Rethinking Fingerprint Evidence Through Integration of Very Large Digital Libraries

Nádia P. Kozievitch[1], Ricardo da S. Torres[1], Sung Hee Park[2], Edward A. Fox[2], Nathan Short[3], A. Lynn Abbott[3], Supratik Misra[3], Michael Hsiao[3]

[1] Institute of Computing, University of Campinas, Campinas, SP, Brazil
{nadiapk,rtorres}@ic.unicamp.br
[2] Department of Computer Science
Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
{shpark,fox}@vt.edu
[3] Department of Electrical and Computer Engineering
Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
{nshort2,abbott,supratik,hsiao}@vt.edu

Abstract. Fingerprints play a key role in biometrics and forensic science because of their uniqueness. Essential is contextual integration of fingerprint evidence from different sources, which involves composing, reusing, and aggregating a large amount of information. Thus, this paper (1) describes different types of fingerprint information from a digital library perspective; (2) investigates compound object concepts as used in connection with fingerprints; and (3) presents a preliminary integration of very large fingerprint digital libraries.

1 Introduction

Fingerprints have been used for identification from the early 1900s. The patterns formed by the ridges are important since they already are formed in the fetus by the fourth month of pregnancy and do not change until death. These patterns cannot be altered, except by accident, mutilation, or very serious skin disease, as they are formed in deep layers of the dermis. The skin consists of two main layers: the outer skin or epidermis, and the inner or true skin, known as the dermis.

The common friction ridge patterns – loops, whorls, and arches – impart class characteristics to a fingerprint [1], pre-aligning algorithms according to these singularities. This is similar to large image retrieval [2] systems, where there is a pre-analysis of quality, direction, ridge flow, angles, etc.

Our contribution is the analysis of fingerprint related activities, unifying different domains, using a digital library (actually, 4 DLs) and compound object (CO) perspective. Those aware of law enforcement activities will know of the first type of DL (DL1), associated with databases of stored fingerprints. Another consideration relates to our BAE Systems funded project to create training materials for fingerprint examiners, which leads to a second type of DL (DL2). A third type of DL relates to the evidence and data describing a crime scene (DL3).

A fourth type of DL relates to our NIJ funded research studies supporting experimentation with fingerprint image analysis techniques, quality measures, and matching methods (DL4). Combining these four into an integrated DL, where compound objects allow us to work across these DLs (see Fig. 1), yields a very interesting and very large DL.

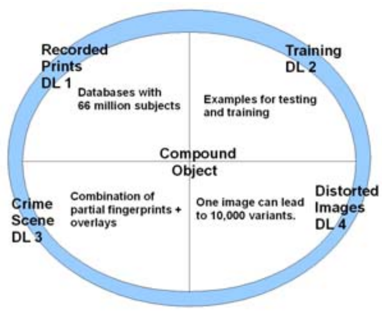

Fig. 1. The integration of fingerprint digital libraries.

In DL1, information is used to identify a person. DL2 has a different purpose: to educate and train users. In DL3, images are used for matching or excluding individuals. In DL4, the focus is on algorithms, varying parameters. Through our integration, digital libraries unify four different communities, allowing each one to see different perspectives, and explore the system as a whole, or focus in a determined area. In addition, we can take advantage of digital library services (e.g., browsing and searching), formalisms, and preservation solutions.

We plan to use compound objects (COs) [3] to facilitate aggregation abstraction, embracing components from different domains, and unifying them with a single concept. COs also can help us achieve benefits arising from the script concept of Schank [4], where components in a CO have the same behavior, or respect the same rules. Finally, solutions to some of the "very large" issues in digital libraries result from using COs, e.g., when specific operations are applied to a set, or when aggregating parts.

This paper explains fingerprint digital libraries for evidence and training in Section 2, integration and CO concepts using the 5S framework in Section 3, and a preliminary integration of the 4 DLs in Section 4.

2 The Different Fingerprint Digital Libraries

2.1 Recorded Prints

Recorded prints are the basis for the matching of images and the distorted images created by experiments. Large law enforcement databases may have millions of people's prints, where each one can come with 10 fingers, 10 toes, palm, pads of feet, etc. There are direct and rolled prints, and sometimes repeated captures, including over time. The largest collections and systems generally are proprietary and not available to the public, or are related to scene analysis [5]. One of the biggest biometric database and fingerprint identification system is from the Federal Bureau of Investigation, at http://www.fbi.gov/hq/cjisd/iafis.htm. It has at least 66 million subjects in the criminal master file, along with more than 25 million civil print images. To determine whether two fingerprints match (Figure 2-A), examiners move beyond the common ridge patterns and focus on the unique and complex details of ridges that divide, cross, and terminate. This classification is based on four classes: terminations, bifurcations, trifurcations (or crossovers), and undetermined. The process of analysis and feature extraction from a single print can produce an enormous amount of information, like quality and direction maps, quality measures, etc. Besides the matching, there is assessment of image quality, e.g., based on NIST Fingerprint Image Quality (NFIQ) [6], considering details like direction, contrast, flow, and curvature.

2.2 Distorted or Synthetic Images

Distorted or synthetic images are created by algorithms that simulate motion and/or skin distortion. To investigate their effects on image quality, two types of distortions were considered: skin distortion [7] and blurring. The skin distortion model used in our initial experiments simulates skin plasticity around the contact point of a finger tip. It has 10 parameters controlling the model, as shown in Figure 2-B. The combination of a single recorded print with the 10 parameters, for example, can synthetically generate about 10,000 images. The blurring distortion model uses an increasing amount of blurring (Figure 2-C). The objective is to simulate several level of distortions, to compare levels of acceptable quality. Here one single image generates three other images.

2.3 Crime Scene

The evidence from a crime scene can come from thousands of people who visited a popular place, or touched an object, as shown in Figure 2-D, creating data which can be later compared with a criminal history record. Each person has ten fingers, and each finger can produce different images depending on the type of distortion. In addition, there are overlays of different prints, i.e., combinations of images from the fingers under the same substrate. The matching can process one fingerprint, multiple fingerprints, or combinations of entire and partial images against one database. Additional details can be present, regarding the fingerprint (location, orientation, size, pressure, distortion, etc.), the object touched (curvature, substrate, etc.), or methods of extraction and preservation. After the sample is collected, there is still the need to document the evidence history or provenance.

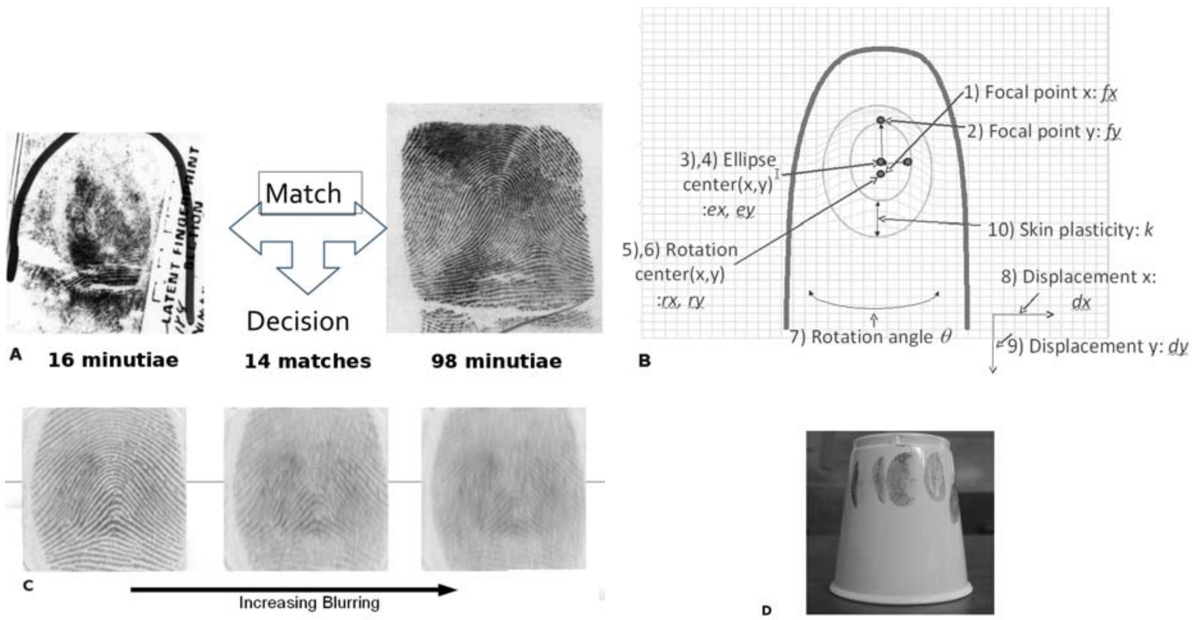

Fig. 2. (A) Matching Images. (B) Skin Distortion. (C) Blurring. (D) Crime Scene.

2.4 Training Materials

A first goal of the training effort is to develop templates for modules that encompass twenty different topics in biometrics, along with the use of combinations of examples to illustrate each of myriad types of situations. Ideally, for testing fingerprint examiners, the combination of examples identified could be used for assessment, so each case in an exam is distinct, reducing opportunities for cheating. The training modules will have examples for instruction, and yet others for exercises and examinations, taken from all of the other DLs. To give a sense of scale: Suppose that one image generates 100 distorted images. Multiply by 25 million possible suspects. Then try to match a crime scene image which has 55 partial fingerprints. Finally, select and link good examples for use in training.

We propose the use of compound objects, detailed in the next section, for connecting, aggregating, and re-using appropriate information in support of such large scale efforts.

3 Integration and Compound Objects in 5S Framework

3.1 COs and DL Integration

Agosti et al. [8] defined a Compound Object (CO) as a digital object that includes information about context, provenance, and relationships between resources. COs are aggregations of different information combined together in order to shape a unique logical object. Several CO formats arise from different communities [9]. Even though there are a number of standards to support the management of COs, there is still incompatibility, motivating solutions for integration and interoperability.

Thus there is a second factor that needs to be analyzed: the process of integration. Kostas and Delis [10] divided the integration process into four steps: (i) discovery: systems "learn" about the existence of each other; (ii) identification: systems unambiguously identify their individual items; (iii) access: systems access their items; and (iv) utilization: systems synthesize their items.

In the case of COs, there is a fifth step, regarding how the objects are aggregated. The Dexter Hypertext Reference Model [11] for example, uses the "hidden structure approach", placing all of the data and all of the data interpretation inside the content portion of a component. The Amsterdam Model [12], on the other hand, uses the "separate structure approach", defining each piece of multimedia information as a separate block.

We propose to connect, reuse, and integrate COs, taking advantage of the 5S (Streams, Structures, Spaces, Scenarios, and Societies) framework, along with the 5S approach to integrate digital libraries; see http://si.dlib.vt.edu/. The integration of archaeological digital libraries has been described from the 5S perspective, but then we considered only digital objects, not their composition. Due to that work, we can build upon a well-documented and validated formal framework describing some of the essential aspects of digital libraries.

3.2 Concepts and Definitions

An **Integrated Digital Library** is a 4-tuple consisting of a union repository, a union catalog, union services, and a union society. The minimal union services of a digital library are represented by mapping and harvesting services, which are necessary to support integration. For the integration of COs, we can use the same definitions, considering the following aspects:

1. each CO has a handle, a structure, its internal components, and a boundary (so we clearly distinguish the CO from other objects) [3];
2. each CO has an interface or description which specifies how its information can be accessed;
3. each CO has a vocabulary to describe it and its internal composition;
4. the same vocabulary that is used to describe the components can be used for labels for the schema in the mapping service;
5. the mapping service is responsible for unique identification of all the objects in the union set;
6. the boundary is represented in the mapping service by what is within the internal structure of each CO;
7. the application should specify which approach is used for the object aggregation: "hidden structure approach", "separate structure approach", etc.

For the harvesting process, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) can be used, defining a mechanism for data providers to expose their metadata. For disseminating the content in concert with a metadata harvesting protocol, some steps are necessary [13]: (i) wrap the data in a packaging format; (ii) include the metadata; (iii) encode the references to the files; and (iv) harvest the package. For this, OAI-ORE [14] or DCC [15] can be used, representing the objects and aggregations.

The complexity of the mapping and updating in the integration process can be affected by several factors, such as knowledge of the application domain, the number of elements in the local schema, and the size of the collection.

In the case of compound object technologies, such as DCC and OAI-ORE, the mapping process also depends on how the components are aggregated, what is their granularity, which vocabulary each technology is using, how the components are identified and structured, and how they are organized in a schema.

4 Integrating Fingerprint Digital Libraries

Building upon the fingerprint digital libraries summarized in section 2, and the compound object concepts described in section 3, this section presents the "discovery" and reuse/integration of the large amount of data present in fingerprint DLs.

We begin with an example of COs and the four initial sub-systems, as in Figure 3: (A) the recorded prints; (B) the distorted images; (C) the crime scene images; and (D) the training material, with suitable sequencing for pedagogy.

Compound Object 1 (CO1) has the following components: a fingerprint image from system A, one distorted image from system B, a crime scene image from system C, and a link to related training material, taken from system D. The components can be identified by CO1.A.1, CO1.B.1, CO1.C.1 and CO1.D.1, respectively. The CO1 structure can be represented by RDF, while the content could be packaged using OAI-ORE or DCC. The interface of CO1 can comprise the union information of its four components, along with the union of their respective vocabularies (individual, fingers, thumb, quality, distortion, parameters, etc.).

Further, DCC could be used to encapsulate the objects, or even OAI-ORE with a RDF parser, in an integrated DL service, providing *the match between latent and recorded fingerprints*, or *a chain of evidence to convince a jury of confidence of match*, for example. Other integrated DL services could consider *the object versions* (with the composition of distortions, for example) or *correspondence of versions with provenance*, in the crime scene application. Due to the amount of information and detail, these analyses would take longer if the services were not integrated.

Our preliminary results include: (i) an Entity-Relationship Diagram design; (ii) the implementation of the skin distortion model (Figure 2-B); (iii) testing of the blurring distortion (Figure 2-C); (iv) the description of internal steps of the NFIQ quality; and (v) an initial exploration of concepts that will be analyzed from the CO perspective. Though our project is in an early development stage, these preliminary results were important to highlight the amount of information

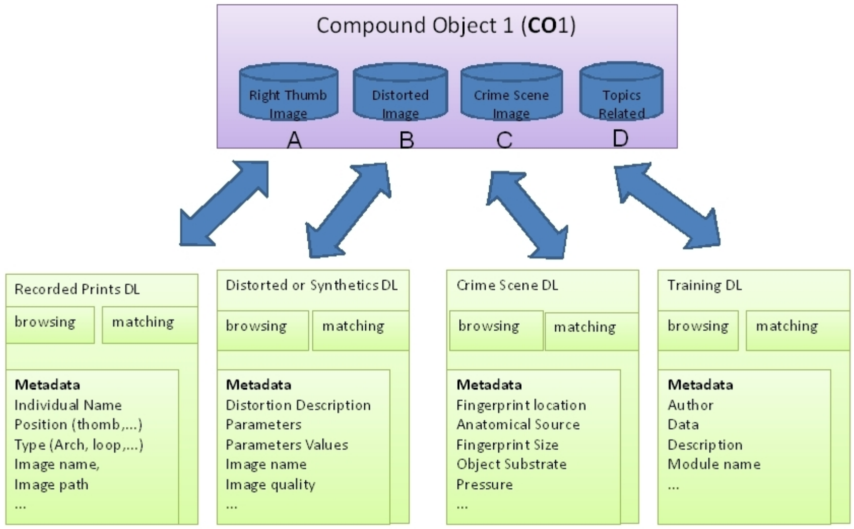

Fig. 3. An example of compound object using four digital libraries: (A) Recorded Prints, (B) Distorted Images, (C) Crime Scene Images, and (D) Training Material.

and details we need to manage, guiding us to explore the overall system by using a very large DL approach.

5 Summary and Conclusions

There are many integrations which relate to addressing large numbers of objects, considering combination, versions, and reuse of information. Our approach takes advantage of volume, concepts, and services already available and manageable in digital libraries.

We presented preliminary results for the integration of fingerprint digital libraries, along with an initial analysis from the compound object perspective. The following items were described: (i) four types of DLs (recorded prints, training materials, crime scenes, and experiments with distorted images); (ii) a summary description about the distortion models accomplished; (iii) examples of services available; and (iv) an initial analysis of CO integration concepts present in the 5S framework, along with minimum services such as harvesting and mapping.

Future work will further address the matching of latent vs. recorded prints, the determination of sufficiency and quality related to the matches, the analysis of other parameters/services for COs, and encapsulation and description using CO technologies.

6 Acknowledgments

We would like to thank CAPES (4479-09-2), FAPESP (2009/18438-7) and CNPq (481556/2009-5 and 306587/2009-2). We are grateful for grants from NIJ (National Institute of Justice) and BAE Systems. Thanks also go to NSF grants IIS-0910183, IIS-0916733, DUE-0840719, and CCF-0722259.

References

1. Jain, A.K., Maltoni, D.: Handbook of Fingerprint Recognition. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2003)
2. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. **40**(2) (2008) 1–60
3. Murthy, U., Kozievitch, N.P., Leidig, J., Torres, R., Yang, S., Goncalves, M., Delcambre, L., Archer, D., Fox, E.A.: Extending the 5S Framework of Digital Libraries to support Complex Objects, Superimposed Information, and Content-Based Image Retrieval Services. Technical Report TR-10-05, Virginia Tech, Department of Computer Science (April 2010)
4. Schank, R.C.: Tell Me a Story: Narrative and Intelligence. Northwestern University Press (1995)
5. Gloe, T., Böhme, R.: The 'Dresden Image Database' for benchmarking digital image forensics. In: SAC '10: Proceedings of the 2010 ACM Symposium on Applied Computing, New York, NY, USA, ACM (2010) 1584–1590
6. Theofanos, M., Micheals, R., Scholtz, J., Morse, E., May, P.: Does habituation affect fingerprint quality? In: CHI '06 Extended abstracts on human factors in computing systems, New York, NY, USA, ACM (2006) 1427–1432
7. Maltoni, D., Cappelli, R.: Advances in fingerprint modeling. Image Vision Comput. **27**(3) (2009) 258–268
8. Agosti, M., Ferro, N., Silvello, G.: The design of a DLS for the management of very large collections of archival objects. First Workshop on Very Large Digital Libraries (2008)
9. Nelson, L., de Sompel, H.V.: IJDL special issue on complex digital objects: guest editors' introduction. International Journal of Digital Libraries **6**(2) (2006) 113–114
10. Saidis, K., Delis, A.: Integrating multi-dimensional information spaces. Second Workshop on Very Large Digital Libraries, Oct. (2009)
11. Grønb, K.: Composites in a Dexter-based hypermedia framework. In: ECHT '94: Proceedings of the 1994 ACM European conference on Hypermedia technology, New York, NY, USA, ACM (1994) 59–69
12. Gruzman, V.A., Senichkin, V.I.: Hypermedia models. Autom. Remote Control **62**(5) (2001) 677–694
13. Maslov, A., Mikeal, A., Phillips, S., Leggett, J., McFarland, M.: Adding OAI-ORE Support to Repository Platforms. Fourth International Conference on Open Repositories (2009)
14. Lagoze, C., Sompel, H.V.: Compound Information Objects: the OAI-ORE Perspective. Open Archives Initiative Object Reuse and Exchange, White Paper, http://www.openarchives.org/ore/documents (2007)
15. Santanchè, A., Medeiros, C.B.: A Component Model and Infrastructure for a Fluid Web. IEEE Transactions on Knowledge and Data Engineering **19**(2) (February 2007) 324–341