Intelligent Information Management, **2010, 2, 499-511**
doi:10.4236/iim.2010.29060 Published Online September 2010 (http://www.SciRP.org/journal/iim)

Intelligent Biometric Information Management

Harry Wechsler
Department of Computer Science, George Mason University, Fairfax, USA
E-mail: *wechsler@gmu.edu*

Received June 10, 2010; *revised July* 29, 2010; *accepted August* 30, 2010.

Abstract

We advance here a novel methodology for robust intelligent biometric information management with inferences and predictions made using randomness and complexity concepts. Intelligence refers to learning, adaptation, and functionality, and robustness refers to the ability to handle incomplete and/or corrupt adversarial information, on one side, and image and or device variability, on the other side. The proposed methodology is model-free and non-parametric. It draws support from discriminative methods using likelihood ratios to link at the conceptual level biometrics and forensics. It further links, at the modeling and implementation level, the Bayesian framework, statistical learning theory (SLT) using transduction and semi-supervised learning, and Information Theory (IY) using mutual information. The key concepts supporting the proposed methodology are a) local estimation to facilitate learning and prediction using both labeled and unlabeled data; b) similarity metrics using regularity of patterns, randomness deficiency, and Kolmogorov complexity (similar to MDL) using strangeness/typicality and ranking p-values; and c) the Cover – Hart theorem on the asymptotical performance of k-nearest neighbors approaching the optimal Bayes error. Several topics on biometric inference and prediction related to 1) multi-level and multi-layer data fusion including quality and multi-modal biometrics; 2) score normalization and revision theory; 3) face selection and tracking; and 4) identity management, are described here using an integrated approach that includes transduction and boosting for ranking and sequential fusion/aggregation, respectively, on one side, and active learning and change/outlier/intrusion detection realized using information gain and martingale, respectively, on the other side. The methodology proposed can be mapped to additional types of information beyond biometrics.

Keywords: Authentication, Biometrics, Boosting, Change Detection, Complexity, Cross-Matching, Data Fusion, Ensemble Methods, Forensics, Identity Management, Imposters, Inference, Intelligent Information Management, Margin gain, MDL, Multi-Sensory Integration, Outlier Detection, P-Values, Quality, Randomness, Ranking, Score Normalization, Semi-Supervised Learning, Spectral Clustering, Strangeness, Surveillance, Tracking, Typicality, Transduction

1. Introduction

Information can be viewed as an asset, in general, and resource or commodity, in particular. Information management [using information technology] stands for the "application of management principles to the acquisition, organization, control, alert and dissemination and strategic use of information for the effective operation of organizations [including information architectures and management information systems] of all kinds. 'Information' here refers to all types of information of value, whether having their origin inside or outside the organization, including data resources, such as production data; records and files related, for example, to the personnel [subject] biometric function; market research data; and competitive intelligence from a wide range of sources. In formation management deals with the value, quality, ownership, use and security of information in the context of organizational performance" [1]. The life-cycle of information includes 1) *creation and acquisition*; 2) *management of information*, e.g., creation of (biometric and forensic) databases, storage, retrieval, sharing and dissemination, leading to and including full-fledged information systems; and purposeful *use of information*. Intelligent information management, according to HP, enables near real-time business intelligence with robust, scalable data management, data-intensive analytics and fusion of structured and unstructured information.

We start with parsing and understanding the meaning of intelligent information management when information refers to biometrics where physical characteristics, e.g., appearance, are used to verify and/or authenticate individuals. Physical appearance and characteristics can include both internal ones, e.g., DNA and iris, and external ones, e.g., face and fingerprints. Behavior, e.g., face expression and gait, and cognitive state, e.g., intent, can further expand the scope of what biometrics stand for and are expected to render. Note that some biometrics, e.g., face expression (see smile) can stand for both appearance, inner cognitive state, and/or medical condition. With information referring here to biometrics one should consider identity management as a particular instantiation of information management. Identity management (IM) is then responsible, among others, with authentication, e.g., (ATM) verification, identification, and large scale screening and surveillance. IM is also involved with change detection, destruction, retention, and/or revision of biometric information, e.g., as people age and/or experience illness. Identity management is most import- ant among others to (homeland) security, commerce, fi nance, mobile networks, and education. A central infrastructure needs to be designed and implemented to enforce and guarantee robust and efficient enterprise-wide policies and audits. Biometric information need to be safeguarded to ensure regulatory compliance with privacy and anonymity best (lawful) practices.

The intelligent aspect is directly related to what biometrics provides us with and the means and ways used to accomplish it. It is mostly about management principles related to robust inference and prediction, e.g., authentication via classification and discrimination, using incremental and progressive evidence ("information") accumulation and disambiguation, learning, adaptation, and closed-loop control. Towards that end, the specific means advocated here include discriminative methods (for practical intelligence) linking the Bayesian framework, forensics, statistical learning, and information theory, on one side, and likelihoods (and odds), randomness, and complexity, on the other side. The challenges that have to be met include coping with incomplete ("occlusion") and corrupt ("disguise") information, image variability, e.g., pose, illumination, and expression (PIE) and temporal change. The "human subject" stands at the center of any IM system. The *subject* interfaces and mediates between *biometric tasks*, e.g., filtering and indexing data, searching for identity, categorization for taxonomy purposes or alternatively classification and discrimination using information retrieval and search engines crawling the web, data mining and business intelligence (for abstraction, aggregation, and analysis purposes) and knowledge discovery, and multi-sensory integration and data fusion; *biometric contents*, e.g., data, information, knowledge, and intelligence / wisdom / meta-knowledge; *biometric organization*, e.g., features, models, and ontologies and semantics; and last but not least, *biometric applications*, e.g., face selection and CCTV surveillance, and mass screening for security purposes. We note here that data streams (subject to exponential growth) and their metamorphosis are important assets and processes, respectively, for each business enterprise. Data and processes using proper context and web services facilitate decision-making and provide value-added to the users. Intelligent information management is further related to autonomic computing and self-managing operations. Intelligent biometric information management revolves mostly around robust predictions, despite interferences in data capture, using adaptive inference. For the remainder of this paper the biometric of interest is the human face with predictions on face identities, and reasoning and inference as the aggregate means to make predictions.

There has been much realization that face recognition is still lacking. The recent call for papers (CFP) for a Special Issue on *Real-World Face Recognition* issued in March 2010 by *IEEE Transactions on Pattern Analysis and Machine Intelligence*, includes as a matter-of-fact the statement "Face recognition in well-controlled environments is relatively mature and has been heavily studied, but face recognition in uncontrolled or moderately controlled environments is still in its early stages." Two significant efforts have been undertaken over the last several years to alleviate the concerns expressed above and to advance the state-of-the art of biometric authentication. One effort is the use of multi-modal but hopefully complementary (relative to authentication) biometrics, while the other effort is geared toward finding better ways and means to fuse the data that such suites of multimodal biometrics acquire and/or derive. The latter effort of fusing data is performed at different levels of functionality and granularity. Data fusion (including multi-sensory integration), however, is just a euphemism for reasoning and sound inferences. As it is practiced today it involves ad-hoc combination rules. The goal for this paper is to advance an integrated, principled and unified methodology for biometric inference, whose realization interfaces between the Bayesian framework (including forensics), Statistical Learning Theory (SLT), and Information Theory (IT) using randomness and complexity concepts. The interface and its implementation are built around discriminative methods whose realization takes place using likelihood ratios (LR) using model-free and non-parametric concepts borrowed from transduction and semi-supervised learning (SSL), e.g., strangeness ("typicality") and p-values (for relative ranking but different from distribution tails). Reasoning and inference are known as iterative processes that accrue evidence to lessen ambiguity and to eventually reach the stage where decisions can be made. Evidence accumulation involves many channels of information and takes place over time and varying contexts. It further requires the means for information aggregation, which are provided in our methodology by boosting methods. Additional expansions to our methodology are expected. In particular, it is apparent that the methodology should involve stage-wise the mutual information, between the input signals and the eventual output labels, and links channel capacity with data compression and expected performance. Compression is after all comprehension (practical inference/intelligence for decision-making) according to Leibniz. The potential connections between the Bayesian framework and SLT, on one side, and Information Theory (IT), on the other side, were recently explored [2].

Complementary to biometric face information is biometric process, in general, and data fusion, in particular. Recent work related to data fusion [3,4] is concerned among others with cross-device matching and device interoperability, and quality dependent and cost-sensitive score level multi-modal fusion. The solutions offered are quality dependent cluster indexing and sequential fusion, respectively. The tasks considered and the solutions proffered can be subsumed by our proposed SLT & IT methodology. In terms of functionality and granularity biometric inference can address multi-level fusion: feature / parts, score (“match”), and detection (“decision”); multi-layer fusion: modality, quality, and method (algorithm); and multi-system fusion using boosting for aggregation and transduction for local estimation and score normalization. This is explained, motivated, and detailed throughout the remaining of the paper as outlined next. Section 2 is a brief on discriminative methods and forensics. Background on randomness and complexity comes in Section 3. Discussion continues in Section 4 on strangeness and p-values, in Section 5 on transduction and open set inference, and in Section 6 on aggregation using boosting. Biometric inference to address specific data fusion tasks is presented in Section 7 on generic multi-level and multi-layer fusion, in Section 8 on score normalization and revision theory, in Section 9 on face selection (tracking mode 1), and in Sect. 10 on identity management (tracking mode 2) using martingale for change detection and active learning. The paper concludes in Section 11 with suggestions for promising venues for future research at the intersection between the Bayesian framework, Information Theory (IT), and Statistical Learning Theory (SLT) using the temporal dimension as the medium of choice.

2. Discriminative Methods and Forensics

Discriminative methods support practical intelligence, in general, and biometric inference and prediction, in particular. Progressive processing, evidence accumulation, and fast decisions are their hallmarks. There is no time for expensive density estimation and marginalization characteristic of generative methods. There are additional philosophical and linguistic arguments that support the discriminative approach. It has to do with practical reasoning and epistemology, when recalling from Hume, that “all kinds of reasoning consist in nothing but a comparison and a discovery of those relations, either constant or inconstant, which two or more objects bear to each other,” similar to non-accidental coincidences and sparse but discriminative codes for association [5]. Formally, “the goal of pattern classification can be approached from two points of view: informative [generative] - where the classifier learns the class densities, [e.g., HMM] or discriminative – where the focus is on learning the class boundaries without regard to the underlying class densities, [e.g., logistic regression and neural networks]” [6]. Discriminative methods avoid estimating how the data has been generated and instead focus on estimating the posteriors similar to the use of likelihood ratios (LR) and odds. The informative approach for 0/1 loss assigns some input $\boldsymbol{x}$ to the class k ε K for whom the class posterior probability P(y = k | **x**)

$$\mathrm{P}\left(\mathrm{y}=k\middle|\mathbf{x}\right)=\mathrm{P}\left(\mathbf{x}\middle|\mathrm{y}=k\right)\mathrm{P}\left(\mathrm{y}=k\right)\Big/\sum\nolimits_{m}^{K}\mathrm{P}\left(\mathbf{x}\middle|\mathrm{y}=m\right)\mathrm{P}\left(\mathrm{y}=m\right)$$

yields the maximum. The MAP decision requires access to the log-likelihood P_θ (**x**, y). The optimal (hyper) parameters $\boldsymbol{\theta}$ are learned using maximum likelihood (ML) and a decision boundary is then induced, which corresponds to a minimum distance classifier. The discriminative approach models directly the conditional log-likelihood or posteriors P_θ(y | **x**). The optimal parameters are estimated using ML leading to the discriminative function

$$\lambda_k\left(\mathbf{x}\right)=\log\left[\mathrm{P}\left(\mathrm{y}=k\middle|\mathbf{x}\right)\Big/\mathrm{P}\left(\mathrm{y}=\mathrm{k}\middle|\mathbf{x}\right)\right]$$

that is similar to the use of the Universal Background Model (UBM) for score normalization and LR definition. The comparison takes place between some specific class membership *k* and a generic distribution (over *K*) that describes everything known about the population at large. The discriminative approach was found [6] to be more flexible and robust compared to informative/generative methods because fewer assumptions are made. One possible drawback for discriminative methods comes from ignoring the marginal distribution P(**x**), which is difficult to estimate anyway. Note that the informative approach is biased when the distribution chosen is incorrect.

The likelihood ratio LR provides straightforward means for discriminative methods using optimal hypothesis testing. Assume that the null "H0" and alternative "H1" hypotheses correspond to impostor ***i*** and genuine ***g*** subjects, respectively. The probability to reject the null hypothesis, known as the false accept rate (FAR) or type I error, describes the situation when impostors are authenticated as genuine subjects by mistake. The probability for correctly rejecting the null hypothesis (in favor of the alternative hypothesis) is known as the hit or genuine acceptance ("hit") rate (GAR). It defines the power of the test $1 - \beta$ with β the type II error when the test fails to reject the null hypothesis when it is false. The Neyman-Pearson (NP) statistical test $\psi(\mathbf{x})$ compares in an optimal fashion the null hypothesis against the alternative hypothesis, e.g., $P(\psi(\mathbf{x}) = H1 \mid H0\} = \alpha$, $\psi(\mathbf{x}) = 1$ when $f_{\mathbf{g}}(\mathbf{x}) / f_{\mathbf{i}}(\mathbf{x}) > \tau$, and $\psi(\mathbf{x}) = H0$ when $f_{\mathbf{g}}(\mathbf{x}) / f_{\mathbf{i}}(\mathbf{x}) < \tau$ with τ some constant. The Neyman-Pearson lemma states that for some fixed FAR = α one can select the threshold τ such that the $\psi(\mathbf{x})$ test maximizes GAR and it is the most powerful test for testing the null hypothesis against the alternative hypothesis at significance level α. Specific implementations for $\psi(\mathbf{x})$ during cascade classification are possible and they are driven by boosting and strangeness (transduction) (see Section 4).

Gonzales-Rodriguez *et al.* [7] provide strong motivation from forensic sciences for the evidential and discriminative use of the likelihood ratio (LR). They make the case for rigorous quantification of the process leading from evidence (and expert testimony) to decisions. Classical forensic reporting provides only "identification" or "exclusion/elimination" decisions and it requires the use of subjective thresholds. If the forensic scientist is the one choosing the thresholds, he will be ignoring the prior probabilities related to the case, disregarding the evidence under analysis and usurping the role of the Court in taking the decision, "… *the use of thresholds is in essence a qualification of the acceptable level of reasonable doubt adopted by the expert*" [8].

The Bayesian approach's use of the likelihood ratio avoids the above drawbacks. The roles of the forensic scientist and the judge/jury are now clearly separated. What the Court wants to know are the posterior odds in favor of the prosecution proposition (*P*) against the defense (*D*) [posterior odds = LR × prior odds]. The prior odds concern the Court (background information relative to the case), while the likelihood ratio, which indicates the strength of support from the evidence, is provided by the forensic scientist. The forensic scientist cannot infer the identity of the probe from the analysis of the scientific evidence, but gives the Court the likelihood ratio for the two competing hypothesis (*P* and *D*). The likelihood ratio serves as an indicator of the discriminating power (similar to Tippett plots) for the forensic system, e.g., the face recognition engine, and it can be used to comparatively assess authentication performance.

The use of the likelihood ratio has been recently motivated by similar inferences holding between biometrics and forensics [9] with evidence evaluated using a probabilistic framework. Forensic inferences correspond to authentication, exclusion, or inconclusive outcomes, and are based on the strength of biometric (filtering) evidence accrued by prosecution and defense competing against each other. The use of the LR draws further support from the US Supreme Court Daubert ruling on the admissibility of scientific evidence [10]. The Daubert ruling called for a common framework that is both transparent and testable and can be the subject of further calibration ("normalization"). Transparency comes from the Bayesian approach, which includes likelihood ratios as mechanisms for evidence assessment ("weighting") and aggregation ("interpretation"). The likelihood ratio LR is the quotient of a similarity factor, which supports the evidence that the query sample belongs to a given suspect (assuming that the null hypothesis is made by the prosecution *P*), and a typicality factor, e.g., UBM (Universal Background Model) which quantifies support for the alternative hypothesis made by the defense *D* that the query sample belongs to someone else (see Sect. 4 for the similarity between LR and the strangeness measure provided by transduction).

3. Randomness and Complexity

Let **#(**z**)** be the length of the binary string z and ***K(z)*** be its Kolmogorov complexity, which is the length of the smallest program (up to an additive constant) that a Universal Turing Machine needs as input in order to output z. The randomness deficiency ***D(z)*** for string z [11] is $D(z) = \#(z) - K(z)$ with D(z) a measure of how random the binary string z is. The larger the randomness deficiency is the more regular and more probable the string z is. Kolmogorov complexity and randomness using MDL (minimum description length) are closely related. Transduction (see Section 4) chooses from all the possible labeling ("identities") for test data the one that yields the largest randomness deficiency, *i.e.*, the most probable labeling. The biometric inference engine is built around randomness and complexity with similarity metrics and corresponding rankings driven by strangeness and p-values throughout the remaining of the paper.

4. Strangeness and p-Values

The strangeness measures the lack of typicality (for a face or face component) with respect to its true or putative (assumed) identity label and the labels for all the other faces or parts thereof. Formally, the strangeness measure α_i is the (likelihood) ratio of the sum of the *k* nearest neighbor (k-nn) distances *d* from the same class *y* divided by the sum of the *k* nearest neighbor (k-nn) distances from all the other classes (y). The smaller the strangeness, the larger its typicality and the more probable its (putative) label *y* is. The strangeness facilitates both feature selection (similar to Markov blankets) and variable selection (dimensionality reduction). One finds empirically that the strangeness, classification margin, sample and hypothesis margin, posteriors, and odds are all related via a monotonically non-decreasing function with a small strangeness amounting to a large margin.

Additional relations that link the strangeness and the Bayesian approach using the likelihood ratio can be observed, e.g., the logit of the probability is the logarithm of the odds, logit (p) = log (p/(1-p)), the difference between the logits of two probabilities is the logarithm of the odds ratio, *i.e.*, log (p/(1-p)/ q/(1-q)) = logit (p) – logit (q) (see also logistic regression and the Kullback-Leibler (KL) divergence). The logit function is the inverse of the "sigmoid" or "logistic" function. Another relevant observation that buttresses the use of the strangeness comes from the fact that unbiased learning of Bayes classifiers is impractical due to the large number of parameters that have to be estimated. The alternative to the unbiased Bayes classifier is logistic regression, which implements the equivalent of a discriminative classifier.

The likelihood-like definitions for strangeness are intimately related to discriminative methods. The p-values suggested next compare ("rank") the strangeness values to determine the credibility and confidence in the putative classifications ("labeling") made. The p-values bear resemblance to their counterparts from statistics but are not the same [12]. p-values are determined according to the relative rankings of putative authentications against each one of the identity classes known to the enrolled gallery using the strangeness. The standard p-value construction shown below, where *l* is the cardinality of the training set *T*, constitutes a valid randomness (deficiency) test approximation [13] for some putative label *y* hypothesis assigned to a *new* sample

$$\mathrm{p}_y(e) = \#\left(i : \alpha_i \geq \alpha_{new}^{y}\right) \Big/ (l+1)$$

P-values are used to assess the extent to which the biometric data supports or discredits the null hypothesis H0 (for some specific authentication). When the null hypothesis is rejected for each identity class known, one declares that the test image lacks mates in the gallery and therefore the identity query is answered with "none of the above." This corresponds to forensic exclusion with rejection. It is characteristic of open set recognition with authentication implemented using *Open Set Transduction Confidence Machine (TCM) – k-nearest neighbor (k-nn)* [14]. TCM facilitates outlier detection, in general, and imposters detection, in particular.

5. Transduction

Transduction is different from inductive inference. It is local inference ("estimation") that moves from particular(s) to particular(s). In contrast to inductive inference, where one uses empirical data to approximate a functionnal dependency (the inductive step [that moves from particular to general] and then uses the dependency learned to evaluate the values of the function at points of interest (the deductive step [that moves from general to particular]), one now directly infers (using transduction) the values of the function only at the points of interest from the training data [15]. Inference now takes place using both labeled and unlabeled data, which are complementary to each other. Transduction incorporates unlabeled data, characteristic of test samples, in the decision-making process responsible for their labeling for prediction, and seeks for a consistent and stable labeling across both (near-by) training ("labeled data") and test data. Transduction seeks to authenticate unknown faces in a fashion that is most consistent with the given identities of known but similar faces (from an enrolled gallery/data base of raw images and/or face templates). The search for putative labels (for unlabeled samples) seeks to make the labels for both training and test data compatible or equivalently to make the training and test error consistent.

Transduction "works because the test set provides a nontrivial factorization of the [discrimination] function class" [16]. One key concept behind transduction (and consistency) is the symmetrization lemma [15], which replaces the true (inference) risk by an estimate computed on an independent set of data, e.g., unlabeled or test data, referred to as 'virtual' or 'ghost samples'. The simplest realization for transductive inference is the method of *k* – nearest neighbors. The Cover – Hart theorem [17] proves that asymptotically, the one nearest neighbor classification algorithm is bounded above by twice the Bayes' minimum probability of error. This mediates between the Bayesian approach and likelihood ratios, on one side, and strangeness / p- values and transduction, on the other side (see below). Similar and complementary to transduction is semi-supervised learning (SSL) [16]. Face recognition requires (for discrimination purposes) to compare face images according to the way they are different from each other and to rank them accordingly. Scoring, ranking and inference are done using the *strangeness* and *p – values*, respectively, as explained below.

Similar to semi-supervised learning, changing the class assignments (characteristic of impostor behavior) provides the bias needed to determine ("infer") the rejection threshold required to make an authentication or to decline making one. Towards that end one re-labels the training exemplars, one at a time, with all the ("impostor") putative labels except the one originally assigned to it. The PSR (peak-to-side) ratio, PSR = $(p_{\max} - p_{\min}) / p_{\mathrm{stdev}}$, traces the characteristics of the resulting p-value distribution and determines, using cross validation, the [a priori] threshold used to identify ("infer") impostors. The PSR values found for impostors are low because impostors do not mate and their relative strangeness is high (and p-value low). Impostors are deemed as outliers and are thus rejected [14]. The same cross-validation is used for similar purposes during boosting.

6. Boosting

The motivation for boosting goes back to Marvin Minsky and Levin Kanal who have claimed at an earlier time that "It is time to stop arguing over what is best [for decision-making] because that depends on context and goal. Instead we should work at a higher level of [information] organization and discover how to build [decision-level] managerial [fusion] systems to exploit the different virtues and evade the different limitations of each of these ways of comparing things" and "No single model exists for all pattern recognition problems and no single technique is applicable for all problems. Rather what we have in pattern recognition is a bag of tools and a bag of problems", respectively. This is exactly what data fusion is expected to do with biometric samples that need to be authenticated. The combination rule for data fusion is now principled. It makes inferences using sequential aggregation (similar to [4]) of different components, which are referred to in the boosting framework as weak learners (see below). Inference takes now advantage of both localization and specialization to combine expertise. This corresponds to an ensemble of method and mixtures of experts.

Logistic regression is a sigmoid function that directly estimates the parameters of P (y|**x**) to learn mappings f : **x** → y or P(y|**x**), e.g., P{y = 1 | **x**} for the case when *y* is Boolean. Logistic regression is behind discriminative methods and likelihood ratios, e.g., label y = 1 if P{y = 1 | **x**} / P{y = 0 | **x**} > 1 (see Section 2). Finally, logistic regression can be approximated by Support Vector Machines (SVM). AdaBoost [18] (see below) minimizes (using greedy optimization) some functional whose minimum defines logistic regression [19], while an ensemble of SVM is functionally similar to AdaBoost [20]. The strangeness is thus quite powerful as it provides alternative but simpler realizations for a wide range of well known discriminative methods for inference, in general, and classification, in particular.

The basic assumption behind boosting is that "weak" learners can be combined to learn any target concept with probability 1 – η. Weak learners, usually built around simple features but here built using the full range of components available for data fusion, learn to classify at better than chance (with probability 1/2 + η for η > 0). AdaBoost [18] works by adaptively and iteratively re-sampling the data to focus learning on samples that the previous weak (learner) classifier could not master, with the relative weights of misclassified samples increased ("refocused") after each iteration. AdaBoost involves choosing *T* effective components h_t to serve as weak (learners) classifiers and using them to construct the separating hyper-planes. The mixture of experts or final boosted (stump) strong classifier *H* is

$$H(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t h_t(\mathrm{x}) > \frac{1}{2}\sum_{t=1}^{T} \alpha_t$$

with α the reliability or strength of the weak learner. The constant 1/2 comes in because the boundary is located mid – point between 0 and 1. If the negative and positive examples are labeled as – 1 and +1 the constant used is 0 rather than 1/2. The goal for AdaBoost is margin optimization with the margin viewed as a measure of confidence or predictive ability. The weights taken by the data samples are related to their margin and explain the AdaBoost's generalization ability. AdaBoost minimizes (using greedy optimization) some risk functional whose minimum defines logistic regression. AdaBoost converges to the posterior distribution of *y* conditioned on ***x***, and the strong but greedy classifier *H* in the limit becomes the log-likelihood ratio test.

The multi-class extensions for AdaBoost are AdaBoost.M1 and .M2, the latter one used to learn strong classifiers with the focus now on both difficult samples to recognize and labels hard to discriminate. The use of features or in our case (fusion) components as weak learners is justified by their apparent simplicity. The drawback for AdaBoost.M1 comes from its expectation that the performance for the weak learners selected is better than chance. When the number of classes is k > 2, the condition on error is, however, hard to be met in practice. The expected error for random guessing is 1 – 1/k; for k = 2 the weak learners need to be just slightly better than chance. AdaBoost.M2 addresses this problem by allowing the weak learner to generate instead a set of plausible labels together with their plausibility (not probability), *i.e.*, $[0, 1]^k$. The AdaBoost.M2 version focuses on the incorrect labels that are hard to discriminate. Towards that end, AdaBoost.M2 introduces a pseudo-loss e_t for hypotheses h_t such that for a given distribution D_t one seeks h_t: **x** × Y → [0,1] that is better than chance. "The pseudo-loss is computed with respect to a distribution over the set of all pairs of examples and incorrect labels. By manipulating this distribution, the boosting algorithm can focus the weak learner not only on hard-to-classify examples, but more specifically, on the incorrect labels *y* that are hardest to discriminate" [18]. The use of Neyman-Pearson is complementary to AdaBoost.M2 training (see Section 2) and can meet pre-specified hit and false alarm rates during weak learner selection.

7. Multi-Level and Multi-Layer Fusion

We discuss here biometric inference and address specific data fusion tasks. The discussion is relevant to both generic multi-level and multi-layer fusion in terms of functionality and granularity. Multi-level fusion involves feature/parts, score ("match"), and detection ("decision"), while multi-layer fusion involves modality, quality, and method (algorithm). The components are realized as weak learners whose relative performance is driven by transduction using strangeness and p-value (see Section 5), while their aggregation is achieved using boosting (see Section 6). Additional data fusion-like tasks are discussed in subsequent sections.

The strangeness is the thread to implement both representation and boosting (learning, inference, and prediction regarding classification). The strangeness, which implements the interface between the biometric representation (including its attributes and/or parts) and boosting, combines the merits of filter and wrapper classification methods. The coefficients and thresholds for the weak learners, including the thresholds needed for open set recognition and rejection are learned using validation images, which are described in terms of components similar to those found during enrollment [21]. The best feature correspondence for each component is sought between a validation and a training biometric image over the component ("parts" or "attributes") defining that component. The strangeness of the best component found during training is computed for each validation biometric image under all its putative class labels *c* (c = 1,…,C). Assuming *M* validation biometric images from each class, one derives *M* positive strangeness values for each class *c*, and M(C – 1) negative strangeness values. The positive and negative strangeness values correspond to the case when the putative label of the validation and training image are the same or not, respectively. The strangeness values are ranked for all the components available, and the best weak learner h_i is the one that maximizes the recognition rate over the whole set of validation biometric images *V* for some component *i* and threshold θ_i. Boosting execution is equivalent to cascade classification [22]. A component is chosen as a weak learner on each iteration (see **Figure 1**).

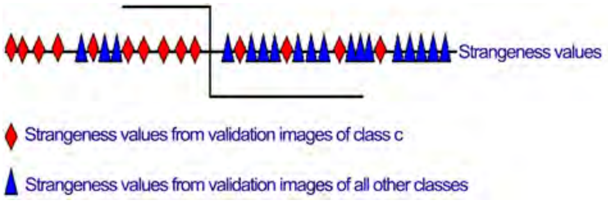
Strangeness values
♦ Strangeness values from validation images of class c
▲ Strangeness values from validation images of all other classes

Figure 1. Learning Weak Learners ("Biometric Components") as Stump Functions.

The level of significance α determines the scope for the null hypothesis H0. Different but specific alternatives can be used to minimize Type II error or equivalently to maximize the power $(1 - \beta)$ of the weak learner [23]. During cascade learning each weak learner ("classifier") is trained to achieve (minimum acceptable) hit rate h = $(1 - \beta)$ and (maximum acceptable) false alarm rate α (see Sect. 2) Upon completion, boosting yields the strong classifier H(**x**), which is a collection of discriminative biometric components playing the role of weak learners. The hit rate after *T* iterations is h^T and the false alarm α^T.

8. Score Normalization, Revision Theory, and CMC Estimation

The practice of score normalization in biometrics aims at countering subject/client variability during verification. It is used (a) to draw sharper class or client boundaries for better authentication and (b) to make the similarity scores compatible and suitable for integration. The emphasis in this section is the former rather than the latter, which was already discussed in Section 7. Score normalization is concerned with adjusting both the client dependent scores and the thresholds needed for decision – making during post-processing. The context for score normalization includes clients *S* and impostors $\neg S$. One should not confuse post processing score normalization with normalization implemented during preprocessing, which is used to overcome the inherent variability in the image acquisition process. Such preprocessing type of normalization usually takes place by subtracting the mean (image) and dividing the result by the standard deviation. This leads to biometric data within the normalized range of [0, 1]. Score normalization during post-processing can be adaptive or empirical, and it requires access to additional biometric data prior and/or during the decision-making process.

The details for empirical score normalization and its effects are as follows [24]. Assume that the PDF of match ("similarity") scores is available for both genuine transactions (for the same client), *i.e.*, P_g, and impostor transactions (between different clients), *i.e.*, P_i. Such information can be gleaned from sets maintained during enrollment or gained during the evaluation itself. One way to calculate the normalized similarity score *ns* for a match score *m* is to use Bayes' rule

$$ns = \mathrm{P}(g|m) = \mathrm{P}(m|g)\mathrm{P}(g)/\mathrm{P}(m)$$

where P(g) is the a priori probability of a genuine event and P($m \mid g$) is the conditional probability of match score *m* for some genuine event *g*. The probability of *m* for all events, both genuine and impostor transactions, is

$$\mathrm{P}(m) = \mathrm{P}(g)\mathrm{P}_g(m) + (1 - \mathrm{P}(g))\mathrm{P}_i(m)$$

The normalized score *ns* is then

$$ns = \mathrm{P}(g)\mathrm{P}_g(m) / \left[\mathrm{P}(g)\mathrm{P}_g(m) + (1 - \mathrm{P}(g))\mathrm{P}_i(m)\right]$$

The accuracy for the match similarity scores depends on the degree to which the genuine and impostor PDF approximate ground truth. Bayesian theory can determine optimal decision thresholds for verification only when the two (genuine and impostor) PDF are known. To compensate for such PDF estimation errors one should fit for the "overall shape" of the normalized score distribution, while at the same time seek to discount for "discrepancies at low match scores due to outliers" [25]. The normalized score serves to convert the match score into a more reliable value.

The motivation behind empirical score normalization using evaluation data can be explained as follows. The evaluation data available during testing attempts to overcome the mismatch between the estimated and the real conditional probabilities referred to above. New (on – line) estimates are obtained for both $\mathrm{P}_g(m)$ and $\mathrm{P}_i(m)$, and the similarity scores are changed accordingly. As a result, the similarity score between a probe and its gallery counterpart varies. Estimates for the genuine and impostor PDF, however, should still be obtained at enrollment time and/or during training rather than during testing. One of the innovations advanced by FRVT 2002 was the concept of virtual image sets. The availability of the similarity (between queries *Q* and targets *T*) matrix enables one to conduct different "virtual" experiments by choosing specific query *P* and gallery *G* sets as subsets of *Q* and *T*. Examples of virtual experiments include assessing the influence of demographics and/or elapsed time on face recognition performance. Performance scores relevant to a virtual experiment correspond to the P × G similarity scores. Empirical score normalization comprromises, however, the very concept of virtual experiments. The explanation is quite simple. Empirical score normalization has no access to the information needed to define the virtual experiment. As a result, the updated or normalized similarity scores depend now on additional information whose origin is outside the specific gallery and probe subsets.

Revision theory expands on score normalization in a principled way and furthers the scope and quality of biometric inference. Semi-supervised learning operates under the smoothness assumption (of supervised learning) that similar patterns should yield similar matching scores; and under the low-density separation assumption for both labeled and unlabeled examples. Training and testing are complementary to each other and one can revise both the labels and matching scores to better accommodate the smoothness assumption. Genuine and imposter individual contributions, ranked using strangeness and p-values, are updated, if there is need to do so, in a fashion similar to that used during open set recognition. This contrasts with the holistic approach where matching scores are re-estimated using the Bayes rule and generative models as described earlier.

The basic tools for revision theory are those of perturbation, relearning, and stability to achieve better learning and predictions and therefore to make better inferences [26]. Perturbations to change labels and/or matching scores in regions of relatively high-density and then re-estimate the margin together with quality measures for the putative assignments made, e.g., credibility and confideence (see Section 10). Gradient-descent and stochastic optimization are the methods of choice for choosing and implementing among perturbations using the regularization framework. Re-learning and stability are relevant as explained next. Transduction seeks consistent labels and matching scores for both training (labeled) and test (unlabeled) data. Poggio *et al.* [26] suggest it is the stability of the learning process that leads to good predictions. In particular, the stability property says that "when the training set is perturbed by deleting one example, the learned hypothesis does not change much. This stability property stipulates conditions on the learning map rather than on the hypothesis space." Perturbations ("what if") should therefore include relabeling, exemplar deletion (s), and updating matching scores. As a result of guided perturbations more reliable and robust biometric inference and predictions become possible.

Identification is different from verification in both functionality and implementation. The closed set recognition case is 1 – MANY rather than 1 – 1 and it retrieves, using repeated 1 – 1 verifications a rank – based list of candidates ordered according to their similarity to the unknown test probe. Rank one corresponds to the gallery image found most similar to the test probe. The percentage of probes for which the top ranked candidate is the correct one defines the probability for rank one. The probabilities for rank ***r*** record the likelihood that the correct gallery image shows at rank ***r*** or lower. The probability points trace the Cumulative Match Curve (CMC). CMC are useful for (ranked) identification only when there is exactly one mate for each probe query, *i.e.*, the gallery sets are singletons. Assume *A* classes enrolled in the gallery, *N* query example, the strangeness ("odds") defined for k = 1 to yield (similar to cross TCM validation) *NA* values, *A* "valid", and *N*(*A-1*) invalid (kind of imposters). Determine for each query and for their correct putative class assignment ***a*,** the corresponding p-value rank ***r*** ε (1, ..., A). The lower the rank ***r***, the more typical the biometric sample is to its true class ***a***. Tabulate the number of queries for each class *A* and normalize by the total number of queries *N*. This yields an estimate for CMC. The presence of more than one mate for some or all of the probes in the gallery, e.g., Smart Gate and FRGC, which employ k > 1 and are thus more in tune with the strangeness and p-values definitions, can be handled in several ways. One way is to declare a probe matched at rank ***r*** if at least one of its mated gallery images shows up at that rank. Similar to the singleton case, tabulate the minimum among the p-values for all samples across their correct mates. Other possibilities would include retrieving all the mates or a fraction thereof at rank ***r*** or better and/or using a weighted metric that combines the similarity between the probe and its mates. There is also the possibility that the test probe set itself consists of several images and/or that both the gallery and the probe sets include several images for each subject. This can be dealt using the equivalent of the Hausdorff distance with the minimum over gallery sets performed in an iterative fashion for query sets or using the minimum over both the gallery and query sets pair wise distances. Last but not least, recall and precision (sensitivity and specificity) and F1 are additional (information retrieval) indexes that can be estimated using the strangeness and p-values in a fashion similar to CMC estimation.

9. Face Selection and Tracking

Face selection expands on the traditional role of face authentication. It assumes that multiple still image sets and/ or video sequences for each enrollee are available during training, and that a data streaming video sequence of face images, usually acquired from CCTV, becomes available during surveillance. The goal is to identify the subset of (CCTV) frames, if any, where each enrolled subject, if any, shows up. Subjects can appear and disappear as time progresses and the presence of any face is not necessarily continuous across (video) frames. Faces belonging to different subjects thus appear in a sporadic fashion across the video sequence. Some of the CCTV frames could actually be void of any face, while other frames could include occluded or disguised faces from different subjects. Kernel k-means and/or spectral clustering [27] using biometric patches, parts, and strangeness and p-values for typicality and ranking are proposed for face selection and tracking. This corresponds to the usual use of tracking during surveillance, while another use of tracking for identity management is deferred to the next section. Face selection counts as biometric inference. Biometric evidence accumulates and inferences on authentication can be made for familiar ("enrolled") faces.

Spectral clustering is a recent methodology for segmentation and clustering. The inspiration for spectral clustering comes from graph theory (minimum spanning trees (MST) and normalized cuts) and the spectral (eigen decomposition) of the adjacency/proximity ("similarity") matrix and its subsequent projection to a lower dimensional space. This describes in a succinct fashion the graph induced by the set of biometric data samples ("patterns"). Minimizing the "cut" (over the set of edges connecting *k* clusters) yields "pure" (homogeneous) clusters.

Similar to decision trees, where information gain is replaced by gain ratio to prevent spurious fragmentation, one substitutes the "normalized cut" (that minimizes the cut while keeping the size of the clusters large) for "cut." To minimize the normal cut (for k = 2) is equivalent to minimize the Raleigh quotient of the normalized graph Laplace matrix L* where $L^* = D^{-1/2}LD^{-1/2}$ with $L = D – W$; W is the proximity ("similarity") matrix and the (diagonal) degree matrix D is the "index" matrix that measures the "significance" for each node. The Raleigh quotient (for k = 2) is minimized for the eigenvector ***z*** corresponding to the second smallest eigenvalue of L*. Given *n* data samples and the number of clusters expected *k*, spectral clustering (for $k > 2$) employs the Raleigh – Ritz theorem and leads among others to algorithms such as Ng, Jordan, and Weiss [28] where one (1) computes W, D, L, and L*; (2) derives the largest *k* eigenvectors $\boldsymbol{z_i}$ of L*; (3) forms the matrix **U** ε $\mathbf{R}^{n \times k}$by normalizing the row sums of $\boldsymbol{z_i}$ to have norm 1; (4) cluster the samples $\boldsymbol{x_i}$ corresponding to $\boldsymbol{z_i}$ using K-means.

An expanded framework that integrates graph-based semi-supervised learning and spectral clustering for the purpose of grouping and classification, *i.e.*, label propagation, can be developed. One takes now advantage of both labeled and mostly unlabeled biometric patterns. The graphs reflect domain knowledge characteristics over nodes (and sets of nodes) to define their proximity ("similarity") across links ("edges"). The solution proposed is built around label propagation and relaxation. The graph and the corresponding Laplacian, weight, and diagonal matrices L, W, and D are defined over both labeled and unlabeled biometric patterns. The harmonic function solution [29] finds (and iterates) on the (cluster) assignment for the unlabeled biometric patterns $\boldsymbol{Y_u}$ as $Y = – (L_{uu})^{-1} L_{ul} Y_l$ with $\boldsymbol{L_{uu}}$ the submatrix of ***L*** on unlabeled nodes and $\boldsymbol{Y_l}$ the group indicator over the labeled nodes. Each row of $\boldsymbol{Y_u}$ reports on the posteriors for the Cartesian product between *k* clusters and *n* biometric samples. Class proportions for the labeled patterns can be estimated and used to scale the posteriors for the unlabeled biometric patterns. The harmonic solution is in sync with a random (gradient) walk on the graph that makes predictions on the unlabeled biometric patterns according to the weighted average of their labeled neighbors.

An application of semi-supervised learning to person identification in low-quality webcam images is described by Balcan *et al.* [30]. Learning takes place over both (few) labeled and (mostly) unlabeled face images using spectral clustering and class mass normalization. The functionality involved is similar to face selection from CCTV. The graphs involved consists of i) time edges for adjacent frames likely to contain the same person (if moving at moderate speed); ii) color edges (over short time interval) assume person's apparel the same; and iii) face edges for similarity over longer time spans. Label propagation and scalability become feasible using ranking [31] and semi-supervised learning and parallel MapReduce [32] (in a fashion similar to Page Rank).

10. Identity Management

Identity management stands for another form of tracking. It involves monitoring the gallery of biometric templates in order to maintain an accurate and faithful rendition of enrolled subjects as times moves on. This facilitates reliable and robust temporal mass screening and it represents yet another for of biometric inference. The surveillance aspect is complementary and its role is to prevent imposters from subverting the security arrangements in place. Open set recognition [14], discussed earlier, is integral to surveillance but it does not provide the whole answer. The effective and efficient proper management of the gallery is the main challenge here and it is discussed next.

There are two main problems with identity management. One is to actively monitor the rendering of biometrics signatures and/or templates and the other is to update them if and when significant changes take place. The two problems correspond to active learning [12] and change detection [25], respectively. Active learning is concerned with choosing the most relevant examples needed to improve on classification both in terms of effectiveness ("accuracy") and efficiency ("number of signatures needed"). The scope for active learning can be expanded to include additional aspects including but not limited to choosing the ways and means to accomplish effectiveness and efficiency, on one side, and adversarial learning, characteristic to imposters, on the other side. Our active learning solution [12] is driven by transduction and it is built using strangeness and p-values. The p-values provide a measure of diversity and disagreement in opinion regarding the true label of an unlabeled example when it is assigned all the possible labels. Let p_i be the p-values obtained for a particular example x_{n+1} using all possible labels $i = 1, \ldots, M$. Sort the sequence of p-values in descending order so that the first two p-values, say, p_j and p_k are the two highest p-values with labels *j* and *k*, respectively. The label assigned to the unknown example is *j* with a p-value of p_j. This value defines the credibility of the classification. If p_j (credibility) is not high enough, the prediction is rejected. The difference between the two p-values can be used as a confidence value on the prediction. Note that, the smaller the confidence, the larger the ambiguity regarding the proposed label. We consider three possible cases of p-values, p_j and p_k, assuming $p_j > p_k$: Case 1. p_j high and p_k low. Prediction "*j*" has high credibility and high-confidence value; Case 2. p_j high and p_k high. Prediction "*j*" has high credibility but low-confidence value; Case 3. p_j low and p_k low. Prediction "*j*" has low credibility and low-confidence value. High uncertainty in prediction occurs for both Case 2 and Case 3. Note that uncertainty of prediction occurs when $p_j \approx p_k$. Define "closeness" as $I(x_{n+1}) = p_j - p_k$ to indicate the quality of information possessed by the example. As $I(x_{n+1})$ approaches 0, the more uncertain we are about classifying the example, and the larger the margin information gain from "advise" is. Active learning will add this example, with its true label, to the training set because it provides new information about the structure of the biometric data model. Extensions to the active learning inference strategy describe above will incorporate error analysis and population diversity characteristic of pattern specific error inhomogeneities (PSEI) [14].

The basics for change detection using martingale are as follows [33]. Assume time-varying multi-dimensional data (stream) matrix $\mathbf{R} = \{R(j) = \mathbf{x}_j\}$ where $R(j)$ are "columns" and stand for time-varying (data stream) biometric vectors. Assume that seeding provides some initial $R(j)$ with $j = 1, \ldots, 10$. K-means clustering will find (in an iterative fashion) center "prototypes" $Q(k)$ for the data stream (seen so far). Define the strangeness corresponding to $R(j)$ using the cluster model (with $R = \{\mathbf{x}_j\}$ standing for data stream and *c* standing for cluster center) and the Euclidean distance $d(j)$ between $R(j)$ and $Q(k)$ for $j >= 10$ and $k = j - 9$ as

$$s\left(\mathrm{R}, \mathbf{x}_j\right) = \left\|\mathbf{x}_j - c\right\|$$

Define p-values as

$$p_i\left\{\left(\mathbf{x}_1, \mathrm{y}_1\right), \ldots, \left(\mathbf{x}_i, \mathrm{y}_i\right), \theta_i\right\}$$
$$= \left[\#\left\{j : s_j > s_i\right\} + \theta_i \#\left\{j : s_j = s_i\right\}\right] \Big/ i$$

where s_j is the strangeness measure for $(\mathbf{x}_j, \mathrm{y}_j)$, $j = 1$, 2, …, I and θ_i is randomly chosen from [0, 1] at instance *i*. Define a family of martingale starting with $M^{\varepsilon}(0) = 1$ and continuing with $M^{\varepsilon}(j)$ indexed by ε in [0, 1]

$$M^{\varepsilon}(j) = \prod_{i=1}^{j} \left\{ \varepsilon (p_i)^{\varepsilon - 1} \right\}$$

The martingale test

$$0 < M^{\varepsilon}(j) < \lambda$$

rejects the null hypothesis H0 "no change in the data stream" for H1 ("change detected in data stream") when M (j) > = λ with the value for λ (empirically chosen to be greater than 2) determined by the FAR one is willing to accept, *i.e.*, $1/\lambda$ = FAR. An alternative (parametric) test, e.g., SPRT, will employ the likelihood ratio (LR) with B < LR < A and decide for H0 as soon as LR(j) < B, decide for the alternative H1 ("change") when LR(j) > A, with B ≈ β (1 – α) and A ≈ (1 – β)/α using α for test significance ("size") and (1 – β) for test power. The changes ("spikes") found can identify critical (transition) states, e.g., ageing, and model an appropriate Hidden Markov Model (HMM) for personal authentication and ID management.

11. Conclusions

This paper proposes a novel all encompassing methodology for robust biometric inference and prediction built around randomness and complexity concepts. The methodology proposed can be mapped to other types of information beyond the biometric modality discussed here. The theoretical framework advanced here for biometric information management is model-free and non-parametric. It draws support from discriminative methods using likelihood ratios to link the Bayesian framework, statistical learning theory (SLT), and Information Theory (IT) using transduction, semi-supervised learning, and mutual information between biometric signatures and/or templates and their labels. Several topics on biometric inference related to i) multi-level and multi-layer data fusion including quality and multi-modal biometrics; ii) cross-matching of capture devices, revision theory, and score normalization; iii) face selection and tracking; and iv) identity management and surveillance were addressed using an integrated approach that includes transduction and boosting for ranking and sequential fusion/aggregation, respectively, on one side, and active learning and change /outlier/intrusion detection using information gain and martingale, respectively, on the other side**.**

One venue for future research would expand the scope of biometric space regarding information contents and processes. Regarding the biometric space, Balas and Sinha [34] have argued that "it may be useful to also employ region-based strategies that can compare noncontiguous image regions." They further show that "under certain circumstances, comparisons [using dissociated dipole operators] between spatially disjoint image regions are, on average, more valuable for recognition than features that measure local contrast." This is consistent with the expectation that recognition-by-parts architectures [21] should learn [using boosting and transduction] "optimal" sets of regions' comparisons for biometric authentication across varying data capture conditions and contexts. The choices made on such combinations for both multi-level and multi-layer fusion amount to "rewiring" operators and processes. Rewiring corresponds to an additional processing and competitive biometric stage. As a result, the repertoire of information available to biometric inference will now range over local, global, and non-local (disjoint) data characteristics with an added temporal dimension. Ordinal rather than absolute codes are feasible in order to gain invariance to small changes in inter-region and temporal contrast. Disjoint and "rewired" patches of information contain more diagnostic information and are expected to perform best for "expression", self-occlusion, and varying image capture conditions. The multi-feature and rewired based biometric image representations and processes together with exemplar-based biometric representations enable flexible matching. The added temporal dimension is characteristic of video sequences and it should lead to enhanced biometric authentication and inference performance using set similarity. Cross-matching biometric devices is yet another endeavor that could be approached using score normalization, non-linear mappings, and revision theory (see Section 8).

Another long-term and needed research venue should consider useful linkages between information theory, the Bayesian framework, and statistical learning theory to advance modes of reliable and robust reasoning and inference with directed application to biometric inference. Such an endeavor will be built around the regularization framework using fidelity of representation, compressive sensing, constraints satisfaction and optimization subject to penalties, and margin for prediction. Biometric dictionaries are also needed for biometric processes to choose from for flexible exemplar-based representation, reasoning, and inference, and to synthesize large-scale databases for biometric evaluations. The ultimate goal is to develop powerful and wide scope biometric language(s) and the corresponding biometric reasoning ("inference") apparatus in a fashion similar to the way language and thought are available for human ("practical") intelligence and inference [35].

12. References

[1] T. D. Wilson, "Information Management," in: J. Feather

and P. Sturges Eds., *International Encyclopedia of Information and Library Science*, Routledge, London, 2003, pp. 263-278.

[2] N. Schmid and H. Wechsler, "Information Theoretical (IT) and Statistical Learning Theory (SLT) Characterizations of Biometric Recognition Systems," *SPIE Electronic Imaging: Media Forensics and Security*, San Jose, CA, Vol. 7541, 2010, pp. 75410M-75410M-13.

[3] N. Poh, T. Bourlai, J. Kittler *et al*., "Benchmark Quality-Dependent and Cost-Sensitive Score-Level Multimodal Biometric Fusion Algorithms," *The IEEE Transaction on Information Forensics and Security*, Vol. 4, No. 4, 2009, pp. 849-866.

[4] N. Poh, T. Bourlai and J. Kittler, "A Multimodal Biometric Test Bed for Quality-dependent, Cost-Sensitive and Client-Specific Score-Level Fusion Algorithms," *Pattern Recognition*, Vol. 43, No. 3, 2010, pp. 1094-1105.

[5] H. B. Barlow, "Unsupervised Learning," *Neural Computation*, Vol. 1, 1989, pp. 295-311.

[6] Y. D. Rubinstein and T. Hastie, "Discriminative Versus Informative Learning," *Knowledge and Data Discovery* (*KDD*), 1997, pp. 49-53.

[7] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano and J. Ortega-Garcia, "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition," *IEEE Transaction on Audio, Speech and Language Processing*, Vol. 15, No. 7, 2007, pp. 2104-2115.

[8] C. Champed and D. Meuwly, "The Inference of Identity in Forensic Speaker Recognition," *Speech Communication*, Vol. 31, No. 2-3, 2000, pp. 193-203.

[9] D. Dessimoz and C. Champod, "Linkages between Biometrics and Forensic Science," in A. K. Jain, Ed., *Handbook of Biometrics*, Springer, New York, 2008.

[10] B. Black, F. J. Ayala and C. Saffran-Brinks, "Science and the Law in the Wake of Daubert: A New Search for Scientific Knowledge," *Texas Law Review*, Vol. 72, No. 4, 1994, pp. 715-761.

[11] M. Li and P. Vitanyi, "An Introduction to Kolmogorov Complexity and Its Applications," 2nd Edition, Springer-Verlag, Germany, 1997.

[12] S. S. Ho and H. Wechsler, "Query by Transduction," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 9, 2008, pp. 1557-1571.

[13] T. Melluish, C. Saunders, A. Gammerman, and V. Vovk, "The Typicalness Framework: A Comparison with the Bayesian Approach," TR-CS, Royal Holloway College, University of London, 2001.

[14] F. Li and H. Wechsler, "Open Set Face Recognition Using Transduction," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 11, 2005, pp. 1686-1698.

[15] V. Vapnik, "Statistical Learning Theory," Springer, New York, 1998.

[16] O. Chapelle, B. Scholkopf and A. Zien (Eds.), "Semi-Supervised Learning," MIT Press, USA, 2006.

[17] T. M. Cover and P. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transaction on Information Theory*, Vol. IT-13, 1967, pp. 21-27.

[18] Y. Freund and R. E. Shapire, "Experiments with a New Boosting Algorithm," *Proceedings of 13th International Conference on Machine Learning* (*ICML*), Bari, Italy, 1996, pp. 148-156.

[19] F. H. Friedman, T. Hastie and R. Tibshirani, "Additive Logistic Regression: A Statistical View of Boosting," *Annals of Statistics*, Vol. 28, 2000, pp. 337-407.

[20] V. Vapnik, "The Nature of Statistical Learning Theory" 2nd Edition, Springer, New York, 2000.

[21] F. Li and H. Wechsler, "Face Authentication Using Recognition-by-Parts, Boosting and Transduction," *International Journal of Artificial Intelligence and Pattern Recognition* (*IJPRAI*), Vol. 23, No. 3, 2009, pp. 545-573.

[22] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Proceedings of the Computer Vision and Pattern Recognition Conference* (*CVPR*), Kauai, Hawaii, 2001, pp. I-511-I-518.

[23] R. O. Duda, P. E. Hart and D. G. Sork, "Pattern Classification," 2nd Edition, Wiley, New York, 2000.

[24] A. Adler, "Sample Images Can be Independently Regenerated from Face Recognition Templates," 2003. http://www.site.uotawa.ca/~adler/publications/2003/adler-2003-fr-templates.pdf.

[25] T. Poggio and S. Smale, "The Mathematics of Learning: Dealing with Data" *Notices of American Mathematical Socity*, Vol. 50, No. 5, 2003, pp. 537-544.

[26] T. Poggio, R. Rifkin, S. Mukherjee and P. Niyogi, "General Conditions for Predictivity of Learning Theory," *Nature*, Vol. 428, No. 6981, 2004, pp. 419-422.

[27] I. S. Dhillon, Y. Guan and B. Kulis, "Kernel k-means, Spectral Clustering and Normalized Cuts," *Proceedings of the Conference on Knowledge and Data Discovery* (*KDD*), Seattle, WA, 2004.

[28] A. Y. Ng, M. I. Jordan and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," *Neural Information Processing Systems* (*NIPS*) 14, MIT Press, Boston, MA, 2002, pp. 849-856.

[29] X. Zhu, Z. Ghahramani and L. Lafferty, "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions," *Proceedings of the 20th International Conference on Machine Learning* (*ICML*), Washington, DC, 2003, pp. 912-919.

[30] M. F. Balcan, A. Blum, P. P. Choi, J. Lafferty, B. Pantano, M. R. Rwebangira and X. Zhu, "Person Identification in Webcam Images: An Application of Semi-Supervised Learning," *Proceedings of 22nd ICML Workshop on Learning with Partially Classified Training Data*, Bonn, Germany, 2005, pp. 1-9.

[31] K. Duh and K. Kirchhoff, "Learning to Rank with Partially Labeled-Data," SIGIR, Singapore, 2008, pp. 20-27.

[32] D. Rao and D. Yarowsky, "Ranking and Semi-Supervised Classification on Large-Scale Graphs Using Map Reduce," *Proceedings of the Workshop on Graph-based*

Methods for Natural Language Processing (*ACL-IJCNLP*), Singapore, 2009, pp. 58-65.

[33] S. S. Ho and H. Wechsler, “A Martingale Framework for Detecting Changes in the Data Generating Model in Data Streams,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, No. 99, 2010 (to appear).

[34] B. J. Balas and P. Sinha, “Region-Based Representations for Face Recognition,” *ACM Transactions on Applied Perception*, Vol. 3, No. 4, 2006, pp. 354-375.

[35] H. Wechsler, “Linguistics and Face Recognition,” *Journal of Visual Languages and Computation*, Vol. 20, No. 3, 2009, pp. 145-155.