Beyond Novelty Detection: Incongruent Events, When General and Specific Classifiers Disagree

Daphna Weinshall, Alon Zweig, Hynek Hermansky, *Fellow*, *IEEE*,
Stefan Kombrink, *Student Member*, *IEEE*, Frank W. Ohl, Jörn Anemüller, *Member*, *IEEE*,
Jörg-Hendrik Bach, *Member*, *IEEE*, Luc Van Gool, *Member*, *IEEE*,
Fabian Nater, *Student Member*, *IEEE*, Tomas Pajdla, *Member*, *IEEE*,
Michal Havlena, *Member*, *IEEE*, and Misha Pavel, *Senior Member*, *IEEE*

Abstract—Unexpected stimuli are a challenge to any machine learning algorithm. Here, we identify distinct types of unexpected events when *general-level* and *specific-level* classifiers give conflicting predictions. We define a formal framework for the representation and processing of incongruent events: Starting from the notion of label hierarchy, we show how partial order on labels can be deduced from such hierarchies. For each event, we compute its probability in different ways, based on adjacent levels in the label hierarchy. An incongruent event is an event where the probability computed based on some more specific level is much smaller than the probability computed based on some more general level, leading to conflicting predictions. Algorithms are derived to detect incongruent events from different types of hierarchies, different applications, and a variety of data types. We present promising results for the detection of novel visual and audio objects, and new patterns of motion in video. We also discuss the detection of Out-Of-Vocabulary words in speech recognition, and the detection of incongruent events in a multimodal audiovisual scenario.

Index Terms—Novelty detection, categorization, object recognition, out-of-vocabulary words.

1 Introduction

Typically, machine learning algorithms build models of the world using training data from the application domain and prior knowledge about the problem. The models are later applied to data in order to estimate the current state of the world. An implied assumption is that the future is stochastically similar to the past. The approach fails when the system encounters situations that are not anticipated from the past experience. In contrast, successful natural organisms quite readily identify new unanticipated stimuli and situations, and frequently generate an appropriate response. How this can be done is one of the questions motivating the current work.

By definition, an unexpected event is one whose probability of confronting the system is low, based on the data that has been observed previously. In line with this observation, much of the computational work on novelty detection focused on the probabilistic modeling of known classes, identifying outliers of these distributions as novel events (see, e.g., [1], [2] for recent reviews).

To advance beyond the detection of outliers, we observe that there are many different reasons why some stimuli could appear novel. In Section 2, we focus on those unexpected events, which are defined by the incongruence between a prediction induced by prior experience (training data) and the evidence provided by the sensory data. To identify an item as incongruent, we use two parallel classifiers. One of them is strongly constrained by specific knowledge (either prior knowledge or data-derived during training), the other classifier is more general and less constrained. Both classifiers are assumed to yield class-posterior probabilities in response to a particular input signal. A sufficiently large discrepancy between posterior

- *D. Weinshall and A. Zweig are with the School of Computer Science and Engineering, Hebrew University of Jerusalem, Jerusalem 91904, Israel. E-mail: daphna@cs.huji.ac.il, alon.zweig@mail.huji.ac.il.*
- *H. Hermansky is with the Center for Language and Speech Processing, The Johns Hopkins University, 3400 N. Charles Street, Hackerman Hall, Baltimore, MD 21218. E-mail: hynek@jhu.edu.*
- *S. Kombrink is with the Department of Computer Graphics and Multimedia (DCGM), Faculty of Information Technology, BUT, Božetěchova 2, Brno 612 66, Czech Republic. E-mail: kombrink@fit.vutbr.cz.*
- *F.W. Ohl is with the Department of Systems Physiology of Learning, Leibniz Institute for Neurobiology (LIN), Brenneckestr. 6, Magdeburg D-39118, Germany. E-mail: frank.ohl@lin-magdeburg.de.*
- *J. Anemüller and J.-H. Bach are with the Department of Physics, Carl von Ossietzky University Oldenburg, Carl-von-Ossietzky-Str. 9-11, Oldenburg 26111, Germany. E-mail: {joern.anemueller, j.bach}@uni-oldenburg.de.*
- *L. Van Gool and F. Nater are with the Computer Vision Laboratory, ETH Zentrum, Sternwartstrasse 7, Zürich CH-8092, Switzerland. E-mail: {vangool, fnater}@vision.ee.ethz.ch.*
- *T. Pajdla and M. Havlena are with the Center for Machine Perception, Department of Cybernetics, FEE CTU in Prague, Technicka 2, 166 27 Praha 6, Czech Republic. E-mail: {pajdla, havlem1}@cmp.felk.cvut.cz.*
- *M. Pavel is with the Department of Biomedical Engineering and Department of Computer Science and Electrical Engineering, Oregon Health and Science University, CHH-13B, 3303 SW Bond Ave, Portland, OR 97239. E-mail: pavel@bme.ogi.edu.*

Manuscript received 7 Feb. 2010; revised 11 Dec. 2011; accepted 15 Dec. 2011; published online 30 Dec. 2011.
Recommended for acceptance by M. Pelillo.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-02-0079.
Digital Object Identifier no. 10.1109/TPAMI.2011.279.

probabilities induced by input data in the two classifiers is taken as evidence that an item is incongruent.

Thus, in comparison with most existing work on novelty detection, one new and important characteristic of our approach is that we look for a level of description where the novel event is sufficiently probable. Rather than simply responding to an event which is rejected by all classifiers, which often requires no special attention (as in pure noise), we construct and exploit a hierarchy of representations. We attend to those events which are recognized (or accepted) at some more abstract levels of description in the hierarchy while being rejected by the classifiers at the more specific (concrete) levels.

Our approach to the detection of incongruent novel events is general and can be applied to many engineering and biological domains. However, depending on the application—the properties of the hierarchy and the type of data—a different algorithm will be called for to implement the approach.

In Section 3, we assume a disjunctive tree-like hierarchy, where each class is linked to a number of more specific subclasses (as in human categorization), and develop algorithms which are designed to detect unexpected audio and visual novel objects given a set of related known objects. We train the more general, less constrained classifier using a larger more diverse set of stimuli (e.g., the facial images of many individuals). The constrained (specific) classifier is built from a family of classifiers, each trained with positive examples from a single object (e.g., the set of one individual's facial images). An incongruous item (e.g., a new individual) could then be identified by some significant discrepancy between the low confidence of the specific classifier and the high confidence of the general classifier. To conclude the treatment of tree-like hierarchies, in Section 4 we discuss a conjunctive hierarchy, where different modalities (visual and audio) provide different part descriptions of the target object.

In Section 5, we discuss general hierarchies. In one case, the hierarchy is less diverse—a simple chain of inclusion relations, where the motivating application is the detection of out-of-vocabulary lexical items in speech. In this case, the more general classifier is engineered to identify a more generic pattern—speech sounds unconstrained by language—while the more constrained classifier is trained to classify a specific pattern—using language-dependent models of phonemes. An incongruent object is detected when some noticeable discrepancy exists between the two classifiers. In the second case discussed in this section, the hierarchy includes both disjunctive and conjunctive nodes, and the motivating application is the detection of new patterns of motion in video.

One motivation for our work is the way biological systems are more adept at handling unexpected events. In Section 6, we return to this point, describing experiments with gerbils that investigate knowledge transfer from known stimuli to a novel modality in a biological system, where the transfer occurs between sensory modalities.

Prior work. Often, novelty is detected based on generative models of known objects, when new data are rejected by all these models. Previous work may estimate a spherically shaped boundary around a single class data set [3], learn a hyperplane which separates the class data set from the rest of the feature space (one class SVM) [4], or utilize the nonparametric Parzen-window density estimation approach [5]. A few methods use a multiclass discriminative approach, as, for example, [6] for the detection of novel objects in videos and [7] for the specific task of face verification. To our knowledge, all novelty detection approaches which do not rely on samples of outliers or otherwise model the outliers distribution detect novelty by rejecting normality (i.e., novelty is detected when all classifiers of known objects fail to accept a new sample). There are many studies on novelty detection in biological systems [8], often focusing on regions of the hippocampus [9].

We assume in this work that the hierarchical relations between categories is given, at least to some extent. There has been some recent interest in the learning of object hierarchies [10], [11], and these algorithms can be effectively integrated into our approach. In addition, the notion of hierarchical organization of object classes has been acknowledged and used in several recent visual object class recognition papers, such as [12], [13], [14]. Otherwise, a number of different methods have been developed to detect and recognize object classes, trained and tested on a wide range of publicly available data sets (see, e.g., [15], [16], [17]). These algorithms are trained to recognize images of objects from a known class.

2 Incongruent Events

We now define the concept of incongruent events as induced by a classification process that can, in general, correspond to partial order on sets of event classes. This partial order, represented by a directed graph (DAG) that captures subset-superset relations among sets (as in algebra of sets), can also be interpreted in terms of two category-forming operations: conjunctive and disjunctive hierarchies, as described in Section 2.1. We provide a unified definition in Section 2.2, and analyze its implication for the two cases.

2.1 Label Hierarchy and Partial Order

The set of labels (or concepts) represents the knowledge base about the stimuli domain, which is either given (by a teacher) or learned. In cognitive systems, such knowledge is hardly ever a set; often, in fact, labels are given (or can be thought of) as a hierarchy.

In general, a hierarchy can be represented by a directed graph where each label (a set of objects) corresponds to a single node in the graph. A directed edge exists from label (specific child concept) a to (general parent concept) b iff a corresponds to a smaller set of events or objects in the world which is contained in the set of events or objects corresponding to label b, i.e., $a \subset b$. In this way, the edges represent a partial order defined over the set of labels or concepts.

Because the graph is directed, it defines for each concept a two distinct sets of concepts (parent-child) related to it: *disjunctive concepts*, which are smaller (subsets) according to the partial order, i.e., they are linked to node a by incoming edges converging on a; and *conjunctive concepts*, which are larger (supersets) according to the partial order, i.e., they are linked to node a by outgoing edges diverging from a. If the DAG of partial order is a tree, only one of these sets is nontrivial (larger than 1).

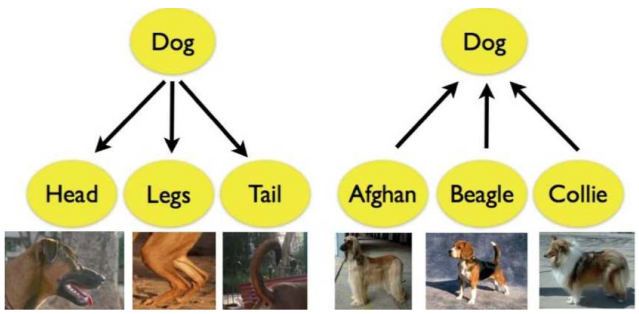


Fig. 1. Examples. Left: *Conjunctive hierarchy*, the concept of a dog requires the conjunction of parts, including head, legs, and tail. Right: *Disjunctive hierarchy*, the concept of a dog is defined as the disjunction of more specific concepts, including afghan, beagle, and collie.

We consider two possible tree-like hierarchies, which correspond to two interesting intuitive cases:

Conjunctive hierarchy. Modeling part membership, as in biological taxonomy or speech. For example, eyes, ears, and nose combine to form a head; head, legs, and tail combine to form a dog (see left panel of Fig. 1); and sequences of phonemes constitute words and utterances. In this case, each node has a single child and possibly many parents.

Disjunctive hierarchy. Modeling class membership, as in human categorization—where objects can be classified at different levels of generality, from subordinate categories (most specific level), to basic level (intermediate level), to superordinate categories (most general level). For example, a beagle (subordinate category) is also a dog (basic level category), and it is also an animal (superordinate category); see the right panel of Fig. 1. In this case, each node has a single parent and possibly many children.

The sets of *disjunctive* and *conjunctive* concepts induce constraints on the observed features in different ways. Accordingly, the set of objects corresponding to a given label (node) is *included* in the *intersection* of the objects in its set of *conjunctive concepts*. Thus, in the example shown in the left panel of Fig. 1, the concept of *Dog* requires the conjunction of parts as in $DOG \subseteq LEGS \cap HEAD \cap TAIL$. To the contrary, the set of objects corresponding to a given label (node) *contains* the *union* of objects in its set of *disjunctive concepts*. In the example shown in the right panel of Fig. 1, the class of dogs requires the disjunction of the individual members as in $DOG \supseteq AFGHAN \cup BEAGLE \cup COLLIE$.

2.2 Definition of Incongruent Events

2.2.1 Multiple Probabilistic Models for Each Concept

For each node a, define $A^s = \{b \in G, b \preceq a\}$—the set of *disjunctive concepts*, corresponding to all nodes more specific (smaller) than a in accordance with the given partial order. Similarly, define $A^g = \{b \in G, a \preceq b\}$—the set of *conjunctive concepts*, corresponding to all nodes more general (larger) than a in accordance with the given partial order.

For each node a and training data $\mathcal{T}$, we hypothesize three probabilistic models which are derived from $\mathcal{T}$ in different ways in order to determine whether a new data point X can be described by concept a:

- $Q_a(X)$: A probabilistic model of class a, derived from training data $\mathcal{T}$ unconstrained by the partial order relations in the graph.
- $Q_a^s(X)$: A probabilistic model of class a which is based on the probability of concepts in A^s, assuming their independence from each other. Typically, the model incorporates a simple *disjunctive* relation between concepts in A^s.
- $Q_a^g(X)$: A probabilistic model of class a which is based on the probability of concepts in A^g, assuming their independence from each other. Here, the model typically incorporates a simple *conjunctive* relation between concepts in A^g.

2.2.2 Examples

To illustrate, consider again the simple examples shown in Fig. 1, where our concept of interest a is *Dog*.

In the Conjunctive hierarchy (left panel), $|A^g| = 3$ (Head, Legs, Tail) while $|A^s| = 1$. We derive two different models for the class *Dog*:

1. Q_{Dog}—Obtained using training pictures of *dogs* and *not dogs* without body part labels.
2. Q_{Dog}^g—Obtained using the outcome of models for Head, Legs, and Tail which have been derived from the same training set $\mathcal{T}$ with body part labels only. If we further assume that concept a is the conjunction of its part member concepts as defined above and assuming that these part concepts are independent of each other, we get

$$Q_{\text{Dog}}^g = \prod_{b \in A^g} Q_b = Q_{\text{Head}} \cdot Q_{\text{Legs}} \cdot Q_{\text{Tail}}. \tag{1}$$

In the disjunctive hierarchy (right panel), $|A^s| = 3$ (afghan, beagle, collie) while $|A^g| = 1$. We therefore derive two models for the class *Dog*:

1. Q_{Dog}—Obtained using training pictures of *dogs* and *not dogs* without breed labels.
2. Q_{Dog}^s—Obtained using the outcome of models for afghan, beagle, and collie which have been derived from the same training set $\mathcal{T}$ with dog breed labels only. If we further assume that class a is the disjunction of its subclasses as defined above and once again assume that these subclasses are independent of each other, we get

$$Q_{\text{Dog}}^s = \sum_{b \in A^s} Q_b = Q_{\text{Afghan}} + Q_{\text{Beagle}} + Q_{\text{Collie}}.$$

2.2.3 Incongruent Events

In general, we expect the different models to provide roughly the same probabilistic estimate for the presence of concept a in data X. A mismatch between the predictions of the different models may indicate that something new and interesting is being observed, unpredicted by the existing knowledge of the system. In particular, we are interested in the following discrepancy:

Definition. *Observation X is incongruent if there exists a concept "a" such that*

$$Q_a^g(X) \gg Q_a(X) \quad \text{or} \quad Q_a(X) \gg Q_a^s(X). \tag{2}$$

In other words, observation X is *incongruent* if a discrepancy exists between the inference of two classifiers, where the more general classifier is much more confident in the existence of the object than the more specific classifier.

Classifiers come in different forms: They may accept or reject, they may generate a (possibly probabilistic) hypothesis, or they may choose an action. For binary classifiers that either accept or reject, the definition above implies one of two mutually exclusive cases: Either the classifier based on the more general descriptions from level g accepts X while the direct classier rejects it, or the direct classifier accepts X while the classifier based on the more specific descriptions from level s rejects it. In either case, the concept receives high probability at some more general level (according to the partial order), but much lower probability when relying only on some more specific level.

2.2.4 Discussion: Why This Definition?

We first note one underlying assumption—that all the models and derived classifiers are veridical and able to capture the state of affairs as seen in the training data. Discrepancies between classifiers may occur due to errors when one of the three classifiers simply fails to recognize an object where it exists or recognizes the object where it does not. We do not provide theoretical analysis of these cases, but rather provide empirical evidence that our method is robust to such errors, which are present in our experiments with real data.

Another underlying assumption is that the assumed hierarchy is correct. Once again, discrepancies may occur if we got the hierarchy wrong when building the three classifiers. Specifically, discrepancies where

$$Q^g_a(X) \ll Q_a(X) \text{ or } Q_a(X) \ll Q^s_a(X) \tag{3}$$

are not considered here as they correspond to some logical contradiction with the partial order, implying errors in the models, the partial order, or both.

If discrepancies occur while our assumptions hold—the models are correct and the hierarchy is veridical—then we may conclude that something new is being observed, not present in the training data. These cases are captured by our definition.

To illustrate, consider the following examples where our definition seems to capture interesting "surprises":

1. In the left panel of Fig. 1, we have

$$Q^g_{\text{Dog}} = Q_{\text{Head}} \cdot Q_{\text{Legs}} \cdot Q_{\text{Tail}} \gg Q_{\text{Dog}}.$$

In other words, while the probability of each part is high (since the multiplication of those probabilities is high), the more specific *Dog* classifier is rather uncertain about the existence of a dog in this data.

How can this happen? Maybe the parts are configured in an unusual arrangement for the object as seen in the training data (e.g., my 3-legged cat), or we may have encountered an unusual part combination (e.g., the donkey with a cat's tail from *Shrek 3*). Those are two examples for the kind of unexpected events we are interested in. But since we assume that our dog classifier is correct and therefore there is no dog in the image, we must conclude that we are seeing something new, possibly as trivial as an occluded dog never seen before in the training data. More interestingly, we may be seeing something whose parts are familiar from other objects, but the whole is new.

2. In the right panel of Fig. 1, we have

$$Q^s_{\text{Dog}} = Q_{\text{Afghan}} + Q_{\text{Beagle}} + Q_{\text{Collie}} \ll Q_{\text{Dog}}.$$

In other words, while the probability of each subclass is low (since the sum of these probabilities is low), the generic *Dog* classifier is certain about the existence of a dog in this data.

How may such a discrepancy arise? Maybe we are seeing a new breed of dog that we haven't seen before—a pointer. The dog model, which is assumed to correctly capture the notion of *dogness*, should be able to identify this new object, while models of previously seen dog breeds (afghan, beagle, and collie) correctly fail to recognize the new object.

2.3 How to Use the Proposed framework

Our framework presented above is rather general and in principle can be applied to any partial order. In order to use it for a given application, one needs first to establish the relevant hierarchical relations between objects or events. Given the type of existing relations, one may proceed as follows:

Disjunctive hierarchies. A concept is defined by its own labeled examples, as well as the disjunction of subconcepts each defined by more specific labels. To implement our framework, we need at least two classifiers, one more general than the other. We may define the two classifiers as follows (see Section 3): First is the vanilla classifier $Q_{concept}$, which is trained using the concept labels as positive examples. Second is a more specific classifier $Q^s_{concept}$, which is trained using the subconcept labels as positive examples (in our implementation in Section 3, this second classifier is trained discriminatively). When $Q_{concept}$ accepts a new example and $Q^s_{concept}$ does not, we identify the example as incongruent.

Conjunctive hierarchies. A concept is defined by its own labeled examples, as well as the conjunction of superconcepts, each defined by more general labels. We may define the two classifiers as follows (see Section 4): $Q_{concept}$ is defined and trained as above. We then train a classifier for each superconcept and define $Q^g_{concept}$ as the conjunction of these classifiers. As before, when $Q^g_{concept}$ accepts a new example and $Q_{concept}$ does not, we identify the example as incongruent.

General hierarchies. When a concept has multiple views, both as the disjunction of subconcepts and the conjunction of superconcepts, we may combine the two schemes as illustrated in Section 5.2.

3 Disjunctive Hierarchies

We now adopt the framework described above to the problem of novel class detection when given a *Disjunctive Hierarchy*. We assume a rich hierarchy, with nontrivial (i.e., of size larger than 1) sets of *disjunctive concepts*; see the right panel of Fig. 1. This assumption enables the use of discriminative classifiers.

As discussed in Section 2.2 and specifically in the second example there, in a *disjunctive hierarchy* we have two classifiers for each label or concept: the more general classifier

$Q_{concept}$ and the specific disjunctive classifier $Q^s_{concept}$. The assumed classification scenario is multiclass, where several classes are already known.

Next, we describe two closely related algorithms for two applications: an algorithm to detect a new visual object in Section 3.1 and a similar algorithm to detect a new auditory object in Section 3.2.

3.1 Novel Subclasses of Visual Objects

In order to identify novel classes, our algorithm detects a discrepancy between $Q_{concept}$ and $Q^s_{concept}$. The classifier $Q_{concept}$ is trained in the usual way using all the examples of the object, while the specific classifier $Q^s_{concept}$ is trained to discriminatively distinguish between the concepts in the set of *disjunctive concepts* of the object. Our approach is general in the sense that it does not depend on the specifics of the underlying object class recognition algorithm.

We tested the algorithm experimentally on the set of motorbike classes from the Caltech256 benchmark data set. We found that discriminative methods which capture distinctions between the related known subclasses perform significantly better than generative methods. We demonstrate in our experiments the importance of modeling the hierarchical relations tightly. Finally, we compare the performance of the proposed approach to results obtained from novelty detection based on one-class SVM outlier detection.

3.1.1 Algorithm for Novel Class Detection

Algorithm 1 is formally described below.

Algorithm 1. Unknown Class Identification

```
Input:
  x            test image
  $C^g$        general level classifier
  $C_j$        specific level classifiers, $j = 1..|$known sub-classes$|$
  $V^c_{C_i}$  average certainty of train or validation examples
               classified correctly as $C_i$
  $V^w_{C_i}$  average certainty of train or validation examples
               classified wrongly as $C_i$ (zero if there are none)

1) Classify $\mathbf{x}$ using $C^g$
2) **if** accept
     Classify $\mathbf{x}$ using all $C_j$ classifiers and obtain a set of
     certainty values $V_{C_j}(\mathbf{x})$
     Let $i = \arg\max_j V_{C_j}(\mathbf{x})$
     Define $S(\mathbf{x}) = (V_{C_i}(\mathbf{x}) - V^w_{C_i})/(V^c_{C_i} - V^w_{C_i})$
     a) **if** $S(\mathbf{x}) > 0.5$
          label $\mathbf{x}$ as belonging to a known class
     b) **else** label $\mathbf{x}$ as belonging to a novel (unknown)
          class
3) **else** label $\mathbf{x}$ as a background image
```

Basic Object Class Classifiers. To verify the generality of our approach, we tested it using two different embedded object class representation methods. For conciseness, we only describe results with method [15]; the results with method [16] are comparable but slightly inferior, presumably due to the generative nature of the method and the fact that it does not use negative examples when training classifiers.

The object recognition algorithm of [15] learns a generative relational part-based object model, modeling appearance, location, and scale. Location and scale are described relative to some object location and scale, as captured by a star-like Bayesian network. The model's parameters are discriminatively optimized (given negative examples during the training phase) using an extended boosting process. Based on this model and some simplifying assumptions, the likelihood ratio test function is approximated (using the MAP interpretation of the model) by

$$F(\mathbf{x}) = \max_{C} \sum_{k=1}^{P} \max_{u \in Q(\mathbf{x})} \log\ p(u|C, \theta^k) - \nu, \tag{4}$$

with P parts, threshold ν, C denoting the object's location and scale, and $Q(\mathbf{x})$ the set of extracted image features.

General Category Level Classifier. In order to learn the general classifier $Q_{concept}$, we consider all the examples from the given subclasses as the positive set of training examples. For negative examples, we use either clutter or different unrelated objects (none of which is from the known siblings). As we shall see in Section 3.1.3, this general classifier demonstrates high acceptance rates when tested on the novel subclasses.

Specific Category Level Classifier. The problem of learning the specific classifier $Q^s_{concept}$ is reduced to the standard novelty detection task of deciding whether a new sample belongs to any of the known classes or not. However, the situation is somewhat unique and we take advantage of this: While there are multiple known classes, their number is bounded by the degree of the graph of partial orders (they must all be subclasses of a single abstract object). This suggests that a discriminative approach could be effective.

The algorithm is formally described in the next section.

3.1.2 Algorithm for Subclass Detection

The discriminative training procedure of the specific level classifier is summarized in Algorithm 2, with details subsequently provided.

Algorithm 2. Train Known versus Unknown Specific Class Classier

```
1) For each specific class, build a discriminative classifier
   with:
   *positive examples:* all images from the specific class
   *negative examples:* images from all sibling classes

2) Compute the Normalized Certainty function, and
   choose classification threshold for novel classes.
3) Accept iff the normalized certainty is larger than the
   fixed threshold.
```

Step 1, Discriminative Multiclass Classification. The specific level object model learned for each known class is optimized to separate the class from its siblings. A new sample $\mathbf{x}$ is classified according to the most likely classification (max decision) C_i. The output of the learned classifier (4) provides an estimate for the measure of classification certainty $V_{C_i}(\mathbf{x})$.

Step 2, Normalized Certainty Score. Given $V_{C_i}(\mathbf{x})$, we seek a more sensitive measure of certainty as to whether or not the classified examples belongs to the group of known

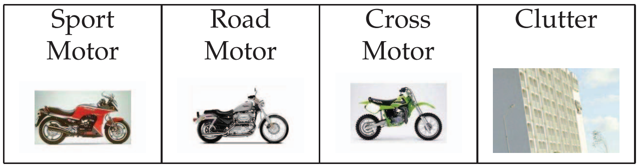


Fig. 2. Examples from the object classes and clutter images used to train and test the different Category level models of the "Motorbikes" hierarchy. The more specific offspring levels are: "Sport-Motorbikes," "Road-Motorbikes," and "Cross-Motorbikes." These images are taken from the Caltech-256 [19] data set. Clutter images are used as negative examples.

subclasses. To this end, we define a normalized score function which normalizes the certainty estimate $V_{C_i}(\mathbf{x})$ relative to the certainty estimates of correct classification and wrong classification for the specific-class classifier, as measured during training or validation.

Specifically, let $V^c_{C_i}$ denote the average certainty of train or validation examples classified correctly as C_i, and let $V^w_{C_i}$ denote the average certainty of train or validation examples from all other subclasses classified wrongly as belonging to class C_i. The normalized score $S(\mathbf{x})$ of $\mathbf{x}$ is calculated as follows:

$$S(\mathbf{x}) = \frac{(V_{C_i}(\mathbf{x}) - V^w_{C_i})}{(V^c_{C_i} - V^w_{C_i})}. \tag{5}$$

If the classes can be well separated during training, that is, $V^c_{C_i} \gg V^w_{C_i}$, and both groups have low variance, the normalized score provides a reliable certainty measure for the multiclass classification.

Step 3, Choosing a threshold. Unlike the typical learning scenario, where positive (and sometimes negative) examples are given during training, in the case of novelty detection no actual positive examples are known during training (since, by definition, novel objects have never been observed before). Thus, it becomes advantageous to set more conservative limits on the learned classifiers, more so than indicated by the training set. Specifically, we set the threshold of the normalized certainty measure, which lies in the range [0..1] to 0.5.

3.1.3 Experiments

Data Sets. In the current set of experiments, we used images from a subset of classes extracted from the Caltech256 data set and corresponding to some crude notion of natural hierarchy.

Specifically, in the chosen hierarchy, the general parent category level is "Motorbikes"; see Fig. 2. Twenty-two object classes, taken from [19], were added in order to serve together with the original data set as a joint pool of object classes from which the unseen-objects are sampled.

Method. All experiments were repeated at least 25 times with different random sampling of test and train examples. We used 39 images for the training of each specific level class. Three conditions were simulated, leaving each of the classes out as the unknown (novel) class.

Basic Results. Fig. 3 shows classification rates for the different types of test examples: *Known*—new examples from all known classes during the training phase; *Unknown*—examples from the unknown (novel) class which belong to the same General level as the Known classes but have been left out during training; *Background*—examples not belonging to the general level which were used as negative examples during the General level classifier training phase; and *Unseen*—examples of objects from classes not seen during the training phase, neither as positive nor as negative examples. The three possible types of classification are: *Known*—examples classified as belonging to one of the known classes; *Unknown*—examples classified as belonging to the unknown class; and *Background*—examples rejected by the General level classifier.

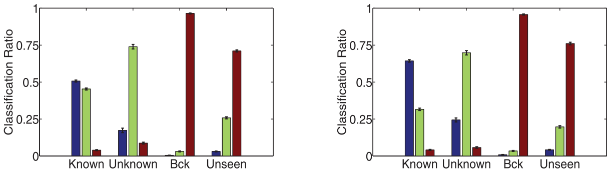


Fig. 3. Classification ratios for four groups of labels: Known Classes, Unknown Class, Background, and unseen classes. Bars corresponding to the three possible classification rates are shown: The left bar shows the known classification rate, the middle bar shows the unknown classification rate, and the right bar shows the background classification rate (rejection by the general level classifier). Left panel—Cross Motorbikes are left out as the unknown class; right panel—Sport Motorbikes are left out as the unknown class (similar results were obtained with the Road Motorbikes class left out).

The results in Fig. 3 show the desired effects: Each set of examples—Known, Unknown, and Background—has the highest rate of correct classification in its own category. As desired, we also see similar recognition rates (or high acceptance rates) of the Known and Unknown classes by the general level classifier, indicating that both are regarded as similarly belonging to the same general level. Finally, examples from the Unseen set are rejected correctly by the general level classifier.

Discriminative specific classifiers improve performance. We checked the importance of using a discriminative approach by comparing our approach for building discriminative specific-level classifiers to nondiscriminative approaches. Note that the general level classifier remains the same throughout.

We varied the amount of discriminative information used when building the specific level classifiers, by choosing different sets of examples as the negative training set: 1) *1vsSiblings—Exploiting knowledge of sibling relations*, the most discriminative variant, where all train examples of the known sibling classes are used as the negative set when training each specific known class classifier. 2) *1vsBck—No knowledge of siblings relations*, a less discriminative variant, where the negative set of examples is similar to the one used when training the general level classifier.

Results are given in Fig. 4, showing that discriminative training with the sibling classes as negative examples significantly enhances performance.

Novel class detector is specific. To test the validity of our novel class detection algorithm, we checked two potential false misclassification as novel subclass when given either low-quality images or totally unrelated novel classes (*unseen* in Fig. 3). For the second case, Fig. 3 shows that by far the most unseen examples are correctly rejected by the general level classifier.

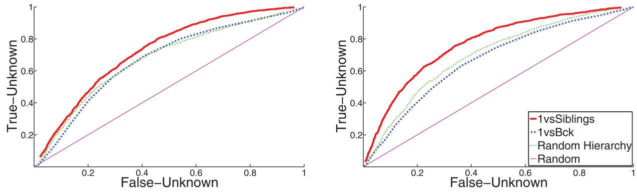


Fig. 4. ROC curves showing True-Unknown classification rate on the vertical axis versus False-Unknown Classification rate on the horizontal axis. We only plot examples accepted by the General level classifier. *1vsSiblings* denotes the most discriminative training protocol, where specific class object models are learned using the known siblings as the negative set. *1vsBck* denotes the less discriminative training protocol where the set of negative examples is the same as in the training of the General level classifier. Random Hierarchy denotes the case where the hierarchy was built randomly, as described in the text. Left panel—Cross Motorbikes are left out as the unknown class; right panel—Sport Motorbikes are left out as the unknown class (similar results were obtained with the Road Motorbikes class left out).

To test the recognition of low-quality images, we took images of objects from known classes and added increasing amounts of Gaussian white noise to the images. With this manipulation, background images continued to be correctly rejected by the general level classifier as in Fig. 3, while the fraction of known objects correctly classified decreased as we increased the noise.

We further examined the patterns of change in the misclassification of examples from the known class with increasing levels of noise—do they get misclassified as novel class or as background? Our experiments show the latter—raised levels of noise increase misclassification as background. As hoped for, our model does not falsely identify these images as coming from a novel class.

How veridical should the hierarchy be. In order to explore the significance of the hierarchy in our proposed scheme, we followed the procedure described in Section 3.1.1 using different hierarchies imposed on the same set of classes, where the different hierarchies are less faithful to the actual similarity between classes in the training set. The least reliable should be the random hierarchy, obtained by assigning classes together in a random fashion. We expect to see reduced benefit to our method as the hierarchy becomes less representative of similarity relations in the data. On the other hand, if our method maintains any benefit with these sloppy hierarchies, it will testify to the robustness of the overall approach.

We therefore compared the reliable hierarchy used above to the random hierarchy, obtained by randomly putting classes together regardless of their visual similarity. As expected, our comparisons show a clear advantage to the reliable hierarchy. In order to gain insight into the causes of the decrease in performance, we separately analyzed the general and specific level classifiers. The comparison of acceptance rate by the general level classifier using the veridical hierarchy versus random hierarchy is shown in Fig. 5. For examples that were accepted by the general level classifier, correct unknown classification versus false unknown classification is shown in Fig. 4, for both the veridical and random hierarchy.

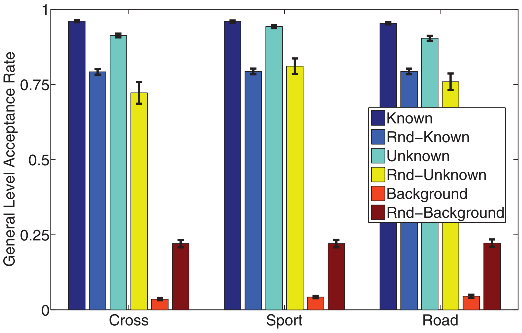


Fig. 5. General level classifier acceptance rates, comparing the use of the reliable and random hierarchies. Six bars show, from left to right, respectively: reliable hierarchy known classes (“Known”), random hierarchy known classes (“Rnd-Known”), reliable hierarchy unknown class (“Unknown”), random hierarchy unknown class (“Rnd-Unknown”), reliable hierarchy background (“Background”), random hierarchy background (“Rnd-Background”). Results are shown for the cases where the Cross-Motorbikes, Sport-Motorbikes, or Road-Motorbikes are left out as the unknown class, from left to right, respectively.

Results are clearly better in every aspect when using the veridical hierarchy. The performance of the learned general level classifier is clearly better (Fig. 5). The distinction between known classes and the unknown class by the specific classifier is improved (Fig. 4). We actually see that when using a discriminative approach based on the random hierarchy, the accuracy of this distinction decreases to the level of the nondiscriminative approach with the veridical hierarchy. Combining both the general level classifier and the specific level classifier, clearly Algorithm 1 for the identification of unknown classes performs significantly better with the veridical hierarchy.

Comparison to alternative methods. Novelty detection is often achieved with single class classifiers. In the experiments above, we used 1vsBck classifiers as proxy to single class classifiers and compared their performance to our approach in Fig. 4. In order to compare our approach to some standard novelty detection method, we chose the one class SVM [4]. Note that this is only a partial comparison since one class SVM (like any novelty detection method which is based on rejecting the known) does not provide the distinction between novel object class and background, as we do.

For technical reasons, in order to conduct this comparison we need to create a single vector representation for each instance in our data set. To achieve this goal, we followed the scheme presented in [12], describing each image by a single vector whose components are defined by the object class model of the general level class. Given this image representation, we modified our algorithm and replaced the specific level classifier with a binary SVM classifier, basing the final decision on a voting scheme.

We conducted this experiment on audiovisual data collected by a single Kyocera camera with fish-eye lens and an attached microphone. In the recorded scenario, individuals walked toward the device and then read aloud identical text; we acquired 30 sequences with 17 speakers. We tested our method by choosing six speakers as members of the trusted group, while the rest were assumed unknown. The comparison was done separately for the audio and visual data. Fig. 6 shows the comparison of our discriminative approach to the one class SVM novelty detection approach using the visual data; clearly, our approach achieves much better results (similar improvement was obtained when using the auditory data).

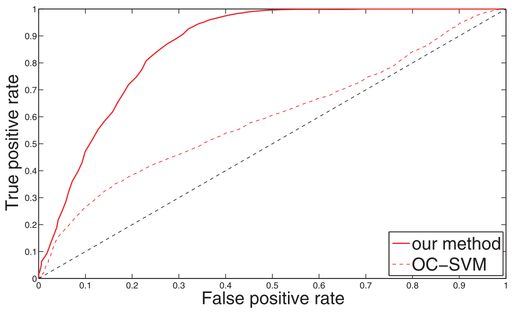


Fig. 6. True Positive versus False Positive rates when detecting unknown versus trusted individuals. The unknown are regarded as positive events. Results are shown for our proposed method (solid line) and one class SVM (dashed line).

3.2 Novel Subclasses of Audio Objects

We use Algorithm 1 with an application from the domain of audio object classification in order to evaluate the proposed framework in a different modality under systematically controlled noise levels. Here, the general classifier C_G detects the presence (or absence) of a general audio object ("acoustic blob") in a background of real recorded environmental noise. Each specific classifier C_j discriminatively detects the identity of a certain object once the general classifier has detected an acoustic blob.

The task is to discriminate known from novel audio objects appearing in an ambient sound background of a typical office environment. Hence, the inputs fall into three broad groups: pure background noise (ambient environmental sounds such as ventilation noise recorded in an office room) with no specific audio object, known audio object embedded in background noise at a certain signal-to-noise ratio (SNR), and novel audio object embedded in the background at some SNR. For each SNR, ranging from -20 to 20 dB on the logarithmic decibel (dB) scale, analysis was carried out considering four classes of objects: door opening and closing, keyboard typing, telephone ringing, and speech. The nonspeech sounds and the noise background were recorded on-site; speech was taken from the TIMIT database [20]. The continuous audio signals were cut into 1 second long frames on which the analysis described below was carried out.

In analogy to the experiments performed in Section 3.1, performance is evaluated in a leave-one-out procedure, i.e., each of the office objects is defined as novel once and left out of the training set. The performance of the proposed approach is compared to results obtained from novelty detection based on one-class SVM outlier detection.

3.2.1 General and Specific Detector Architecture

The architecture of the general "acoustic blob" detector is based on the observation that the most general feature of sound objects are fluctuations in sound pressure level (amplitude modulations) that separate them from a less variable background noise floor. The implementation is based on psychophysically motivated RASTA-PLP amplitude modulation features [21] that are combined with temporal and spectral integration to yield a single confidence score for audio object presence. Subsequent to the ROC analysis, the detector is tuned to 5 percent false positive rate [22], cf. Fig. 7.

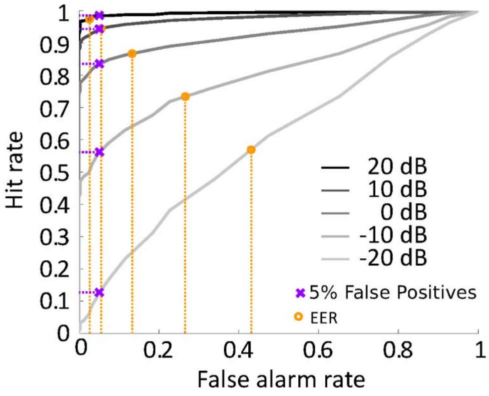


Fig. 7. ROC performance of the general acoustic model ("blob detector"). False alarm rate of 5 percent is selected for subsequent experiments.

The specific detectors use amplitude modulation spectrograms (AMS, [23]) as input features. The AMS extracts the temporal modulation of the signal by applying a short-term Fourier transformation in each of 17 Bark-scaled frequency bands. This results in a 493-dimensional feature space (17 frequency bands, 29 modulation frequencies from 2 to 30 Hz). Using these features, a radial basis function SVM was trained with a 1-versus-all approach for each of the known object subclasses.

3.2.2 Detection of Novel Events

Novelty of an acoustic event is detected when the general classifier accepts the input data as containing an object and the confidence score (signed distance from margin) of the best-matching specific classifier remains below a threshold γ. By varying γ, ROC curves are obtained that display the tradeoff between correct novelty detection and false alarms of a known event being classified as novel. Evaluating the performance at the equal error rate (EER) point of the ROC curves, novel event detection accuracy as a function of SNR is obtained.

Outlier detection based on one-class support vector machines is used as a baseline algorithm to compare the performance of our algorithm with. In the same leave-one-out fashion used for the proposed method, one SVM model is trained for each of the known classes. Varying a threshold on the best matching (highest confidence score) SVM model, ROC curves and performance at equal error rate for detection of novel-class objects are derived. Results reported here correspond to the best post hoc choice of one-class SVM parameters, in effect giving an upper bound estimate on their outlier detection performance.

3.2.3 Results

Performance of novelty detection by the specific classifiers, assuming an ideal general classifier that perfectly detects the presence of acoustic objects, is shown in Fig. 8. Since the physical characteristics of the various signals differ considerably, also the degree to which novelty is detected reliably shows a corresponding variability. For beneficial signal-to-noise ratios of 20 dB, novelty is detected reliably. When the SNR decreases, performance degrades gracefully until it reaches chance level at about -20 dB.

Combining the general classifier with the specific classifiers for the full implementation of Algorithm 1, the

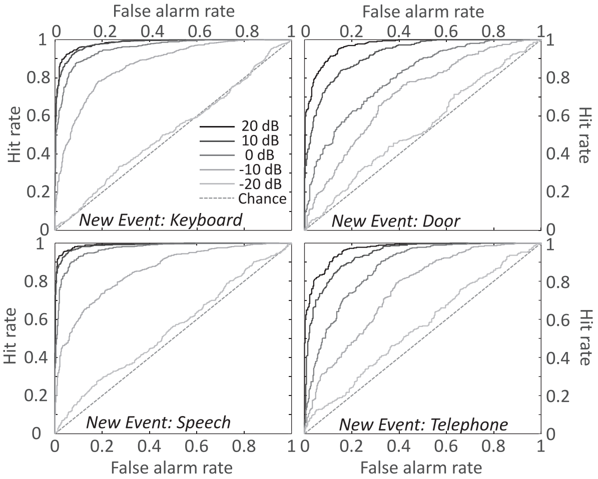


Fig. 8. ROC performance of the specific object classifiers for detecting novel audio objects at different signal-to-noise ratios. The hit and false alarm rates when classifying an object as "new" are displayed for four different novel objects.

resulting performance levels at equal error rate are displayed in Fig. 9. Here, the performance is bounded from above by the (arbitrary) choice of 5 percent false positive rate for the tuning of the general classifier. At SNR levels of -5 dB or better, the performance of the reference one-class SVM approach is lower than the proposed method's for all investigated conditions. The results of the proposed approach at -10 dB SNR (and below, not shown) are predominantly influenced by the large performance drop of the general classifier at -10 dB and below (cf. Fig. 7), effectively rendering the general classifier unable to detect the presence of objects embedded in the corresponding adverse noise levels. Results demonstrate that novelty detection based on a hierarchy of classifiers is possible in the acoustic domain and its performance depends on the type of novel signal and SNR.

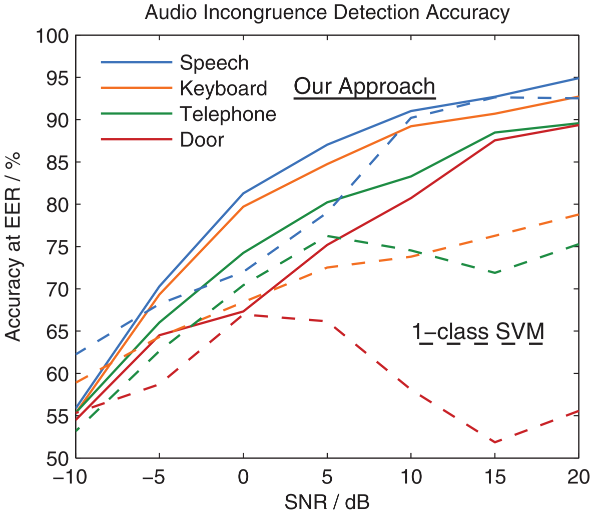


Fig. 9. Accuracy of novelty detection (Algorithm 1) for our approach (solid lines) and one-class SVM (dashed lines). One curve per type of novel audio object (see legend). The accuracy is taken at the EER point (equal false alarm and miss rates). Note that the accuracy of our approach has an upper bound due to the choice of 5 percent false positive rate for the general classifier.

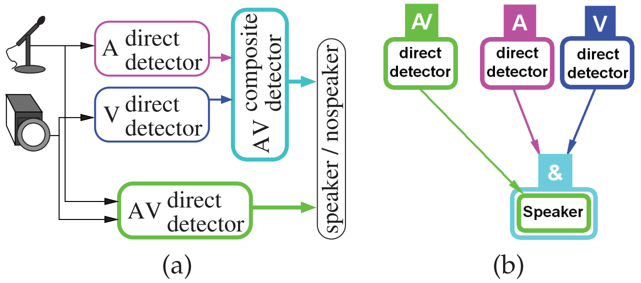


Fig. 10. (a) *Direct* and *general* (composite) audio-visual detectors provide alternatives for the speaker event modeling in (b).

4 Conjunctive Hierarchies

We present an example of a conjunctive hierarchy in audio-visual processing. As before, alternative detectors (i.e., discriminative classifiers) are used to model events in a hierarchical manner; see Fig. 10.

We concentrate on the single audiovisual event of a *human speaker* in a scene and model it in two alternative ways. We assume a scene observed by a camera with wide view-field and two microphones. Visual processing detects the presence and position of a human. Sound processing detects the intensity of sound and its direction of arrival (see [24] for technical details).

$Q_{concept}$ is the *direct* human speaker detector Q_{AV}; it is obtained by training a discriminative RBF SVM classifier on audio-visual features extracted from manually labeled training data of human speakers versus background. The detector Q_{AV} is evaluated on all spatial windows of meaningful size across the view-field, thus implicitly providing the positions of its decisions; see Fig. 11.

$Q^g_{concept}$, the more general classifier, is the conjunctive human speaker detector Q^g_{AV}; it is a composite detector, obtained by the conjunction of the *direct visual detector* Q_V and the *direct audio detector* Q_A, i.e., $Q^g_{AV} = Q_V \cdot Q_A$ (see (1)). Unlike Q_{AV}, Q^g_{AV} does not exploit the information about where Q_A and Q_V are active in the view-field. In effect, it looks to see whether they are active, irrespective of position; see Fig. 12.

By construction, Q^g_{AV} returns a positive outcome when observing a human body and human sound in different positions in the view-field. Detector Q_{AV}, on the other hand, is passive in this situation since it has been trained only on

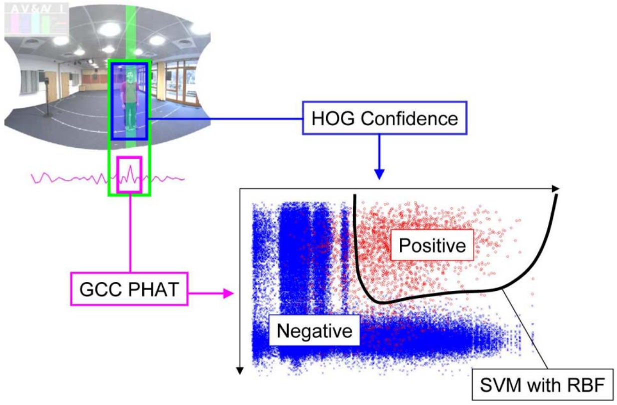


Fig. 11. Direct audio-visual detector Q_{AV}.

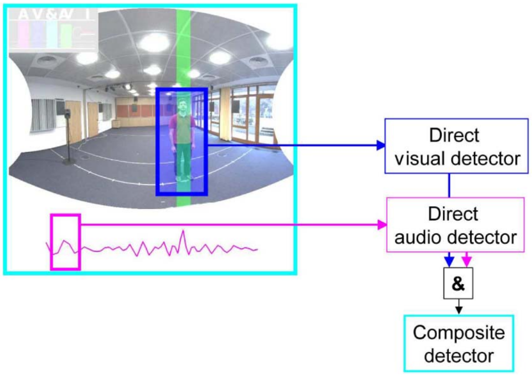


Fig. 12. *General* (composite) audio-visual detector Q^g_{AV}.

colocated human sound and visual examples. Thus, there is an incongruence where $Q^g_{AV} \gg Q_{AV}$ appears; see Fig. 13.

Table 1 presents the quantitative evaluation of the incongruence detection on real data shown in Fig. 13, consisting of $M = 462$ images. The threshold on incongruence detection was set to the smallest value achieving zero falsely detected incongruences, i.e., $FP = 0$. With this setting, we obtained recall $(TP/(TP + FN))$ of 96.2 percent and accuracy $((TP + TN)/M)$ of 97.6 percent.

The concept of incongruence, as defined in Section 2, also signifies the insufficiency of the composite general classifier. This is an interesting functionality in systems which build and maintain a structured model of events and behaviors. The composite model aims at explaining observations by a combination of simple processing blocks. It can be viewed as modeling our ability to explain the observation in terms of simple known concepts. Direct detectors, on the other hand, can be seen as a mechanism for efficient memorization and outlier rejection. Incongruence in this case may serve to signify that the model of the environment should be updated.

5 General Hierarchies

Here, we discuss two applications where general hierarchies are involved. In Section 5.1, we detect Out-Of-Vocabulary words in speech recognition, and in Section 5.2 we detect new patterns of motion from video. In the first application, the hierarchy is very simple (a chained list). In the second application, the hierarchy is richer than we have seen before, including both disjunctive and conjunctive nodes.

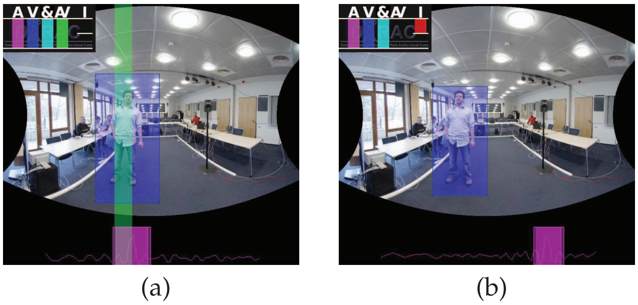


Fig. 13. A scene with a human speaker and a loudspeaker. (a) The congruent event is detected when the human speaker speaks and the loudspeaker is silent, i.e., the direct audio detector Q_A (the magenta bar under A), the direct visual detector Q_V (blue bar under V), the composite human speaker detector Q^g_{AV} (cyan bar under &), and the direct human speaker detector Q_{AV} (green bar under AV), are all active. (b) The incongruent event (red bar under I) appears when the person is silent and the loudspeaker emits speaker sounds. Notice that the direct human speaker detector Q_{AV} is not active, i.e., $Q^g_{AV} \gg Q_{AV}$ (no green bar under AV).

TABLE 1
Contingency Table for Detecting Incongruent Audio-Visual Events in the Sequence Shown in Fig. 13 Consisting of 462 Images (TP—True Positive, FP—False Positive, FN—False Negative, TN—True Negative)

	incongruent event	congruent event
detection incongruent	TP = 281	FP = 0
detection congruent	FN = 11	TN = 170

5.1 Out-of-Vocabulary Words

In speech recognition, one may consider a simple minimal hierarchy where, in effect, only a single class is given for each of the specific and general concepts. Thus, the hierarchy is really a chained list of more and more general concepts, rather than a tree as we have seen above. In this case, when trying to adapt the framework described in Section 2 to the problem of novel class detection, a generative approach proves more useful. Specifically, we build two generative classifiers for each concept: $Q_{concept}$—a classifier trained with examples of the concept, and $Q^g_{concept}$—a classifier trained with examples of the more general concept according to the partial order.

5.1.1 OOV Detection Method

We define an OOV word as a word whose pronunciation does not match the pronunciation dictionary. An OOV word can contain an incongruent phoneme—e.g., when someone has mispronounced a word or, more generally, the pronunciation of the word is not present in the dictionary. The specific classifier $Q_{concept}$ introduces constraints from top-down knowledge by using a longer ranging word language model, forcing constraints on the possible phoneme strings and using the context of the neighboring words. The general classifier $Q^g_{concept}$ computes event probability based on the independent parts, e.g., not conditioned on the language and not considering the order of parts in time. We realize the general classifier as a phoneme recognizer using only short acoustic context, which allows a wide range of (or all possible) phoneme combinations. Thus, the strongly constrained classifier in our case is a Large Vocabulary Continuous Speech Recognizer (LVCSR).

Initially [25], we performed the comparison of posterior probability vectors on a frame-by-frame basis, using vectors of phoneme posteriors from both classifiers. We used the Kullback-Leibler divergence between the posterior probability vectors from the respective systems as a measure of the agreement between the models. Later [26], we introduced another classifier based on a multilayer perceptron that was trained on phoneme posterior data labeled with in-vocabulary and out-of-vocabulary word labels. In [27], we trained a neural network to learn to distinguish different classes of events, and this latter classifier is extended and used here.

Fig. 14 shows an example when processing a speech sample containing the Out-of-Vocabulary word "BELGIUM"

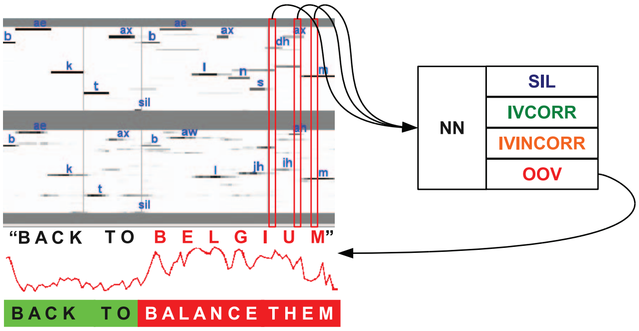


Fig. 14. Input and output of the OOV word detection system. *Left top:* Phoneme posteriors from strongly and weakly constrained recognizers, used as input features for the neural net classifier. *Left bottom:* Output score of neural net classifier—recognized words and their classification (green: correct, red: OOV). *Right:* Schematic illustration of the neural network classifier, with one hidden layer and four outputs for the estimation of frame level class posteriors.

[27]. In correctly recognized speech segments, we generally find

- agreement between the general and the specific recognizers,
- high certainty about predicted phonemes in the specific recognizer.

In the part covered by OOV input, we tend to find

- disagreement between the specific and the general recognizers,
- low posterior probability reflected in high entropy of posterior distribution from the constrained recognizer.

5.1.2 Experiments

Both specific (constrained by language model) and unconstrained (acoustic) recognizers were trained on conversational telephone speech and tested on the Call Home English (CHE) corpus.[1] To introduce OOV words, the vocabulary was restricted to the 2,860 most frequent words from the language model training texts, leaving the remaining words unknown to the specific recognizer. The evaluation set consists of 1.33 hours.

In our experiments, we used a realistic spontaneous speech telephone data (CHE) because most real-world data has similar characteristics like distortions, noise, low audio quality or sloppy speakers. When using these data, it turned out to be beneficial to train the neural network in a four class paradigm (instead of the binary OOV versus rest), as follows:

- *sil*—silence, no speech at all,
- *ivcorr*—correctly recognized speech (in-vocabulary),
- *ivincorr*—misrecognized speech (in-vocabulary),
- *oov*—misrecognized speech due to out-of-vocabulary content.

Table 2 shows the improvement in equal error rate of OOV detection gained on CHE and WSJ data using three or four classes in our neural net classifier. The classifier, trained on one database (CHE), clearly generalizes very well to the WSJ data.

TABLE 2
EER (in Percent) of NN-Based MISREC/OOV Detection

Test Data	Training Data	3 Classes *oov*	4 Classes *oov*
CHE	WSJ	27.07	25.80
CHE	CHE	27.60	**21.73**
WSJ	WSJ	11.90	**11.41**
WSJ	CHE	13.63	14.35

In addition, we tested the generalization of our system by replacing the complex and powerful LVCSR system that we used before with a faster and more rudimentary LVCSR system (without adaptations, using one pass decoding). We saw that it only resulted in a modest decrease in OOV detection performance of about 1 percent EER on CHE.

In the initial experiments, roughly 40 percent of the OOV word types appeared in both the training and testing data at least once, possibly allowing the classifier to learn to detect specific OOV words. In order to address this issue, we composed a new test set based on 10 hours of Fisher data[2] (conversational telephone speech), containing only OOV types which neither appeared in the training text of the language model nor in the training set of the neural net classifier. In this experiment, the OOV token rate was about 4 percent and the EER in OOV detection was about 22.5 percent, a result which lies in the same range as the previously reported results (21.73 percent). While the numbers are not directly comparable since different test sets were used, the results still show that the learned patterns are generalizable, i.e., our technique is able to detect OOV words which have never been seen during training.

5.2 Novel Patterns of Motion

In this application of video surveillance, we are interested in monitoring the well-being of elderly people in their homes and show how the notion of incongruence helps greatly in the detection of surprising or unusual events. As before, the general idea is to arrange a set of trackers in a hierarchical structure. The output of each tracker acts as a classifier and is analyzed as described in Section 2. Here, we rely on a fixed hierarchy of motion patterns, which is richer than we have seen before, including both disjunctive and conjunctive nodes.

5.2.1 Tracker Tree

Visual trackers in general incorporate a certain amount of information about the normal situations they are applied to. For example, an articulated body motion tracker is highly tuned to a walking person and exploits strong priors for successful tracking, whereas a simple blob tracker relies on very weak assumptions. We propose to arrange multiple different trackers in a tree-like hierarchy, where the location of each tracker is based on the information it relies on. Trackers further up in the tree have been trained for a narrow set of actions—e.g., specific to the walking style of one person, whereas trackers closer to the root node are able to track a broad variety of motion patterns. The implemented tracker tree for elderly care applications is visualized in Fig. 15.

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97S42.

[2] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T19.

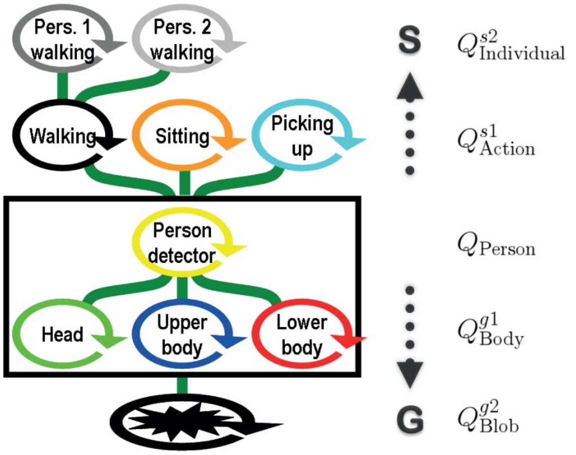


Fig. 15. The *tracker tree* with a different tracker at each node. Each tracker acts as concept Q_a, where the Q_{Person} is the concept of interest in our application. More specific (*disjunctive*) concepts lie in the top of the tree, more general (*conjunctive*) ones rest toward the bottom.

At the root of the tree lies the most general concept, a foreground blob tracker Q^{g2}_{Blob}. It will track any object moving through the scene. Since we want to monitor humans, a person detector Q_{Person} is the stage that immediately follows the generic blob tracker, one level up. Since a more detailed analysis of the actual visibility of body parts is important for the interpretation of incongruences, this tracker tree also contains, apart from the disjunctive hierarchy structures higher up in the tree, a conjunctive part (black box) near the bottom. We now describe these disjunctive and conjunctive parts of the tree in a bit more detail.

As a disjunctive example from the tree, a person (general level) can perform different actions, which are modeled by different action trackers (specific level; here, walking, sitting, and picking object up from floor actions). Thus, the more specific concept is

$$Q^{s1}_{\text{Action}} = Q_{\text{Walk}} + Q_{\text{Sit}} + Q_{\text{Pickup}}. \quad (6)$$

In the same vein, individual walking trackers $Q^{s2}_{\text{Individual}}$ provide subconcepts to the generic walking tracker and model the gait pattern of individual people known to the system.

On the other hand, when moving down in the tree, the trackers inside the black box in Fig. 15 form a conjunctive hierarchy that considers the person as composed of body parts. Separate detectors check whether legs, upper body, and body shoulder patterns are found. In case the person is fully visible and in an expected pose, all three parts should be detected. From the conjunctive perspective, the indication strength of finding a person amounts to

$$Q^{g1}_{\text{Body}} = Q_{\text{Head}} \cdot Q_{\text{UpperBody}} \cdot Q_{\text{LowerBody}}. \quad (7)$$

One advantage of such a tree with multiple hierarchical levels is the possibility of semantic reasoning. From the location in the tree where the novel pattern appears, we can deduce an interpretation on the nature of the incongruence. For instance, if walking is detected, but the gait does not correspond to any of the known individuals, an intruder—or at least someone not observed before—seems to be in the house. As elderly people often are the victims of scams, this would indeed be noteworthy and a sufficient condition to activate some remote attention by an assistant.

Occlusion handling. Partial occlusions occur frequently in in-house surveillance scenarios, e.g., furniture partially blocking the view of a person. In the tracker tree, this means that Q_{Person} is valid, but at least one of Q^{g1}_{Body} fails. As discussed in Section 2.2.4, without proper training (including training images of occlude objects) occlusion leads to rejection.

To prevent an irregular classification of this situation in the in-house scenario, we propose a different interpretation which considers the body part trackers as conditioners for the action trackers. Since the actions (Q^{s1}_{Action}) are trained on examples of fully visible people, the validity of any of these trackers cannot be expected to hold when the person is only partly visible. In the case of occlusion by a sofa, for example, the lower body part is missing and therefore no action is expected to be classified as valid as all action-specific trackers are critically dependent on the visibility of relevant body parts. For instance, the walking detector needs to see the legs.

To address this problem, occlusions are learned from the training data and incorporated into the model, and are therefore not detected as incongruent activity patterns. The detected incongruence of observing a person (yellow detector) but not all of his parts blocks incongruences higher up in the tree from being signaled if the absence of that body part precludes action detectors from functioning properly.

5.2.2 Experiments

The tracker tree as depicted in Fig. 15 is constructed with different state-of-the-art trackers. The root node tracker is a simple color-based blob tracker [28], whereas the person tracker (yellow) is based on a tracking-by-detection approach [29]. The body part trackers and the action trackers all rely on generative low-dimensional representations, as described in [30]. These trackers were trained in an offline procedure with approximately 3,000 images in which the different actions were segmented manually. The footage was recorded with a static camera at 15 frames per second in *VGA* resolution. The images are background subtracted and silhouettes serve as input features. Actions were segmented manually for each action tracker.

We evaluate the tracker tree on a video sequence which was recorded in a living-room environment. A single person is monitored and incongruent events are spotted. The test video of about 1,000 images contains diverse "everyday" actions such as walking, walking behind occluding objects, sitting on different chairs, or picking up small objects. It also contains abnormal events, e.g., when the person falls, limps, jumps over the sofa, or when an intruder enters the room.

In Fig. 16, we present the output scores of the different trackers for a short piece of the video sequence. The plotted curves depict the confidence of the individual trackers. The horizontal line indicates the threshold that is used for classification of the tracker scores. The reasoning in the tree is then performed and the detected incongruent events are highlighted in red.

In Fig. 17, we show exemplary result frames where the active trackers are visualized as bounding boxes in corresponding colors. As long as the person behaves according to expectation (walks, picks up an object, or

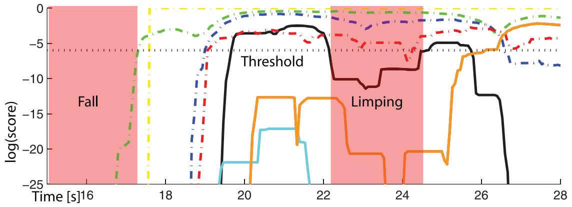


Fig. 16. Segment from a tracked sequence: The tracker output scores are plotted over time; the color code of Fig. 15 is used. The individual walking trackers are omitted and the indicated threshold is applied for classification. For illustration, incongruent patterns are highlighted.

walks behind the sofa), the tracker tree accepts the situation. When an incongruence in the motion pattern is detected, an abnormal event is detected and the frame is marked in red (fall, limping, intruder).

In the following, we analyze the performance of the tracker tree for abnormal event detection and compare it to state-of-the-art methods. To this end, we sweep the threshold that is applied to the tracker confidence scores and compare the tree's output with the ground truth annotation of the test sequence. As baseline comparison, we learn a Gaussian Mixture Model (GMM) from the training data using the EM algorithm [31]. Similarly to most of our trackers (cf. [30]), GMMs are used for tracking and outlier detection, but with no hierarchical structure.

The results are displayed as ROC curve in Fig. 18. Note that the ROC curve for the tracker tree has a particular shape and does not reach full recognition since the nonlinear classifier reasoning is applied after fixing the threshold. Due to the reasoning in the hierarchy, the tracker tree outperforms GMM outlier detection regardless of the number of mixture components.

6 Biological Evidence

Our approach to the detection of incongruent events is motivated in part by evidence from biological systems indicating the existence of special mechanisms mediating top-down incongruence detection that can be differentiated from mere novelty detection. Neuronal mechanisms underlying novelty detection are hypothesized to be based on the increased neuronal responses to deviant stimuli presented in the context of repeated, so-called "standard," stimuli. This phenomenon is fundamentally a consequence of stimulus-specific adaptation, i.e., the decaying neuronal response strengths with repeated presentation of identical stimuli (e.g., [32]) that is now known from various brain systems. To dissociate such bottom-up mechanisms of novelty detection from top-down mechanisms that may underlie incongruence detection, we have designed an experiment investigating "semantic" deviants, i.e., incongruence based on the attributed meaning to stimuli as opposed to incongruence resulting from the presentation statistics.

In this experiment, two groups of rodents (gerbils) were trained to categorize four vowels from human speech in two orthogonally different ways, using a Go/NoGo procedure, thereby establishing different semantic contexts for identical features. The Go/NoGo procedure is a standard technique in behavioral and cognitive neuroscience in which a subject signals a binary response by either performing a predefined trained action (Go) or refraining from doing so (NoGo). Fig. 19a schematically shows the positions of the four vowels in the feature space spanned by the first formant (F1) and the spectral distance between the first and second formant (F2-F1). This space

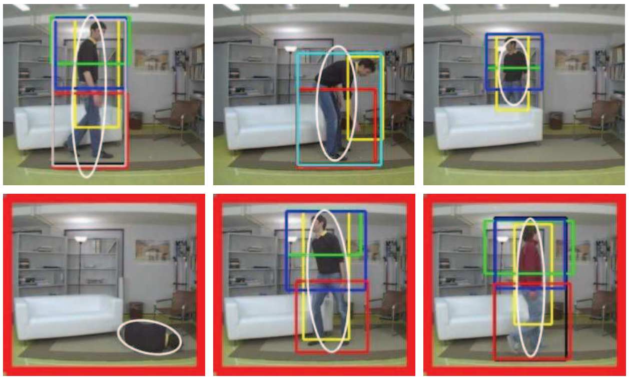


Fig. 17. Selected frames from one sequence. The active trackers are visualized by the bounding box using the color code of Fig. 15. Incongruent events (falling, limping, intruder) are marked in red.

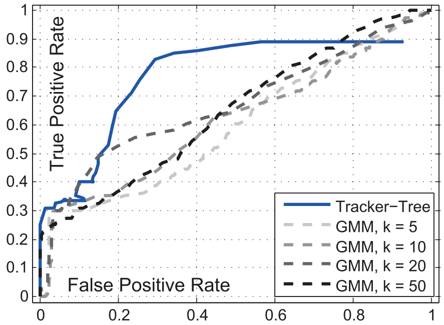
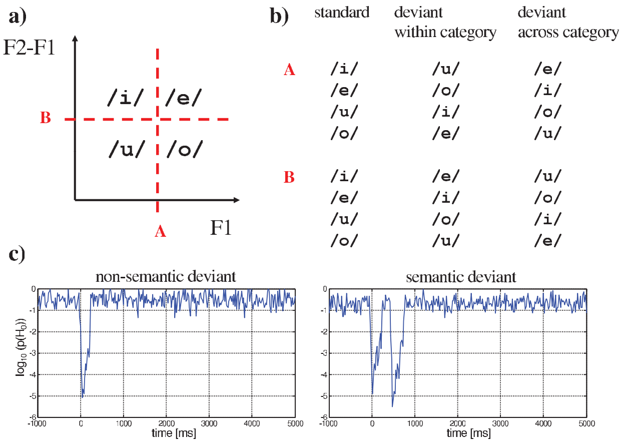


Fig. 18. ROC curve evaluation of abnormal action detection using the tracker tree. Due to the hierarchical reasoning, tracker trees outperform comparable state-of-the-art methods based on GMM.



	standard	deviant within category	deviant across category
A	/i/	/u/	/e/
	/e/	/o/	/i/
	/u/	/i/	/o/
	/o/	/e/	/u/
B	/i/	/e/	/u/
	/e/	/i/	/o/
	/u/	/o/	/i/
	/o/	/u/	/e/

Fig. 19. Dissociation of bottom-up and top-down incongruence detection in rodent psychophysics. a) Schematic depiction of the positions of four vowels in a suitable feature space formed by the first two formants of the vowels, F1 and F2 (cf. [33]). b) Table showing the various tested combinations of standard stimuli and two types of corresponding deviants. c) Detection of significantly different spatiotemporal activity patterns in the auditory cortex electrocorticogram for semantic and nonsemantic deviants using a classification technique. Note the occurrence of significant ($p < 10^{-3}$) classifications in the time window 0.3-0.5 s after stimulus onset only in the case of semantic deviants.

basically conforms to the classical vowel feature space described by [33], but was shown to be physiologically realized in mammalian auditory cortex [34].

One group of gerbils was trained to categorize these stimuli according to category boundary A, while the other group was trained according to boundary B. After training, classical novelty-detection or odd-ball experiments were performed on both groups by presenting one vowel repeatedly as the standard stimulus and a second vowel as the infrequent deviant. Note that, given the previous training, this second vowel could be selected either to be a member of the same category as the standard stimulus or to be a member of the opposite category (associated with the opposite meaning for the required Go/NoGo behavior). Fig. 19 shows the different combinations of standard and deviants used in the posttraining tests. We recorded multichannel electrocorticograms from auditory cortex as these signals have been demonstrated to provide physiological correlates of category formation during learning [35].

Spatial patterns of electrocorticograms were used to classify vowel identity, and classification performance was analyzed in consecutive time bins of 120 ms (stepped in 20 ms steps) by comparing the number of correct classifications across all experimental trials with the expected number of correct classifications by chance (for details of the method see [36]). For each empirically found number of correct classifications, Fig. 19c shows the probability of observing this number of correct classifications by chance (H0), separately for deviants being a member of the same meaning class (nonsemantic deviants) and for deviants being a member of the opposite meaning class (semantic deviants).

Significantly ($p < 10^{-3}$) different electrocorticogram patterns were found for both types of deviants at stimulus onset, but only for semantic deviants during an additional time window 0.3-0.5 s after stimulus onset. This latter result may indicate the existence of a well-separable physiological process mediating the detection of a top-down outlier because of a deviant meaning context, in addition to the bottom-up outlier because of a deviant with respect merely to stimulus occurrence statistics. Moreover, this result is in accordance with our general scheme in that it may be brought about by the mismatch between general level classifiers (vowel detectors) and more specific level classifiers (vowels of class A or B).

7 Summary and Discussion

Unexpected novel events are typically identified by their low posterior probability. In this paper, we employed a hierarchy of generality to obtain a few probability values for each event, which allowed us to discriminate and identify different types of unexpected events. We described how our approach can be used to design new algorithms, which detect "interesting" unexpected situations in a variety of applications and data types, including real audio, speech, image, and video data.

Incongruent events are characterized by some discrepancy between the response of two classifiers, which can occur for a number different reasons. *Out of Context* is one such example. In a given context such as the English language, a sentence containing a Czech word is assigned low probability. In the visual domain, in a given context such as a street scene, an elephant is unlikely to appear. Another example is the recognition of *novel objects*, when a new object is encountered of some known generic type but unknown specifics.

Acknowledgments

This study was supported by the European Union under the DIRAC integrated project IST-027787.

References

[1] M. Markou and S. Singh, "Novelty Detection: A Review-Part 1: Statistical Approaches," *Signal Processing,* vol. 83, no. 12, pp. 2499-2521, 2003.

[2] M. Markou and S. Singh, "Novelty Detection: A Review-Part 2: Neural Network Based Approaches," *Signal Processing,* vol. 83, no. 12, pp. 2481-2497, 2003.

[3] D. Tax and R. Duin, "Support Vector Data Description," *Machine Learning,* vol. 54, no. 1, pp. 45-66, 2004.

[4] B. Scholkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support Vector Method for Novelty Detection," *Proc. Advances in Neural Information Processing Systems,* 2000.

[5] D. Yeung and C. Chow, "Parzen-Window Network Intrusion Detectors," *Proc. Int'l Conf. Pattern Recognition,* 2002.

[6] C.P. Diehl and J.B. Hampshire II, "Real-Time Object Classification and Novelty Detection for Collaborative Video Surveillance," *Proc. IEEE Int'l Joint Conf. Neural Networks,* 2002.

[7] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a Similarity Metric Discriminatively, with Application to Face Verification," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2005.

[8] G.S. Berns, J.D. Cohen, and M.A. Mintun, "Brain Regions Responsive to Novelty in the Absence of Awareness," *Science,* vol. 276, no. 5316, pp. 1272-1275, 1997.

[9] B. Rokers, E. Mercado, M.T. Allen, C.E. Myers, and M.A. Gluck, "A Connectionist Model of Septohippocampal Dynamics during Conditioning: Closing the Loop," *Behavioral Neuroscience,* vol. 116, no. 1, pp. 48-62, 2002.

[10] M. Marszałek and C. Schmid, "Constructing Category Hierarchies for Visual Recognition," *Proc. 10th European Conf. Computer Vision,* 2008.

[11] J. Sivic, B.C. Russell, A. Zisserman, W.T. Freeman, and A.A. Efros, "Unsupervised Discovery of Visual Object Class Hierarchies," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.

[12] A. Bar-Hillel and D. Weinshall, "Subordinate Class Recognition Using Relational Object Models," *Proc. Advances in Neural Information Processing Systems,* vol. 19, 2006.

[13] A. Zweig and D. Weinshall, "Exploiting Object Hierarchy: Combining Models from Different Category Levels," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[14] M. Marszałek and C. Schmid, "Semantic Hierarchies for Visual Object Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.

[15] A. Bar-Hillel, T. Hertz, and D. Weinshall, "Efficient Learning of Relational Object Class Models," *Proc. IEEE Int'l Conf. Computer Vision,* 2005.

[16] B. Leibe, A. Leonardis, and B. Schiele, "Robust Object Detection with Interleaved Categorization and Segmentation," *Int'l J. Computer Vision,* vol. 77, no. 1, pp. 259-289, 2008.

[17] R. Fergus, P. Perona, and A. Zisserman, "Weakly Supervised Scale-Invariant Learning of Models for Visual Recognition," *Int'l J. Computer Vision,* vol. 71, no. 3, pp. 273-303, 2007.

[18] J. Matas et al., "Comparison of Face Verification Results on the XM2VTS Database," *Proc. Int'l Conf. Pattern Recognition,* 2000.

[19] G. Griffin, A. Holub, and P. Perona, "Caltech-256 Object Category Dataset," Technical Report UCB/CSD-04-1366, California Inst. of Technology, http://www.vision.caltech.edu/Image_DataSets/Caltech256, 2007.

[20] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, and V. Zue, "Timit Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium, Philadelphia, 1993.

[21] H. Hermansky and N. Morgan, "Rasta Processing of Speech," *IEEE Trans. Speech and Audio Processing,* vol. 2, no. 4, pp. 578-589, Oct. 1994.

[22] J.-H. Bach and J. Anemüller, "Detecting Novel Objects in Acoustic Scenes through Classifier Incongruence," *Proc. Int'l Conf. Spoken Language Processing,* 2010.

[23] J. Anemüller, D. Schmidt, and J.-H. Bach, "Detection of Speech Embedded in Real Acoustic Background Based on Amplitude Modulation Spectrogram Features," *Proc. Int'l Conf. Spoken Language Processing,* 2008.

[24] T. Pajdla, L. Van Gool, M. Havlena, J. Heller, A. Torii, A. Ess, J.-H. Bach, H. Kayser, J. Anemüller, and P. Van Hengel, "Incongruence Detection in Audio-Visual Processing," Research Report CTU-CMP-2008-28, Center for Machine Perception, K13133 FEE Czech Technical Univ., Prague, Czech Republic, Dec. 2008.

[25] H. Ketabdar, M. Hannemann, and H. Hermansky, "Detection of Out-of-Vocabulary Words in Posterior Based ASR," *Proc. European Conf. Speech Comm. and Technology,* 2007.

[26] L. Burget, P. Schwarz, P. Matějka, M. Hannemann, A. Rastrow, C. White, S. Khudanpur, H. Hermansky, and J. Černocký, "Combination of Strongly and Weakly Constrained Recognizers for Reliable Detection of OOVs," *Proc. 33rd IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* pp. 4081-84, 2008.

[27] S. Kombrink, L. Burget, P. Matějka, M. Karafiát, and H. Hermansky, "Posterior-Based Out of Vocabulary Word Detection in Telephone Speech," *Proc. Int'l Conf. Spoken Language Processing,* pp. 80-83, 2009.

[28] G.R. Bradski, "Computer Vision Face Tracking for Use in a Perceptual User Interface," *Intel Technology J.,* vol. 2, no. Q2, pp. 12-21, 1998.

[29] P. Felzenszwalb, D. Mcallester, and D. Ramanan, "A Discriminatively Trained Multiscale, Deformable Part Model," *Proc. IEEE Conf. Computer Vision and Pattern Recognitio,* 2008.

[30] F. Nater, H. Grabner, T. Jaeggli, and L. Van Gool, "Tracker Trees for Unusual Event Detection," *Proc. IEEE Int'l Conf. Computer Vision Workshop Visual Surveillance,* 2009.

[31] C.M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2007.

[32] N. Ulanovsky, L. Las, D. Farkas, and I. Nelken, "Multiple Time Scales of Adaptation in Auditory Cortex Neurons," *The J. Neuroscience,* vol. 24, pp. 10440-10453, 2004.

[33] F.W. Ohl and H. Scheich, "Orderly Cortical Representation of Vowels Based on Formant Interaction," *Proc. Nat'l Academy of Sciences,* vol. 94, pp. 9440-9444, 1997.

[34] G.E. Peterson and H.L. Barney, "Control Methods Used in a Study of the Vowels," *J. Acoustical Soc. Am.,* vol. 24, pp. 175-184, 1952.

[35] F.W. Ohl and H. Scheich, "Change in Pattern of Ongoing Cortical Activity with Auditory Category Learning," *Nature,* vol. 412, pp. 733-736, 2001.

[36] M. Deliano, H. Scheich, and F.W. Ohl, "Auditory Cortical Activity after Intracortical Microstimulation and Its Role for Sensory Processing and Learning," *The J. Neuroscience,* vol. 29, pp. 15898-15909, 2009.

Daphna Weinshall received the BSc degree in mathematics and computer science from Tel-Aviv University in 1982. She received the MSc and PhD degrees in mathematics and statistics from Tel-Aviv University in 1985 and 1986, respectively, working on models of evolution and population genetics. Between 1987 and 1992, she visited the Center for Biological Information Processing at MIT and the IBM T.J. Watson Research Center. In 1993, she joined the Institute of Computer Science at the Hebrew University of Jerusalem, where she is now a full professor. She is currently an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence*. In the last 10 years, she has served as an area chair for a number of computer vision and learning conferences, including CVPR, ECCV, and NIPS. Her research has been supported by grants from the EC (BeNoGo and DIRAC), as well as some Israeli and binational foundations. She has published more than 100 papers in peer reviewed conferences and journals. Her research interests include computer and biological vision, as well as machine and human learning. Her current interests include the learning of distance function, object class recognition, and novelty detection.

Alon Zweig received the BSc degree in computer and cognitive science and the MSc degree in computer science from the Hebrew University of Jerusalem. He is currently working toward the PhD degree in the Computer Science Department at the Hebrew University of Jerusalem. His research interests are computer vision and machine learning. He is the recipient of the Intel student excellence award (2010).

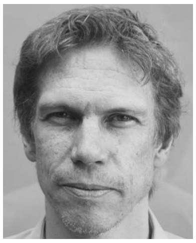
Hynek Hermansky received the DrEng degree from the University of Tokyo, and Dipl Ing degree from Brno University of Technology, Czech Republic. He is the Julian S. Smith Professor of Electrical and Computer Engineering and the director of the Centre for Language and Speech Processing at The Johns Hopkins University in Baltimore, Maryland. He is also a professor at the Brno University of Technology, Czech Republic, and an adjunct professor at the Oregon Health and Sciences University, Portland, Oregon, and an external fellow at the International Computer Science Institute in Berkeley, California. He is a fellow of the IEEE for "invention and development of perceptually-based speech processing methods," and a fellow of the International Speech Communication Association for pioneering bio-inspired approaches to processing of speech. He was a member of the Commitee at the 2011 ICASSP in Prague, technical chair at the 1998 ICASSP in Seattle, and an associate editor for the *IEEE Transactions on Speech and Audio*. Further, he is a member of the editorial board of *Speech Communication*, holds nine US patents, and has authored or coauthored more than 200 papers in reviewed journals and conference proceedings. He has been working in speech processing for more than 30 years, previously as a director of research at the IDIAP Research Institute, Martigny, and an adjunct professor at the Swiss Federal Institute of Technology in Lausanne, Switzerland, a professor and director of the Center for Information Processing at OHSU Portland, Oregon, a senior member of the research staff at US WEST Advanced Technologies in Boulder, Colorado, a research engineer at Panasonic Technologies in Santa Barbara, California, and a research fellow at the University of Tokyo. His main research interests are in acoustic processing for speech recognition.

Stefan Kombrink graduated in computer science in 2005 and linguistics in 2008 from Stuttgart University, Germany. He is currently working toward the PhD degree and is a junior researcher at Brno University of Technology (BUT), Faculty of Information Technology (FIT), Czech Republic. In 2007, he spent six months with the speech group at FIT BUT to work on his master's thesis "Out-of-Vocabulary Word Detection Using Neural Networks." In the autumn of 2008, he returned to Brno, and has since significantly contributed to the EC-sponsored project DIRAC. He is currently involved in the EC-sponsored project GLOCAL. He is specializing in word/sub-word-based speech recognition, OOV detection and processing, and advanced techniques for language modeling. He is the author or coauthor of six papers in peer-reviewed conferences. He is student member of the IEEE and ISCA.

Frank W. Ohl received the Diploma degree in zoology in 1991 and PhD degree in zoology/neuroscience in 1995 from the Technical University of Darmstadt, Germany. In 1998-1999, he was a postdoctoral fellow at the University of California, Berkeley. In 2006, he accepted a W2-professorship for neurobiology/neuroprosthetics at the Otto-von-Guericke University of Magdeburg. Since 2011, he has been a W3-professor on the Faculty of Sciences, University of Magdeburg. He is also head of the department "Systems Physiology of Learning" and codirector at the Leibniz Institute for Neurobiology, Magdeburg. His research interests include neurophysiology and theory of animal learning, computational neuroscience, neuroprostheses research, and cognitive neuroscience.

Jörn Anemüller studied physics at the University of Oldenburg, Germany, and information processing and neural networks at King's College, University of London, where he received the MSc degree in 1996. He received the PhD degree in physics from the University of Oldenburg in 2001 with a dissertation on "Across Frequency-Processing in Convolutive Blind Source Separation." From 2001 to 2004, he conducted work on biomedical signal analysis as a postdoctoral fellow at the Salk Institute for Biological Studies and at the University of California, San Diego. Since 2004, he has been a member of the scientific staff of the Department of Physics, University of Oldenburg, currently leading the statistical signal models research group. His interests include statistical signal processing and machine learning techniques with application to acoustics, speech, and biomedical signals. He is a member of the IEEE.

Jörg-Hendrik Bach received the Diploma in physics in 2007 from the Karlsruhe Institute of Technology, Germany. Since 2007, he has been working toward the PhD degree in the medical physics group at the University of Oldenburg, Germany. His research interests include environmental sound classification and perceptually motivated acoustic feature extraction for robust classification of speech and nonspeech signals. He is a member of the IEEE.



Luc Van Gool received the degree in electromechanical engineering from the Katholieke Universiteit Leuven in 1981. He became an assistant professor at the Katholieke Universiteit Leuven in Belgium in 1992, an associate professor in 1994, a professor in 1996, and a full professor in 1998. That year, he also became a full professor at the ETH in Zurich, Switzerland. He leads computer vision research at both sites, where he also teaches on the subject. He has authored more than 200 papers in this field. He has been a program committee member of several major vision conferences. He was a program chair of ICCV 2005 and general chair of ICCV 2011 and ECCV 2014. His main interests include 3D reconstruction and modeling, tracking and gesture analysis, and object recognition. He has received several best paper (David Marr Prize in 1998, CVPR Best Paper in 2007, ACCV in 2007, ICRA in 2009, etc.). He is a cofounder of the companies Eyetronics, GeoAutomation, kooaba, eSaturnus, and Procedural. He is a member of the IEEE

Fabian Nater received the MSc degree in electrical engineering from the Swiss Federal Institute of Technology in Lausanne in 2006. After two years in industry at Phonak and Baumer Electric, he returned to academia and is currently with the Computer Vision Lab at ETH Zurich. Under the supervision of Professor Luc Van Gool, he is a research assistant and is working toward the PhD degree, interested in human behavior analysis and abnormal event detection in video. He is a student member of the IEEE.

Tomas Pajdla received the MSc and PhD degrees from the Czech Technical University in Prague. He works in geometry and algebra of computer vision and robotics with emphasis on nonclassical cameras, 3D reconstruction, and industrial vision. He contributed to introducing epipolar geometry of panoramic cameras, noncentral camera models generated by linear mapping, generalized epipolar geometries, and to developing solvers for minimal problems in structure from motion. His coauthored works awarded the best paper prize at OAGM '98 and BMVC '02. He is a member of the IEEE.

Michal Havlena received the diploma in computer science from Charles University in Prague in 2005. He has been working toward the PhD degree at the Czech Technical University since then. His research interests are large-scale structure from motion, omnidirectional vision, and robot navigation using vision. He is a member of the IEEE.

Misha Pavel received the BS degree in electrical engineering from the Polytechnic Institute of Brooklyn, the MS degree in electrical engineering from Stanford University, and the PhD degree in experimental psychology from New York University. He is currently a program director at the US National Science Foundation (NSF) in charge of a program called Smart Health and Wellbeing. Concurrently, he has an appointment as a professor in the Department of Biomedical Engineering, with a joint appointment in the Department of Medical Informatics and Clinical Epidemiology, at Oregon Health and Science University. Previously, he was a chair of the Department of Biomedical Engineering and the director of the Point of Care Laboratory, which focuses on unobtrusive monitoring, neurobehavioral assessment, and computational modeling. This fundamental research is focused on technology that would enable transformation of healthcare to being proactive, distributed, and patient-centered. Prior to his academic career, he was a member of the technical staff at Bell Laboratories, where his research included network analysis and modeling. His current research is at the intersection of computational modeling of complex behaviors of biological systems, engineering, and cognitive science with a focus on information fusion, pattern recognition, augmented cognition, and the development of multimodal and perceptual human-computer interfaces. He developed a number of quantitative and computational models of perceptual and cognitive processes, eye movement control, and a theoretical framework for knowledge representation; the resulting models have been applied in a variety of areas, ranging from computer-assisted instruction systems, to enhanced vision systems for aviation, to augmented cognition systems. He is a senior member of the IEEE.