unseen documents, or with document creation (starting with a blank page). Early systems that follow this paradigm include DigitalDesk [4] and CamWorks [19]. In such cases, targeted analysis (e.g. OCR) of specific regions of the document is usually employed. An alternative is to attempt to detect and rectify the unknown page in the video frame. A few DIAR approaches have been proposed for camera-captured documents based on the detection of page borders [20][21], but their application to Human-Document Interaction scenarios is not straightforward and has not been attempted. The authors have recently proposed an approach for real-time document detection and rectification designed for mobile platforms [22] which can be used in the system presented here when "transient paper" functionality is desired.

Typically, the information required to support the interaction process comes from the digital document. On-demand image analysis of the paper document to extract new information rarely takes place, and when it does it is usually restricted to the recognition of annotations. Digital pens have been extensively used for annotation capture. These approaches require that documents are printed on special paper. For example Guimbretiere [8], Liao *et al.*[5], StripTic [12] all use Anoto pens to capture annotations and transfer to the digital counterpart. Alternatively, an efficient way to obtain annotations is by differencing the captured document image from previous versions of the document. In this line Mazzei *et al.*[23] use Nakai's approach [24] to register the document and image differentiation to obtain the annotations. In certain scenarios requiring detecting differences between subsequent versions of documents, recent methods by Jain *et al.*[25] could be used.

Apart from the extraction of annotations, very few cases exist where on-demand information extraction is taking place. In the original Digital Desk setup, a zoomed in camera was used to capture a high resolution version of a limited area of the paper, that was adaptively thresholded and OCRed. We recently proposed a dynamic information extraction approach for digital mailroom applications [26] that can used in conjunction with the system presented here for business documents applications.

Gestural interaction is necessary for any Human-Document Interaction system. A complete review falls outside the scope of this paper, but we offer here a short overview for completeness. When hand gestures are to be detected extra sensors are usually employed, such as a laser field, a 3D sensor [9] or an touch surface on the desk. Alternatively, pen gestures might be employed, usually detected and digitised by a digital pen. In the system presented here, we use either hand gesture recognition using a visible camera, or infrared pens coupled with an extra infrared camera. The use of infrared pens compared to digital pens has the advantage that it works over any part of the surface and no special preparation of the documents is necessary.

A different, interesting direction is leveraging the physical properties of the paper for interaction and defining gestures based on the manipulation of the paper itself. Holman *et al.*[27] defined an interaction grammar motivated by the natural manipulation of paper, including gestures such as collocating, collating (stacking), flipping, rubbing, and stapling, while Tarun

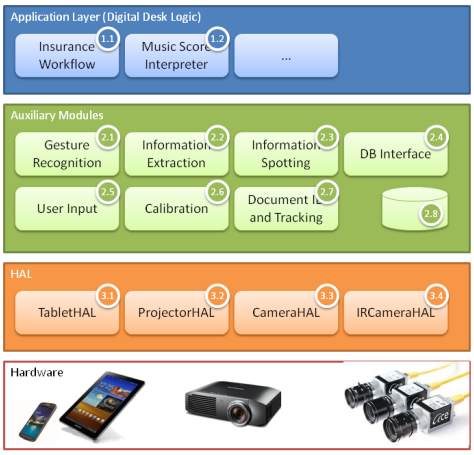

Fig. 2: Software architecture.

et al.[28] uses touch sensitive flexible displays, and touch and bend gestures to navigate content. It might be interesting to see how working with the document content can be achieved in consonance to such gestures – this probably requires real-time dewarping and rectification of document images.

III. System Description

A. Overview

The Human-Document Interaction system presented here, uses a projector-camera setup to create an augmented interactive space on a desk surface. Our goal was to construct a practical, extendible system that can be easily deployed in various real-life scenarios. The design principles that guided us are as follows:

Portable setup. The system should be possible to be deployed in various fixed or nomadic setups, and work with different hardware as needed.

Extendible system. The system should be easy to extend and specialise without invalidating existing fuctionality. No special paper or reading devices. The system should not make use of markers or special micro-optical patterns or reading devices such as digital pens.

Support for general document content. The system should not be designed for a specific type of content, language or script, but should work for any document type.

In the following section an overview of the software architecture is given as well as detailed description of important modules. In section III-C the different hardware incarnations of the system are described along with an appraisal of the different design decisions.

B. Software Architecture

The system is based on a modular architecture. Modules communicate using a message passing interface based on ZeroMQ, a multi-platform and easy to scale framework. Each module subscribes to a set of messages and publishes its results as another set of messages. These modules can be physically running on one or multiple devices (any common platform). The message passing framework permits us to announce each new camera frame to any image analysis module and run several analyses in parallel. When run on the same computer, passing frames is instantaneous and very efficient memory wise. When run on different platforms the speed depends on the network.

Modules performing image analysis such as the gesture recognition or document identification ones act directly on the visible camera feed, which means that the image content might be overlaid by projected information. This is a difficult problem that can nevertheless be tackled in various ways, ranging from compensating for the projected image in software to synchronizing the capture frame rate with the projector refresh rate and take advantage of an off-time (no projection) of the projector to capture. For most of the applications we have developed with our system, the projected information is little, so we were never presented with the necessity to compensate for the projected image.

In the following sub-sections we detail the system modules, giving particular emphasis on the document analysis ones.

1) Hardware Abstraction Layer: A hardware abstraction layer (HAL) is defined with modules that encapsulate different hardware devices (cameras, projector, linked mobile devices). Introducing new hardware devices is thus reduced to defining a corresponding HAL module. This makes it easy to deploy the system in different ways and change hardware as necessary in a way transparent to higher level logic and application layers.

2) Database and DB Interface: The persistence of system data is centralized trough a database module, this database stores configuration and document data. Our system makes use of a database abstraction module that provides an abstraction level that separates the particular database architecture from the system calls, making the system impervious to the particular database implementation chosen.

3) Document Detection and Tracking: The document detection and tracking module is responsible for locating a known document in the image captured and passing its coordinates to higher level modules. We use extracted SIFT keypoints to represent document images and we use the FLANN [29] library for indexing and retrieval. Our choice of SIFT features over spatial layout features typically used for indexing document images (e.g. [15]) is due to the shortcomings of spatial layout features when it comes to images that do not predominantly contain text – this is the case with a number of applications we have implemented, including mixed documents such as administrative forms and music scores. We implement two stage detection. If no document was identified in previous frames, then a full retrieval is launched against the document database. This initial detection might take time, and depends on the size of the dataset. We experimented with up to a few hundred pages, therefore the detection time is not noticeable, but this could be a hurdle when scaling up. Once detected, subsequent frames are first searched for the known document. This is a very fast operation of stable duration as matching against a limited number of keypoints is performed. A full scale retrieval is only needed when document changes, which is detected after a predetermined number of frames for which the matching with the known document has failed. A Kalman filter tracker is employed to smooth out any erratic movement of the detection and create a fluid sensation to the user. As mentioned before, in the case that a "transient paper" functionality is desired

instead of an explicit link to an a-priori known document, then an alternative implementation for this module can be used, based on the real-time page border detection and document rectification method proposed by the authors [22].

4) Information Extraction: Targeted information extraction from administrative documents can be performed using the method proposed by the authors in [26]. This is typically used for administrative documents, and is pre-trained on the form class we are interested to process. This functionality is usually combined with a "transient paper" mode and document classification, for which we use the method we detail in [30].

5) Information spotting: At the current implementation, we use the techniques we detailed at [31] for spotting graphical symbols such as logos or stamps in documents. It would be interesting to incorporate segmentation free word spotting functionality in the near future.

6) User input: At the current implementation of the system, user input is supported by typing on a linked device, such as a mobile phone or tablet. Linking another device to the system is seamless, due to the use of a platform agnostic message passing architecture. It is nevertheless worth noting that using a separate device is distracting as it requires the user to switch context. A more suitable approach would be using document differentiation of subsequent captures to extract user annotations on the paper document.

7) Gesture Recognition: The user is able to interact with the system through a dictionary of one-point or two-point gestures (e.g. tap, flick, scroll or drag and drop). We have created two alternative implementations for the gesture recognition module, The first is based on hand gesture recognition using the visible light camera. This is particularly intuitive and easy to use, but only in scenarios where projected augmentation is limited, as projections over the hands hinders segmentation and can affect detection. Using an extra 3D sensor in a similar fashion as [9] offers a solution to this problem, at an increased cost. Instead of an extra 3D sensor, as an alternative we have set up a pen based interaction technique, using infra-red pens and an infra-red camera. Using a pen is a natural way to interact with the document. The two modules are interchangeable, and use the same dictionary of gestures. The particular implementation used is transparent for the rest of the system and they can be interchanged at run time or used in parallel - an advantage afforded by the messaging system we emply.

8) Visualisation: The visualisation module is responsible for generating the visual information through the system projector that is projected on the physical document and its surrounding area and includes the augmentation information and user controls.

9) Calibration: The Calibration module is responsible for calibrating between the different viewports: visible light camera, infrared camera and projector. For fixed setups (see section III-C) calibration is performed once and then stored in the database. For nomadic setups it is performed by default once at the beginning of a session and be can be repeated on demand. Decalibration can be a problem in nomadic setups - the current implementation of our system does not handle decalibration automatically. The calibration module apart from performing calibration is also responsible for communicating homography matrices to other modules upon request. There

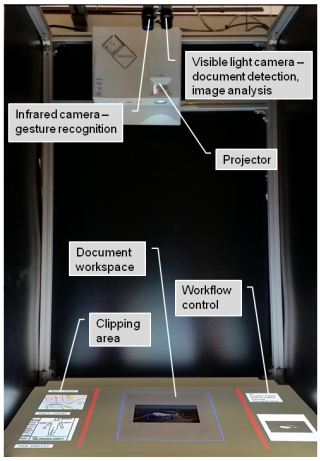

Fig. 3: The key hardware components of the system in the laboratory setup.

are two calibration routines. First, for calibrating between the visible light camera and the projector an automatic process is used involving projecting a set of geometric patterns which are detected by the camera and used to calculate the homography matrix. Second, for calibrating between the infra-red camera and the projector a manual process is used, where a cross shape is projected at a number of locations and the user is asked to click on it with the infra-red pen, which is then detected by the infra-red camera. When all devices have been mutually calibrated, the system defines a workspace as the common area between the viewports of the two cameras and the projector. In other words, the workshops is the area visible by both cameras that the projector can also fully cover.

10) Digital Desk Logic: The Digital Desk Logic module refers to the running application, and defines the behaviour of the digital desk system. See section IV for example applications.

C. System Setup

Given the software architecture described above, different setups are possible simply by providing new hardware abstraction layer modules, corresponding to the devices used. We have tried both fixed setups and nomadic ones. Some design decisions and technical details are given next. Figure 3 shows our laboratory setup, while Figure 4 shows a desktop lamp based setup and the fixed setup within the Library Living Lab in a public library in Barcelona.

The projector type to use depends on the type of preferred setup. For movable setups like the desktop one, a pocket projector is used. The main disadvantage is the limited contrast available, which defines the distance of the projector from the desk in order for the interface to be comfortably visible. In terms of projection technology, laser projectors make good sense for these setups as they are always-in-focus. In our case we use a MicroVision laser projector. For fixed setups, any projector technology can be used, although LCD and LCoS projectors are better suited as the image is fully projected during all the time and thus it is not necessary to adjust the camera exposure time in order to capture a full frame. In reality, since we do not process the projected information, this is actually of

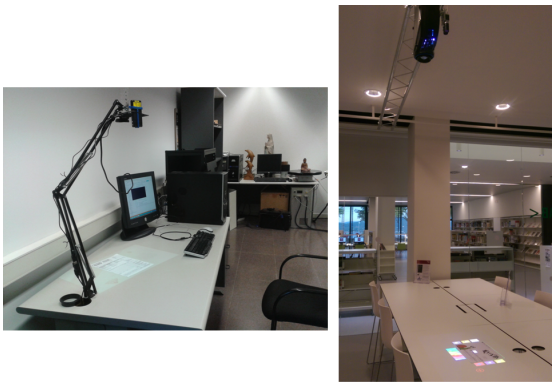

Fig. 4: (Left) a nomadic desk-lamp setup based on a Web cam and a pocket projector, (right) a fixed ceiling setup within the Lirary Living Lab based on ethernet cameras and a 2 lumen projector

little importance and we use DLP technnology projectors as they offer good quality at a reasonable price.

As for the camera type, the best choice would be a CMOS global shutter camera, as it provides the most stable image and does not suffer from distortions when the paper or the user hands are moving. In reality though, the type of movements performed are relatively slow and infrequent to justify the price, and cheaper technologies work equally well in practice.

For the fixed setups we use two Basler AC2500 rolling shutter cameras (2590x1944, 14fps), one of them mounted with an infra-red longpass filter (850nm), and an InFocus 3138HDA, DLP technology projector (1920x1080, 4000 lumen). The price of the fixed setup is around 3000 EUR. For the desk-lamp setup, we use two Logitech HD C510 rolling shutter Web cams (1280x720, 25fps), one of them has had its infra-red filter removed and fit with a longpass filter (850nm). The hardware price of this nomadic setup is around 500 EUR.

IV. Case Studies

We used our Human-Document Interaction system to implement two distinct scenarios, one addressing the optimisation of a business work flow and a second looking into the use of our system in an educational context. In both cases we used fixed setups, because the low contrast of the desk-top setup made it difficult to use.

A. Business: Automating an insurance company workflow

The aim in this particular case is to simulate a business workflow that depends on paper document input at various stages, and examine how a Human-Document Interaction system can facilitate the process. We chose to simulate a workflow of filing a car insurance claim over the desk, where different paper documents have to be handed over and processed by the clerk. In this scenario, we defined the central part of the augmented surface as the interaction area where documents should be placed, the left part displays the summary of information about the claim as documents are being processed, while the right part is acting as a control panel.

A reduced summary of the process is shown in Figure 5. We have not tested this scenario with experts, but we did test it with naive users in the lab who gave us positive feedback about the experience. They had no problem with the pen gestures they had to perform, which they found intuitive and responsive.

B. Education: Music score interpretation for children

The aim of this second use case was to test the system's usefulness in assisting children understand special types of documents they are not familiar with. In this particular case, we chose to automatically interpret musical scores, and offer the possibility to children to play the piece using the same system. A photograph of the system can be seen in Figure 1.

As soon as a score is placed on the reading bench, the score is identified and a virtual keyboard is displayed below it. The first note in the score is highlighted and its name appears next to it, while a symbol is shown indicating which key in the keyboard corresponds to this note. As soon as the child hits the note, the next note and corresponding key is highlighted.

The application was made available to the public at the Library Living Lab, an open participatory, experimentation space setup within a public library in Barcelona, Spain. Although designed for children, at least 50 users or all ages, including children, tried out the application. All users found the interface intuitive and engaging, while certain users that had no prior knowledge of reading music found the automatic interpretation offered useful as an learning instrument.

V. Conclusion

Human-Document Interaction (HDI) is an interesting area of application for document image analysis, that has generally been neglected. Contrary to typical DIAR applications, designed to work in an off-line non-time critical fashion, under more or less controlled conditions, HDI presents new challenges and opportunities for the field. For interaction to be effective, a real-time, fluid interface with the user has to be established. Fast and stable DIAR algorithms are required for enabling Human-Document Interaction in a variety of real-life conditions.

Human-Document Interaction as a concept has been around for the best part the last three decades, but only recently it is becoming viable to build efficient augmented reality smart reading spaces, based on affordable technology. This paper intents a call for action for the document analysis community, and attempts to establish the area of HDI as an important application area in the near future that will require a certain mentality shift in the way our community approaches solutions and real-life working systems intended to interact with users.

As far as future work is concerned for the system presented here, we consider that adding functionality for extracting and recognising annotations is a key improvement, as well as word spotting functionality. Starting with a blank page and following through the cognitive process of taking notes or drawing is an interesting challenge. Finally we consider it very interesting to combine this technology with eye-tracking, both as an extra means of interaction and as useful input modality about the way the user reads.

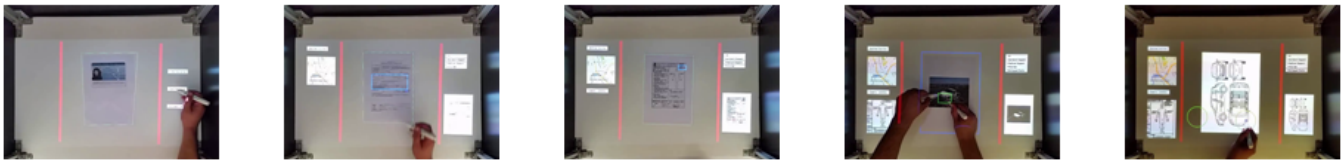

Fig. 5: Different stages in the automation of an insurance workflow.

VI. Acknowledgements

This work was supported by Yooz (France) and the Spanish projects TIN2014-52072-P and RYC-2009-05031.

References

[1] R. Walker, "White Paper: Achieving The Paperless Office," Efficient Technology, inc, Tech. Rep., 08 2009. 1

[2] R. Harper and A. J. Sellen, *The myth of the paperless office*. MIT Press, 2001. 1

[3] H. S. Baird, "Digital libraries and document image analysis," in *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*. IEEE, 2003, pp. 2–14. 1

[4] P. Wellner, "Interacting with paper on the digitaldesk," *Communications of the ACM*, vol. 36, no. 7, pp. 87–96, 1993. 1, 2

[5] C. Liao, F. Guimbretière, and K. Hinckley, "Papiercraft: a command system for interactive paper," in *Proceedings of the 18th annual ACM symposium on User interface software and technology*. ACM, 2005, pp. 241–244. 1, 2

[6] Q. Liu and C. Liao, "Paperui," in *Camera-Based Document Analysis and Recognition*. Springer, 2011, pp. 83–100. 1

[7] J. Smith, J. Long, T. Lung, M. M. Anwar, and S. Subramanian, "Paperspace: a system for managing digital and paper documents," in *CHI'06 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2006, pp. 1343–1348. 1

[8] F. Guimbretière, "Paper augmented digital documents," in *Proceedings of the 16th annual ACM symposium on User interface software and technology*. ACM, 2003, pp. 51–60. 1, 2

[9] N. Linder, "Luminar: a compact and kinetic projected augmented reality interface," Ph.D. dissertation, Massachusetts Institute of Technology, 2011. 1, 2, 4

[10] P. Mistry and P. Maes, "Sixthsense: a wearable gestural interface," in *ACM SIGGRAPH ASIA 2009 Sketches*. ACM, 2009, p. 11. 1, 2

[11] K. D. Willis, T. Shiratori, and M. Mahler, "Hideout: mobile projector interaction with tangible objects and surfaces," in *Proceedings of the 7th International Conference on Tangible, Embedded and Embodied Interaction*. ACM, 2013, pp. 331–338. 2

[12] C. Hurter, R. Lesbordes, C. Letondal, J.-L. Vinot, and S. Conversy, "Strip'tic: exploring augmented paper strips for air traffic controllers," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, 2012, pp. 225–232. 2

[13] J. J. Hull, B. Erol, J. Graham, Q. Ke, H. Kishi, J. Moraleda, and D. G. Van Olst, "Paper-based augmented reality," in *Artificial Reality and Telexistence, 17th International Conference on*. IEEE, 2007, pp. 205–209. 2

[14] B. Erol, E. Antúnez, and J. J. Hull, "Hotpaper: multimedia interaction with paper using mobile phones," in *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 399–408. 2

[15] T. Nakai, K. Kise, and M. Iwamura, "Use of affine invariants in locally likely arrangement hashing for camera-based document image retrieval," in *Document Analysis Systems VII*. Springer, 2006, pp. 541–552. 2, 3

[16] X. Liu and D. Doermann, "Mobile retriever-finding document with a snapshot," in *Int. Workshop on Camera-Based Document Analysis and Recognition*, 2007, pp. 29–34. 2

[17] C. Liao, H. Tang, Q. Liu, P. Chiu, and F. Chen, "Fact: fine-grained cross-media interaction with documents via a portable hybrid paper-laptop interface," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 361–370. 2

[18] Q. Liu, H. Yano, D. Kimber, C. Liao, and L. Wilcox, "High accuracy and language independent document retrieval with a fast invariant transform," in *IEEE International Conference on Multimedia and Expo, ICME 2009*. IEEE, 2009, pp. 386–389. 2

[19] W. Newman, C. Dance, A. Taylor, S. Taylor, M. Taylor, and T. Aldhous, "Camworks: a video-based tool for efficient capture from paper source documents," in *Multimedia Computing and Systems, 1999. IEEE International Conference on*, vol. 2. IEEE, 1999, pp. 647–653. 2

[20] F. Shafait, J. Van Beusekom, D. Keysers, and T. M. Breuel, "Document cleanup using page frame detection," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 11, no. 2, pp. 81–96, 2008. 2

[21] N. Stamatopoulos, B. Gatos, and A. Kesidis, "Automatic borders detection of camera document images," in *2nd International Workshop on Camera-Based Document Analysis and Recognition, Curitiba, Brazil*, 2007, pp. 71–78. 2

[22] M. Rusiñol, J. Chazalon, and J.-M. Ogier, "Normalisation et validation d'images de documents capturées en mobilit'e," in *Colloque International Francophone ser lÉcrit et le Document, CIFED 2014*, 2014. 2, 4

[23] A. Mazzei, F. Kaplan, and P. Dillenbourg, "Cognitive and social effects of handwritten annotations," in *Red-conference, rethinking education in the knowledge society*, no. EPFL-TALK-158761, 2011. 2

[24] T. Nakai, K. Kise, and M. Iwamura, "A method of annotation extraction from paper documents using alignment based on local arrangements of feature points," in *Ninth International Conference on Document Analysis and Recognition*, vol. 1. IEEE, 2007, pp. 23–27. 2

[25] R. Jain and D. Doermann, "Visualdiff: Document image verification and change detection," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 40–44. 2

[26] M. Rusinol, T. Benkhelfallah, and V. P. d'Andecy, "Field extraction from administrative documents by incremental structural templates," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 1100–1104. 2, 4

[27] D. Holman, R. Vertegaal, M. Altosaar, N. Troje, and D. Johns, "Paper windows: interaction techniques for digital paper," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2005, pp. 591–599. 2

[28] A. P. Tarun, P. Wang, A. Girouard, P. Strohmeier, D. Reilly, and R. Vertegaal, "Papertab: an electronic paper computer with multiple large flexible electrophoretic displays," in *CHI13 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2013, pp. 3131–3134. 3

[29] M. Muja and D. G. Lowe, "Flann, fast library for approximate nearest neighbors," in *International Conference on Computer Vision Theory and Applications (VISAPP'09)*, 2009. 3

[30] M. Rusinol, J. Chazalon, J.-M. Ogier, and J. Lladós, "A comparative study of local detectors for mobile document classification," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 2015. 4

[31] M. Rusinol, D. Karatzas, and J. Lladós, "Spotting graphical symbols in camera-acquired documents in real time," in *Graphics Recognition. Current Trends and Challenges*. Springer, 2014, pp. 3–10. 4