EURECOM
Multimedia Communications Department
and
Mobile Communications Department
2229, route des Crêtes
B.P. 193
06904 Sophia-Antipolis
FRANCE

Research Report RR-11-255

# Search Pruning with Soft Biometric Systems: Efficiency-Reliability Tradeoff

June 1st, 2011
Last update June 1st, 2011

Antitza Dantcheva, Arun Singh, Petros Elia and Jean-Luc Dugelay

[1] EURECOM's research is partially supported by its industrial members: BMW Group, Cisco, Monaco Telecom, Orange, SAP, SFR, Sharp, STEricsson, Swisscom, Symantec, Thales.

Tel : (+33) 4 93 00 81 00
Fax : (+33) 4 93 00 82 00
Email : {Antitza.Dantcheva, Arun.Singh, Petros.Elia,
Jean-Luc.Dugelay}@eurecom.fr

# Search Pruning with Soft Biometric Systems: Efficiency-Reliability Tradeoff

Antitza Dantcheva, Arun Singh, Petros Elia and Jean-Luc Dugelay

## Abstract

In the setting of computer vision, algorithmic searches often aim to identify an object inside large sets of images or videos. Towards reducing the often astronomical complexity of this search, one can use pruning to filter out sufficiently distinct objects, thus resulting in a *pruning gain* of an overall reduced search space.

Motivated by practical scenarios such as time-constrained human identification in biometric-based video surveillance systems, we analyze the stochastic behavior of time-restricted search pruning, over large and unstructured data sets which are furthermore random and varying, and where in addition, pruning itself is not fully reliable but is instead prone to errors. In this stochastic setting we explore the natural tradeoff that appears between pruning gain and reliability, and proceed to first provide average-case analysis of the problem and then, using large deviations and informational divergence techniques, to study the atypical gain-reliability behavior, giving insight on how often pruning might fail to substantially reduce the search space. The simplicity of the obtained expressions allows for rigorous and insightful assessment of the pruning gain-reliability behavior, as well as for intuition into designing general object recognition systems.

## Index Terms

database pruning, search pruning, biometrics, soft biometrics, person recognition, reliability–error tradeoff

# Contents

# List of Figures

# 1 Introduction

In recent years we have experienced an increasing need to structure and organize an exponentially expanding volume of data that may take the form of, among other things, images and videos. Crucial to this effort is the often computationally expensive task of algorithmic search for specific elements placed at unknown locations inside large data sets. To limit computational cost, pre-filtering such as pruning can be used, to fast eliminate a portion of the initial data, an action which is then followed by a more precise and complex search within the smaller subset of the remaining data. Such methods can substantially speed up the search, at the risk though of missing the target, thus reducing the overall reliability. Common pre-filtering methods include video indexing and image classification with respect to color [1], patterns, objects [2], or feature vectors [3].

## 1.1 Categorization-based pruning of time-constrained searches over error-inducing stochastic environments

Our interest in analyzing this speed vs. reliability tradeoff, focuses on the realistic setting where the search is time-constrained and where, as we will see later on, the environment in which the search takes place is stochastic, dynamically changing, and can cause search errors. We note here that there is a fundamental difference between search in unstructured versus structured data, where the latter can be handled with very efficient algorithms, such as the sphere decoding algorithm. One widely known practical scenario that adheres to the above stochastic setting, is the scenario of biometric-based video surveillance. In this setting a time constrained search seeks to identify a subject from within a large set of individuals that may consist of, for example, the people surrounding the subject in a specific instance at a specific location. In the language of biometrics we provide analysis on the general speed-reliability behavior in search pruning. In this scenario, a set of subjects can be pruned by means of categorization that is based on different combinations of soft biometric traits such as facial color, shapes or measurements. The need for such biometrically-based search pruning often comes to the fore, such as in the case of the 2005 London bombing where a sizeable fraction of the police force worked for days to screen a fraction of the available surveillance videos relating to the event.

We stay focused on search pruning based on soft-biometrics but remind the reader that this analysis can be generally applied to several domains of computer vision or other disciplines that adhere to the setting of categorization-based pruning in time-constrained searches over error-prone stochastic environments.

## 1.2 Search pruning based on soft biometrics

Soft biometrics are human physical, behavioral or adhered characteristics, which carry information about the individual, are computationally efficient, easy to acquire, but which are generally not sufficient to fully authenticate an individual (cf. [**?**, 4, 5]). Scientific work on using soft biometrics for pruning the search can be found in [6, 7], where a multitude of attributes, like age, gender, hair and skin color were used for classification of a face database, as well as in [8, 9] where the impact of pruning traits like age, gender and race was identified in enhancing the performance of regular biometric systems.

In the setting of human authentication, we consider the scenario where we search for a specific *subject of interest,* denoted as $v'$, belonging to a large and randomly drawn *authentication group* $\boldsymbol{v}$ of $n$ subjects, where each subject belongs to one of $\rho$ *categories*. The elements of the set (authentication group) $\boldsymbol{v}$ are derived randomly from a larger population, which adheres to a set of population statistics. A category corresponds to subjects who adhere to a specific combination of soft biometric characteristics, so for example one may consider a category consisting of blond, tall, females. We henceforth refer to a system which extracts features and classifies them in pre-defined categories, as a soft biometric system (SBS).

With $n$ being potentially large, we seek to simplify the search for subject $v'$ within $\boldsymbol{v}$ by *algorithmic pruning* based on categorization, i.e., by first identifying the subjects that potentially belong to the same category as $v'$, and by then pruning out all other subjects that have not been estimated to share the same traits as $v'$. Pruning is then expected to be followed by careful search of the remaining unpruned set. Such categorization-based pruning allows for a search speedup through a reduction in the search space, from $\boldsymbol{v}$ to some smaller and easier to handle set $\mathcal{S}$ which is the subset of $\boldsymbol{v}$ that remains after pruning, cf. Figure 1. This reduction though happens in the presence of a set of categorization error probabilities $\{\epsilon_f\}$, called confusion probabilities, that essentially describe how easy it is for categories to be confused, hence also describing the probability that the estimation algorithm erroneously prunes out the subject of interest, by falsely categorizing it. This confusion set, together with the set of population statistics $\{p_f\}_{f=1}^{\rho}$ which describes how common a certain category is inside the large population, jointly define the statistical performance of the search pruning, which we will explore. The above aspects will be precisely described later on.

**Example 1** *An example of a sufficiently large population includes the inhabitants of a certain city, and an example of a randomly chosen authentication group ($n$-tuple) $\boldsymbol{v}$ includes the set of people captured by a video surveillance system in the aforementioned city between 11:00 and 11:05 yesterday. An example SBS could be able to classify 5 instances of hair color, 6 instances of height and 2 of gender, thus being able to differentiate between $\rho = 5 \cdot 6 \cdot 2 = 60$ distinct categories. An example search could seek for a subject that was described to belong to the first category of, say, blond and tall females. The subject and the rest of the authentication group of $n = 1000$ people, were captured by a video-surveillance system at approximately the same time and place somewhere in the city. In this city, each SBS-based category appears with probability $p_1, \cdots, p_{60}$, and each such category can be confused for the first category with probability $\epsilon_2, \cdots, \epsilon_{60}$. The SBS makes*

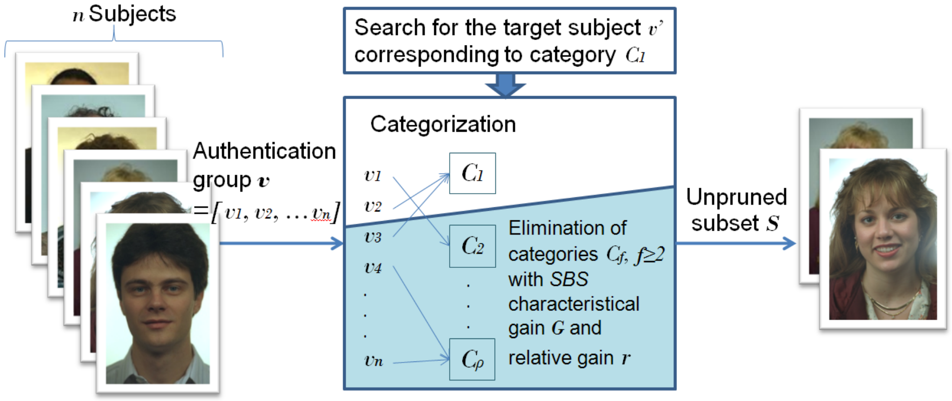
n Subjects
Authentication group v =[ v1, v2, … vn]
Search for the target subject v' corresponding to category C1
Categorization
v1
v2
v3
v4
vn
C1
C2
Cρ
Elimination of categories Cf, f≥2 with SBS characteristical gain G and relative gain r
Unpruned subset S

Figure 1: System overview.

*an error whenever $v'$ is pruned out, thus it allows for reliability of $\epsilon_1$. To clarify, having $p_1 = 0.1$ implies that approximately one in ten city inhabitants are blond-tall-females, and having $\epsilon_2 = 0.05$ means that the system (its feature estimation algorithms) tends to confuse the second category for the first category with probability equal to* $0.05$.

What becomes apparent though is that a more aggressive pruning of subjects in $\boldsymbol{v}$ results in a smaller $\mathcal{S}$ and a higher pruning gain, but as categorization entails estimation errors, such a gain could come at the risk of erroneously pruning out the subject $v'$ that we are searching for, thus reducing the system reliability.

Reliability and pruning gain are naturally affected by, among other things, the distinctiveness and differentiability of the subject $v'$ from the rest of the people in the specific authentication group $\boldsymbol{v}$ over which pruning will take place that particular instance. In several scenarios though, this distinctiveness changes randomly because $\boldsymbol{v}$ itself changes randomly. This introduces a stochastic environment. In this case, depending on the instance in which $v'$ and its surroundings $\boldsymbol{v} - v'$ were captured by the system, some instances would have $\boldsymbol{v}$ consist of bystanders that look similar to the subject of interest $v'$, and other instances would have $\boldsymbol{v}$ consist of people who look sufficiently different from the subject. Naturally the first case is generally expected to allow for a lower pruning gain than the second case.

The pruning gain and reliability behavior can also be affected by the system design. At one extreme we find a very conservative system that prunes out a member of $\boldsymbol{v}$ only if it is highly confident about its estimation and categorization, in which case the system yields maximal reliability (near-zero error probability) but with a much reduced pruning gain. At the other extreme, we find an effective but unreliable system which aggressively prunes out subjects in $\boldsymbol{v}$, resulting in a potentially much reduced search space ($|\mathcal{S}| << n$), at a high risk though of an error. In the above, $|\mathcal{S}|$ denotes the cardinality of set $\mathcal{S}$.

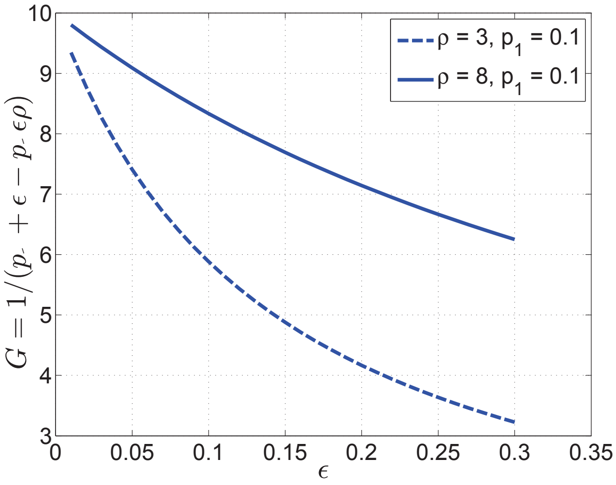

Figure 2: Pruning gain, as a function of the confusability probability $\epsilon$, for the uniform error setting, and for $p_1 = 0.1$. Plotted for $\rho = 3$ and $\rho = 8$.

### 1.3 Contributions

In the next section we elaborate on the concept of *pruning gain* which describes, as a function of pruning reliability, the multiplicative reduction of the set size after pruning: for example a pruning gain of 2 implies that pruning managed to halve the size of the original set. Section 3 provides average case analysis of the pruning gain, as a function of reliability, whereas Section 4 provides atypical-case analysis, offering insight on how often pruning fails to be sufficiently helpful. In the process we try to provide some intuition through examples on topics such as, how the system gain-reliability performance suffers with increasing confusability of categories, or on whether searching for a rare looking subject renders the search performance more sensitive to increases in confusability, than searching for common looking subjects.

Before proving the aforementioned results we hasten to give some insight, in the language of biometrics, as to what is to come. In the setting of large $n$, Section 3 easily tells us that the average pruning gain takes the form of the inverse of $\sum_{f=1}^{\rho} p_f \epsilon_f$, which is illustrated in an example in Figure 4 for different (uniform) confusability probabilities, for the case where the search is for an individual that belongs to a category that occurs once every ten people, and for the case of two different systems that can respectively distinguish 3 or 8 categories. The atypical analysis in Section 4 is more involved and is better illustrated with an example, which asks what is the probability that a system that can identify $\rho = 3$ categories, that searches for a subject of the first category, that has 80 percent reliability, that introduces confusability parameters $\epsilon_2 = 0.2, \epsilon_3 = 0.3$ and operates over a population with statistics $p_1 = 0.4, p_2 = 0.25, p_3 = 0.35$, will prune the search to only

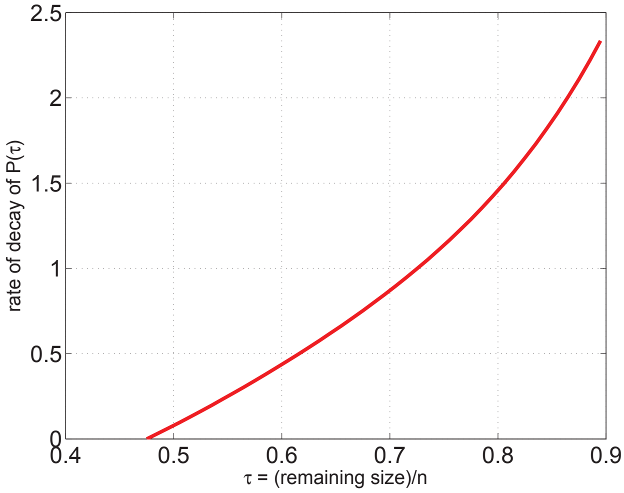

Figure 3: Asymptotic rate of decay of $P(|\mathcal{S}| > \tau n)$, for $\rho = 3$, reliability $0.8$, population statistics $p_1 = 0.4, p_2 = 0.25, p_3 = 0.35$ and confusability parameters $\epsilon_2 = 0.2, \epsilon_3 = 0.3$.

a fraction of $\tau = |\mathcal{S}|/n$. We note that here $\tau$ is the inverse of the pruning gain. We plot in Figure 3 the asymptotic rate of decay for this probability,

$$J(\tau) := -\lim_{n\to\infty} \frac{\log}{n/\rho} P(|\mathcal{S}| > \tau n) \tag{1}$$

for different values of $\tau$. From the $J(\tau)$ in Figure 3 we can draw different conclusions, such as:

- Focusing on $\tau = 0.475$ where $J(0.475) = 0$, we see that the size of the (after pruning) set $\mathcal{S}$ is typically (most commonly - with probability that does not vanish with $n$) $47.5\%$ of the original size $n$. In the absence of errors, this would have been equal to $p_1 = 40\%$, but the errors cause a reduction of the average gain by about $15\%$.

- Focusing on $\tau = 0.72$, we note that the probability that pruning removes less than $1 - 0.72 = 28\%$ of the original set, is approximately given by $e^{-n}$, whereas focusing on $\tau = 0.62$, we note that the probability that pruning removes less than $1 - 0.62 = 38\%$ of the original set, is approximately given by $e^{-n/2}$. The probability that pruning removes less than half the elements is approximately $P(\tau > 0.5) \approx e^{-n/10}$.

The expressions from the above graphs will be derived in detail later on.

# 2 Gain v.s. reliability in soft biometric systems

As an intermediate measure of efficiency we consider the (instantaneous) *pruning gain*, defined here as

$$\mathcal{G}(\boldsymbol{v}) := \frac{n}{|\mathcal{S}|}, \tag{2}$$

which simply describes[1] the size reduction, from $\boldsymbol{v}$ to $\mathcal{S}$, and which can vary from 1 (no pruning gain) to $n$. In terms of system design, one could also consider the *relative gain*,

$$r(\boldsymbol{v}) := 1 - \frac{|\mathcal{S}|}{n} \in [0, 1], \tag{3}$$

describing the fraction of people in $\boldsymbol{v}$ that were pruned out.

It is noted here that $\mathcal{G}(\boldsymbol{v})$, and by extension $r(\boldsymbol{v})$, vary randomly with, among other things, the relationship between $\boldsymbol{v}$ and $v'$, the current estimation conditions as well as the error capabilities of the system. For example, we note that if $\boldsymbol{v}$ and $v'$ are such that $v'$ belongs in a category in which very few other members of $\boldsymbol{v}$ belong to, then the SBS-based pruning is expected to produce a very small $\mathcal{S}$ and a high gain. If though, at the same time, the estimation capabilities (algorithms and hardware) of the system result in the characteristics of $v'$ being easily confusable with the characteristics of another populous category in $\boldsymbol{v}$, then $\mathcal{S}$ will be generally larger, and the gain smaller.

As a result, any reasonable analysis of the gain-reliability behavior must be of a statistical nature and must naturally reflect the categorization refinement, the corresponding estimation error capabilities of the system, as well as the statistics of the larger population.

## 2.1 SBS categorization, estimation capabilities and population statistics

### 2.1.1 Categorization and population statistics

In the setting of interest, we consider that $\boldsymbol{v}$ is chosen at random from a large population, and that everyone in $\boldsymbol{v}$ belongs to one specific category $C_f \subset \boldsymbol{v},\ f = 1, \cdots, \rho$ with probability equal to

$$p_f := \mathbb{E}_{\boldsymbol{v}} \frac{|C_f|}{n},\ \ f = 1, \cdots, \rho. \tag{4}$$

Hence the set of $p_f,\ f = 1, \cdots, \rho$ describes the categorization-based population statistics, i.e., the statistics of the population from where $\boldsymbol{v}$ is randomly picked.

Without loss of generality it is assumed that the subject of interest belongs to the first category, i.e., that $v' \in C_1$. The fact that $v' \in C_1$ is also assumed to be known to the system. In this setting, pruning employs a categorization/estimation algorithm which, correctly or incorrectly, assigns each subject $v \in \boldsymbol{v}$ to a specific

[1] We here assume that the SBS is asked to leave at least one subject in $\mathcal{S}$.

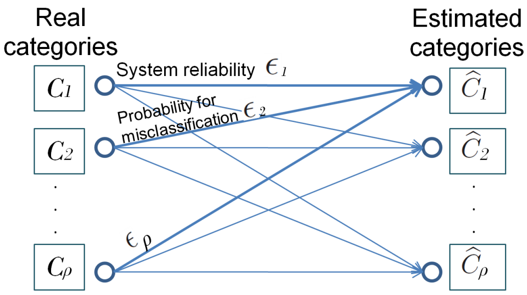

Figure 4: Confusion parameters $\{\epsilon_f\}$.

category $\widehat{C}(v) \in [1, \rho]$. Errors may originate from say, algorithmic failures or reduced image and sensing quality. A subject $v \in \boldsymbol{v}$ is pruned out if and only if $\widehat{C}(v) \neq 1$, i.e., when it is estimated that $v$ is not a member of $C_1$, whereas if $\widehat{C}(v) = 1$ then the subject $v$ is not pruned out and is instead added into the selected set $\mathcal{S}$ of remaining candidates.

### 2.1.2 Error performance capabilities of the SBS

In concisely describing the error performance capabilities of the SBS, we here adopt the simplifying assumption that the confusion probability is a function of the category, i.e., that for any subject $v \in C_f$, the probability that the categorization algorithm does not prune $v$, is a constant denoted as

$$\epsilon_f := P(\widehat{C}(v) = 1), \quad v \in C_f. \tag{5}$$

It becomes clear that $\epsilon_1$ describes the system reliability (1 minus the probability of error), and also that for $f \geq 2$, $\epsilon_f$ describes the probability that any member of $C_f$ is misidentified to share the same characteristics as $v'$, and is thus incorrectly not pruned out. We note that the adopted error measure, albeit an approximation, successfully reflects the fact that different categories may be easier to confuse than others[2]. Specifically having $\epsilon_f > \epsilon_{f'}$ means that people in category $C_f$ can be more easily confused to belong to category $C_1$ of $v'$, than people in category $C_{f'}$.

**Scaling the error with increasing number of categories** Defining the error behavior set $\{\epsilon_f\}$ is an interesting research task in its own right, but it is beyond the scope of this work. We do though provide a few brief thoughts on this, and suggest a simple error measure that can reflect the fact that generally, an increasing refinement of categorization can come at the cost of more erroneous detections. As we are interested in simple insightful expressions, we consider the affine case that

[2] We note that the set $\{\epsilon_f\}$ is simply the first row of what is commonly known as a *confusion matrix*.

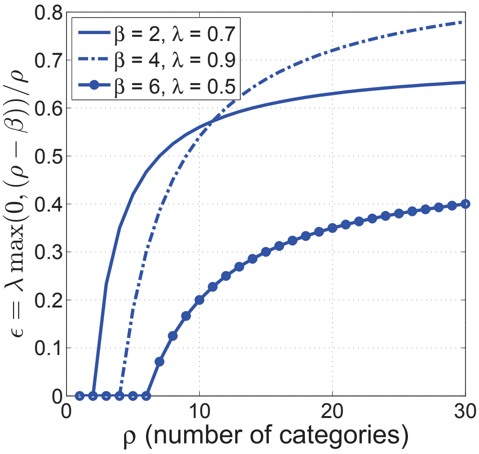

Figure 5: Scaling error probability, $\epsilon_f = \frac{\max(\rho - \beta_f, 0)}{\rho}\lambda_f$.

assigns $\epsilon_f$ the form

$$\epsilon_f = \frac{\max(\rho - \beta_f, 0)}{\rho}\lambda_f, \quad f = 2, \cdots, \rho \tag{6}$$

where $\lambda_f \in [0, 1]$ and $\beta_f \geq 1$ are tuned to fit the categorization capabilities of the system. Specifically the above error measure is suitable for an SBS that introduces negligible probability of categorization error for subjects in $C_f$ whenever $\rho \leq \beta_f$, i.e., whenever it is asked to distinguish between fewer than $\beta_f$ categories. Furthermore in such a system, as the categorization refinement $\rho$ increases, the probability of categorization error for subjects in $C_f$ asymptotically reaches some fixed value $\lambda_f \leq 1$. Figure 5 illustrates this.

## 3 Performance analysis: average behavior of SBS-based pruning

In the following we analyze the pruning gain, and then proceed to suggest another simple metric that combines both gain and reliability.

### 3.1 Pruning gain

We here average $\mathcal{G}(\boldsymbol{v})$ over all possible authentication groups $\boldsymbol{v}$, and over the randomness of the categorization errors $\boldsymbol{w}$, to get the (average) pruning gain

$$\mathcal{G} := \mathbb{E}_{\boldsymbol{v},\boldsymbol{w}}\mathcal{G}(\boldsymbol{v}) \tag{7}$$

as described in the following.

**Proposition 1** *The average pruning rate of categorization-based pruning in SBS is given by*

$$\mathcal{G} = \frac{n}{\mathbb{E}_{\boldsymbol{v},\boldsymbol{w}}|\mathcal{S}|} = \big(\sum_{f=1}^{\rho} p_f \epsilon_f\big)^{-1}. \tag{8}$$

Similarly one can see that the relative gain is averaged to be $r := \mathbb{E}_{\boldsymbol{v},\boldsymbol{w}} r(\boldsymbol{v}) = \sum_{f=1}^{\rho} p_f(1-\epsilon_f)$. We recall that reliability is given by $\epsilon_1$.

The proof of the proposition is simple and is briefly described in the Appendix. We proceed with a clarifying example that follows directly from Proposition 1.

**Example 2 (uniform error setting)** *In the* uniform error setting *where the probability of erroneous categorization of subjects is assumed to be constant and equal to $\epsilon$ for all categories, i.e., where $\epsilon_f = \epsilon = \frac{1-\epsilon_1}{\rho-1}$, $\forall f = 2, \cdots, \rho$, then*

$$\mathcal{G} = \big(p_1 + \epsilon - p_1\epsilon\rho\big)^{-1}, \tag{9}$$

*This was already illustrated in Figure 4. We quickly note that for $p_1 = 1/\rho$, then the gain is equal to $1/p_1$ irrespective of $\epsilon$ and irrespective of the rest of the population statistics.*

*Now considering the case where the uniform error increases with the refinement of the pruning, we set $\epsilon = \frac{\max(\rho-\beta,0)}{\rho}\lambda$, and for any set of population statistics we have*

$$\mathcal{G}(\lambda) = \left(p_1[1 + (\rho-\beta)\lambda] + \frac{\rho-\beta}{\rho}\lambda\right)^{-1}, \tag{10}$$

*which approaches $\mathcal{G}(\lambda) = (p_1[1+(\rho-\beta)\lambda] + \lambda)^{-1}$ as $\rho$ increases. We briefly note that, as expected, in the regime of very high reliability ($\lambda \to 0$), and irrespective of $\{p_f\}_{f=2}^{\rho}$, the pruning gain approaches $\frac{1}{p_1}$. In the other extreme of low reliability ($\lambda \to 1$), the gain approaches $\big(1 - P_{err}\big)^{-1}$.*

### 3.2 Goodput of search pruning

We here identify a simple utility measure, to be referred to as the (average) *goodput* and to be denoted as $\mathcal{U}$, that can be readily used to simultaneously rank the gain and reliability worth of SBS-based pruning. For the sake of simplicity the measure takes the following concise form of a weighted product between reliability and gain,

$$\mathcal{U} := (1 - P_{\text{err}})^{\gamma_1}\mathcal{G}^{\gamma_2} \tag{11}$$

for some chosen positive $\gamma_1, \gamma_2$ that respectively describe the importance paid to reliability and to pruning gain.

We proceed with a clarifying example that focuses on the uniformly scaling error case.

**Example 3 (goodput under uniform error scaling)** *In the uniformly scaling error case where erroneous categorization happens with probability $\epsilon$, and for $\gamma_1 = \gamma_2 = 1$, the goodput is equal to*

$$\mathcal{U}(\epsilon) = \frac{\epsilon + (1 - \epsilon\rho)}{\epsilon + p_1(1 - \epsilon\rho)}. \tag{12}$$

*The goodput starts at a maximum of $\mathcal{U} = \frac{1}{p_1}$ at near zero $\epsilon$, and decreases at a rate of*

$$\frac{\delta\mathcal{U}}{\delta\epsilon} = \frac{p_1 - 1}{[\epsilon + p_1(1 - \rho\epsilon)]^2}, \tag{13}$$

*which as expected*[3] *is negative for all $p_1 < 1$. We here see that $\frac{\delta}{\delta p_1}\delta\frac{\mathcal{U}}{\delta\epsilon}|_{\epsilon\to 0} \to \frac{2-p_1}{p_1^3}$ which is positive and decreasing in $p_1$. Within the context of the example, the intuition that we can draw is that, for the same increase in $\epsilon$*[4]*, a search for a rare looking subject ($p_1$ small) can be much more sensitive, in terms of goodput, to outside perturbations (fluctuations in $\epsilon$) than searches for more common looking individuals ($p_1$ large).*

# 4 How often soft-biometric systems fail in pruning: rare event behavior

In the previous section we analyzed the typical behavior of a system, endowed with the ability to distinguish between $\rho$ categories, having certain estimation error capabilities $\{\epsilon_f\}_{f=1}^{\rho}$ and operating in a general population with statistics given by $\{p_f\}_{f=1}^{\rho}$. Such analysis described how the system behaves *most of the time*.

Let us consider though a scenario where a search for a subject $v'$ turned out to be extremely ineffective, and fell below the expectations, due to a very unfortunate matching of the subject with its surroundings $\boldsymbol{v}$. The natural question is then how often will a system that was designed to achieve a certain average gain-reliability behavior, fall short of the expectations, providing an atypically small pruning gain and leaving its users with an atypically large and unmanageable $\mathcal{S}$.

We begin by recalling that for a given authentication group $\boldsymbol{v}$, the categorization algorithm identifies set $\mathcal{S}$ of all unpruned subjects, defined as

$$\mathcal{S} = \{v \in \boldsymbol{v} \ : \ \widehat{C}(v) = 1\}. \tag{14}$$

We are here interested in the size of the search after pruning, specifically in the parameter

$$\tau := \frac{|\mathcal{S}|}{n/\rho}, \ 0 \leq \tau \leq \rho, \tag{15}$$

[3] We also note here that from the condition that pruning returns at least one element in $\mathcal{S}$, then (cf.(9)) $\epsilon + p_1(1 - \rho\epsilon) > \frac{1}{n}$ which guarantees that $\frac{\delta\mathcal{U}}{\delta\epsilon}$ is finite.

[4] Example for such deterioration can be a reduction in the luminosity around the image-capture cameras.

which represents a relative deviation from a fixed size $n/\rho$ of $\mathcal{S}$. Proposition 1 gave the typical, i.e., common, value of $\tau$ to be

$$\tau_0 := \mathbb{E}_{\boldsymbol{v}} \frac{|\mathcal{S}|}{n/\rho} = \rho \sum_{f=1}^{\rho} p_f \epsilon_f, \tag{16}$$

and we are now interested in the atypical behavior, i.e., we are interested in understanding the probability of having an authentication group $\boldsymbol{v}$ that results in atypically unhelpful pruning ($\tau > \tau_0$), or atypically helpful pruning ($\tau < \tau_0$).

Towards this let

$$\alpha_{0,f} := \frac{|C_f|}{n/\rho}, \tag{17}$$

let $\boldsymbol{a}_0 = \{\alpha_{0,f}\}_{f=1}^{\rho}$ describe the *instantaneous* normalized distribution (histogram) of $\{|C_f|\}_{f=1}^{\rho}$ for the specific, randomly chosen and fixed authentication group $\boldsymbol{v}$, and let

$$\boldsymbol{p} := \{p_f\}_{f=1}^{\rho} = \{\mathbb{E}_{\boldsymbol{v}} \frac{|C_f|}{n}\}_{f=1}^{\rho}, \tag{18}$$

describe the *normalized statistical* distribution of $\{|C_f|\}_{f=1}^{\rho}$.

Furthermore, for a given $\boldsymbol{v}$, let

$$\alpha_{1,f} := \frac{|C_f \cap \mathcal{S}|}{n/\rho}, \ 0 \leq \alpha_{1,f} \leq \rho, \tag{19}$$

let $\boldsymbol{\alpha}_1 := \{a_{1,f}\}_{f=1}^{\rho}$, and for $\boldsymbol{\alpha} := \{\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1\}$, let

$$\mathcal{V}(\tau) := \big\{0 \leq \alpha_{1,f} \leq \min(\tau, \alpha_{0,f}), \sum_{f=1}^{\rho} \alpha_{1,f} = \tau\big\}, \tag{20}$$

denote the set of valid $\boldsymbol{\alpha}$ for a given $\tau$.

Given the information that $\boldsymbol{\alpha}_1$ has on $\boldsymbol{\alpha}_0$, given that $\tau$ is implied by $\boldsymbol{\alpha}_1$, and given that the algorithms here categorize a subject independently of other subjects, it can be seen that for any $\boldsymbol{\alpha} \in \mathcal{V}(\tau)$ then

$$\begin{aligned} P(\boldsymbol{\alpha}, \tau) &= P(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1) = P(\boldsymbol{\alpha}_0) P(\boldsymbol{\alpha}_1 | \boldsymbol{\alpha}_0) && (21) \\ &= \prod_{f=1}^{\rho} P(\alpha_{0,f}) \prod_{f=1}^{\rho} P(\alpha_{1,f} | \alpha_{0,f}). && (22) \end{aligned}$$

The following lemma describes the asymptotic behavior of $P(\boldsymbol{\alpha}, \tau)$, for any $\boldsymbol{\alpha} \in \mathcal{V}(\tau)$, i.e., it describes the (asymptotic rate of decay of the) probability that, under a specific set of population statistics, a specific authentication group with a specific histogram, described by $\boldsymbol{\alpha}$, will cause the pruning algorithm to allow for an unpruned set of size

$$|\mathcal{S}| = \tau \frac{n}{\rho} \tag{23}$$

for some $0 \leq \tau \leq \rho$. This answer will be given below as a concise function of a binomial rate-function (cf. [10])

$$I_f(x) \;=\; \begin{cases} x\log(\frac{x}{\epsilon_f}) + (1-x)\log(\frac{1-x}{1-\epsilon_f}) & f \geq 2 \\ x\log(\frac{x}{1-\epsilon_1}) + (1-x)\log(\frac{1-x}{\epsilon_1}) & f = 1. \end{cases} \tag{24}$$

The lemma follows.

**Lemma 1**

$$\begin{aligned} &-\lim_{n\to\infty} \frac{\log}{n/\rho} P(\boldsymbol{\alpha}, \tau) \\ &\qquad = \rho \sum_{f=1}^{\rho} \alpha_{0,f} \log\big(\frac{\alpha_{0,f}}{p_f}\big) + \sum_{f=1}^{\rho} \big(\alpha_{0,f} - \alpha_{1,f}\big) I_f\big(\frac{\alpha_{1,f}}{\alpha_{0,f} - \alpha_{1,f}}\big). \end{aligned} \tag{25}$$

The proof follows soon after. We now proceed with the main result, which averages the outcome in Lemma 1, over all possible authentication groups.

**Theorem 1** *In SBS-based pruning, the size of the remaining set $|\mathcal{S}|$, satisfies the following:*

$$\begin{aligned} J(\tau) &:= -\lim_{n\to\infty} \frac{\log}{n/\rho} P(|\mathcal{S}| \approx \tau\frac{n}{\rho}) \\ &= \inf_{\boldsymbol{\alpha}\in\mathcal{V}} \rho \sum_{f=1}^{\rho} \alpha_{0,f} \log\frac{\alpha_{0,f}}{p_f} + \sum_{f=1}^{\rho} \big(\alpha_{0,f} - \alpha_{1,f}\big) I_f\big(\frac{\alpha_{1,f}}{\alpha_{0,f} - \alpha_{1,f}}\big). \end{aligned} \tag{26}$$

Furthermore we have the following.

**Theorem 2** *The probability that after pruning, the search space is bigger (resp. smaller) than $\tau\frac{n}{\rho}$, is given for $\tau \geq \tau_0$ by*

$$-\lim_{n\to\infty} \frac{\log}{n/\rho} P(|\mathcal{S}| > \tau\frac{n}{\rho}) = J(\tau) \tag{27}$$

*and for $\tau < \tau_0$*

$$-\lim_{n\to\infty} \frac{\log}{n/\rho} P(|\mathcal{S}| < \tau\frac{n}{\rho}) = J(\tau). \tag{28}$$

The above describe the probability of atypical events (authentication groups) $\boldsymbol{v}$ that deviate from the common behavior described in (16) and that accept atypically ineffective or atypically effective pruning. We offer the intuition that the atypical behavior of the pruning gain is dominated by a small set of authentication groups, that minimize the expression in Theorem 1. Such minimization was presented in Figure 3, and in examples that will follow after the proofs.

We now proceed with the proofs.

*Proof of Lemma 1:*

We first note that

$$P(\boldsymbol{\alpha}_0) \doteq e^{-nD(\boldsymbol{\alpha}_0/\rho||\boldsymbol{p})} = e^{-\frac{n}{\rho}D(\boldsymbol{\alpha}_0||\rho\boldsymbol{p})} \tag{29}$$

where

$$D(\boldsymbol{\alpha}_0||\boldsymbol{p}) = \sum_f \alpha_{0,f} \log \frac{\alpha_{0,f}}{p_f} \tag{30}$$

is the *informational divergence* between $\boldsymbol{\alpha}_0$ and $\boldsymbol{p}$ (cf. [10]). We use $\doteq$ to denote exponential equality, i.e., we write $f(n) \doteq e^{-nd}$ to denote $\lim_{n\to\infty} \dfrac{\log f(n)}{n} = d$ and $\dot{\leq}$, $\dot{\geq}$ are similarly defined. In establishing $P(\boldsymbol{\alpha}_1|\boldsymbol{\alpha}_0)$, we focus on a specific category $f$, and look to calculate

$$P\biggl(|\mathcal{S} \cap C_f| = \frac{n}{\rho}\alpha_{1,f} \;\;|\;\; |C_f| = \frac{n}{\rho}\alpha_{0,f}\biggr), \tag{31}$$

i.e., to calculate the probability that pruning introduces $\frac{n}{\rho}\alpha_{1,f}$ new elements, from $C_f$ to $\mathcal{S}$, given that there are $\frac{n}{\rho}\alpha_{0,f}$ elements of $C_f$. Towards this we note that there is a total of

$$|C_f| = \frac{n}{\rho}\alpha_{0,f} \tag{32}$$

possible elements in $C_f$ which may be categorized, each with probability $\epsilon_f$, to belong to $C_1$ by the categorization algorithm. The fraction of such elements that are asked to be categorized to belong to $C_1$, is defined by $\boldsymbol{\alpha}$ to be

$$x_f := \frac{|\mathcal{S} \cap C_f|}{|C_f|} = \frac{\frac{n}{\rho}\alpha_{1,f}}{|C_f|} = \frac{\alpha_{1,f}}{\alpha_{0,f}}, \tag{33}$$

an event which happens with probability

$$P(x_f) = P\biggl(|\mathcal{S} \cap C_f| = \frac{n}{\rho}\alpha_{1,f} \;\;|\;\; |C_f| = \frac{n}{\rho}\alpha_{0,f}\biggr) \\ \doteq e^{-n_f I_f(x_f)}, \tag{34}$$

where in the above, $I_f(x_f) = x_f \log(\frac{x_f}{\epsilon_f}) + (1-x_f)\log(\frac{1-x_f}{1-\epsilon_f})$ is the rate function of the binomial distribution with parameter $\epsilon_f$ (cf. [11]). Now given that

$$P(\boldsymbol{\alpha}_1|\boldsymbol{\alpha}_0) = \prod_{f=1}^{\rho} P\biggl(|\mathcal{S} \cap C_f| = \frac{n}{\rho}\alpha_{1,f} \;\;|\;\; |C_f| = \frac{n}{\rho}\alpha_{0,f}\biggr) \tag{35}$$

then

$$-\lim_{n\to\infty} \frac{\log}{n/\rho} \log P(\boldsymbol{\alpha}_1|\boldsymbol{\alpha}_0) = (\alpha_{0,f}) I_f(\frac{\alpha_{1,f}}{\alpha_{0,f}}). \tag{36}$$

Finally given that $P(\boldsymbol{\alpha}, \tau) = P(\boldsymbol{\alpha}_0)P(\boldsymbol{\alpha}_1|\boldsymbol{\alpha}_0)$, we conclude that $-\lim_{n\to\infty} \frac{\log}{n/\rho} \log P(\boldsymbol{\alpha}, \tau) = D(\boldsymbol{\alpha}_0||\rho\boldsymbol{p}) + (\alpha_{0,f})I_f(\frac{\alpha_{1,f}}{\alpha_{0,f}})$.

*Proof of Theorem 1:* The proof is direct from Varadhan's lemma (cf. [11]), which applies after noting that $|\mathcal{V}(\tau)| \leq n^{2\rho} \dot{\leq} e^{n\delta}\, \forall \delta > 0$, and that $\sup_{\boldsymbol{\alpha} \in \mathcal{V}(\tau)} P(\boldsymbol{\alpha}) \leq P(\tau) \leq |\mathcal{V}(\tau)| \sup_{\boldsymbol{\alpha} \in \mathcal{V}(\tau)} P(\boldsymbol{\alpha})$.

*Proof of Theorem 2:* The proof is direct by noting that for any $\delta > 0$, then for $\tau \geq \tau_0$ we have

$$-\lim_{n\to\infty} \frac{\log}{n/\rho} P(|\mathcal{S}| > (\tau+\delta)\frac{n}{\rho}) > -\lim_{n\to\infty} \frac{\log}{n/\rho} P(|\mathcal{S}| > \tau\frac{n}{\rho}), \tag{37}$$

and similarly for $\tau < \tau_0$ we have

$$-\lim_{n\to\infty} \frac{\log}{n/\rho} P(|\mathcal{S}| < (\tau-\delta)\frac{n}{\rho}) > -\lim_{n\to\infty} \frac{\log}{n/\rho} P(|\mathcal{S}| < \tau\frac{n}{\rho}). \tag{38}$$

In the following we gain some insight with clarifying examples. We are particularly interested in understanding the behavior of the search pruning in the rare cases where the instances $\boldsymbol{v}$ substantially deviate in distribution $\{|C_f|\}_{f-1}^{\rho}$ from the expected $\boldsymbol{p}$.

**Example 4 (system behavior for rare groups)** *Consider the case where an SBS has $\rho = 2$ categories, distribution probabilities $\boldsymbol{p} = [p,\ 1-p]$, and cross-over probabilities (estimation errors) $\boldsymbol{\epsilon} = [1-\epsilon,\ \epsilon]$. Then, the typical authentication groups $\boldsymbol{v}$ that result in a remaining set of $|\mathcal{S}| \approx \tau\frac{n}{2}$, have distribution given by $\boldsymbol{\alpha}_0' = [\alpha_{0,1},\ \alpha_{1,1} = 1-\alpha_{0,1}]$ which is the solution to $\alpha_{0,1}(2\alpha_{0,1} - \alpha_{1,1}) = (1-\alpha_{0,1})(2 - 2\alpha_{0,1} - \tau + \alpha_{1,1})$.*

*To see this, directly from Lemma 1, we have that*

$$\begin{aligned} I(\alpha,\tau) := -\lim_{n\to\infty} \frac{\log}{n/\rho} P(\boldsymbol{\alpha},\tau) \\ = 2\alpha_{0,1}\log(\frac{\alpha_{0,1}}{p}) + 2\alpha_{0,2}\log(\frac{\alpha_{0,2}}{1-p}) \\ + \alpha_{0,1} I_1(\alpha_{1,1}/\alpha_{0,1}) + \alpha_{0,2} I_2(\alpha_{1,2}/\alpha_{0,2}). \end{aligned} \tag{39}$$

*After some algebra we see that $\inf_{\boldsymbol{\alpha}_0} I(\boldsymbol{\alpha},\tau) = I(\boldsymbol{\alpha}_0', \boldsymbol{\alpha}_1, \tau)$, where $\boldsymbol{\alpha}_0' = [\alpha_{0,1},\ 1-\alpha_{0,1}]$ is the solution to $\alpha_{0,1}(2\alpha_{0,1} - \alpha_{1,1}) = (1-\alpha_{0,1})(2 - 2\alpha_{0,1} - \tau + \alpha_{1,1})$.*

A further clarifying example focuses on the case of a statistically symmetric population.

**Example 5** *Consider the simplifying scenario where the general population is uniformly distributed over the $\rho = 2$ categories, and for simplicity of notation let $y := a_{0,1}$ and $\delta := \alpha_{1,1}$. Then for*

$$I(y,\delta,\tau) := \lim_{n\to\infty} \frac{\log}{n/2} P(|\mathcal{S}| = \frac{n}{2}\tau, y, \delta), \tag{40}$$

*and setting $\boldsymbol{\alpha}_0 = [y,\ 1-y]$, the behavior of $\tau$ and $\boldsymbol{\alpha}_1$ on populations $\boldsymbol{v}$ that are potentially skewed, can be described as follows.*

$$\begin{aligned}\inf_{\delta} I(y,\delta,\tau) &= I(y,\tau,\delta=y\tau)\\ &= 2y\log 2y + 2(1-y)\log 2(1-y) + \tau\log\tau + (2-\tau)\log(2-\tau). \quad (41)\end{aligned}$$

*To see the above, simply calculate the derivative of $I$ with respect to $\delta$. For the behavior of $\tau$ we see that*

$$\begin{aligned}I(\tau) = \inf_{y}\inf_{\delta} I(y,\delta,\tau) &= I(y=p,\delta=p\tau,\tau)\\ &= \tau\log\tau + (2-\tau)\log(2-\tau), \quad (42)\end{aligned}$$

*which can be seen by calculating the derivative of $\inf_\delta I(y,\delta,\tau)$ with respect to $y$.*

# 5 Conclusions

The work provided, in the language of biometrics, statistical analysis of the gain from pruning when applied to searches over large data sets, where these sets are random and where there is a possibility that the pruning may entail errors. In this setting, pruning plays the role of pre-filtering, similar to techniques such as video indexing. The average-case analysis presented here, described the typical assistance that pruning provides in reducing the search space, whereas large-deviations based analysis provided insight as to how often pruning can behave in an atypically unhelpful, or atypically helpful manner. This insight may help in better designing pre-filtering algorithms for different search settings.

# A Proof of Proposition 1

For a given randomly drawn authentication group $\boldsymbol{v}$, let

$$\widehat{S}_f := \{v \in C_f : \widehat{C}(v) = 1\} \quad (43)$$

denote the set of subjects of $C_f$ that were not pruned out by the categorization algorithm, i.e., the subjects that were estimated by the algorithm to have the traits corresponding to $C_1$, and that were thus added to $\mathcal{S}$. Now note that

$$\mathbb{E}_{\boldsymbol{v},\boldsymbol{w}}|\widehat{S}_f| = np_f\epsilon_f, \quad (44)$$

and conclude that $\mathbb{E}_{\boldsymbol{v},\boldsymbol{w}}|\mathcal{S}| = \sum_{f=1}^{\rho}\mathbb{E}_{\boldsymbol{v},\boldsymbol{w}}|\widehat{S}_f| = \sum_{f=1}^{\rho} np_f\epsilon_f$ which results in $\mathcal{G} = \frac{n}{\mathbb{E}_{\boldsymbol{v},\boldsymbol{w}}|\mathcal{S}|} = \left(1-\sum_{f=1}^{\rho}p_f(1-\epsilon_f)\right)^{-1}$.

# References

[1] M.J. Swain and D.H. Ballard, "Color Indexing," *International Journal of Computer Vision*, 7(1): 11–32, 1991. 1

[2] S. Agarwal, A. Awan and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(11): 1475–1490, 2004. 1

[3] E. Nowak, F. Jurie and B. Triggs "Sampling strategies for bag-of-features image classification," *in Proc. of ECCV,* 2006. 1

[4] A. K. Jain, S. C. Dass and K. Nandakumar, "Can soft biometric traits assist user recognition?," *in Proc. of SPIE,* 5404: 561–572, 2004. 2

[5] A.K. Jain, S.C. Dass and K. Nandakumar, "Soft biometric traits for personal recognition systems," *in Proc. of ICBA,* 2004. 2

[6] N. Kumar, P. N. Belhumeur and S. K. Nayar, "FaceTracer: a search engine for large collections of images with faces," *in Proc. of ECCV,* 2008. 2

[7] N. Kumar, A. C. Berg, P. N. Belhumeur and S. K. Nayar, "Attribute and simile classifiers for face verification," *in Proc. of IEEE ICCV,* 2009. 2

[8] G. Givens, J. R. Beveridge, B.A. Draper and D. Bolme, "A statistical assessment of subject factors in the pca recognition of human faces," *in Proc. of IEEE CVPR,* 2003. 2

[9] E. Newham, "The biometric report," SJB Services, New York, 1995. 2

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd edition, ISBN: 0-471-24195-4, Wiley, 2006. 12, 13

[11] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, New York: 2nd edition, Springer-Verlag, 1998. 13, 14