This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE

Domain anomaly detection in machine perception: A system architecture and taxonomy

Josef Kittler *Member, IEEE*[1*], William Christmas *Member, IEEE*[1], Teo de Campos[1], David Windridge[1], Fei Yan[1], John Illingworth[1] and Magda Osman[2]

***Abstract*—We address the problem of anomaly detection in machine perception. The concept of *domain anomaly* is introduced as distinct from the conventional notion of anomaly used in the literature. We propose a unified framework for anomaly detection which exposes the multifacetted nature of anomalies and suggest effective mechanisms for identifying and distinguishing each facet as instruments for domain anomaly detection. The framework draws on the Bayesian probabilistic reasoning apparatus which clearly defines concepts such as outlier, noise, distribution drift, novelty detection (object, object primitive), rare events, and unexpected events. Based on these concepts we provide a taxonomy of domain anomaly events. One of the mechanisms helping to pinpoint the nature of anomaly is based on detecting incongruence between contextual and noncontextual sensor(y) data interpretation. The proposed methodology has wide applicability. It underpins in a unified way the anomaly detection applications found in the literature. To illustrate some of its distinguishing features, in here the domain anomaly detection methodology is applied to the problem of anomaly detection for a video annotation system.**

***Index Terms*—Domain anomaly, anomaly detection framework, machine perception, anomaly detection mechanisms**

I. Introduction

Machine perception systems are invariably designed to deliver a specific functionality and consequently, they are domain dependent. Their design involves collecting a lot of training data and all the modules needed to accomplish a required task are trained as part of the design exercise. If the application domain changes, such systems are paralysed. They cannot adapt to a new scenario, even if there is a considerable degree of commonality between the existing competence and the desired new competence.

When the system is exposed to a new experience, some or all of the current models used by the system will fail to relate observed sensor(y) data to a correct meaning. This will be reflected in the support for various hypotheses allowed by each model becoming weak. We shall refer to this phenomenon as anomaly, which should trigger other mechanisms to initiate transfer of learning, so that the system can regain its useful functionality.

We address the problem of anomaly detection in machine perception. Building on the current state of the art in detecting anomalous events [39], [50], [58], the main goal of the paper is to develop a general framework for anomaly detection. We introduce the concept of domain anomaly, which differs from the conventional meaning of anomaly in the sense that it relates to a set of models characterising a domain. By a domain anomaly we understand a situation when none of the existing models can explain observed data. Using a mathematical apparatus drawing on Bayesian probabilistic reasoning, existing anomaly detection approaches are presented in a unified way and novel detection mechanisms are proposed. The innovative feature of the framework is that it exposes the multifacetted nature of anomaly and makes it possible to identify the diverse causes that can give rise to anomalous events, as well as corresponding detection mechanisms. The proposed extension of known anomaly detection mechanisms in the literature is very important as it enables the anomaly detection system to select an appropriate response. In particular, we shall distinguish between measurement outliers, distribution contamination, distribution model drift, new objects composed of known primitives, and new primitive model vocabularies. These nuances will allow us to introduce a taxonomy of domain anomaly events.

The contributions in this paper can be summarised as follows:

- We develop a unified framework for anomaly detection. This framework is a major extension of the conventional anomaly detection approaches reviewed in the papers of Markou and Singh [35], [36] and encompasses the recent important contributions to the anomaly detection problem presented in [58].
- We identify the concept of sensor data quality and model drift as essential elements of anomaly detection, facilitating understanding of its underlying causes.
- We argue that anomaly can also be caused by a model drift which is not necessarily observable in terms of outliers, and suggest mechanisms for model drift detection and classification.
- We propose a novel methodology for anomaly detection which draws on these criteria. The methodology uses jointly i) the concept of observation likelihood, ii) decision reject option, iii) congruence [58] of multiple (e.g. noncontextual and contextual [60], [33]) interpretations, iv) sensor data quality [34], v) and model drift to detect, identify and categorise different anomalies.
- We argue that Bayesian surprise [25] is not an ideal concept to measure incongruence of multiple interpretations and propose an alternative which obviates the pitfalls of Bayesian surprise.

The authors are with [1]the Centre for Vision, Speech and Signal Processing,University of Surrey, Guildford, GU2 7XH, UK. E-mail: {J.Kittler,W.Christmas,T.deCampos,D.Windridge,F.Yan,J.Illingworth} @surrey.ac.uk, and [2] Department of Psychology, Queen Mary, University of London, London E1 4NS, UK. E-mail: M.Osman@qmul.ac.uk

Digital Object Indentifier 10.1109/TPAMI.2013.209

- We identify and distinguish a number of different anomaly scenarios based on the proposed approach.

Some aspects of the proposed methodology are illustrated on the problem of anomaly detection in the context of transfer learning from automatic interpretation of videos of tennis singles to tennis doubles. We show that even in this relatively simple case, more than one model of the application domain triggers anomaly. We demonstrate that the proposed methodology successfully identifies the nature of these anomalies.

The paper is organised as follows. In the next section we review the literature on anomaly detection. However, in machine perception there are only a few examples of anomaly detection, mainly dealing with the discovery of new objects. Section III introduces the concept of domain anomaly and discusses various mechanisms for anomaly detection categorised by the type of model (generative in Section III-A or nongenerative in Section III-B) adopted for automatic sensor(y) data interpretation. We discuss the role of incongruence between the interpretations generated by multiple experts as an anomaly flagging mechanism in Section III-C. Most commonly multiple sensor(y) data interpretations are derived by contextual and non-contextual experts. A typical example of contextual decision schemes is presented in Section IV. A unified framework for anomaly detection is introduced in Section V where we elaborate some of the nuances of anomaly and how they relate to concepts such as unexpected event, rare event, outlier, out of vocabulary object, and out of vocabulary object primitive. The framework and anomaly detection methodology are applied to the task of domain anomaly detection in a sports video annotation system in the context of transfer of learning in Section VI. The paper is drawn to conclusion in Section VII.

II. Related work

The problem of anomaly detection has received considerable interest in the literature because of its practical potential. Our aim is to look at anomaly detection in the context of complex machine perception systems performing reasoning using multiple hierarchical models where the notion of anomaly assumes new levels of complexity. We shall draw on the existing surveys to define a baseline for anomaly detection and a platform from which more complex notions can be developed.

The early interests in abnormality, see e.g. [15], recorded in the statistical literature in the nineteen century, were motivated by problems of normal distribution parameter estimation caused by discordant observations. This seminal work eventually led to the theory of robust estimation [24]. Although solving a different problem, the byproduct of robust estimation methodology is the identification of outliers, which can be used for anomaly detection [44], [7].

The classical view of anomaly as an outlier from some known distribution [6], [1] which represents normality is referred to as *point anomaly*. The basic classification of anomaly detection approaches applicable to point anomaly, which has been introduced in preceding surveys [35], [36], [22], [2], identifies the following categories:

- statistical [7], [21], [41], [37]
- nearest neighbour [31]
- classification [42], [55], [26], [38], [49], [27], [12]
- clustering [19], [20]

A recent comprehensive and influential review [10] augments this classification by two other categories of methods, namely

- information theoretic [4]
- spectral [61]

These approaches use different criteria to define abnormality but basically they relate to the same notion of anomaly.

Learning of normality depends on the training data available. It can be based on samples representing the mundane (normal) process, or both normal data and samples of abnormal observations. The statistical approaches normally model the distribution functions, whereas the classification methods strive to delineate normal observations by a boundary of normality. This can be learnt from one class training data (set of positive training instances) [52], [53], [54], [46], [40] or using negative samples as well (negative, anomalous instances) [43], [50]. The learning process can be supervised, semi-supervised or unsupervised. Often the training data is corrupted by anomalies. A learning scheme that takes labelling impurities into account has been proposed e.g. in [16]. The relative merits of learning a positive instances detector rather than a negative instances detector has been investigated by [17].

The point anomaly does not capture anomalous situations such as those where individual observations may be consistent with normal data, but collectively, behaviourally or in context, the observations deviate from normality. In their survey, Chandola et al. [10] and [45] do identify this notion of anomaly and review the existing literature as a separate category, with one of the solutions being a conversion of these notions of anomaly into a point anomaly detection problem. A typical example of anomaly in context is an ordered sequence of observations, such as time series, where any single observation in the sequence may appear normal, but as a group, or jointly with its neighbours, the observation is an outlier [28], [8], [47], [33]. Anomalies in sequences of symbolic data have been studied in [13] and spatial outliers in [51]. A Markov chain model has been applied in [60].

More complex situations arise in multisensor systems where it is important to discriminate between corrupted data, faulty sensor nodes, and interesting events such as intrusion [18], [14], [11], [48], [62]. The various scenarios cannot be distinguished by simple point anomaly detection, but more sophisticated reasoning is required [39]. Often the detection of anomaly is motivated by the need to adapt to new environments [63].

In systems with multilevel representation of knowledge, each phenomenon will have more than one model (reference), depending on the number of levels of knowledge representation. This gives rise to a completely new notion of anomaly, a *compound anomaly*. The recent paper on rare events detection [23] is tackling such a problem, but under the assumption that some examples of rare classes are available for learning. Our approach draws on the fact that in the case of compound anomalies the respective interpretations of observations based

on the models at the different levels of representation disagree. This disagreement is referred to as *incongruence*. There is very little work in this emerging problem area, with the exception of speech recognition.

Incongruence detection is the focus of a European Union project Dirac, concerned with the detection of rare events. The idea advocated in [57], [56], [58] is to compare the outputs of weak and strong classifiers. A discrepancy in their output is flagged as incongruence. The approach follows the efforts in out-of-vocabulary word detection [9]. In this case the weak classifier, i.e. the phoneme detector, may be delivering phoneme hypotheses with confidence, but the sequence of detected phonemes is rejected by the strong, contextual classifier, because the word they correspond to does not exist in the system vocabulary. This discrepancy would suggest that an out-of-vocabulary word has been encountered, rather than a noisy speech segment which would produce low confidence phoneme hypotheses. Other examples include anomaly detection in multi-modal systems, where discord is manifest in the inconsistency of evidence provided by different data channels (modalities) [5].

The work in [57] and [65] uses the notion of compound anomaly detectable via incongruence for the detection of new subcategories of objects by measuring the disparity between a generalised context classifier (when giving a low confidence output) and a combination of 'specific-level' classifiers (generating a high confidence output).

These pioneering efforts in detecting anomalous observations in perception systems have identified new problems that require novel notions of anomaly and the corresponding formulations of the anomaly detection problem. It is our aim in this paper to develop a comprehensive framework for anomaly detection which will expose the deficiencies of the existing solutions. More positively, the framework will identify all the mechanisms needed for determining the true nature of anomaly and its detection. This framework is developed in the next section.

III. Anomaly detection mechanisms

In general, any domain will be characterised by a set of models M,

$$M = \{M_i | i = 1, ..., N_D\} \tag{1}$$

where M_i is a specific model relating to an element of domain D, and N_D represents the number of models characterising the domain. The set, M, will be referred to as the domain model and it will be assumed that it has been loaded into the system operational memory to enable the interpretation system to function. It should be noted that each element of domain D may consist of multiple submodels, thus forming a subdomain. For instance, one of the elements of the tennis video domain is a set of objects pertinent to the domain. Recognising these objects will require object appearance models and the set of such models will form a subdomain.

We are interested in detecting *domain anomaly*, by which we understand the failure of the domain models to explain the observed data. The functional form of a model depends on the modelled phenomenon. In very broad terms, all models used in machine perception can be categorised into *generative* and *nongenerative*. In the case of *generative* models, there is a transparent relationship between observations and models.

Nongenerative models lose the direct link to observations. This is exemplified by discriminative models which aim to identify the class identity of a sensory stimulus. However, a class identity is not sufficient to synthesise any specific observation which conveyed the class identity information in the first instance. Nongenerative models transform the interpretation problem from modelling observations to partitioning the observation space. The latter invariably introduces extrapolation which makes it difficult, if not impossible, to detect anomalies.

In the following we shall look at these two types of models from the anomaly detection point of view in more detail. Most importantly, neither generative, nor nongenerative methods directly detect *domain anomaly*, which arises when none of the domain models is able to explain the observed data. Nevertheless, they are the key instruments in domain anomaly detection and after their overview in the rest of the section, their role in domain anomaly detection will be discussed in Section V.

A. Generative models

Generative methods link model identity and measurements in a direct manner. In general the measurements will be derived from the sensor(y) data in some fashion. In computer vision, at the lowest level we may be dealing directly with image pixels, or with some higher level representations, such as image descriptors, or shape primitives. A generative model specifies how measurements are generated. By the same token, given a measurement, we can hypothesise a model and verify whether the measurement could have possibly been generated by the model by computing the likelihood of the observation. Typically, especially when dealing with signals captured by a sensor, the assumed generative process will be probabilistic. However, there are other generative models, e.g. grammatical models, which also link a model directly to observations but in this case via a set of deterministic generative rules. For the moment, we shall confine our discussion to probabilistic generative models which are defined by probability distributions. We shall refer to them as distributional models.

A distributional model $p(x)$ applies to a phenomenon where the process, generating members of a population, is characterised by a probability distribution over all its possible multidimensional outcomes, x, i.e. x is a random vector variable. An anomaly is an observation that is not consistent with our model. In general, an anomaly is manifest in a very low likelihood value of observation x, and can be detected by measuring $p(x)$.

An outlier relates to a single observation. In many interpretation tasks, instantiation of a model involves multiple observations. We can still apply the notion of outlier to all the observations $\{x_1, ..., x_k\}$ jointly, by using the joint distribution $p(x_1, ..., x_k)$ as our model. By assuming that our observations are independent, identically distributed (i.i.d.) random variables, we we can measure the log likelihood of

their occurrence using

$$\log p(x_1, ..., x_k) = \int \hat{p}(x) \log p(x) dx \tag{2}$$

where $\hat{p}(x)$ is the empirical distribution modelling the observations, while $p(x)$ is the hypothesised model distribution.

or using the Kulback-Leibler divergence

$$\Delta_{KL} = \int \hat{p}(x) \log \frac{\hat{p}(x)}{p(x)} dx \tag{3}$$

Comparing the two measures in (2) and (3) we note that they are related. The advantage of the Kulback-Leibler divergence is that it goes to zero when the empirical and model distributions are identical. In contrast, the optimum value of log likelihood, which will be achieved when the two distributions are identical, will be distribution dependent. This may cause some problems in setting anomaly detection threshold.

For multiple observations, the test in (3), which we shall refer to as *Distribution anomaly*, is more powerful than the likelihood test in (2), as each observation individually may be consistent with the model distribution and therefore, it would not be flagged as outlier. However, together the observations define an empirical distribution $\hat{p}(x)$, which may deviate from the model distribution.

It cannot be over-emphasised that neither outlier anomaly nor distribution anomaly necessarily imply domain anomaly. They simply indicate whether one or more observations are consistent with a hypothesised model. Observations that are anomalous with respect to a given model may be perfectly consistent with another model. Thus observations are anomalous with respect to a domain (subdomain) if and only if they cannot be explained by any of the models characterising the domain (subdomain).

B. Nongenerative Models

Nongenerative models do not explicitly estimate the measurement distributions. Consequently they do not facilitate any testing for measurement consistency with a hypothesised model. This renders anomaly detection rather difficult. A typical scenario where nongenerative models are used for sensory data interpretation is data classification. Nongenerative models are favoured in pattern classification because they focus on the classification task, rather than on modelling the class conditional measurement distributions. Owing to the emphasis on classification, rather than on generative modelling, the resulting solutions tend to yield better classification performance.

To formalise the discussion, consider a domain Ω with elements $\omega_i, i = 1, r$ each representing a class. Suppose the elements of Ω are not directly observable, i.e. they are observable only indirectly via a vector of measurements x. Then the interpretation of an observation becomes a standard pattern recognition problem where x is assigned to that class which is most probable, i.e.

$$x \rightarrow \omega_i \quad \text{if } P(\omega_i|x) = \max_l P(\omega_l|x) \tag{4}$$

It has been suggested in [58] that instead of working directly with the a posteriori class probabilities, it may be preferable to use a normalised version, $\Delta_c(x)$, referred to as decision confidence, which is defined as

$$\Delta_c(x) = \frac{P(\omega_i|x) - e_i}{1 - 2e_i} \tag{5}$$

where e_i is the average probability of objects belonging to class ω_i being misclassified. However, either measure, (eq. (4) or (5)), may suggest false confidence as an aposteriori class probability can be high even when $p(x) \rightarrow 0$, i.e. when the measurement is an outlier. This explains why discriminative classification methods cannot detect an anomaly reliably. They will always identify the most probable hypothesis whether they are competent to make a decision or not. Thus alternative solutions are required, as suggested in [58] (see the paragraph on incongruence below).

It is evident that if discriminative models are to be used to get better classification performance, they need a gating channel that will use one of the observation anomaly detection methods described in Subsection III-A to establish whether the output of a discriminative model procedure can be accepted or rejected. Alternatively, this gating could be accomplished using a discriminative method such as [52], [53], [54], [46]. However, these one-class classifiers would have to learn the domain of the measurement distribution $p(x)$, rather than the classification task itself. Thus even in this case one would need a separate method for anomaly detection and for classification. This is an important conclusion which contributes to the understanding of the anomaly detection problem. We shall return to this point in Section V.

Note also that although the decision confidence measure discussed above cannot be used alone for anomaly detection, it is a useful measure for characterising the sensory data interpretation landscape, especially helping to distinguish anomaly from labelling errors due either to genuine ambiguity, or noisy, or otherwise corrupted measurements.

C. Incongruence

Although a single expert does not have the capacity to detect and or qualify unexpected events, the ability to detect anomaly improves dramatically when more than one expert is involved in decision making [58]. In the past decade or so, we have seen the tendency to engage more than one expert for sensory data interpretation for a multitude of reasons. Multiple experts improve performance by exploiting

- multiple modalities of sensing
- multiple representations
- contextual information
- interpretation process structuring

If a domain is characterised by more than one model type, the chance of two models reacting in exactly the same way to an anomaly is quite low.

Let $\tilde{P}(\omega_j|x)$ and $P(\omega_j|x)$ denote the aposteriori probabilities associated with the hypothesis that model ω_j explains the input data, which have been generated by two experts. The idea of measuring incongruence emerged in the context of speech recognition where one of the challenges is to detect out-of-vocabulary words, and this is achieved by comparing

noncontextual and contextual phoneme classifiers [29]. This idea has been considerably developed within the European project DIRAC where it has been extended and applied to other multiple classifier scenarios to detect incongruence of multimodal experts [5] and to the problem of detecting novel subclasses of objects [57]. By considering the aposteriori class probability distribution output by one of the experts as a reference, one can detect incongruence by measuring the Kulback-Leibler divergence between the two distributions [29] as

$$\Delta_{BS} = \sum_{j=1}^{r} \tilde{P}(\omega_j|x) \log \frac{\tilde{P}(\omega_j|x)}{P(\omega_j|x)} \quad (6)$$

which is known as the Bayesian surprise [25]. A close inspection of the measure reveals that it goes to infinity for any hypothesis ω for which $P(\omega|x) \to 0$ while $\tilde{P}(\omega|x) \neq 0$. This can occur even for insignificant hypotheses and result in producing false alarms of incongruence. To avoid the problems associated with the Bayesian surprise measure we propose an alternative which focuses on the dominant hypotheses flagged by the two experts. Let us denote these hypotheses by $\tilde{\mu} = \arg\max_\omega \tilde{P}(\omega|x)$ and $\mu = \arg\max_\omega P(\omega|x)$. Then an incongruence indicator can be defined as

$$\Delta_{max} = \frac{1}{2}[|\tilde{P}(\tilde{\mu}|x) - P(\tilde{\mu}|x)| + |\tilde{P}(\mu|x) - P(\mu|x)|] \quad (7)$$

though other norms could be used as well. This incongruence measure has several advantageous characteristics. It is symmetric, i.e. its value does not depend on an arbitrary choice of one of the experts as a reference. It eliminates the noise injected by the nondominant classes. Its values are not driven to infinity, but are confined to the interval $(0, 1)$.

It should be noted that any incongruence detected between the outputs of experts only flags potential anomalies, rather than pinpointing their origin and nature. For that a follow-up analysis using observational anomaly detectors would have to be carried out. This will further be explored in Section V.

D. Overview of findings

It is pertinent to summarise the key points of the discussion so far. The classical methods of anomaly detection are concerned with *observational* anomalies, which are of two types: *likelihood anomaly* (outliers) or *distributional anomaly*, depending on whether we are dealing with single or multiple measurements. Detectors based on these notions of anomaly do not directly flag a domain anomaly, but are the means of detecting domain anomaly.

Commonly, data interpretation processes make use of nongenerative models which are often preferred to generative ones because of their focus on decision boundaries and their speed of execution. However, they also have a disadvantage; they lack the capacity to detect domain anomalies. However, when the interpretation process involves multiple experts for each decision, the situation changes. In particular, incongruence between the outputs of multiple nongenerative models is indicative of potential domain anomaly. Incongruence can also help to qualify the type of anomaly, even in the case of generative classifiers. It can be measured using Bayesian surprise. Once incongruence is detected, the cause of anomaly and its nature must be analysed using supplementary techniques based on observational anomaly detection.

IV. Contextual classifiers

In the previous section we showed that nongenerative models do not have the capacity to detect domain anomaly, with the exception of the cases when more than one expert is involved in the instantiation of a hypothesis. Incongruence of the expert outputs is a sufficient condition for anomaly, the nature of which has to be established by further processing. In the list of scenarios where multiple experts might be engaged in interpretation, the contextual classifier category is a particularly important family. It encompasses classification approaches where sensor(y) data is represented hierarchically in the process of deriving a symbolic representation of the sensor(y) signals, as in the application discussed in Section VI. At each level of representation we then have two opinions on the class identity of a segment of data, voiced by a noncontextual expert, using only the measurements relating to the data segment, and a contextual expert which bases the decision on the information drawn from both the segment and its neighbours. The noncontextual classifiers are often referred to as *weak* classifiers and contextual ones are known as *strong* classifiers.

Contextual classifiers [60], [33] are important not only because of their prevalence in machine perception, but also because they provide information that facilitates a deeper analysis of anomalous situations. From the methodological point of view they are interesting because contextual decision making can be formulated in many different ways.

Hierarchical models are composed of objects (object primitives) which are combined at the next level to construct higher level concepts. Let a meaningful group be constituted by k components with associated measurements $x_i, \quad i = 1, ..., k$ and their labels θ_i. Then apart from noncontextual interpretation of each component based on $P(\omega_i|x_i)$, we can also interpret objects in context by computing $P(\theta_i|x_1, .., x_k, \theta_1, ..., \theta_{i-1}, \theta_{i+1}, .., \theta_k)$, which can be be expressed as

$$P(\theta_i|x_1, .., x_k, \theta_1, ..., \theta_{i-1}, \theta_{i+1}, .., \theta_k) = \frac{p(x_1,...,x_k|\theta_1,..,\theta_k)P(\theta_1,..,\theta_k)}{\sum_\lambda p(x_1,...x_k|\theta_1,..,\theta_i=\lambda,...,\theta_k)P(\theta_1,..,\theta_i=\lambda,...,\theta_k)} \quad (8)$$

where $P(\theta_1, ...\theta_i, .., \theta_k)$ is the prior world model of object configurations, and $p(x_1, ..., x_k|\theta_1, .., \theta_k)$ denotes the joint measurement distribution. We can see that anomaly detection becomes quite complex. First of all, individual object detectors can produce anomalous results either because of the associated measurement is an outlier, or the primitive concept is missing from the list of primitives. We can also have an anomaly caused by a missing item in the world model. Low values of the joint measurement distribution could also be flagging an outlier. Thus there are *four* drivers behind anomalous situations and, to understand the meaning of anomaly, these have to be properly differentiated. We shall explore these different situations further in Section V.

As anomalies of different kinds can occur jointly, methods for pinpointing their cause is required. The observation

anomaly and incongruence measures discussed in Section III can be used as anomaly detection mechanisms, and help to identify different anomaly scenarios from the combinatorial list of possibilities generated by the four main drivers of anomaly. There are various ways this can be accomplished. For instance, for each primitive, i, we would expect an agreement between the probability, $P(\theta_i|x_i)$, assigned by a weak (non contextual) expert and $\tilde{P}(\theta_i|x_i) = P(\theta_i|x_1, .., x_k, \theta_1, ..., \theta_{i-1}, \theta_{i+1}, .., \theta_k)$ assigned by a strong (contextual) classifier. The sequence of incongruence measure values $\Delta(i), i = 1, ..., k$ would need to be further analysed to discriminate between different scenarios. There are some simple scenarios which are indicative of, for instance, the existence of an unknown class, or (the case) of noisy measurements. However, the complexity of the anomaly landscape is quite high, especially taking into account the role of measurement likelihoods in identifying outliers. There are no comprehensive anomaly measures at the moment which could, in a systematic way, reflect all the different anomaly scenarios.

When the labels on the neighbouring primitives are not available, the computation of the strong classifier probability $\tilde{P}(\theta_i|x_i) = P(\theta_i|x_1, .., x_k)$ involves compounding the supporting evidence for a particular hypothesis over all contextual interpretations. This will be illustrated in Section VI-A where we discuss an example application exercising features of the proposed framework.

As already pointed out in Section III-A, in many cases objects or phenomena constituted by primitives do not have a fixed structure, as for instance, do words in a vocabulary, which are defined in terms of specific sequences of characters. The structure will be determined by a grammar, or a probabilistic model capable of generating different structures. There are many models that fall into this category, with a Markov model being the most common. Under the Markovian assumption the computation of the contextual probabilities of the class identity of the successive primitives can be considerably simplified.

V. Anomaly detection framework

One of the key mechanisms of anomaly detection exploited in many of the scenarios is outlier detection. In almost all cases *outlier detection* relates to probability distributions, but domain anomaly arises when one or more observations cannot be explained by our world models. As a result of the unified treatment of the anomaly detection problem in different scenarios, a number of important conclusions emerge, which facilitate the development of an appropriate anomaly detection methodology for machine perception applications.

The various output states of anomaly assessment are summarised in Table I. They can be identified by analysing the relevant factors which include the likelihoods of measurements made on objects (components), the distribution of aposteriori probabilities for the various object/component hypotheses, aposteriori probabilities of contextual labelling of components, or unconditional likelihood of joint observations on multiple components. Anomaly can also be caused by distribution drift, and this can in principle happen without any individual observation being an outlier. An anomaly can be also a manifestation of some measurement corrupting processes such as noise. This situation should be recognised by means of auxiliary measurements such as image (sensor data) quality measures.

Referring to the multiplicity of the factors that can lead to anomaly, the identification of the different situations is far from trivial. We can clearly conclude that none of the papers reviewed in Sections II and III are capable of detecting and distinguishing all the nuances of anomaly. The key approach used in the literature, which effectively defines a reject class, makes use of only one of the measures listed in Table I, i.e. the distributions of posteriori class probabilities. The work in [32] supports a more sophisticated detection and analysis of anomaly, but it does not take into account all the cues identified in Table I. In principle it is extendible to identify certain other anomaly cases. However, even this approach is limited in scope, as it does not take the measurement distributions into account and has no mechanism for independent assessment of the quality of observational data to avoid generating false anomaly positives.

We propose a comprehensive methodology for anomaly detection which builds on the evaluation measures suggested for the various anomaly factors in the literature. The key contribution here is that all the relevant factors have to be evaluated jointly. Thus for single entities, we have to assess measurement likelihood, decision ambiguity and sensory data quality. For structures, in addition, we have to measure incongruence between noncontextual and contextual interpretations of the structural primitives, as well as the likelihood of joint observations of these primitives. The bag of tools therefore comprises:

1) **Observation anomaly (outlier) detector-** using likelihood (e.g. $p(x_i|\theta_i), p(x_i), \forall i,\ p(x_1, ..., x_k)$. An outlier can be identified using any of the standard methods suggested in the literature (viz a comprehensive review in [35]), such as likelihood falling below a certain threshold.
2) **Reject option detector-** The lack of convincing support for any of the hypotheses associated with an application domain, whether relating to single entities or structures. A reject option can be flagged by measures defined in terms of aposteriori probabilities of the various hypotheses, such as the decision confidence measure [58] introduced formally in (5) or an entropy measure. Note that a lack of confidence in a decision may be due entirely to genuine ambiguity, and can be observed even in the case of good quality sensor data.
3) **Incongruence detector-** Any inconsistency between the interpretations suggested by two experts in general, and by noncontextual and contextual labelling processes relating to the components of a structure in particular, is potentially indicative of anomaly. The nature of an anomaly flagged by incongruence will depend on the respective confidences in these two decision outcomes. Incongruence can be measured as suggested, for instance, in (7) in Section III, or using machine learning [32]. The anomaly types associated with incongruence include

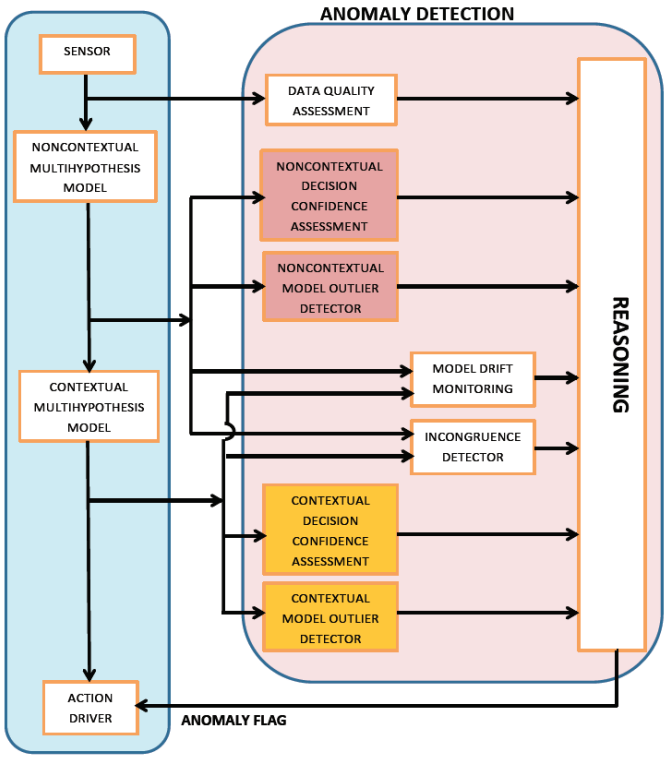

Fig. 1. Domain anomaly detection system architecture. The sensor(y) data to be interpreted feeds into a discriminative object/primitive (noncontextual) classification system. The output of the noncontextual decision making system is then channelled to a contextual classifier. The "noncontextual model" stands for object/primitive/component model, and the contextual model captures the "scene" (configuration) information. Both classifier outputs are fed into decision incongruence detector and model drift monitor.The sensor(y) data is also processed by generative models to measure the marginal and joint likelihoods of the observations. These likelihoods help to detect and qualify anomalous situations, with the help of data quality estimates and decision confidence measurements.

unexpected event, where a given component is out of context, a *rare event*, where a given configuration of components occurs very infrequently, and an unknown structure (out of vocabulary word).

4) **Sensory data quality gauge-** An independent assessment of sensory data quality is necessary to disambiguate some of the anomaly cases where the above detectors alone would not be able to judge whether, for instance, incongruence between noncontextual and contextual labelling of structural components is caused by model inadequacy or by measurement errors. This is a largely unexplored area, but it is evident that sensory data quality is a multifacetted concept. Standard measures of, e.g., signal to noise ratio, resolution, bandwidth, contrast, etc. and their combination, will be indicative of various aspects of quality and can be used for this purpose (see e.g. [34]).

A schematic diagram of the overall anomaly detection system is shown in Figure 1. The system is illustrative of the case when the multiple (two) discriminative models relate to noncontextual and contextual labelling of individual objects/primitives. For other scenarios, such as multimodal experts, the system would have to be suitably adapted. The anomaly detection subsystem engages five different mechanisms: Observation outlier detection, and decision confidence estimation for the contextual and non contextual classifiers, data quality gauging, decision incongruence detection, and model drift monitoring. The anomaly detection system diagram subsumes the noncontextual decision-making scenario where the scene model would not exist, and congruency would not be measurable. It cannot be over-emphasised that the detection of an anomaly and its comprehensive qualification cannot be successfully accomplished without all these sources of information contributing to the final inference. We do not detail the actual anomaly analysis processes as they will be problem specific. The anomaly analysis stage of the system in Figure 1 can identify different states of the sensor(y) data interpretation (model instantiation) process and can detect different types of anomaly, depending on the information provided by the different gauging systems, which have been identified and suitable measures suggested. In the following the various output states will be briefly elaborated.

a) **No anomaly:** This refers to the normal mode of operation when a good quality observation supports a distinct hypothesis from the available set of possible interpretations.

b) **Noisy measurement:** When the measurements are affected by noise, the interpretation of a single entity will inevitably become more ambiguous. The ambiguity will be reflected in the entropy of the aposteriori class probability distribution. This case should be flagged by a quality measurement extracted from the sensory data.

c) **Unknown object:** When sensory data relates to an object which has no model in the existing model database, the likelihoods of the unconditional measurement distributions will be low or even report outliers. This will inevitably lower the entropy of aposteriori class probabilities and consequently the decision confidence. However, these two indicators alone cannot differentiate between the noisy measurement scenario discussed in *Paragraph b* above and the case of an unknown object. While the latter should trigger a learning mode during which the model database is augmented by a new object model, the former should simply issue a warning about the low confidence in interpretation, as a result of noisy observation. These two types can be discriminated with the help of a suitable measure of sensory data quality. The case of *unknown object* would be reflected in the sensory data quality measure indicating good quality signal.

d) **Measurement model drift:** Unrepresentative training data, or changes in environmental conditions, may result in measurement model drift when the designed system is deployed operationally. Such a drift will not necessarily be manifest in the detection of outliers. These changes can be detected by monitoring the measurement distributions over time and by comparing them with the learnt models. Potentially there are two main situations of interest. Either the underlying models remain conceptually the same and a drift simply signifies that the measurement model should be adapted to accommodate the range of operational conditions. Alternatively, the drift is a result of semantic domain changes which call for learning new domain models.

e) **Measurement ambiguity:** Genuinely ambiguous measurement will give rise to low confidence decisions.

f) **Congruent labelling:** A structure is formed by its components (primitives). Different configurations of components define different structures. If noncontextual and contextual labelling of the components are congruent, then the observations are deemed to be consistent with the domain models and the conduct of the interpretation process is considered to be normal. Some configurations of components may occur less frequently than others. Such configurations are congruent but correspond to *rare events*.

g) **Unknown structure:** If the component labelling is performed with confidence, but the resulting configuration does not exist in the domain model base, the observations most likely relate to an unknown structure. A typical example of this situation is out-of-vocabulary word detection in speech recognition. In this application words are composed of phonemes, and the world of all the possible utterances is modelled by a vocabulary, i.e. a list of valid words. If a speech utterance contains a word which is not included in the vocabulary, such as proper names of people and places, the phoneme recogniser may function with confidence but fail to output a sensible interpretation. However, the contextual interpretation of the components will be incongruent with the noncontextual interpretation. This incongruence observed in the context of good quality sensory data will be indicative of the configuration of components forming no known structure. The model base will have to be updated to make it complete, or potentially a new domain model will have to be created (e.g. vocabulary for another language).

h) **Unexpected structural component:** Here the most likely cause of the measurements on some components being outliers is the absence of a relevant object/component model. Although in this scenario the sensory data quality would be high, the observational evidence would fail to support any component model in the model base and the event would be deemed to be *unexpected*, signifying a domain anomaly. Note that *unexpected event* could also arise for instance when, for computational expediency, only a subset of object models is in active use. However, if the observed data cannot be interpreted congruently using the active section of the model base, but is interpretable using an extended or complete model base, then the relevant event would not be anomalous. In fact it would be a *rare event*. *Unexpected event* could also be caused by spurious noise which affects only the measurements on a single object/component. Such an event would be *unexpected* by virtue of the prevalent context. The reasoning mechanism (not elaborated herein) that analyses the various anomaly qualifying measures would have to allow for all the possible outcomes and, if necessary, instigate a follow up exploration to disambiguate the various options.

i) **Unexpected structure and structural components:** When the application domain of a machine perception system is changed, neither component models, nor structure models are relevant to observations. A simple example is an optical character recognition system designed for automatic reading and understanding (word level) text in English presented with a text in Arabic. In such a case neither the world model (vocabulary), nor the set of Latin character measurement models will be relevant to the task. The domain change will be characterised by most observations being classified as outliers for all class conditional measurement distributions, accompanied by a systematic failure of component interpretation. If, at the same time, the sensory data quality is high, these anomaly detection measures will be indicative of a major change in the sensory data content and the system will have to switch to a training phase to learn the new domain models.

j) **Noisy joint measurements:** If the anomaly detection tools discussed in the previous paragraph exhibit similar symptoms, but the sensory data quality is deemed to be low, the most likely interpretation of the situation is that more than one observation are severely corrupted by noise or changes in the sensory data acquisition conditions. The first corrective step in this situation is to initiate a system diagnosis and environment monitoring check to eliminate any malfunction.

k) **Component model drift:** Referring to our discussion in *Paragraph d* above, the class conditional measurement distributions relating to structural components may be subject to drift. Again, this would not necessarily become obvious from individual observations as these may perfectly well be distribution inliers. However, monitoring these distributions over time would give an opportunity to detect any drift that requires adaptation, or alternatively, that may be indicative of the underlying models being rendered irrelevant by a change of sensory data content. The techniques suggested in *Paragraph d* would be applicable to the problem of model drift and its identification.

l) **Ambiguous measurements:** Ambiguous interpretation of components may give rise to false positive incongruence. In such situations, the anomaly detection mechanism should be disabled.

The domain anomaly cases discussed in the preceding paragraphs are identified in Table I. Their taxonomy derives from the type of subdomain they relate to, namely a component subdomain or a configuration subdomain, resulting in the following three categories:

1) Component Domain Anomaly *CpntDomAn*
2) Configuration Domain Anomaly *CfgDomAn*
3) Component and Configuration Anomaly *Cpnt&CfgDomAn*

The observational and distributional anomalies are merely some of the detection tools that are needed to flag and identify a domain anomaly.

In Table I we cite one or two examples for each category of anomaly, as well as the cases where data quality or decision ambiguity measures are used to disable anomaly detection mechanisms so as not to generate false positives. The examples in the next section arise in interpreting a video of tennis doubles using a system trained on tennis singles. The only exceptions (no examples given) are Case i which would arise in e.g. speech recognition where a change of language would potentially involve both new component and new configuration models, and Case l, where a detected drift of component distributions could be accompanied with a change of high level rules. For instance, analysing a badminton video with a tennis game interpretation system would be a case in point.

TABLE I

AN OVERVIEW OF THE VARIOUS STATES DETECTED BY THE ANOMALY ANALYSIS STAGE OF THE ANOMALY DETECTION SYSTEM IN FIGURE 1. THE TABLE IDENTIFIES THE DOMAIN ANOMALY STATES. THE DOMAIN ANOMALY TYPES LABELLED BY *CpntDomAn*, *CfgDomAn* AND *Cpnt&CfgDomAn* IN THE LAST COLUMN DENOTE *component domain anomaly*, *configuration domain anomaly*, AND JOINT *component and configuration domain anomaly* RESPECTIVELY. THE MEASURABLE CHARACTERISTICS ASSOCIATED WITH THE STATES AND DOMAIN ANOMALY TYPES ARE ALSO INDICATED IN THE TABLE.

Case	Confidence of noncontextual labelling of components $\Delta_c(x_i), \forall i$	Confidence of contextual labelling of components $\Delta_c(x_1, .., x_k)$	Unconditional likelihood $p(x_i), \forall i$	Drift Δ_{KL}	Incongruence of labelling components $\Delta_{max}(x_i),\ \forall i$	Probability of joint labelling of components $P(\theta_1, ..., \theta_k))$	Joint likelihood $p(x_1, ..., x_k)$	Data quality	Domain anomaly	*Domain anomaly type* / Comment
a	High	k=1: no context	High	No				Good	No	Routine single object (component) labelling
b	Any	k=1: no context	Any	No				Low	Disabled	Noisy measurement (see Section VI).
c	Any	k=1: no context	Low	No				Good	Yes	*CpntDomAn* / Novelty object (component), new model required [64]
d	Any	k=1: no context	Any	Yes				Good	Yes	*CpntDomAn* / Model drift (observable over time): Model innovation or adaptation required (see Section VI).
e	Low	k=1: no context	High	No				Good	No	Ambiguous measurement (see Section VI).
f	All high	All high	All high	No	No	High	High	Good	No	Routine operation
g	All high	Some Low	All high	No	Some high	Zero	Low	Good	Yes	*CfgDomAn* / Scene model not available (e.g. out of vocabulary word) [9], [56]
h	Some high, some low	Some low	Some low	No	Some high	Low $\rightarrow$ zero	Low	Good	Yes	*CpntDomAn* / Unexpected component: Component model correction/revision restores congruency (see Section VI).
i	All any	All low	All low	No	Some high	Low	Low	Good	Yes	*Cpnt&CfgDomAn* / All components unexpected: Different alphabet and vocabulary required
j	Any	Any	Any	No	Some high	Low	Low	Low	Disabled	Noisy joint measurements on more than one primitive [32]
k	High	High	Any	Yes	Some high	Any	Any	Good	Yes	*CpntDomAn* / Component model drift (observable over time): New component and possibly world model or adaptation required
l	Low	Low	All high	No	Some high	Any	High	Good	Disabled	Ambiguous measurements (see Section VI).

VI. Anomaly detection in tennis video interpretation

We shall demonstrate some elements of the architecture discussed in the previous section on the problem of anomaly detection in the domain of sports video annotation. The anomaly detection problem arises in the context of an autonomous system which has the ability to interpret video of tennis singles, and the long term aim is to transfer this competence to a new domain, such as game of badminton, volleyball, or table tennis.

The tennis annotation system we use as a basis is quite complex, comprising more than 15 modules, with each module realising its functionality with the help of multiple models [30], [59]. The modules perform, for instance, tennis court localisation, the detection and tracking of players and the ball, and detecting ball events that are identified by a rapid change in the ball direction caused by a bounce, hit or net. The high level modules of the system process the ball events and player information to make decisions about the match score.

An integral part of the system operation is the ability to flag anomalous situations and thereby identify when some of the modules and/or models no longer have the competence to interpret the incoming data. Thus every module is equipped with an anomaly detection system engaging some or all the elements of the general architecture introduced in Figure 1. For simplicity we shall limit our discussion to a simple scenario where the system, which has been designed to process videos of tennis singles, is suddenly presented with a video of tennis doubles. Clearly, in this simple situation many of the system modules will function normally. The exceptions are the player detection module which should report the most apparent change between the domain of tennis singles and that of tennis doubles, that is the number of players present. More subtle is the change of rules relating to the definitions of the play area in singles and doubles respectively. The detection of these two anomalies will now be discussed in turn.

A. Number of players anomaly

As we use a very simple motion-based blob detector, the number of people detected will vary from frame to frame, as it will be affected by the presence of other agents (line judges, ball boys). Thus, normality has to be modelled in terms of a distribution of the number of players over time in a video shot, rather than as an instantaneous count.

Anomaly will be manifest as a deviation from the distribution learnt during the system design. This will be detected by the model drift monitoring module in Figure 1. For tennis singles and doubles, examples of the respective distributions are shown in Figure 2. A number of similarity measures could be used to compare a test histogram $p(x)$ with a model histogram $\hat{p}(x)$, both with D bins, but we use the simple mode difference $MD(p(x), \hat{p}(x)) = \arg\max_x p(x) - \arg\max_x \hat{p}(x)$ We find the upper and lower thresholds for which none of the shots in videos of singles is rejected. an upper thresholds are required for anomaly detection. Two videos of tennis singles are used; one for training and one for validation. In order to estimate the thresholds, comparisons on validation sets of *normal* (singles) videos were performed and the maximum differences determined. By measuring the similarity of the learnt model with the histogram of the player count for a test video, any deviation from the norm can be detected.

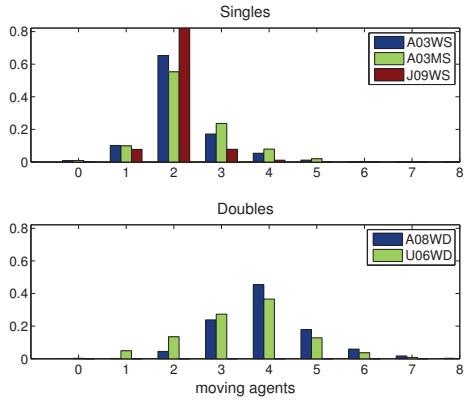

Fig. 2. Normalised histograms of the number of moving agents detected per frame in the games of singles (top) and doubles (bottom).

We conducted evaluation experiments with the videos in Table II. With each play shot lasting up to 2 minutes, we tested the anomaly detector on more than 5 hours of footage.

TABLE II
Tennis videos used in experiments and their duration

Label	Tennis match	# play shots	noise mean $\pm$std
A03WS	Australia03 Women's Singles	76	1.9 ± 0.1
A03MS	Australia03 Men's Singles	143	3.1 ± 2.3
J09WS	Japan09 Women's Singles	100	1.6 ± 0.5
A08WD	Australia08 Women's Doubles	164	1.3 ± 1.0
U06WD	USA06 Women's Doubles	66	1.5 ± 1.9

The results obtained with test video histograms computed from the shot frames as a function of the number of play shots are shown in Figure 3. The results show that even from the frames of one shot the system can detect anomalies most of the time. We see that with a temporal integration, i.e. accumulating the statistics over several play shots, a perfect detection of player count anomaly can be achieved. However, for the training configuration used in Figure 3(c), the zero positive rate on test singles is recovered only for video segments exceeding 9 shots.

A closer analysis revealed the importance of data quality estimation in anomaly detection. At each pixel the intensity standard deviation, estimated using robust statistics over a sequence of motion compensated frames in one shot, is averaged over all scene pixels and shots. Table II shows the mean noise measure and its standard deviation for each of the videos used in this paper. Note that the amount of noise in some of the training and validation videos is very different, which leads to relatively high thresholds to eliminate false positives. This causes under-detection of true anomalies in doubles for low levels of temporal integration in Figures 3 (a) and (b). The temporal integration over several shots improves the anomaly detection performance. Note that in Figure 3(c) the quality of the two videos used for training is comparable, which yields tighter thresholds and the need for much shorter temporal integration to achieve a perfect anomaly detection performance. However, the temporal averaging initially increases the false

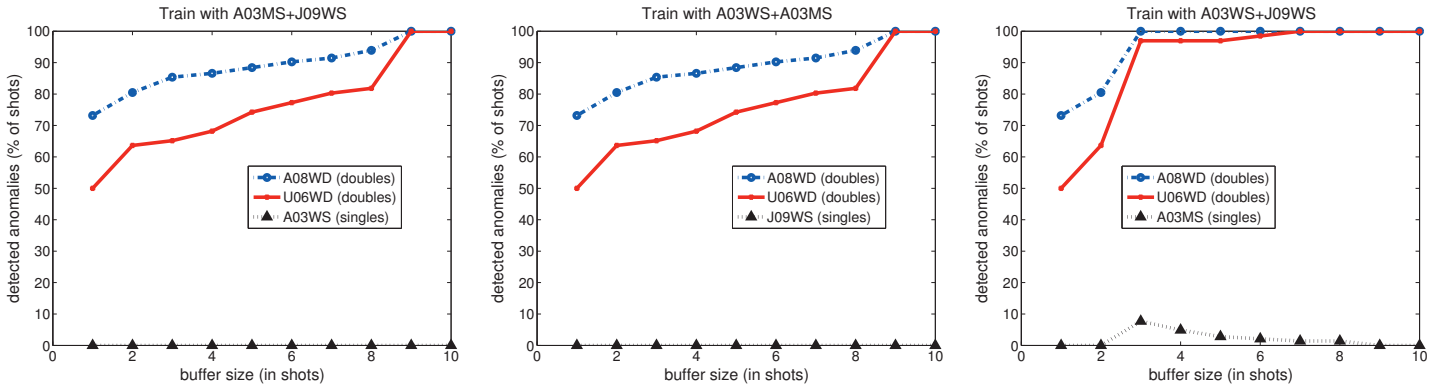

Fig. 3. Percentage of anomalies detected as a function of the number of shots used for the analysis on games of *singles* and *doubles*, using different combinations of two games of singles for training and to set thresholds.

positive rate on singles, before it recovers for long temporal integration periods. The conclusion from this study is that "normality" can be defined reliably only when the quality of data used for the system design (both training and validation) and in operation (testing) is comparable. Thus the training configurations A03MS+J09WS and A03WS+A03MS should not be used for the system design, and no anomaly detection should be attempted with A03MS on a system trained with A03WS+J09WS. The data quality estimation module in the anomaly system architecture in Figure 1 is absolutely essential to flag any discrepancy between the quality of the sources of data used for design and to inactivate the anomaly detection module for inputs corrupted by noise.

B. Out-of-play area anomaly

The evolution of a tennis game can be described entirely in terms of tennis ball events. These are the points in a tennis ball trajectory where the direction of motion changes dramatically, caused either by a player action, or by the ball bouncing off the ground or hitting the net. At the end of a normal play, the exchanges between the players may continue for a little while out of the inertia of behaviour. The ball activity may also be driven by other agents, such as ball boys, before finally stopping.

Depending on where the ball events take place (side of the court) and their type (hit, bounce), they can be classified into the following categories:

Notation	Event type
hit_A	ball hit by player_A
bounce_in_A	ball bounces in play area_A
bounce_out_A	ball bounces outside play area_A
net_A	ball played by player_A hitting the net
serve_A	serve delivered by player_A

These states are duplicated for $\text{player}_{\neg A}$.

These ball events are measurable and detectable, with uncertainties, using cues such as ball event vicinity to each player, and their relationship to the court. For a ball event, i, this information is conveyed by the measurement vector $\mathbf{x}_i$, with its measurement distribution for event type θ_i given by $p(\mathbf{x}_i|\theta_i)$. Note, that the interpretation of ball events is dependent on the previous state. In other words, the label, θ_i of ball event i, is given by $\theta_i = \arg\max_\omega P(\theta_i = \omega|\mathbf{x}_i, \theta_{i-1})$where $P(\theta_i = \omega|\mathbf{x}_i, \theta_{i-1})$ denotes the aposteriori event class probability function. Using the above measurement distributions, this aposteriori probability for label θ_i is given as

$$P(\theta_i|\mathbf{x}_i, \theta_{i-1}) = \frac{p(\mathbf{x}_i|\theta_i)P(\theta_i|\theta_{i-1})}{\sum_{\theta_i} p(\mathbf{x}_i|\theta_i)P(\theta_i|\theta_{i-1})} \tag{9}$$

In spite of its dependence on the previous state, this ball event labelling process can be considered as noncontextual, as it lacks the capacity to capture the complete picture of the tennis game evolution. The full understanding of the game is provided by a contextual model that processes the complete sequence of events from the initial ball event, i.e. the serve. The admissible sequence of ball events is modelled using a Markov chain with learnt state transition probabilities. This Markov model is used to monitor the state of play and to decide which player should be awarded a point. This is described in detail in [30].

When a ball event signals the end of play, by being classified as a bounce_out or hit twice by the same player, both the noncontextual and the Markov models detect illegal evolution of the game and the play is expected to terminate. Due to the inertia of the player action, a few normal exchanges between the players may follow the game terminating event, followed by other ball events associated with the ball(s) being collected from the court by the ball boys/girls. While the first exchanges between players (typically not more than 3) may comply with the rules of the game, the latter ball events will not, and will form an illegal sequence.

Let the ball event, i, be a game terminating event, ω. Then we would expect the last ball event in the sequence to have index $i+n$ where n is low. For $n > 3$ the sequence of observed ball events would contain illegal transitions from the point of view of the rules of a tennis game proper. Thus a contextual check on the decision that θ_i is an end of play event can be made by looking ahead at events $\theta_{i+1}, ..., \theta_{i+n}$, and computing $P(\theta_i = \omega|\mathbf{x}_{i+1}, ..., \mathbf{x}_{i+n})$. We compute $P(\theta_i = \omega|\mathbf{x}_{i+1}, ..., \mathbf{x}_{i+4})$ as

$$P(\theta_i = \omega|\mathbf{x}_{i+1}, ..., \mathbf{x}_{i+4}) = \begin{cases} 1 & n \leq 3 \text{ legal exch.} \\ 0 & n \geq 4 \text{ legal exch.} \\ 1 & \text{illegal} \end{cases} \tag{10}$$

A measure of incongruence between the non-contextual and contextual probabilities of a fault event is used to signal a potential anomaly. We adopt the measure introduced in (7). For our two class problem, the measure can be shown easily to simplify to

$$\Delta_{max} = |P(\theta_i = \omega|\mathbf{x}_i, \theta_{i-1}) - P(\theta_i = \omega|\mathbf{x}_{i+1}, ..., \mathbf{x}_{i+4})| \quad (11)$$

The anomaly detection mechanism for this ball event interpretation process also uses quality information, but as it works with an intermediate representation (ball events, player positions), rather than raw pixels, measuring the noise level of the video is inappropriate. The notion of quality is task-dependent and it must be defined for each process. As the ball event detection and classification processes use local contextual information, which to a large degree mitigates the effect of noise, the data quality is satisfactory and the quality assessment module does not report any quality issues. However, here the domain anomaly detection system engages the confidence assessment modules as the ball event classification is ambiguous close to bounce_in and bounce_out boundaries. The ambiguity is the consequence of attempting to extract 3D measurements from 2D projections. These measurement inaccuracies lead to overlapping class conditional measurement distributions, resulting in ambiguous noncontextual decisions. These ambiguities could then cause incongruence between contextual and non-contextual aposteriori class probabilities, as measured in eq. (11). This is avoided by filtering out incongruences associated with ambiguously determined class labels. The decision confidence is determined by applying measure (5) in Section III.B to the noncontextual probabilities in (9). An experimentally determined confidence threshold on the noncontextual posteriors carves out a 30 pixels wide incongruence exclusion zone around the court boundaries. The noncontextual decision-making threshold is determined by training on matches of tennis singles, so that no anomaly is detected in any singles videos. For any ball event in tennis doubles, with quality measurement less than the threshold, no anomaly is flagged either. Only ball events of "good quality" are analysed for incongruence. The details of the interpretation process can be found in [3].

The game evolution module was evaluated on the same set of videos discussed earlier. The training of the system was carried out using tennis singles matches (A03MS, A03WS and J09WS). The three videos were used in rotation as follows: The first video was used to learn the module models. The second video was used to set the confidence thresholds. The third video was used for testing on unseen singles. The experiment was performed three times for different combinations of the videos. For each configuration, the system was then run on two tennis doubles matches (A08WD,U06WD). The anomaly detection results on unseen singles and doubles were averaged over the three configurations. They are shown in Table III where TP denotes true positive detections, FN false negatives and FP false positive detections respectively. Note that the resulting system detected no anomalies in the unseen test tennis singles videos.

TABLE III
OUT-OF-PLAY ANOMALY DETECTION RESULTS.

Test video:	TP	FN	FP
A08WD	7	37	0
U06WD	3	4	0.33

Ideally we would like to detect all the anomalies, i.e. the sum of TP and FN. Unfortunately the decision confidence filter set on tennis singles results in a relatively large number of undetected anomalies, because many anomalous events fall close to the inner tramline of the doubles play area. With more sophisticated video processing techniques, or using a 3D measurement system such as Hawkeye, the decision ambiguity would be reduced considerably and a lower false negative rate achieved. Nevertheless, the system detects a significant number of anomalies which clearly indicate a domain change.

VII. CONCLUSIONS

We addressed the problem of anomaly detection in machine perception. We argued that the conventional notions of anomaly such as outlier or distribution drift alone cannot detect all anomalous events of interests in machine perception where the key objective is to instantiate models to explain observations. The inability to detect anomalies is aggravated by the common use of nongenerative models for decision making, which is motivated by their speed of processing and better classification performance. However, such models lack the inherent capacity to detect anomalous situations.

In order to clarify the anomaly landscape, we introduced the concept of domain anomaly, which refers to the situation when none of the models characterising a domain are able to explain the data. We showed that a number of mechanisms are required to detect a domain anomaly. They include detectors of outliers of noncontextual and contextual measurement distributions, detectors of incongruence of contextual and noncontextual sensor(y) data interpretations, decision confidence estimation and sensor(y) data quality assessment. These gauging mechanisms jointly facilitate not only the detection of domain anomaly, but also its identification. A taxonomy of domain anomalies, which distinguishes between component, configuration, and joint component and configuration domain anomaly events, has been introduced.

We developed a unified framework for domain anomaly detection. The framework draws on the Bayesian probabilistic reasoning apparatus which clearly defines the concepts such as outlier, noise, distribution drift, novelty detection (object, object primitive), rare events, and unexpected events. The proposed methodology has wide applicability and it underpins in a coherent way the anomaly detection applications found in the literature.

The proposed anomaly detection system architecture includes a mechanism for detecting incongruence between the decisions of multiple classifiers, a measurement distribution drift detector, data quality assessment and a decision ambiguity monitor. The outputs from these modules are processed by a reasoning mechanism to identify anomalies and their meaning. Incongruence is gauged by a criterion related to the Bayesian

surprise measure. The architecture is applied to two different interpretation processes within a tennis video annotation system to demonstrate the role of incongruence in domain anomaly detection, and to emphasise the importance of data quality and decision ambiguity assessment in distinguishing genuine anomalies from false positives caused by noise or ambiguous measurements.

According to the anomaly taxonomy introduced in Section V, both applications in Section VI demonstrate *Component Domain Anomaly* detection. In contrast to "novelty" detection studied in [56], the player count anomaly detects "innovation". The out of play application flagged by incongruence detects unexpected component in the context of the tennis game. Both anomaly detectors distinguish and respond appropriately to noise and ambiguity. We plan to demonstrate the detection of *Component and Configuration Domain Anomaly* in the context of transfer learning from tennis to badminton.

ACKNOWLEDGEMENTS

This work was carried out as part of EPSRC projects ACASVA (Adaptive cognition for automated sports video annotation) under contract EP/F069421/1 and Signal processing in a networked battlespace under contract EP/K014307/1. The EPSRC financial support is gratefully acknowledged. One of the authors (JK) would like to thank the DIRAC Project consortium for the discussions that inspired this work.

REFERENCES

[1] D Agarwal. Detecting anomalies in cross-classified streams: a bayesian approach. *Knowledge and Information Systems*, 11:1:29–44, 2006.

[2] M Agyemang, K Barker, and R Alhajj. A comprehensive survey of numeric and symbolic outlier mining techniques. *Intelligent Data Analysis*, 10:6:521–538, 2006.

[3] I. Almajai, J. Kittler, T. de Campos, W. Christmas, F. Yan, D. Windridge, and A. Khan. Ball event recognition using HMM for automatic tennis annotation. In *ICIP*, pages 1509–1512, 2010.

[4] S Ando. Clustering needles in a haystack: An information theoretic analysis of minority and outlier detection. In *Proceedings of 7th International Conference on Data Mining*, pages 13–22, 2007.

[5] J. Anemüller, J.-H. Bach, B. Caputo, M. Havlena, L. Jie, H. Kayser, B. Leibe, P. Motlicek, T. Pajdla, M. Pavel, A. Torii, L. V. Gool, A. Zweig, and H. Hermansky. The DIRAC AWEAR audio-visual platform for detection of unexpected and incongruent events. In *Proc. International Conference on Multimodal Interaction (ICMI)*, pages pp. 289–293, 2008.

[6] F J Anscombe and I Guttman. Rejection of outliers. *Technometrics*, 2:2:123–147, 1960.

[7] V Barnett and T Lewis. *Outliers in statistical data*. John Wiley and sons, 1994.

[8] S Basu and M Meckesheimer. Automatic outlier detection for time series: an application to sensor data. *Knowledge and Information Systems*, 11:2:137–154, 2007.

[9] L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rastrow, C. White, S. Khudanpur, H. Hermansky, and J. Cernocky. Combination of strongly and weakly constrained recognizers for reliable detection of oovs. In *Proceedings 33rd International Conference on Acoustics, Speech, and Signal Processing ICASSP*, pages 4081–4084, 2008.

[10] V Chandola, A Banerjee, and V Kumar. Anomaly detection : A survey. *ACM Computing Surveys*, 41:15:1–15:58, 2009.

[11] V Chatzigiannakis, S Papavassiliou, M Grammatikou, and B Maglaris. Hierarchical anomaly detection in distributed large-scale sensor networks. In *ISCC'06: Proceedings of the 11th IEEE Symposium on Computers and Communications*, pages 761–767. IEEE Computer Society, Washington, DC, USA, 2006.

[12] P A Crook, S Marsland, G Hayes, and U Nehmzow. A tale of two filters: Online novelty detection. In *Proceedings of International Conference on Robotics and Automation*, pages 3894–3899, 2002.

[13] K Das and J Schneider. Detecting anomalous records in categorical datasets. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Press, 2007.

[14] C P Diehl and J B Hampshire. Real-time object classification and novelty detection for collaborative video surveillance. In *Proceedings of IEEE International Joint Conference on Neural Networks*, pages 2620–2625, 2002.

[15] F Y Edgeworth. On discordant observations. *Philosophical Magazine*, 23:5:364–375, 1887.

[16] E Eskin. Anomaly detection over noisy data using learned probability distributions. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 255–262. Morgan Kaufmann Publishers Inc., 2000.

[17] F Esponda, S Forrest, and P Helman. A formal framework for positive and negative detection schemes. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 34:1:357–373, 2004.

[18] W Fan, M Miller, S J Stolfo, W Lee, and P K Chan. Using artificial anomalies to detect unknown and known network intrusions. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 123–130. IEEE Computer Society, 2001.

[19] Z He, X Xu, and S Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24:9-10:1641–1650, 2003.

[20] Z He, X Xu, J Z Huang, and S Deng. Mining class outliers: Concepts, algorithms and application in crm. *Expert Systems and Applications*, 27:4:681–697, 2004.

[21] P Helman and J Bhangoo. A statistically based system for prioritizing information exploration under uncertainty. *IEEE Transactions on Systems, Man, and Cybernetics*, 27:449–466, 1997.

[22] V Hodge and J Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:2:85–126, 2004.

[23] T M Hospedales, J Li, S Gong, and T Xian. Identifying rare and subtle behaviours: A weakly supervised joint topic model. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2011.

[24] P Huber. *Robust Statistics*. Wiley, New York, 1974.

[25] L Itti and P F Baldi. A principled approach to detecting surprising events in video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 631–637, 2005.

[26] N Japkowicz, C Myers, and M A Gluck. A novelty detection approach to classification. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 518–523, 1995.

[27] M V Joshi, R C Agarwal, and V Kumar. Predicting rare classes: can boosting make any weak learner strong? In *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining*, pages 297–306. ACM, New York, NY, USA, 2002.

[28] E Keogh, J Lin, S-H Lee, and H V Herle. Finding the most unusual time series subsequence: algorithms and applications. *Knowledge and Information Systems*, 11:1:1–27, 2006.

[29] H. Ketabdar, M. Hannemann, and H. Hermansky. Detection of out-of-vocabulary words in posterior based asr. In *Proceedings European Conference on Speech Communication and Technology: Interspeech 2007*, pages 1757–1760, 2007.

[30] J. Kittler, W J Christmas, F Yan, I Kolonias, and D Windridge. A memory architecture and contextual reasoning for cognitive vision. In *Proc. SCIA*, pages 343–358, 2005.

[31] E M Knorr, R T Ng, V, and Tucakov. Distance-based outliers: algorithms and applications. *The VLDB Journal*, 8:3-4:237–253, 2000.

[32] S. Kombrink, L. Burget, P. Matejka, M. Karafiat, and H. Hermansky. Posterior-based out of vocabulary word detection in telephone speech. In *Proc. Interspeech 2009*, pages 80–83. Available: http://www.fit.vutbr.cz/research/view pub.php?id=9037, 2009.

[33] J D Lafferty, A McCallum, and F C N Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001.

[34] X Li. Blind image quality assessment. In *Proc. IEEE Int. Conf. Image Processing*, volume 1, pages 449–452, 2002.

[35] M Markou and S Singh. Novelty detection: A review-part 1: Statistical approaches. *Signal Processing*, 83:12:2481–2497, 2003.

[36] M Markou and S Singh. Novelty detection: A review-part 2: Neural network based approaches. *Signal Processing*, 83:12:2499–2521, 2003.

[37] A McNeil. Extreme value theory for risk managers. *Internal Modelling and CAD II*, pages 93–113, 1999.

[38] A Nairac, T Corbett-Clark, R Ripley, N Townsend, and L Tarassenko. Choosing an appropriate model for novelty detection. In *Proceedings of the 5th IEEE International Conference on Artificial Neural Networks*, pages 227–232, 1997.

[39] A Patcha and J-M Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Comput. Networks*, 51:12:3448–3470, 2007.

[40] G Ratsch, S Mika, B Scholkopf, and K-R Muller. Constructing boosting algorithms from svms: An application to one-class classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:9:1184–1199, 2002.

[41] S Roberts. Novelty detection using extreme value statistics. *Proceedings of IEE - Vision, Image and Signal processing*, 146:124–129, 1999.

[42] S Roberts and L Tarassenko. A probabilistic resource allocating network for novelty detection. *Neural Computing*, 6:2:270–284, 1994.

[43] V Roth. Kernel fisher discriminants for outlier detection. *Neural Computation*, 18:4:942–960, 2006.

[44] P J Rousseeuw and A M Leroy. *Robust regression and outlier detection*. John Wiley & Sons, Inc., 1987.

[45] V Saligrama, J Konrad, and P-M Jodoin. Video anomaly identification. *IEEE Signal Processing Magazine*, 27:18–33, 2010.

[46] B Schalkopf, J C Platt, J C Shawe-Taylor, A J Smola, and R C Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:7:1443–1471, 2001.

[47] X Song, M Wu, C Jermaine, and S Ranka. Conditional anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 19: 5:631–645, 2007.

[48] A Soule, K Salamatian, and N Taft. Combining filtering and statistical methods for anomaly detection. In *IMC '05: Proceedings of the 5th ACM SIGCOMM conference on Internet measurement*, pages 1–14. ACM, New York, NY, USA, 2005.

[49] C Stefano, C Sansone, and M Vento. To reject or not to reject: that is the question - an answer in case of neural classifiers. *IEEE Transactions on Systems, Man and Cybernetics Part C*, 30:1:84–94, 2000.

[50] I Steinwart, D Hush, and C Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6:211–232, 2005.

[51] P Sun and S Chawla. Slom: a new measure for local spatial outliers. *Knowledge and Information Systems*, 9:4:412–429, 2006.

[52] D Tax and R Duin. Data domain description using support vectors. In M Verleysen, editor, *Proceedings of the European Symposium on Artficial Neural Networks*, pages 251–256, 1999.

[53] D Tax and R Duin. Support vector data description. *Pattern Recognition Letters*, 20:11-13:1191–1199, 1999.

[54] D M J Tax. *One-class classification; concept-learning in the absence of counter-examples*. Ph.D. thesis, Delft University of Technology, 2001.

[55] G C Vasconcelos, M C Fairhurst, and D L Bisset. Investigating feedforward neural networks with respect to the rejection of spurious patterns. *Pattern Recognition Letters*, 16:2:207–212, 1995.

[56] D Weinshall, J Anemuller, and L van Gool. *Detection and identification of rare audiovisual cues*. Springer, 2012.

[57] D Weinshall, H Hermansky, A Zweig, J Luo, H Jimison, F Ohl, and M Pavel. Beyond novelty detection: Incongruent events, when general and specific classifiers disagree. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1–8, Dec 2009.

[58] D Weinshall, A Zweig, H Hermansky, S Kombrink, F W Ohl, J Anemeuller, J-H Bach, L Van Gool, F Nater, T Pajdla, M Havlena, and M Pavel. Beyond novelty detection: Incongruent events, when general and specific classifiers disagree. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34:1886–1901, 2012.

[59] F Yan, W Christmas, and J Kittler. Layered data association using graph-theoretic formulation with application to tennis ball tracking in monocular sequences. *Transactions on Pattern Analysis and Machine Intelligence*, 30:10:1814–1830, 2008.

[60] N Ye. A markov chain model of temporal behavior for anomaly detection. In *Proceedings of the 5th Annual IEEE Information Assurance Workshop*, 2004.

[61] J Zhang and H Wang. Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance. *Knowledge and Information Systems*, 10:3:333–355, 2006.

[62] K Zhang, S F Shi, H Gao, and J Li. Unsupervised outlier detection in sensor networks using aggregation tree. In *International Conference on Advanced Data Mining and Applications*, pages 158–169, 2007.

[63] K. Zimmermann, T. Svoboda, and J. Matas. Adaptive parameter optimization for real-time tracking. In *Workshop on Non-rigid Registration and Tracking through Learning (in proc. ICCV)*, 2007.

[64] A Zweig, D Eshar, and D Weinshall. Identification of novel classes in object class recognition. In *Detection and identification of rare audiovisual cues*. Springer, 2012.

[65] A. Zweig and D. Weinshall. Exploiting object hierarchy: Combining models from different category levels. In *IEEE 11th International Conference on Computer Vision (ICCV 2007)*, 2007.

Josef Kittler is Professor of Machine Intelligence at the Centre for Vision, Speech and Signal Processing at the University of Surrey, U.K. He conducts research with a focus on biometrics, video and image database retrieval, automatic inspection, medical data analysis, and cognitive vision. He published a Prentice-Hall textbook on Pattern Recognition: A Statistical Approach and more than 170 journal papers. He serves as Series Editor for Springer Verlag Lecture Notes in Computer Science.

Bill Christmas received his BA degree from the University of Oxford and the PhD degree from the University of Surrey. After graduating, he held posts with BBC and BP Research International, working on hardware and software aspects of parallel processing, real-time image processing, and computer vision. He is currently a University Fellow in Technology Transfer at the University of Surrey. His research interests include the integration of machine vision algorithms to create complete applications.

Teófilo de Campos is a research fellow at the University of Surrey since July 2009. He finished DPhil at the University of Oxford in 2006, and his award winning MSc thesis at the University of São Paulo in 2001. He has worked at the research laboratories of Sharp, Microsoft and Xerox. He is currently a site manager for the PASCAL project (EU). His research interests include object recognition, tracking, event detection in video, and transfer learning for action classification.

David Windridge (BSc (Hons) 1993, MSc 1995, PhD 1999) has research interests in Multiple Classifiers Systems, Pattern Recognition and Cognitive Systems. He has played a leading role in a range of cognitive systems projects (including EPSRC ACASVA and EU FP7 DIPLECS). He has authored more than 50 peer-reviewed publications.

Fei Yan Dr. Fei Yan is a research fellow in the Centre for Vision, Speech and Signal Processing (CVSSP) at the University of Surrey. His research interests include machine learning and computer vision, in particular, kernel methods, distance metric learning, object recognition, and object tracking. He has publications in major machine learning and computer vision conferences and journals, including ICDM, ECML, CVPR, CVIU, PR, PAMI, JMLR.

John Illingworth is Professor of Machine Vision at the University of Surrey. He has been actively researching image processing, computer vision and pattern recognition since the 1980s. His interests cover segmentation, shape analysis, 3D data and vision systems. He has published widely and is a Fellow of the International Association of Pattern Recognition and a Distinguished Fellow of the British Machine Vision Association.

Magda Osman is a Lecturer in Experimental Cognitive Psychology in Queen Mary University of London. She has published in a wide range of journals on topics ranging from problem solving in abstract, social, and applied settings; motor learning; conscious and unconscious reasoning and decision making; and the cognitive basis of deception. .