Delta Divergence: A Novel Decision Cognizant Measure of Classifier Incongruence

Josef Kittler, *Life Member, IEEE*, and Cemre Zor

***Abstract*—In pattern recognition, disagreement between two classifiers regarding the predicted class membership of an observation can be indicative of an anomaly and its nuance. Since, in general, classifiers base their decisions on class *a posteriori* probabilities, the most natural approach to detecting classifier incongruence is to use divergence. However, existing divergences are not particularly suitable to gauge classifier incongruence. In this paper, we postulate the properties that a divergence measure should satisfy and propose a novel divergence measure, referred to as *delta divergence*. In contrast to existing measures, it focuses on the dominant (most probable) hypotheses and, thus, reduces the effect of the probability mass distributed over the non dominant hypotheses (clutter). The proposed measure satisfies other important properties, such as symmetry, and independence of classifier confidence. The relationship of the proposed divergence to some baseline measures, and its superiority, is shown experimentally.**

***Index Terms*—Classifier incongruence, divergence clutter, *f*-divergences, total variation distance.**

I. Introduction

DIVERGENCE in information theory has been intensively studied and researched over the last six decades. On the one hand, the massive interest in the subject has been driven by the diversity of applications where divergence plays the key role as an objective function. On the other hand, the investigation of the underlying theoretical properties of divergence has motivated the discovery of new measures with tailor made characteristics that are fine tuned for specific applications. This dual drive has produced extensive families of divergences which are encapsulated in the generic expressions presented, e.g., in [15], with many specific examples listed in [27]. We shall provide a very brief overview of these developments in Section II and give representative examples in Section III-A.

The key designation of divergences is to measure differences between two probability distributions. These distributions can be related to discrete random variables such as symbols in communication systems, or continuous random variables when comparing, for example, two density functions. The differences can also stem from comparing an empirical distribution of some data, and its parametric model. In decision making applications the two distributions could be *a posteriori* class probability functions of observations to be classified. The nuances of these different applications call for divergences of different properties and the existing spectrum of divergence measures bears witness to the endeavors in the field reported over the decades.

In this paper, we focus on the use of divergences to measure incongruence of two classifiers. The problem arises in complex decision making systems which often perform sensor data classification tasks using multiple classifiers. Examples of such systems include classifiers processing different modalities of data, ensemble of classifiers aiming to improve classification performance, or hierarchical classification systems where the base classifiers at one level feed their outputs to a contextual classification level. At this decision level, the context provided by neighboring objects is used to improve performance, or derive structural interpretation of the input data. These multiple classifiers voice their opinions about a given set of hypotheses, expressed in terms of *a posteriori* class probability for each possible outcome.

In decision making systems engaging multiple classifiers, one would normally expect all the classifiers to support the same hypothesis. A classifier disagreement usually signifies something abnormal; a subsystem malfunction, a sensor data modality being absent, or some anomalous event or situation in the observed scene. It is therefore desirable to monitor classifier outputs with the aim of detecting "surprising" classifier incongruence as a trigger for a deeper investigation of its possible causes.

In information theory, the magnitude of surprise is intimately linked to the probabilities of the outcome of an experiment. In the decision making context considered in this paper, the experiment outcome is the true class membership of a given observation (i.e., finding out which class hypothesis is correct). For outcomes of low probability the surprise is huge, whereas for events that are certain (with probability approaching unity) the surprise is null. The conventional way of measuring the amount of information learned from an outcome with probability P is using the logarithm of the inverse of P. The information gain from an experiment is then measured by averaging over all the possible outcomes.

In the case of classifier incongruence we are interested in measuring the information gain from an outcome involving

Manuscript received September 7, 2017; revised January 8, 2018; accepted March 19, 2018. This work was supported in part by the U.K. Defence Science and Technology Laboratory, and in part by the Engineering and Physical Research Council (collaboration between U.S. DOD, U.K. MOD, and U.K. EPSRC through the Multidisciplinary University Research Initiative) under Grant EP/K014307/1 and Grant EP/R018456/1. This paper was recommended by Associate Editor R. Tagliaferri. *(Corresponding author: Cemre Zor.)*

The authors are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: j.kittler@surrey.ac.uk; c.zor@surrey.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2018.2825353

two or more classifiers. For the sake of simplicity, in this paper we shall consider two classifiers only. More specifically, we have two random variables representing class identities, with their distributions, and the question is whether the classifiers agree in supporting the various class hypotheses, or disagree. The nature of information gain from an experiment changes to a comparison of the respective probabilities of possible outcomes. Congruent classifiers would have similar probability distributions over classes, whereas for incongruent the distributions would be different.

Measuring the information gain from an experiment involving two classifiers is different from quantifying the gain from learning the outcome involving a single classifier. What matters in the case of two classifiers is their comparison. Even if the information gain associated with an experiment involving a single classifier is huge, if two classifiers have the same *a posteriori* class probability distribution, they will be congruent.

A common criterion used for comparison of the distributions of two random variables is divergence. The most popular divergence measure is the Kullback–Leibler (KL) divergence, referred to by Itti and Baldi [18] as *Bayesian surprise* measure. It has been used as a measure of classifier incongruence by Weinshall *et al.* [44], but it is not ideal for a number of reasons.

1) If the distributions are different, the value of incongruence will depend on the actual class probability distributions, rather than on probability differences only.
2) Measure is asymmetric, i.e., its value depends on which of the two classifier distributions is used as a reference.
3) Its values are unbounded, which makes it difficult to set a threshold on congruence.
4) In multiclass problems the nondominant classes contribute to clutter, which makes the divergence very noisy.

Some of the above drawbacks have been addressed by alternative divergence measures discussed in Section II. The symmetrized KL divergence recovers the symmetry property. The Jensen–Shannon divergence [28] is both, symmetric, and bounds the range of its values to the interval [0, 1]. However, neither of these measures address properties 1) and 4). In search for more suitable candidates, one can consider the general family of f-divergences [27]. It includes, the divergences based on the Rényi [39] α-entropies, of which the commonly used Shannon entropy—the basis of KL divergence—is a special case for $\alpha = 1$. Another interesting member is, for instance, the α-entropy, for $\alpha = \infty$, defined entirely by the probability of the most likely hypothesis, which is used for decision making by each classifier. This choice would avoid the problem of clutter in property 4), but this particular property migrates to the associated α-divergence family in an undesirable way by focusing on the maximum ratio of *a posteriori* probabilities, which can emanate from nondominant hypotheses. This can potentially provide a highly misleading information about classifier incongruence. The only measure that addresses the problem of clutter is the decision cognizant KL (DC-KL) divergence [38], but, as in the case of KL divergence, its values are unbounded.

In this paper, we address the problem of measuring classifier incongruence by first introducing the mathematical framework and our baseline—the KL divergence. This classical information theory divergence is critically assessed in the context of classifier incongruence detection. The critical analysis allows us to identify the properties that a divergence should possess to be able to serve as a measure of classifier incongruence effectively. A brief overview of the options offered by existing tools, and their ability to satisfy the incongruence measure properties identified provides the motivation for a new measure, called *delta divergence*. Its basis is total variation distance, but we eliminate the clutter by noting that classifier congruence assessment involves only at most three outcomes of material interest: the two classes predicted by the two classifiers, plus the possibility that the true class is neither of the two. The proposed divergence is a function of the absolute value of the difference of the *a posteriori* class probabilities estimated by the respective classifiers for the dominant hypotheses. It is shown to exhibit all the required properties, i.e., being bounded, symmetric, decision cognizant, and decision confidence independent. The relationship of the proposed divergence with state-of-the-art classifier incongruence measures highlight its advantages which are also confirmed experimentally by showing the effect of clutter on the KL divergence, as well as on other baseline measures.

This paper is organized as follows. The related literature is briefly reviewed in Section II. Section III-A introduces the mathematical framework and analyses the properties of KL divergence from the point of view of detecting classifier incongruence. As an outcome of this analysis the properties required by any measure of classifier incongruence are postulated in Section III-B. After a brief discussion of the properties of other existing tools for measuring classifier incongruence, a new divergence is proposed in Section III-C and its properties established in Section III-D. The novel, *decision cognizant* formulation of the classifier incongruence detection problem mitigates the clutter generated by nondominant class hypotheses. This is first shown analytically and later experimentally in Section III-D. In Section IV, we discuss the relationship of the proposed divergence with some baseline criteria as well as with the recently advocated heuristic measures of classifier incongruence. Section V presents illustrative examples of applications of classifier incongruence measures and demonstrates the advantages of delta divergence on real data relating to the problem of detecting incongruence of face and fingerprint modalities in a multimodal biometric system. Section VI draws this paper to conclusions.

II. Related Work

The introduction of the concept of divergence is attributed to Jeffreys [19] who proposed it as a measure for comparing the likelihood of two competing hypotheses in statistical hypothesis testing. Jeffreys' [19] divergence is defined as the difference between the means of the log likelihood ratio computed, respectively, under the two hypotheses. However, earlier references to the notion of divergence can be traced back to Mahalanobis [30] in his work on measures for comparing two statistical populations, and Bhattacharyya [4] who proposed to measure the distance between two distributions using the cosine of the angle between the vectors whose components are constituted by the square root of the values of the associated two probability distributions. The Bhattacharyya coefficient is closely related to the Hellinger distance (see [33]) which dates as far back as 1909.

In spite of the above credits, the key impetus of the intensive study of the topic over the last six decades was the information theoretic notion of divergence proposed by Kullback and Leibler [26]. Inspired by the seminal work of Shannon [41] on information theory, Kullback and Leibler [26] conceived divergence as the relative gain in information received from an experiment involving two probability distributions relating to the same random variable.

In their original paper, the authors define divergence as the mean information for discrimination between two competing hypotheses. They point out a link between divergence and Fisher's [13] information, and therefore the relevance of the information theoretic notion of divergence to statistical estimation theory. This paper also establishes basic properties of KL divergence, including its nonnegativity and the conditions that would need to be satisfied for divergence to exhibit the property of transformation invariance.

One of the factors constraining the use of the KL divergence involving probability densities is the requirement that the probability distributions are absolutely continuous. To overcome this problem, Lin [28] proposed an alternative, the Jensen–Shannon divergence, which mitigates this problem and renders his measure more generally applicable.

The information theoretic framework inspired immense interest in theoretical properties of KL divergence and led to its generalization using other entropy functions such as α-entropy of Rényi [39], which includes the KL divergence as a special case. An even broader generalization was proposed by Csiszár [9] under the name of f-divergences. The family of f-divergences is defined by various choices of convex functions of the likelihood ratio of the respective probability distribution values associated with the alternative hypotheses [27], [33]. The family is included in the class of yet more general divergences known as Bregman divergences (see [42]).

The properties of the numerous divergences have been intensively studied by [9], [27], and [37]. The studies investigate divergence measure characteristics such as boundedness, finiteness, additivity for independent observations, behavior under transformation [36], symmetry, sensitivity to outliers, treatment of inliers, uniqueness, range, behavior in the case of the two distributions being orthogonal [14], convergence of quantized divergences [17], and relationships between divergences and their mutual bounds. For instance, some divergence measures are less amenable to analytical simplification, and mutual bounds are useful to compare them with those measures that can be analytically developed for certain types of distributions (e.g., KL divergence for normal distributions). There is interest in establishing the existence of metric properties, as well as topological and geometric properties. The study of topological and geometric properties of f-divergences by Csiszár [10], [11] led to the advocation of perimeter divergences [32] and their generalization proposed by Österreicher and Vajda [34].

An interesting overview of the properties of f-divergences is presented in [27]. The authors provide elegant derivations of the well known properties based on the Taylor expansion of f-divergences, rather than by resorting to the commonly adopted approach based on Jensen's inequality. The subject of properties of divergence measures continues to generate interest even now, especially in the context of specific applications [40].

Divergence measures have been used for diverse applications in pattern recognition and related problems. Kailath [20] investigated the relative merits of divergence and Bhattacharyya distance as surrogate criteria for error probability in signal selection for signal detection. In a similar vein, Boekee and van der Lubbe [6] studied divergence as a criterion for feature selection in pattern recognition and Toussaint [43] advocated its use instead of error probability for pattern classification. The use of divergence instead of classification error probability may have computational advantages. Most of all, the results in the literature are normally applicable to two class pattern recognition problems only, but some of the divergences, such as the Jensen–Shannon divergence [28] support extension to multiclass cases, including error bounds. In [35], KL divergence is used for local image content clustering to reduce the complexity of processing images of large resolution and in [1] for sensor validation. Bregman divergences have also been used for nonsupervised pattern classification and for data analysis based on clustering [2].

In communication systems, divergence is used to measure, for example, communication channel distortion rates and to optimize channel and source coding (see [40]). Similarly, divergences play a role in optimizing the quality of audio and video material compression for storage and archival purposes. In statistics, divergence measures have been used for the analysis of contingency tables [16] and for estimating the parameters of model distributions [19], gauging the consistency of observations with a hypothesized probability distribution model [14], and comparing true distributions with their approximations [7]. Bian *et al.* [5] used KL divergence for regularization of an objective function for action recognition learning. Zhang *et al.* [45] compared a range of divergence measures, including KL and Rényi divergences, in the context of sensor planning for target classification. Most recently, Lin *et al.* [29] employed KL divergence to search for efficient approximation of a hash code distribution in a nearest neighbor retrieval problem.

In this paper, our focus is on the application of divergences for detecting classifier incongruence. Closest to this particular interest is the use of KL divergence for gauging classifier incongruence by Weinshall *et al.* [44]. They adopted KL divergence following Itti and Baldi [18] who used it as an objective measure of surprise experienced by subjects reacting to a stimulus induced by the content of a test video. In their experiments, divergence was used to compare prior belief captured in terms of a prior distribution, with a new stimulus represented by a posterior distribution. They referred to the KL divergence in this context as "Bayesian surprise" measure. Some of the deficiencies of the KL divergence as a measure of classifier incongruence were addressed in [38] and by the heuristic measures proposed in [21] and [24]. In the next section, we provide a more principled basis for classifier incongruence detection and develop a novel measure, referred to as delta divergence, which satisfies the set of desirable properties identified for this specific application.

III. Delta Divergence

We start the discussion by introducing the necessary mathematical notation. We then revisit the classical KL divergence to establish a baseline and to point out some of its deficiencies from the point of view of measuring classifier incongruence. This will allow us to define the notion of classifier incongruence and postulate the properties a divergence measure should possess to support this particular application. We then consider the spectrum of available divergences to identify a suitable candidate and develop it to a novel classifier incongruence measure that is classifier decision cognizant and reflects the specified properties.

A. Baseline

Let us consider a pattern recognition problem where the object or phenomenon to be recognized is represented by a pattern vector $\mathbf{x}$ belonging to one of mutually exclusive classes $\omega_i, i = 1, \ldots, m$. Given observation $\mathbf{x}$, we shall denote the *a posteriori* probability of its membership in class ω_i as $P(\omega_i|\mathbf{x})$. The automatic assignment of pattern vector $\mathbf{x}$ to one of the classes is carried out by a classifier employing an appropriate decision function. Regardless of the type of machine learning solution, we shall assume that the classifier effectively computes the *a posteriori* class probabilities $P(\omega_i|\mathbf{x}), \forall i$ and engages a Bayesian decision rule to effect the class assignment.

Let us assume that for the same object or phenomenon there is another classifier which is basing its opinion about the object's class membership on its set of *a posteriori* class probabilities $\tilde{P}(\omega_i|\mathbf{y}), \forall i$, this time based on observation $\mathbf{y}$. The observation could be the same as $\mathbf{x}$ but, in general, $\mathbf{y}$ can be distinct. We are concerned with the problem of measuring the congruence of these two classifiers in supporting the respective hypotheses given the observations $\mathbf{x}$ and $\mathbf{y}$. In essence, we have two probability distributions, and the classifiers would be deemed congruent if the two probability distributions agree, and incongruent, if the two probability distributions are different. For the sake of simplicity and notational clarity, in the following we shall focus on a specific instance $\mathbf{x}, \mathbf{y}$ and drop referring to these observations explicitly, using a shorthand notation for the class probabilities as P_i and $\tilde{P}_i$, i.e.,

$$P_i = P\big(\omega_i|\mathbf{x}) \quad \tilde{P}_i = \tilde{P}(\omega_i|\mathbf{y}\big) \quad \forall i.$$

The basic concept in information theory is the notion of self-information. It conveys the amount of information we gain by observing an event ω which occurs with probability $P(\omega)$. If the probability of occurrence is high, i.e., close to one, we learn very little when the event occurs. However, when the probability $P(\omega)$ is low, the amount of information we gain is huge. Accordingly, self-information $I(\omega)$ is defined as

$$I(\omega) = -\log P(\omega)$$

which takes values from the interval $[0, \infty]$. $I(\omega)$ is referred to as "surprisal," as it quantifies the surprise of seeing a particular outcome.

In general, when an experiment has a number of possible outcomes $\omega_i, i = 1, \ldots, m$, the uncertainty associated with the experiment is expressed in terms of the average information gain from observing the outcome. Let P_i be the probability distribution over the events ω_i. The information gain $h(P)$ is defined as

$$h(P) = -\sum_{i=1}^{m} P_i \log P_i.$$

$h(P)$ is known as entropy. It is interesting to note that, as a result of the averaging process, the contribution to entropy made by events with small probability values is low, as

$$\lim_{x \to 0} x \log x = 0. \tag{1}$$

Rather than measuring the information gained from an experiment, here we are interested in assessing the degree of agreement between two probability distributions P and $\tilde{P}$ estimated over a set of hypotheses $\Omega = \{\omega_i, i = 1, \ldots, m\}$ by two different classifiers to gauge whether the classifiers agree in supporting a particular hypothesis or not. This can be achieved by comparing relative uncertainties associated with the two probability distributions P and $\tilde{P}$. A disagreement in their opinion about the identity of an object being classified would be considered surprising. We therefore need a measure of surprise which compares these two distributions. The classical measure suggested for this purpose is the KL divergence

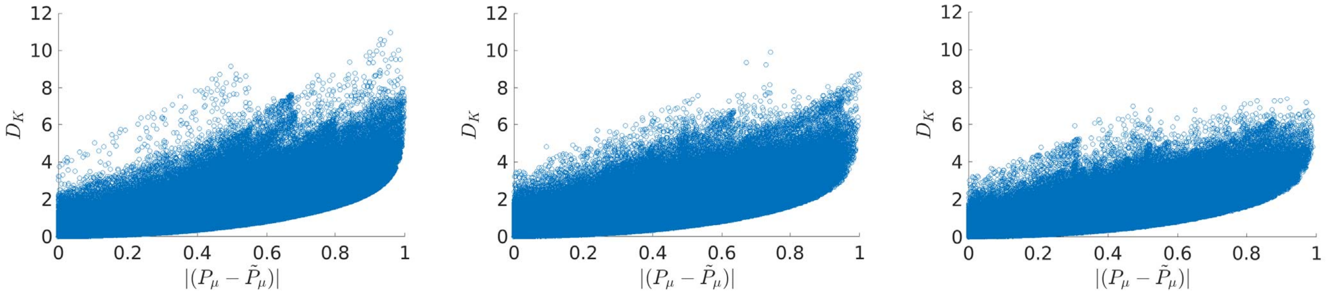
$$D_K = \sum_i \tilde{P}_i \log \frac{\tilde{P}_i}{P_i} \tag{2}$$

coined Bayesian surprise by Itti and Baldi [18], and used for measuring classifier incongruence by Weinshall [44].

B. Notion of Classifier Incongruence

We know that classifiers compute class *a posteriori* probabilities to make a decision, and that these probabilities must be involved in the definition of classifier incongruence. However, the notion of classifier incongruence is far from self evident. It is not crisply defined as, for instance, classifier error, or a particular shade of color. If the class probabilities output by two classifiers are similar, then we would agree that the classifiers are congruent. However, by how much can they differ before they cease to be congruent? If incongruence is like "distance," then the concept is clearly a continuum, rather than a discrete property, and the dichotomy between congruence and incongruence can only be defined by an appropriate threshold. However, what gauge should be used as an incongruence measure?

To answer these questions and to develop a suitable metric, we shall consider the classical KL divergence, as given



(a)

(b)

(c)

Fig. 1. Scatter plots of KL divergence against the difference of the posterior class probabilities computed for the dominant class selected by one of the classifiers. Given for (a) three classes, (b) six classes, and (c) ten classes.

in (2), in more details by elaborating a few special cases that should give us insight regarding the essence of congruence/incongruence.

1) *Identical Probability Distributions:* First of all, let us start with the simplest case when all the *a posteriori* class probabilities generated by the two classifiers are identical. In such a scenario the KL divergence (D_K) will be zero, flagging the status of congruence of the two decision making experts.

2) *Identical Dominant Hypotheses and Their Probabilities:* Next, we consider the case when the classifiers agree on the dominant hypothesis, i_{dom}. First of all, by *dominant*, we understand the most probable hypothesis, i.e., the class, indexed by i_{dom}, satisfying

$$P_{i_{\text{dom}}} = \max_i P_i.$$

In addition, we expect this class to dominate all the other hypotheses by a reasonable margin between $P_{i_{\text{dom}}}$ and $P_j = \max_{i, i \neq i_{\text{dom}}} P_i$. Let us assume that the classifiers support the common dominant hypothesis with identical strength, i.e., $P_{i_{\text{dom}}} = \tilde{P}_{i_{\text{dom}}}$. Clearly, the contribution to the KL divergence due to the dominant class would be zero. We would probably all agree that in such situation the classifiers would be congruent. Yet the support for the nondominant hypotheses, which we shall refer to as clutter, given by the two classifiers

$$D_K = \sum_{i, i \neq i_{\text{dom}}} \tilde{P}_i \log \frac{\tilde{P}_i}{P_i}$$

could be substantially different from zero, giving potentially a high value to the KL divergence. It is apparent, that for a given threshold, the KL divergence may give rise to false rejections of congruent classifier outputs.

3) *Different Dominant Hypotheses:* As the next scenario, we shall investigate the case when the two classifier disagree on the dominant hypothesis, but support the nondominant hypotheses in an identical way. Denoting the respective dominant hypotheses by i_{dom} and $\tilde{i}_{\text{dom}}$, the KL divergence in this case will be

$$D_K = P_{i_{\text{dom}}} \log \frac{P_{i_{\text{dom}}}}{\tilde{P}_{i_{\text{dom}}}} + P_{\tilde{i}_{\text{dom}}} \log \frac{P_{\tilde{i}_{\text{dom}}}}{\tilde{P}_{\tilde{i}_{\text{dom}}}}. \tag{3}$$

Note that in (3) the value of KL divergence in this "zero clutter" case will depend on the actual dominant class probabilities, reflecting the surprisal value in the relative information gained.

We shall now observe some of these properties on artificially generated data, where the *a posteriori* class distributions P and $\tilde{P}$ are sampled, without loss of generality, as follows.

Step 1: Draw P_1 from a uniform distribution defined on the interval [0, 1] quantized to N values. If $P_1 = 1$, then set $P_j = 0, \forall j > 1$ and break.

Step 2: For $\forall i = 2, \ldots, m$, draw P_i from the uniform distribution defined on $[0,\ 1 - \sum_{j=1}^{i-1} P_j]$. If $\sum_{j=1}^{i} P_j = 1$ then set $P_j = 0, \forall j > i$ and break.

Step 3: $\mu = \arg\max_i P_i$.

Step 4: If $P_\mu \leq \epsilon$ where ϵ is the minimum probability that the dominant class should assume in order to make a decision, then discard the sample distribution.

Step 5: Repeat steps 1–4 for $\tilde{P}_i$.

We set $N = 10\,000$, and create 1000 different P and $\tilde{P}$ distributions for three, six and ten class cases. Using all P and $\tilde{P}$ combinations, we end up with a total of one million pairs for each case. It is important to mention that values of $P_i \leq 0.0001$ in the denominator of KL divergence are replaced by $P_i = 0.0001$ to avoid overflow and the results plotting problems.

Having computed D_K values for each P and $\tilde{P}$ pair, we plot them, in Fig. 1, as a function of the difference between the *a posteriori* probabilities corresponding to the dominant hypothesis output by the reference classifier (μ). We can see that for every choice of the difference, the range of KL divergence values is large, even for $\Delta = P_\mu - \tilde{P}_\mu = 0$. The scatter plots make it clear that KL divergence cannot naturally distinguish the state of classifier incongruence from classifier congruence. This is primarily due to the contribution to KL divergence made by the nondominant hypotheses.

There are a number of conclusions that can be drawn from this analysis. First of all we can see that while "perfect" congruence is independent of the actual values of *a posteriori* class probabilities of the two distributions, as they are identical, in the case of general congruence and incongruence scenarios, the magnitude of the D_K measure will exhibit strong dependence on the probability distribution values. The clutter induced by nondominant classes will create ambiguity, that will degrade the separability of notionally congruent and incongruent classifier cases. It should also be noted that the value of KL divergence will depend on the class probability distributions used as a reference. If we choose $\tilde{P}_i$ instead of P_i, the observed incongruence value will be different. This is not a useful property for applications where the notion is conceptually symmetric. Also the values of the incongruence measure should be confined to a bounded interval to facilitate the setting of a suitable threshold to dichotomize congruence and incongruent cases.

From these observation the following desirable properties of the ideal measure of classifier incongruence are beginning to emerge.

1) Overriding focus on dominant hypotheses.
2) Independence of surprisal content.
3) Minimum clutter effect.
4) Symmetry (independence of the choice of distribution as a reference).
5) Bounded range of incongruence measure values.

Properties 1) and 3) are linked, and suggest that the required measure should concentrate on the dominant hypotheses, and suppress the effect of nondominant classes. Thus the measure we seek should be decision cognizant as the DC-KL divergence of Ponti *et al.* [38]. Property 2) suggests that classifier incongruence should be a function of differences in *a posteriori* class probabilities, rather than some function of their respective values. The choice of a divergence measure should exhibit symmetry property 4) and yield values which are bounded, as specified by property 5). In the following section, we shall identify a suitable starting point and develop a novel divergence measure which satisfies the above postulated properties.

C. Delta Divergence Measure

Our aim is to develop a divergence that will have all the above stated properties when used as a classifier incongruence measure, namely boundedness, symmetry, being clutter free, and ideally also of low sensitivity to probability estimation errors. Heuristic attempts at finding incongruence gauging measures satisfying these properties were presented in [21] and [24]. The key idea in these two papers is to focus on dominant classes as identified by the two classifiers and ignore all the other hypotheses. More specifically, let $\omega = \arg\max_i P_i$ and $\tilde{\omega} = \arg\max_i \tilde{P}_i$. These decision dependent measures are defined in [21] and [24], respectively, as

$$\Delta^* = \frac{1}{2}\left[\left|P_\omega - \tilde{P}_\omega\right| + \left|\tilde{P}_{\tilde{\omega}} - P_{\tilde{\omega}}\right|\right]$$

and

$$\begin{aligned}\Delta_{\max} = \tfrac{1}{2}\max\Big\{&\left|P_\omega - \tilde{P}_\omega\right| + \delta\{\omega, \tilde{\omega}\}\left|\tilde{P}_{\tilde{\omega}} - \tilde{P}_\omega\right| \\ &\left|\tilde{P}_{\tilde{\omega}} - P_{\tilde{\omega}}\right| + \delta\{\omega, \tilde{\omega}\}\left|P_\omega - P_{\tilde{\omega}}\right|\Big\}\end{aligned}$$

where $\delta\{\omega, \tilde{\omega}\}$ is defined as

$$\delta\{\omega, \tilde{\omega}\} = \begin{cases} 0 \text{ if } \omega = \tilde{\omega} \\ 1 \text{ if } \omega \neq \tilde{\omega}. \end{cases}$$

In contrast to these heuristic techniques, our objective is to develop a classifier incongruence measure with a solid theoretical underpinning by demanding that it is a proper divergence. The appropriate toolbox for measuring incongruence between two discrete probability distributions is the family (h, ϕ) of functions

$$h\left[\sum_i \phi\left(P_i, \tilde{P}_i\right)\right] \tag{4}$$

with h and ϕ being polynomial, logarithmic, polylogarithmic, quasi-polynomial, quasi-polylogarithmic functions [15], or convex functions [9]. This family includes Bregman divergences [42]

$$D_B = \sum_i \left[f(P_i) - f\left(\tilde{P}_i\right) - \left(P_i - \tilde{P}_i\right)f'\left(\tilde{P}_i\right)\right]$$

the Cziszar f-divergences [9] reviewed in [27] and [33]

$$D_C = \sum_i P_i f\left(\frac{\tilde{P}_i}{P_i}\right) \tag{5}$$

and the Rényi [39] divergences parameterized by α

$$D_R = \frac{1}{\alpha - 1}\log\left[\sum_i P_i\left(\frac{\tilde{P}_i}{P_i}\right)^\alpha\right].$$

For an overview the reader is referred to [15].

Armed with the toolbox, the key question of interest to us is which member of the family would exhibit the properties that reflect the notion of classifier incongruence discussed in Section III-B. We already established in Section III-B that the KL divergence does not. The Jensen–Shannon divergence [28]

$$D_J = \frac{1}{2}\sum_i\left[P_i \log\frac{2P_i}{P_i + \tilde{P}_i} + \tilde{P}_i \log\frac{\tilde{2}P_i}{P_i + \tilde{P}_i}\right]$$

confines its values to a bounded interval, and is symmetric. However, the contributions to divergence generated by a difference in probabilities for a particular hypothesis are a function of the probabilities themselves, which does not satisfy property 2) in Section III-B. Most importantly, all the measures, including Jensen–Shannon divergence, are affected by the divergence clutter injected by weakly supported hypotheses. This clutter is also likely to aggravate the sensitivity of these divergence measures to noise.

Herein we set to develop an incongruence measure which is a member of the family of divergences in (4). This is the most general family of divergences which has the potential to source the starting point of our development. For the sake of simplicity, we start by choosing

$$h(z) = z$$

in (4) and opting for the family of f-divergences in (5). In this family, the bounded measures [required by property 5)] are the ones whose functional form in the denominator terms of the convex function $f(\tilde{P}_i/P_i)$ approaches zero as a function of P_i at a linear rate, at most. These include, for instance, the Cziszár and Fisher [12] and Matusita [31] divergences. Note that the Jensen–Shannon divergence in [28] does not ensure boundedness through the properties of the convex function of the two probability distributions, but instead by measuring divergence between one of the probability distribution functions and the average of the two. However, none of these bounded divergences meets the requirement that the contribution to divergence is dependent purely on differences in probabilities, rather than their actual values [property 2)]. By virtue of the ℓ_1 norm, this characteristic is exhibited only by the *total variation distance*, defined as

$$D_T = \frac{1}{2}\sum_i P_i \left| \frac{\tilde{P}_i}{P_i} - 1 \right| = \frac{1}{2}\sum_i \left| \tilde{P}_i - P_i \right|. \tag{6}$$

This measure is symmetric [property 4)] and bounded, taking values from the interval [0, 1].

The measure in (6) is still affected by clutter of nondominant classes. The effect of clutter can significantly be reduced by the following argument: when we compare the outputs of two classifiers, there are only three outcomes of interest: the dominant class ω identified by the classifier with probability distribution P, the dominant class $\tilde{\omega}$ identified by the other classifier, and neither of the two, in other words $\hat{\omega} = \Omega - \omega - \tilde{\omega}$. We thus define a new decision cognizant divergence D_Δ, which we name delta divergence, as

$$D_\Delta = \frac{1}{2}\left[\sum_{i \epsilon \{\omega, \tilde{\omega}\}} \left| \tilde{P}_i - P_i \right| + \left| \tilde{P}_{\hat{\omega}} - P_{\hat{\omega}} \right| \right] \tag{7}$$

which parallels the DC-KL divergence [38]

$$D_D = \left[\sum_{i \epsilon \{\omega, \tilde{\omega}\}} \tilde{P}_i \log \frac{\tilde{P}_i}{P_i} \right] + \tilde{P}_{\hat{\omega}} \log \frac{\tilde{P}_{\hat{\omega}}}{P_{\hat{\omega}}}.$$

Noting that the outcome $\hat{\omega}$ arises with the complement probabilities, we can further analyze delta divergence, D_Δ, in (7) further by considering the cases when the labels of the dominant classes identified by the two classifiers agree and when they disagree.

1) Label Agreement: When the labels agree, i.e., $\omega = \tilde{\omega}$, the complement probabilities for the event that the true class is not ω are $1 - \tilde{P}_\omega$ and $1 - P_\omega$. Then the delta divergence in (7) can be expressed

$$\begin{aligned} D_\Delta &= \tfrac{1}{2}\left[\left| \tilde{P}_\omega - P_\omega \right| + \left| 1 - \tilde{P}_\omega - 1 + P_\omega \right| \right] \\ &= \left| \tilde{P}_\omega - P_\omega \right|. \end{aligned}$$

In other words, the classifier incongruence can be measured simply by comparing the probabilities of the dominant hypothesis output by the two classifiers.

2) Label Disagreement: When the dominant labels identified by the two classifiers disagree, the probabilities of the event $\hat{\omega}$ that neither of the two dominant classes is the true class are given as

$$\begin{aligned} P_{\hat{\omega}} &= 1 - P_\omega - P_{\tilde{\omega}} \\ \tilde{P}_{\hat{\omega}} &= 1 - \tilde{P}_\omega - \tilde{P}_{\tilde{\omega}}. \end{aligned}$$

In this scenario, the delta divergence becomes

$$\begin{aligned} D_\Delta = &\frac{1}{2}\left[\left| \tilde{P}_{\tilde{\omega}} - P_{\tilde{\omega}} \right| + \left| P_\omega - \tilde{P}_\omega \right| \right. \\ &\left. + \left| \tilde{P}_{\tilde{\omega}} - P_{\tilde{\omega}} + \tilde{P}_\omega - P_\omega \right| \right] \\ = &\frac{1}{2}[|A| + |B| + |A - B|]. \end{aligned} \tag{8}$$

Note that the terms A and B can either be both positive, or one of them positive and the other negative. It can be easily shown that it is impossible for both terms to be negative. Consider, for instance, the case $A < 0$, i.e., $\tilde{P}_{\tilde{\omega}} - P_{\tilde{\omega}} < 0$. Then, since $P_{\tilde{\omega}} < P_\omega$ (ω being the dominant class for classifier with distribution P), and $\tilde{P}_{\tilde{\omega}} > \tilde{P}_\omega$ ($\tilde{\omega}$ being the dominant class for classifier $\tilde{P}$) we have

$$0 < P_{\tilde{\omega}} - \tilde{P}_{\tilde{\omega}} < P_\omega - \tilde{P}_\omega.$$

The positivity of A when B is negative can be shown in the same way.

Now suppose A is negative. Then $A - B$ in (8) is also negative, and its absolute value is equal $|A + B| = |A| + |B|$. If, on the other hand, B is negative, then $-B$ is positive, and the absolute value of $A - B$ will again equal $|A| + |B|$. Thus when one of the terms, A and B is negative, delta divergence will be

$$D_\Delta = [|A| + |B|] = \left[\left| \tilde{P}_{\tilde{\omega}} - P_{\tilde{\omega}} \right| + \left| P_\omega - \tilde{P}_\omega \right| \right].$$

When both A and B are positive, the term $A - B$ is either positive, or negative, depending on the relationship of A and B. If $A > B$, then the difference will be positive and we can ignore the absolute value operation, i.e., $|A - B| = A - B$. If $A < B$, then the difference will be negative and $|A - B| = B - A$. Thus we can write for D_Δ in (8)

$$D_\Delta = \begin{cases} A & \text{if} \quad A \geq B \\ B & \text{if} \quad A < B. \end{cases}$$

3) Delta Divergence Overview: Combining the results for these scenarios yields a surprisingly simple divergence measure for gauging classifier incongruence, i.e.,

$$\begin{aligned} &D_\Delta \\ &= \begin{cases} \left| \tilde{P}_\omega - P_\omega \right| & \omega = \tilde{\omega} \\ \max\left\{ \left| \tilde{P}_{\tilde{\omega}} - P_{\tilde{\omega}} \right|, \left| P_\omega - \tilde{P}_\omega \right| \right\} & \omega \neq \tilde{\omega} \; A \geq 0, B \geq 0 \\ \left[\left| \tilde{P}_{\tilde{\omega}} - P_{\tilde{\omega}} \right| + \left| P_\omega - \tilde{P}_\omega \right| \right] & \omega \neq \tilde{\omega} \; \begin{cases} A < 0, B \geq 0 \\ A \geq 0, B < 0. \end{cases} \end{cases} \end{aligned} \tag{9}$$

In other words, the incongruence measure is defined either by the maximum absolute value difference between the probabilities output by the two classifiers for the respective dominant hypotheses or by the sum of these differences.

The measure has attractive properties. It is zero, whenever the *a posteriori* probabilities for the shared dominant class are identical, regardless of the differences in the distribution of the residual probability mass over all the other classes. As it always involves the difference of two probability values, it is symmetric. Also, its sensitivity to estimation errors should be very low. It has a monotonic transition between the function values for the label agreement and label disagreement cases. In fact, as we move from the label agreement to the label disagreement case, when another class for the second classifier begins to assume the dominant role, delta divergence will continue increasing (potentially by a step change) by virtue of the growing difference between the dominant class probability of the first classifier and the support for this hypothesis voiced by the second classifier.

4) Two Class Case: In the specific two class case, when the classifiers agree on the dominant hypothesis ω, delta divergence is given as

$$D_{\Delta} = \frac{1}{2}\left[\left|P_{\omega} - \tilde{P}_{\omega}\right| + \left|1 - P_{\omega} - 1 + \tilde{P}_{\omega}\right|\right] = \left|P_{\omega} - \tilde{P}_{\omega}\right|. \tag{10}$$

In the label disagreement case, the set of nondominant hypotheses is empty. Hence the delta divergence has just two terms that are identical to those in (10). Thus the general formula for D_{Δ} in the case of agreement and disagreement is as given in (10).

D. Properties of Delta Divergence

In this section, we briefly review the properties of delta divergence and verify that it satisfies the characteristics specified in Section III-B. In addition, we shall determine the conditions under which the proposed divergence measure is a metric. This particular property is interesting in the context of assessing incongruence of more than two classifiers.

1) *Decision Cognizance Property:* The delta divergence proposed in (9) is defined in terms of the *a posteriori* class probabilities associated with the dominant hypotheses identified by the two classifiers. The measure therefore focuses only on the dominant class hypotheses as required by property 1) in Section III-B.
2) *Surprisal Independence:* The proposed divergence is defined in terms of differences in *a posteriori* class probabilities of the dominant hypotheses, rather than their respective values. Thus the value of delta divergence is independent of the base level of these probabilities, and consequently of the surprisal values.
3) *Robustness to Clutter:* The advantage of delta divergence over total variation distance can be demonstrated by comparing the contributions of the nondominant hypotheses to these two measures. In the case of delta divergence, the implicit contribution to "clutter" is given by $(1/2)|P_{\hat{\omega}} - \tilde{P}_{\hat{\omega}}|$ where $\hat{\omega}$ represents the set of nondominant classes. In the case of total variation distance, the clutter contribution becomes

$$\frac{1}{2}\sum_{i \epsilon \hat{\omega}} \left|P_i - \tilde{P}_i\right|.$$

Rearranging the clutter contribution to delta divergence we have

$$\frac{1}{2}\left|P_{\hat{\omega}} - \tilde{P}_{\hat{\omega}}\right| = \frac{1}{2}\left|\sum_{i \epsilon \hat{\omega}}\left[P_i - \tilde{P}_i\right]\right| \leq \frac{1}{2}\sum_{i \epsilon \hat{\omega}}\left|P_i - \tilde{P}_i\right|. \tag{11}$$

Thus the sensitivity of delta divergence to clutter is significantly lower than that of total variation distance. It is interesting to note that if the first two terms in (8) are considered as "pure incongruence measure" (PIM) and the last term as a group clutter, $D_{\Delta_{\text{clutter}}}$, then from (9) we conclude

$$D_{\Delta_{\text{clutter}}} = \begin{cases} \frac{1}{2}\left|\tilde{P}_{\tilde{\omega}} + \tilde{P}_{\omega} - P_{\omega} - P_{\tilde{\omega}}\right| & \begin{cases} \tilde{P}_{\tilde{\omega}} - P_{\tilde{\omega}} \geq 0 \\ P_{\omega} - \tilde{P}_{\omega} \geq 0 \end{cases} \\ \frac{1}{2} \times \text{PIM} \quad \text{elsewhere.} \end{cases}$$

This shows that the contributed group clutter is equal to the magnitude of PIM in most cases. When the labels of the dominant hypotheses selected by the classifiers disagree, and the difference between the probability for the top ranking hypothesis rendered by the supporting classifier relative to the other classifier is nonnegative, the group clutter equals one half of the difference of the two differences. Alternatively, the clutter is equal to the difference between the support for the union of the two hypotheses. Thus, in this particular case the clutter is proportional to the difference between the residual probability masses associated with the nondominant classes. The superiority of D_{Δ} over D_T from the clutter point of view is also evident from the experimental results shown in Fig. 2. After generating the *a posteriori* class probability distributions of the two classifiers as described in Section III-B, we compute and record the clutter injected in D_{Δ} and D_T as defined on the left- and right-hand side of (11), respectively. The figure presents the scatter plots of the clutter associated with D_T against the clutter of D_{Δ}, for three, six, and ten class problems. The plots show clearly that the D_T clutter is almost always greater than that of D_{Δ}. It also be should be noted that the effect of clutter on D_T is less severe for smaller number of classes, because the scope for cluttering is considerably more limited. In the two class case, the clutter disappears altogether. By the same token, in pattern recognition problems involving a large number of classes, the induced clutter can dominate the value of total variation divergence and make it impossible to detect classifier incongruence reliably.
4) *Symmetry:* As (9) involves only differences of *a posteriori* class probabilities, D_{Δ} is symmetric in compliance with property 4) of Section III-B.
5) *Bounded Range:* Inspecting (9), it is evident that its values satisfy $0 \leq D_{\Delta} \leq 1$. Hence delta divergence is bounded to interval [0, 1] in compliance with property 5 of Section III-B.
6) *Metric Property:* The total variation distance, D_T, from which the proposed divergence has been developed is a metric. This can easily be checked by considering three classifiers A, B, C with probability distributions P, $\tilde{P}$, and $\hat{P}$, respectively. The sum of variation distances D_{AB} and D_{BC} can be written as

$$\begin{aligned} D_{AB} + D_{BC} &= \textstyle\sum_i\left[\left|P_i - \tilde{P}_i\right| + \left|\tilde{P}_i - \hat{P}_i\right|\right] \\ &\geq \textstyle\sum_i\left[\left|P_i - \tilde{P}_i + \tilde{P}_i - \hat{P}_i\right|\right] \\ &= \textstyle\sum_i\left[\left|P_i - \hat{P}_i\right|\right] = D_{AC}. \end{aligned}$$

The metric property does not extend to D_{Δ} because of the clutter reducing operation of merging all nondominant hypotheses into a single event, as the resulting sets for the three classifiers can be different. However, in the

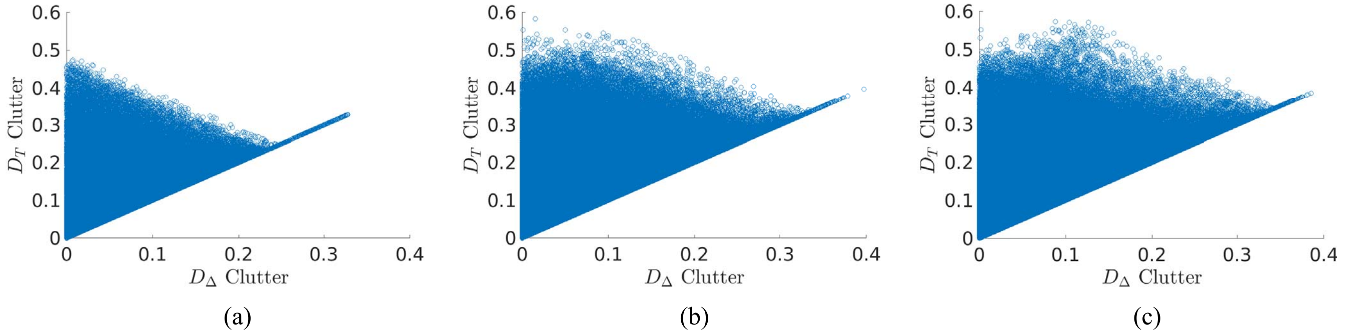


Fig. 2. Scatter plots of the clutter contribution to D_T versus the clutter contribution to D_Δ. The values are computed for samples from a population of probability distributions defined over (a) three classes, (b) six classes, and (c) ten classes.

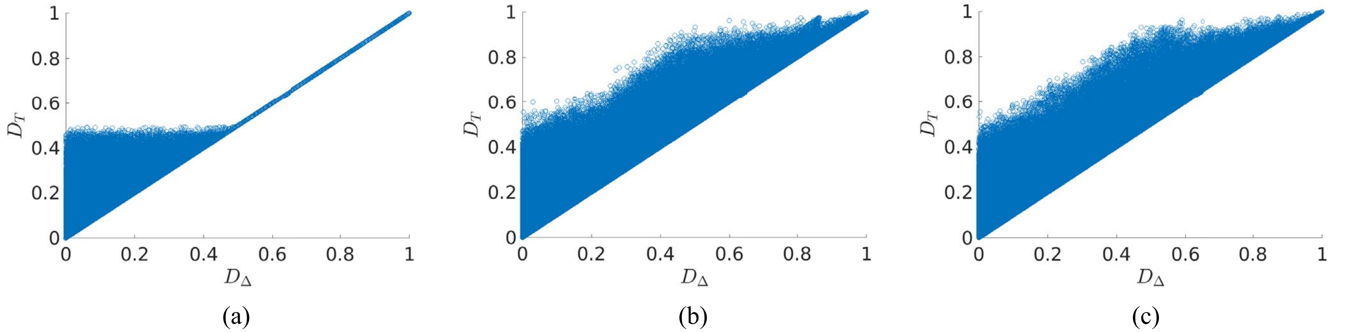


Fig. 3. Scatter plots of values of total variation distance (D_T) against delta divergence (D_Δ) obtained in the simulation experiment involving (a) three classes, (b) six classes, and (c) ten classes.

two class case when the set of nondominant hypotheses is empty, the delta divergence (7) will degenerate to the total variation distance (6), and the incongruence measure will become a metric.

7) *Sensitivity to Estimation Errors:* An important factor in selecting a tool is its robustness to noise, which in the current context means robustness to probability estimation errors. An excessive sensitivity may render a tool ineffective, even if its theoretical foundations are sound and strong. An example of this is the brittleness of the product fusion rule as compared with the sum fusion rule in multiple classifier fusion [22]. An extensive experimental study of delta divergence in [25] showed that it retains its favorable properties even in the presence of estimation errors.

IV. Relationship of D_Δ to Other Measures

The aim of the simulation studies reported in this section is to show the relationship between delta divergence and two baseline divergences, namely the total variation distance D_T and the KL divergence D_K, over the full spectrum of scenarios captured by sampling the classifier probability distributions as described in Section III-B. Due to space limitations, we only show the results for three, six, and ten class problems. However, even this sparse sample is sufficient to demonstrate the trend in the relationships as the number of classes increases.

A. Total Variation Distance

As we have developed delta divergence from the total variation distance it is pertinent to elaborate the key differences between these two divergences. The main distinguishing feature of delta divergence is the way it deals with clutter. Let us denote by Ω^+ the set of dominant hypotheses identified by the two classifiers, which will have a single element for label agreement and two elements for label disagreement. The complement set Ω^- is constituted by all the nondominant hypotheses, i.e., $\Omega^- = \Omega - \Omega^+$ and the probability of one of its members being the true class is $P_{\Omega^-} = \sum_{i\epsilon\Omega^-} P_i$ and $\tilde{P}_{\Omega^-} = \sum_{i\epsilon\Omega^-} \tilde{P}_i$, respectively, for the two classifiers. Referring to (6) and (7), we can bound D_Δ as

$$\begin{aligned} D_\Delta &= \tfrac{1}{2}\Big[\textstyle\sum_{i\epsilon\Omega^+}\big|P_i - \tilde{P}_i\big| + \big|P_{\Omega^-} - \tilde{P}_{\Omega^-}\big|\Big] \\ &\le \tfrac{1}{2}\Big[\textstyle\sum_{i\epsilon\Omega^+}\big|P_i - \tilde{P}_i\big| + \sum_{i\epsilon\Omega^-}\big|P_i - \tilde{P}_i\big|\Big] = D_T. \end{aligned}$$

Thus $D_\Delta \le D_T$, with equality only for the two class case $m = 2$. Even for $m = 3$ the total variation distance will be greater than delta divergence because the set of nondominant hypotheses will contain more than one element in the case of label agreement.

The relationship between these two divergences is demonstrated experimentally in Fig. 3 which plots values of D_T against D_Δ. It should be noted that for every value of D_Δ there are many possible values of D_T and vice versa, as already shown in Section III-D. These points are identified by sampling the probability distributions P and $\tilde{P}$ with the procedure described in Section III-B, and plotting the corresponding divergence values against each other. It is apparent from the

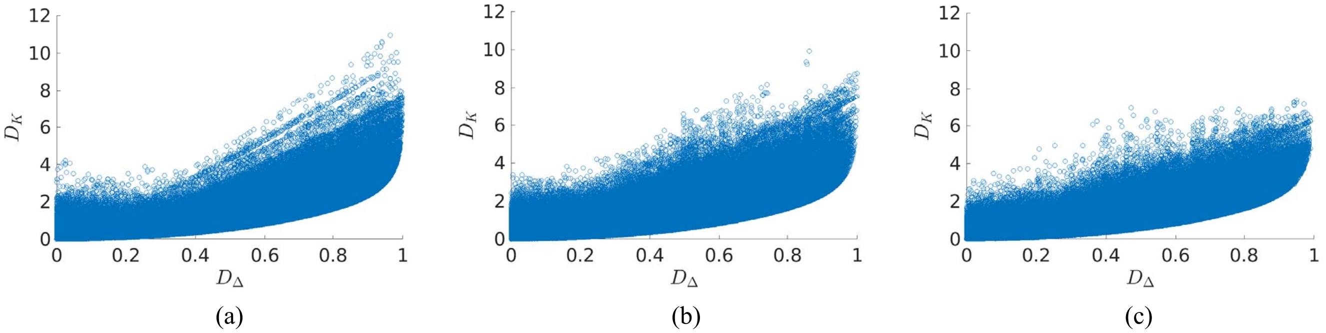


Fig. 4. Scatter plots of values of KL divergence D_K against delta divergence (D_Δ) obtained in the simulation experiment involving (a) three classes, (b) six classes, and (c) ten classes.

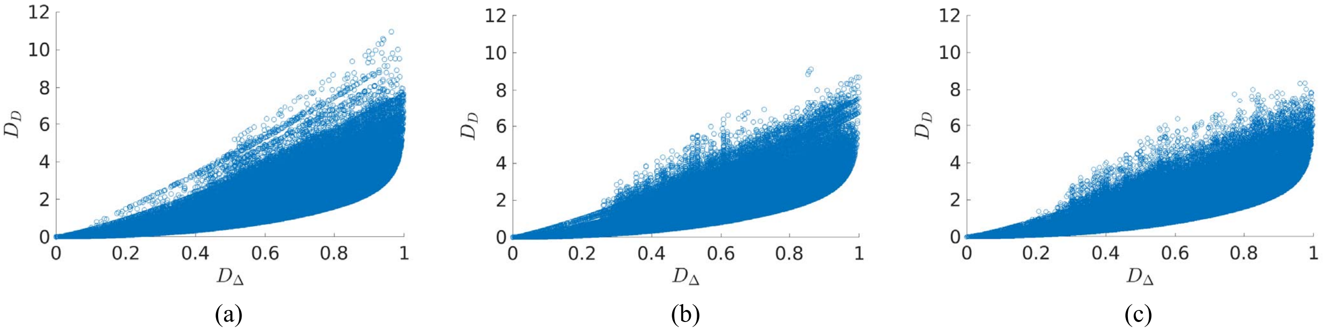


Fig. 5. Scatter plots of values of DC-KL divergence D_D against delta divergence (D_Δ) obtained in the simulation experiment involving (a) three classes, (b) six classes, and (c) ten classes.

plots that for higher number of classes, the distribution scenarios are much less heavily constrained, and this results in much greater differences in the values of D_T and D_Δ.

Let us consider thresholds $D_{D_T} = 0.5$ and $D_{D_\Delta} = 0.5$ that could potentially be used to separate the states of congruence and incongruence. In the case of $m = 3$ in Fig. 3, both thresholds would achieve a good separation. However, for $m = 6$ and $m = 10$, the threshold D_{D_T} fails to dichotomizing congruence and incongruence adequately. Among the samples falling in the incongruence category there are many with low value of D_Δ. By virtue of the close link between delta divergence and the difference of posteriors of the dominant classes, these cases should clearly be deemed congruent. Thus, the clutter makes it difficult for D_T to discriminate between congruent and incongruent cases.

B. Kullback–Leibler and Decision Cognizant Kullback–Leibler Divergences

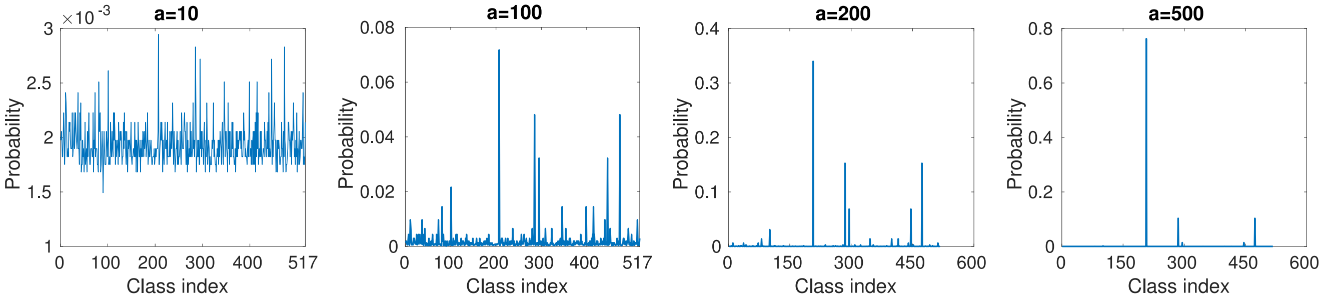
Next we compare delta divergence with KL divergence and its decision cognizant variant, DC-KL. The same experiment, involving the sampling of the space of probability distributions P and $\tilde{P}$ is conducted for three, six and ten class problems. In Fig. 4, D_K is plotted against D_Δ. We note that the range of values exhibited by the KL divergence is much greater, which makes it more difficult to set a suitable threshold between classifier congruence and incongruence. The unbounded range reflects the dependence of KL divergence on the surprisal values of the additive terms in the expression for the KL divergence.

The clutter and the surprisal value dependence are jointly responsible for a significant overlap of KL divergence values for the classifier congruence and classifier incongruence cases. This can be seen by drawing horizontal lines cutting the scatter plots at different KL divergence thresholds and noting the resulting distributions (data scatters). For instance setting the threshold to $D_K = 3$ will retain many cases with a high value of D_Δ in the congruent category, leading to underdetection of incongruence. Lowering the threshold to, say, 0.75 will miss many cases with low value of delta divergence, resulting in a high proportion of false positives.

In Fig. 5, we plot DC-KL divergence values against delta divergence. Comparing Figs. 4 and 5, the effect of the suppression of clutter in DC-KL is evident from the scatter plots. However, the unboundedness and the dependence of DC-KL on surprisal values still compromise the separability of the states of congruence and incongruence.

V. Relevance of Classifier Incongruence in General and of D_Δ in Particular

The relevance of classifier incongruence measures was demonstrated in [21] in the context of tennis video interpretation. The output of a detector of visual events (player actions, tennis ball hit, and tennis ball bounce) was monitored and compared with the output of high level tennis game interpretator to detect incongruences between noncontextual and contextual decision making processes. Incongruence was indicative of different types of anomalies, such as the deployment of an incorrect scene evolution model (game of singles



(a) (b) (c) (d)

Fig. 6. *A posteriori* probability distributions belonging to the fingerprint modality of a test subject for all classes, computed for different values of a.

TABLE I

False Negative Rates for Given True Negative Rates and "a" Values Using (a) Delta Divergence, (b) KL Divergence, and (c) DC-KL Divergence

(a)

	TNR=90%	TNR=95%	TNR=99%
a=0.1	0.9037	0.9597	0.9919
a=1	0.9435	0.9677	0.9919
a=10	0.9677	0.9758	0.9919
a=80	0.0726	0.1371	0.2984
a=120	0.0081	0.0323	0.0806
a=160	0	0.0081	0.0323
a=200	0	0	0.0242
a=250	0	0	0
a=500	0	0	0

(b)

	TNR=90%	TNR=95%	TNR=99%
a=0.1	0.8548	0.9194	0.9839
a=1	0.8790	0.9435	0.9919
a=10	0.9597	0.9677	0.9919
a=80	0.1048	0.5161	0.8548
a=120	0	0.2581	0.6935
a=160	0	0.0323	0.5645
a=200	0	0	0.5
a=250	0	0	0.4597
a=500	0	0	0.2661

(c)

	TNR=90%	TNR=95%	TNR=99%
a=0.1	0.8952	0.9597	0.9919
a=1	0.9113	0.9597	0.9919
a=10	0.9597	0.9677	0.9919
a=80	0.1129	0.4435	0.7984
a=120	0.0161	0.1774	0.5806
a=160	0	0.0081	0.4677
a=200	0	0	0.4274
a=250	0	0	0.3145
a=500	0	0	0.1694

instead of doubles). Coppi *et al.* [8] used incongruence between generic and specific object classifiers arranged in a hierarchical structure to flag novel (unknown) subclasses of object categories such as motorbikes, flowers, etc.

The effectiveness of delta divergence was shown in [25] where KL and DC-KL divergences failed to detect incongruence between noncontextual and contextual classifiers detecting actions and activities in breakfast preparation videos. The incongruences successfully detected by delta divergence flagged anomalies such as missing steps due to occlusion and the simultaneous presence of multiple actions (in the background and foreground). Similarly the deficiency of KL and DC-KL divergences was observed in [23] in the context of analyzing videos recording breakfast preparation activities. This paper involved detecting incongruences between multiple modalities (audio and visual). The use of delta divergence produced much lower rate of false positives than the KL-based alternatives.

Here we provide further experimental evidence of the advantages of delta divergence using real data in the application domain of multimodal biometric person recognition. We analyze the scores of two independent biometric modalities for incongruence to inform operational decision making. We use the NIST-BSSR1 dataset of raw matching scores for the face and fingerprint modality. The data involves 517 subjects, whose biometric traits are matched against gallery templates [3]. As the scores of the two modalities have vastly different ranges, they are first normalized to the [0, 1] interval and then converted to *a posteriori* class probabilities. Let the normalized matching score for subject i for one of the modalities be $\mathbf{x}_i$. Then the corresponding *a posteriori* class probability is given by

$$Q_i = \frac{\exp\{a\mathbf{x}_i\}}{\sum_j \exp\{a\mathbf{x}_j\}} \tag{12}$$

where a is a parameter of the score to probability conversion.

The probabilities $P_i, \tilde{P}_i, \forall i$ computed for all the subjects based on the face and fingerprint modalities, respectively, provided an input to the KL, DC-KL and delta divergence measures. Note that the ground truth labels (congruent, incongruent) for the classifier outputs are available for the dataset. The aim of the experiment is to measure the overlap of the true congruent and false incongruent distributions. This is accomplished by setting the confidence level for detecting true congruences at 90%, 95%, and 99%, respectively, and measuring the corresponding false incongruence rates.

The experiment was repeated for different values of the parameter a. Note that when $a = 0$ the *a posteriori* class probability distribution is uniform. At the other extreme, when $a = \infty$, $Q_j = 1$ for $j = \arg\max_i \mathbf{x}_i$ and zero for all the others. Thus a controls the relative magnification of the scores. Most importantly, different values of a represent scenarios with different levels of clutter, that is the probability mass distributed over the nondominant classes. These scenarios are illustrated in Fig. 6 which shows the *a posteriori* class probability distributions for the fingerprint biometric trait of a single subject for different values of parameter a. The false negative rates corresponding to the three confidence levels for the different scenarios are given in Table I. We can see in Table I that for practically uniform probability distribution corresponding to $a \epsilon [0.1, 10]$ the delta divergence values for congruences and

TABLE II
Overview of the Properties of Selected Divergence Measures

	KL Div.	Jensen-Shannon	DC-KL Div.	Delta Div.
Decision Cognizance			✓	✓
Surprisal Independence				✓
Robostness to Clutter			✓	✓
Symmetry		✓		✓
Boundedness		✓		✓

incongruences overlap almost 100% as the concept of dominance effectively breaks down. In the range $a\epsilon[80, 160]$ where the concept of dominance begins to apply, but the clutter is still high, the overlap drops significantly and gradually diminishes. When there is no clutter ($a \geq 200$), the proposed incongruence measure separates the categories perfectly, as expected. In comparison, the overlap of the distributions of KL divergence values obtained for the congruent and incongruent classifier outputs for all values of $a \geq 80$ is much greater, especially for the high levels of the confidence threshold. The overlap is lower for the DC-KL divergence, but still considerably worse than that achieved by delta divergence. This demonstrates the merit of the proposed classifier incongruence measure.

It is pertinent to ask, whether the proposed divergence would also find applicability in the context of training deep neural networks for measuring incongruence between the target and achieved probability distributions, and displace the KL divergence (cross entropy). Interestingly, this is unlikely, as an important consideration in adopting a loss function are the characteristics of the loss function gradients. In this use case, the KL divergence is preferable as it has the capacity to drive the nondominant class probabilities to zero much more forcefully than the delta divergence. The proposed measure is appropriate for monitoring and comparing classifier outputs with the aim of using the incongruence measure values in subsequent reasoning, rather than for machine learning.

VI. Conclusion

The problem of detecting classifier incongruence was addressed in this paper. It involves comparing the output of two classifiers to gauge the level of agreement in their support for a particular decision. As, in general, the output of a classifier is a probability distribution over the admissible hypotheses, classifier incongruence detection basically involves a comparison of these distributions. The existing classifier incongruence measures advocated in the literature include the Bayesian surprise (KL divergence) [18] and the DC-KL divergence [38], or the heuristic delta measures (Δ^* and $\Delta_{\max}$) introduced in [21] and [24]. Unfortunately, the former two have undesirable properties and the latter two are heuristic.

Measuring differences between two probability distributions is a standard problem in information theory and statistics. The key tool for this purpose is divergence. Many different divergence functions have been proposed in the literature, each exhibiting different properties. In order to adopt or develop a suitable measure for detecting classifier incongruence it is of paramount importance to understand the properties required for this particular application. We argued that a classifier incongruence measure should focus on differences in the classifier support for the dominant hypotheses, be bounded, symmetric, insensitive to surprisal, and insensitive to clutter induced by nondominant hypotheses.

The list of required properties postulated in this paper can be considered as an important contribution in its own right. However, in the context of this paper, this was just a prerequisite for the main task of developing a principled method of measuring classifier incongruence. A review of existing divergences established that none of them fully satisfied the list of requirements. We adopted the total variation divergence as a starting point, because of its insensitivity to surprisal values. We then reformulated the problem of comparing two probability distributions by grouping all the nondominant classes into a single event. This allowed us to develop the total variation measure into a novel divergence, called *delta divergence*, which is classifier decision cognizant. As a result of this reformulation, the proposed measure is less sensitive to clutter induced by nondominant hypotheses. By studying the characteristics of the proposed measure we demonstrated that it satisfied all the required properties. An overview of the adherence of various classifier incongruence measures to these properties is presented in Table II.

Finally, we conducted a number of experiments on real and synthetically generated data to show the relationship of the proposed delta divergence to baseline classifier incongruence measures, and its robustness to clutter. The experiments confirmed its superiority as a measure of classifier incongruence.

References

[1] P. Aarabi, “Localization-based sensor validation using the Kullback–Leibler divergence,” *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 2, pp. 1007–1016, Apr. 2004.

[2] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, “Clustering with Bregman divergences,” *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, Dec. 2005.

[3] J. Basak, K. Kate, V. Tyagi, and N. K. Ratha, “QPLC: A novel multimodal biometric score fusion method,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, San Francisco, CA, USA, Jun. 2010, pp. 46–52.

[4] A. Bhattacharyya, “On a measure of divergence between two statistical populations defined by their probability distributions,” *Bull. Calcutta Math. Soc.*, vol. 65, no. 2, pp. 99–109, Feb. 1977.

[5] W. Bian, D. Tao, and Y. Rui, “Cross-domain human action recognition,” *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 298–307, Apr. 2012.

[6] D. E. Boekee and J. C. A. van der Lubbe, “Some aspects of error bounds in feature selection,” *Pattern Recognit.*, vol. 11, nos. 5–6, pp. 353–360, 1979.

[7] C. K. Chow and C. N. Liu, “Approximating discrete probability distributions with dependence trees,” *IEEE Trans. Inf. Theory*, vol. IT-14, no. 3, pp. 462–467, May 1968.

[8] D. Coppi, T. de Campos, F. Yan, J. Kittler, and R. Cucchiara, “On detection of novel categories and subcategories of images using incongruence,” in *Proc. Int. Conf. Multimedia Retrieval*, Glasgow, U.K., Apr. 2014, p. 337.

[9] I. Csiszár, “Information-type measures of difference of probability distributions and indirect observation,” *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 299–318, 1967.

[10] I. Csiszár, “On topological properties of f-divergences,” *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 329–339, 1967.

[11] I. Csiszár, “*I*-divergence geometry of probability distributions and minimization problems,” *Ann. Probability*, vol. 3, no. 1, pp. 146–158, 1975.

[12] I. Csiszár and J. Fischer, "Informationsentfernungen im raum der wahrscheinlichkeitsverteilungen," *Magyar Tud. Akad. Mat. Kutató Int. Kösl.*, vol. 7, pp. 159–180, 1962.

[13] R. A. Fisher, "Theory of statistical estimation," in *Proc. Cambridge Philosoph. Soc.*, vol. 22, 1925, pp. 700–725.

[14] I. Frýdlová, I. Vajda, and V. Kůs, "Modified power divergence estimators in normal models—Simulation and comparative study," *Kybernetika*, vol. 48, no. 4, pp. 795–808, 2012.

[15] V. Girardin and L. Lhote, "Rescaling entropy and divergence rates," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 5868–5882, Nov. 2015.

[16] D. V. Gokhale and S. Kullback, *Information in Contingency Tables*. New York, NY, USA: Marcel Dekker, 1978.

[17] P. Harremoes and I. Vajda, "On the Bahadur-efficient testing of uniformity by means of the entropy," *IEEE Trans. Inf. Theory*, vol. 54, no. 1, pp. 321–331, Jan. 2008.

[18] L. Itti and P. F. Baldi, "A principled approach to detecting surprising events in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 631–637.

[19] H. Jeffreys, "An invariant form for the prior probability in estimation problems," in *Proc. Royal Soc. A*, vol. 186, no. 1007, pp. 453–461, 1946.

[20] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Technol.*, vol. COM-15, no. 1, pp. 52–60, Feb. 1967.

[21] J. Kittler *et al.*, "Domain anomaly detection in machine perception: A system architecture and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 845–859, May 2014.

[22] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[23] J. Kittler *et al.*, "Intelligent signal processing mechanisms for nuanced anomaly detection in action audio-visual data streams," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Apr. 2018, pp. 6563–6567.

[24] J. Kittler and C. Zor, "A measure of surprise for incongruence detection," in *Proc. Intell. Signal Process.*, London, U.K., 2015, pp. 1–6.

[25] J. Kittler, C. Zor, I. Kaloskampis, Y. Hicks, and W. Wang, "Error sensitivity analysis of delta divergence—A novel measure for classifier incongruence detection," *Pattern Recognit.*, vol. 77, pp. 30–44, May 2018.

[26] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, 1951.

[27] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4394–4412, Oct. 2006.

[28] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.

[29] Z. Lin, G. Ding, J. Han, and J. Wang, "Cross-view retrieval via probability-based semantics-preserving hashing," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4342–4355, Dec. 2017.

[30] P. C. Mahalanobis, "On test and measures of group divergence, part I: Theoretical formulae," *J. Proc. Asiatic Soc. Bengal New*, vol. 26, pp. 541–588, 1930.

[31] K. Matusita, "Decision rules based on the distance for problems of fit, two samples and estimation," *Ann. Math. Stat.*, vol. 26, no. 4, pp. 631–640, 1955.

[32] F. Österreicher, "On a class of perimeter-type distances of probability distributions," *Kybernetika*, vol. 32, no. 4, pp. 389–393, 1996.

[33] F. Österreicher, "Csiszár's f-divergence-basic properties," Res. Rep. Collection, Victoria Univ., Melbourne, VIC, Australia, Tech. Rep., 2002.

[34] F. Österreicher and I. Vajda, "A new class of metric divergences on probability spaces and its applicability in statistics," *Ann. Inst. Stat. Math.*, vol. 55, no. 3, pp. 639–653, 2003.

[35] N. R. Pal and J. C. Bezdek, "Complexity reduction for 'large image' processing," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 32, no. 5, pp. 598–611, Oct. 2002.

[36] M. C. Pardo and I. Vajda, "About distances of discrete distributions satisfying the data processing theorem of information theory," *IEEE Trans. Inf. Theory*, vol. 43, no. 4, pp. 1288–1293, Jul. 1997.

[37] M. C. Pardo and I. Vajda, "On asymptotic properties of information-theoretic divergences," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1860–1868, Jul. 2003.

[38] M. Ponti, J. Kittler, M. Riva, T. E. de Campos, and C. Zor, "A decision cognizant Kullback–Leibler divergence," *Pattern Recognit.*, vol. 61, pp. 470–478, Jan. 2017.

[39] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Stat. Probability*, 1961, pp. 547–561.

[40] I. Sason, "Tight bounds for symmetric divergence measures and a refined bound for lossless source coding," *IEEE Trans. Inf. Theory*, vol. 61, no. 2, pp. 701–707, Feb. 2015.

[41] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[42] W. Stummer and I. Vajda, "On Bregman distances and divergences," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1277–1288, Mar. 2012.

[43] G. T. Toussaint, "Probability of error, expected divergence, and the affinity of several distributions," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-8, no. 6, pp. 482–485, Jun. 1978.

[44] D. Weinshall *et al.*, "Beyond novelty detection: Incongruent events, when general and specific classifiers disagree," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1886–1901, Oct. 2012.

[45] G. Zhang, S. Ferrari, and C. Cai, "A comparison of information functions and search strategies for sensor planning in target classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 1, pp. 2–16, Feb. 2012.

Josef Kittler (LM'12) received the B.A. degree in electrical engineering, the Ph.D. degree in pattern recognition, and the Sc.D. degrees from the University of Cambridge, Cambridge, U.K., in 1971, 1974, and 1992, respectively.

He is a Professor of Machine Intelligence with the Centre for Vision, Speech and Signal Processing, Department of Electronic Engineering, University of Surrey, Guildford, U.K. He published a textbook entitled *Pattern Recognition: A Statistical Approach* (Prentice-Hall), and over 200 journal papers. His current research interests include biometrics, video and image database retrieval, and cognitive vision.

Dr. Kittler serves on the editorial board of several journals in pattern recognition and computer vision.

Cemre Zor received the M.S. and Ph.D. degrees in signal processing and machine intelligence from the Centre for Vision, Speech, and Signal Processing, University of Surrey, Surrey, U.K., in 2008 and 2014, respectively.

She is currently a Research Associate with the Centre for Vision, Speech, and Signal Processing, University of Surrey. Her current research interests include anomaly detection, multiple classifier systems, as well as bias and variance theory of classification.