FUSION OF BIOMETRIC AND BIOGRAPHIC DATA

FOR LARGE IDENTIFICATION PROJECTS

WHITE PAPER

Introduction

In the past ten years development and implementation of biometrics in identification projects has taken great strides. By now, biometrics are extensively used in military operations, law enforcement, and border control, as well as in commercial applications such as access control and laptop protection. Much of the focus until now has been on performance improvements related to accuracy and speed of single-modality biometrics as well as acceptance of proofs of concept, and now biometrics is primed for widespread use in large-scale identification projects.

However, a growing awareness is emerging that this development brings a new set of challenges, evolving around the notion that biometrics are no longer isolated standalone operations nor that the performance of single biometrics – despite significant progress - is adequate for large-scale identification operations.

Changed conditions force the industry to address and overcome three fundamental challenges:

- **Biographic Legacy Data**
  Identification systems have improved dramatically through the use of biometrics, but must also accommodate large existing repositories of legacy biographic data.
- **Biometrics Accuracy and Interoperability**
  Large-scale identification projects amplify the impact of potentially insufficient accuracy of individual biometric modalities, and interoperability among and between modalities is limited.
- **Cultural name anomalies and practices:** "von", "bin", and other practices that are applied and recorded inconsistently;
- **Large-Scale Infrastructure**
  Demands on large-scale identification system are huge in terms of scalability, reliability, flexibility and performance.

An effective method for addressing all these challenges – simultaneously or successively – needs to be found.

Multi-modal fusion – the technique of combining results of individual biometric modalities into one single one – is often cited as the answer. However, as it is exclusively focused on biometrics many challenges remain. Prime amongst those are the interaction and integration with legacy biographic data, the need to achieve interoperability between different biometric modalities, and the infrastructure that supports the development and operation of large-scale identification systems.

ELISE is a multi-modal fusion platform incorporating advanced search and match algorithms to find and resolve duplicates in legacy biographic databases, multiple fusion algorithms to arrive at a single representation from multiple biometric and biographic identification methods, and the advanced database and infrastructure technologies to achieve the scalability, speed, and reliability required for large-scale operations.

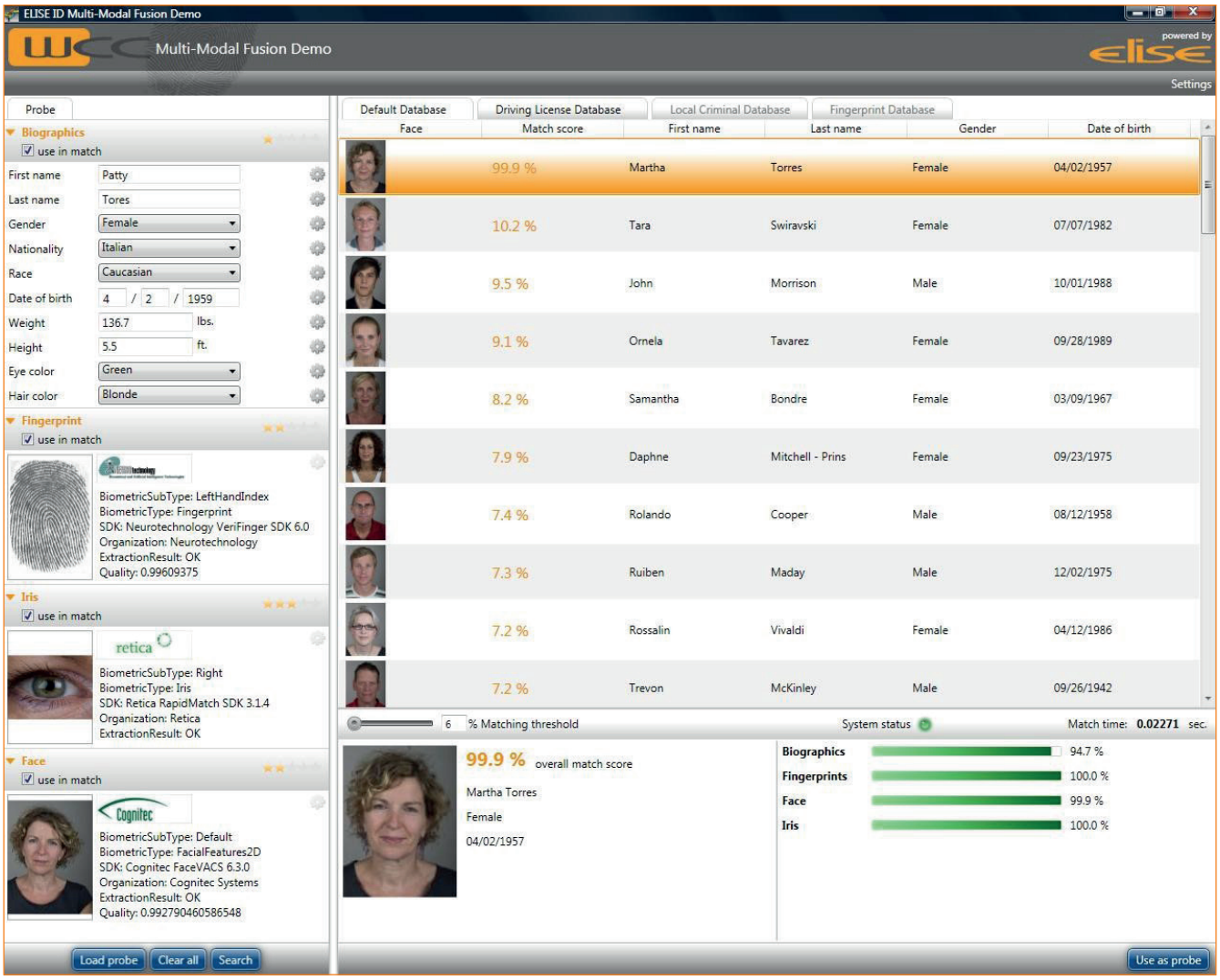
FUSION OF BIOMETRIC & BIOGRAPHIC DATA FOR LARGE IDENTIFICATION PROJECTS

This white paper explores the challenges of large-scale identification and describes the benefits that can be achieved by fusing biographic and multiple biometric identification methods.

Biographic Legacy Data

At the same time that widespread use of biometrics is becoming more and more a reality, a renewed awareness is developing that biometrics is a critical but not the only component of the larger identity puzzle; most if not all situations involving identification of people include biographic data; most IT systems are built around a collection of textual and numeric data that identify a person.

In order to take maximum advantage of the benefits offered by biometrics, it is imperative that organizations not only include pre-existing biographic data in their identification programs, but also that these legacy records are as reliable as possible.



Large-scale identity projects often involve matching both biographic data and biometrics. This demo screen shows how ELISE handles this complex task with ease

FUSION OF BIOMETRIC & BIOGRAPHIC DATA FOR LARGE IDENTIFICATION PROJECTS

Any organization considering enhancing its identification programs with biometrics – either voluntary or mandated (such as may be the case as a result of the HSPD-24 directive) would therefore be wise to ascertain the accuracy of its existing records prior to embarking on a biometrics project. Such efforts start with advanced name search and match technologies to detect and resolve duplicates, and extend into analytical applications to help uncover linkages between people that may assist in alerts for fraudulent or dangerous situations.

At the simplest level, text and numbers are used to name and describe both personal and physical attributes to identify a person. Even at this level, there are many issues that affect the accuracy of identity data. For example, when attributes such as name, address, birth date, and social security number are used, they are subject to erroneous input (e.g. typing errors) and changes over time (e.g. moving to a new address). This is also true of simple physical and weight can both change, eye color can be somewhat subjective, and all can be erroneously recorded.

Additional issues arise with traditional searches when data is missing, partially correct, or out-of-date. In fact, a combination of these issues is quite normal for identity data-bases. The issue becomes even more serious when, for example, several databases from different government agencies must be combined into a single identity search database. Because different databases typically store different attributes, the combined data population is likely to contain missing attributes.

Biometric Accuracy and Interoperability

So far, we have described situations where two names should be treated as equivalent despite data-entry errors. Another class of challenges includes name variations due to natural processes. for a single Arabic name.

The current state of the biometric industry is often a collection of isolated programs, made up of matured pilot programs and powered by vendor-proprietary biometric algorithms. Each of these programs may work exceedingly well within the boundaries of the project identified and executed by a particular organization, but by and large obstruct the emergence of an integrated and interoperable infrastructure required for large-scale identification programs.

Identification programs using multiple modalities offer many advantages: They provide a better coverage of the total population, overcoming issues such as failure to enroll; they discourage deliberate spoofing attempts because it is exponentially harder for imposters to fake more than one biometric trait; and they provide better system resilience in case one system modality fails to operate. They also allow for easier migration as new modalities and/or methods become available. However, attempting to accurately identify an individual by using two or more different modalities could be like comparing apples and oranges. Even comparing results taken from the same modality but with algorithms supplied by different vendors could be problematic.

Finally, rapid development of new biometric modalities that eclipse performance, reliability, speed, and/or ease of use of current ones can quickly result in today's cutting edge biometric identity solution becoming obsolete. For example, many of today's biometric identity solutions are based on fingerprints and facial scans, with iris scans being increasingly deployed. However, even newer systems, such as vascular pattern scans, may offer vastly improved performance, better ease of use, or other benefits not found in current biometrics. Such rapid evolution can present a serious obstacle for today's hardware and software that is hardwired for a particular technology.

Large-Scale Identification Systems

Large-scale identification systems incorporating biometrics face some major challenges. They must be robust to handle high-load demands, scalable and extensible in anticipation of future growth, highly reliable with built-in redundancy to ensure continuous operation, fast, secure and cost-effective.

FUSION OF BIOMETRIC & BIOGRAPHIC DATA FOR LARGE IDENTIFICATION PROJECTS

Specific challenges applicable to biometrics include:

- Will the large number of individuals to be identified affect the time required to make the match? Is there a performance difference if the identification process is applied to 1 million people instead of 10,000? A two second differential may, for instance, result in unacceptable long lines at border control stations or driver license issuance offices.

- Will large-scale identification result in an intolerable number of people falsely accepted or rejected? In other words, in a small, controlled environment of 10,000 people a false acceptance rate (FAR) of 0.1% (or 10 people) might be acceptable, but the same rate applied to a population of 1 million people – resulting in 1,000 people falsely accepted – may pose an unacceptable risk.

- How will national systems enroll and process entire populations to deliver services to their citizens? In a project such as the Unique ID project in India, the government must implement a system that can enroll data from many disparate databases to capture as many citizens as possible in a short period of time. The ability to create unique identities from these disparate databases while maximizing citizen coverage is a challenge that can be conquered by using fusion and multi-modalities.

Each of these are challenges by themselves – in combination they are even more daunting.

Systems for large-scale identification involving biometrics must therefore be designed to minimize these potential shortcomings. This can be achieved by using multiple biometric modalities, merged in a single system, with software that combines the output of each individual modality into one measurement to arrive at a single determination about the identity of an individual. This is known as multi-modal fusion. Results must be delivered fast to accommodate large sets of individuals, and must be accurate to avoid too many erroneously approved or rejected people.

Finally, such systems must be proven in large-scale environments, scalable to accommodate future growth and demands, ultra-reliable to ensure continuous operation, and flexible enough to easily incorporate and adjust for yet unknown advances in biometrics and fusion technologies. And, they must be able to effectively handle input from disparate databases located on- or off-premises or in different geographical locations.

ELISE: The Platform for Multi-modal Fusion

ELISE is a proven software matching system that easily incorporates multiple biometrics from third party biometric vendors and provides a versatile platform for performing multi-modal fusion of biometric and biographic data. With an open, standards-based approach, ELISE offers not only the capability to integrate any existing biometric algorithms, but also to select – or even create – the normalization and/or fusion algorithms best suited for customer-specific conditions.

Multi-modal fusion in ELISE is achieved through robust third-party and proprietary algorithms for score level fusion and score normalization, as well as cascaded fusion. ELISE is versatile and vendor-neutral, meaning that any type of biometric or fusion algorithms can be deployed, depending on the preferences and specific operational conditions of the client. The WCC white paper "Multi-Modal Biometric Fusion Methods in ELISE" provides a comprehensive overview of standard supported third-party and proprietary fusion algorithms.

ELISE is built upon a robust and flexible database architecture that features unlimited scalability to accommodate both large populations and large numbers of match requests, high-availability and complete redundancy with no single point of failure, and an open, extensible infrastructure for integration with virtually any existing IT operation. ELISE is based on non-proprietary, open standards and fully supports the SOA or web services model. ELISE can operate in central or federated mode, thus facilitating infrastructures with large, disparate databases in multiple geographic locations.

FUSION OF BIOMETRIC & BIOGRAPHIC DATA FOR LARGE IDENTIFICATION PROJECTS

Traditional search and match solutions rely on indexed Boolean search of biographic data. These solutions are less suitable to large-scale, multi-layer, multiple-criteria processes, because – like Google – they will return too many results if too few criteria are used and too few usable results if too many criteria are applied.

With ELISE the confidence and accuracy of the match increases when more criteria or layers of criteria are applied so instead of getting fewer results, you get better results. ELISE evaluates each criterion separately, assigns an individual match score, and calculates a combined, fused match score for the search subject based on the individual scores, their weights, and the fusion algorithm used.

Matching

Fuzzy matching: finds duplicate identities

Fuzzy matching in ELISE can be used to create a uniform, complete and reliable database by consolidating identities – including text and numeric data – from several sources. It extracts and normalizes data from each source, performs a match between each record in a particular database, identifies duplicate records by means of scoring multiple factors and helps create a new database of consolidated identities.

Core matching capabilities include:

- Application of fuzzy logic to each criterion
- Differential weighting of each criterion
- Integration of new matching algorithms
- True score level fusion
- Use of metadata to enhance scoring
- Analytics to understand match scores and result ranking
- Semantic matching to extract meaning from unstructured text

The fuzzy searching concept offers more flexibility than traditional keyword searching. With fuzzy searching techniques, it is possible to define ranges and approximations for each search criterion instead of specifying exact values. Not all search criteria are necessarily required, and criteria may be given a relative weight. This results in a greater flexibility and increased quality over traditional searching.

Multi-modal fusion

Score level fusion: delivers optimal accuracy

While there are several types of fusion (decision level, score level, feature level), academic and industry research has shown that score level fusion is the most effective in delivering increased accuracy, because it can reduce the false Acceptance Rate (FAR) as well as the False Reject Rate (FRR).

Score level fusion is the process of calculating one final score from a set of input scores. By analyzing fusion scores as one would with single algorithm outputs, it is possible to construct the regular precision measures.

Because ELISE calculates match scores based on specific algorithms, weight factors, and fusion models, results are deterministic and repeatable. ELISE provides score analytics to understand match scores and result ranking by showing the calculations behind the match scores.

In addition, ELISE can assign weights to the included biometric modalities, which helps to mitigate errors introduced by poor quality input data or lesser performance of one modality versus another. An additional benefit of this functionality is that the system can be fine-tuned to a particular situation or population.

Cascaded fusion: optimizes speed

Cascaded fusion is conceptually similar to score normalization and fusion, but has a different purpose than score level fusion. Instead of attempting to increase accuracy, it tries to optimize computational performance. This, in turn, will lead to a higher throughput of matches. The ELISE implementation of cascaded fusion allows the user to select a fast and a slow but accurate algorithm to work on the same modality.

During a match the fast algorithm is used to calculate initial match scores. All match scores above a certain threshold are passed on to the slow but more accurate algorithm, which then performs its own matching operation on a much smaller fraction of the original number of matches to be made. When the parameters are chosen correctly, the trade-off between accuracy and computational performance can be optimized.

Normalization: prepares for decision-making

Score normalization converts scores of single biometric algorithms into a range where they possess a common meaning in terms of biometric performance. ELISE supports a wide variety of score normalization methods, which are typically deployed depending on specific customer conditions. More information on score normalization is available in the WCC white paper "Multi-Modal Biometric Fusion Methods in ELISE".

System infrastructure

Proven

The ELISE platform has been implemented in many large-scale projects globally, serving government and enterprise clients in identity, cargo, health insurance, and employment applications.

Scalable and Fast

ELISE uses grid computing techniques to achieve linear scalability and to leverage performance of industry standard hardware. ELISE is installed in environments holding millions of data objects and executing hundreds of match requests per second. And, because disk access is kept to an absolute minimum, transactions on large data volumes can be processed very quickly.

ELISE returns match results on biographic and contextual data in the sub-second time frame by applying a combination of industry-standard and proprietary algorithms, while response times for biometric algorithms will be dependent on the execution speed of the third party biometric algorithm.

Reliable

The ELISE architecture includes full redundancy for maximum uptime. It supports multiple nodes, which may be co-located or hosted in physically different locations on different network infrastructures. Its high-availability architecture means that maintenance and upgrades of the ELISE system, associated biometric algorithms, or hardware and software components can be performed without system downtime. There is no single point of failure in the architecture.

Multi-modal fusion in action

So far this paper has looked at the challenges involved with multi-modal biometrics and biographic data in large-scale environments, and described a solution to overcome these challenges. Multi-modal fusion with biometric and biographic data is applicable to many situations and can be deployed in a very flexible manner. Below we will examine some of these situations.

Adding biometrics to large biographic legacy databases

While there are many high-profile examples of successful identification programs involving biometrics, there are many more programs whose identification efforts rely solely on biographic data residing in large and proprietary repositories – for instance driver license and voter registration databases. Identity information is kept in textual and numerical format, covering name, address, date of birth, and physical characteristics such as height and weight, etc.

Over time these repositories of identifying data experience loss of data quality, meaning for instance that people have moved, changed their name through marriage, or changed their physical appearance. It could also mean that data have been erroneously entered upon reissuance of a driver license (Johnson vs Jonnson), or that people have fraudulently applied for a driver license under a different name. The net result is that identity repositories are not 100% accurate and may contain in excess of 5% duplicate entries. The existence of duplicate has a profound effect on organizations relying on accurate driver license information, for instance law enforcement officers during a traffic violation or 911 call, and financial institutions when they are about to extend a mortgage.

Biometrics – single or multiple – are by now a proven and reliable method to gain better insight in the true identity of person applying for or holding a driver license, but prior to embarking on a biometric program it is essential to find, analyze, and resolve duplicate entries in the existing legacy database.

FUSION OF BIOMETRIC & BIOGRAPHIC DATA FOR LARGE IDENTIFICATION PROJECTS

ELISE provides a platform that allows driver license agencies a phased path to more accurate identification. Through fuzzy matching it can find duplicate entries faster and with more certainty than with standard search functionality. ELISE helps to create single records of identity, which form the basis for any future biometric plan under consideration. When these plans materialize, ELISE can easily incorporate the selected biometric algorithm(s) and – again in a phased manner – add the biometric to the biographic record to create a single, fused identity record.

Scaling small biometric pilot projects up to large programs

When biometrics emerged from the research labs in the late 1990s and became a commercially viable method to check the identities, the most common method of implementation was in relatively small programs and pilots. Some of these have grown up, others have not.

The challenge facing successful pilots or small-scale programs is how to replicate the presumably satisfactory performance (speed and accuracy) when they change from a controlled environment to an open, large-scale one.

As shown above, the use of multiple biometrics and multi-modal fusion – when executed well – can be an excellent method to retain both speed and accuracy required in large-scale programs.

ELISE is well-suited to evaluate the performance of currently used biometrics when applied to large-scale environments. It can add a second or third biometric algorithm to enhance the performance of the currently used one, or easily replace the currently used one with one or more biometric algorithms or modalities that perform better in large-scale environments. ELISE provides an open, extensible and robust system infrastructure that helps organizations to confidently scale-up to larger operations.

Managing, controlling, and optimizing existing programs

For organizations already deploying multiple biometric modalities in large-scale identification programs, ELISE provides a single platform for uniting the results of each individual biometric algorithm, instead of running separate disjoint systems.

In addition to obtaining required accuracy and speed, a single platform reduces the costs associated with having to manage separate biometric systems, and allows organizations to deploy a single security regime to control system access to all modalities.

Furthermore, ELISE helps organizations to optimize system performance through fine-tuning or easy replacement of biometric and fusion algorithms.

ELISE, the proven flagship product of WCC Smart Search and Match for large-scale identification, delivers:

- True score level fusion of biometric, biographic, and contextual identifying data, using unlimited criteria;
- Ultra-high speed in searching through large amounts of structured and unstructured data;
- High accuracy with large number of criteria by fusing individual match scores;
- Analytics to understand match scores and results ranking;
- Powerful, real-time matching capabilities for finding similar words, names, concepts, numbers, with complete control over rules, weights, value ranges, and other factors;
- Transparent, explainable, repeatable results for adjudication, query tuning, system tuning;
- Extendable architecture to include any third party matching algorithm (e.g. graphics matchers, alternative phonetic matchers, biometrics, etc.);
- Prevents technology lock-in and obsolescence by allowing the easy integration of future biometric and fusion algorithms;
- Scalability to handle multiple, very large disparate databases in either consolidated or federated fashion.

About WCC

Our vision

People in organizations make decisions. In the markets we focus on, those decisions profoundly impact people's lives. To make the right decisions in an increasingly complex world, it is necessary to have excellent software. That is what drives us at WCC: enabling people to make better decisions.

Our mission & strategy

WCC wants to give people the answers they need, not just the ones they asked for. We thrive on developing software that can connect, combine, and make sense of large amounts of data stored in different systems. Software that can communicate with the users in a human way, and that delivers superior results so our customers can make a difference. We call this "software that matters". But great software alone is not enough to get the best results. What sets WCC apart is the combination of remarkable software with in-depth knowledge of our customers' business. That is why business and implementation consultancy is an important part of our strategy. We focus on two markets: Employment and Identity.

Our products and services

The core of the Employment market is matching people with sustainable jobs effectively and efficiently. WCC has proven to be unequalled in doing just that. Our Employment Platform, which combines unique search and match capability with advanced gap analysis and referral to the right measures, delivers superior strategic value to our customers. Many of the world's largest employment and staffing organizations use our products and expertise, including Randstad, Robert Half, and the public employment services of Germany, France, and the Netherlands.

The security needs of the Identity market are stringent. Border management and law enforcement agencies face the challenge of quickly and accurately identifying people from huge amounts of data spread over many different databases and formats. WCC's software incorporates the necessary evidence-based algorithms, such as multi-cultural name matching, to make correct identifications. HERMES, our API/PNR solution, adheres to industry standards and is easy to implement and operate. Our customers include UNHCR and the European Union.

WCC Smart Search & Match
Zonnebaan 19
3542 EA Utrecht
The Netherlands
T: +31 (0)30 750 32 00

info@wcc-group.com
www.wcc-group.com

WCC Services US Inc.
228 Hamilton Avenue
Suite 300, Palo Alto
CA 94301, USA
T: +1-888-922 9224

WPFOBABD02