# New learning paradigms for real-world environment perception

Alexander Gepperth



Habilitation à diriger des recherches

présentée à l'université Pièrre et Marie Curie
Spécialité ingénierie

par
Alexander Gepperth

# New learning paradigms for real-world environment perception

Rapporteurs:
Fabien MOUTARDE
Antoine CORNUÉJOLS
Rolf WÜRTZ

soutenue le 27 juin 2016 devant le jury composé de:
Fabien MOUTARDE
Antoine CORNUÉJOLS
David FILLIAT
Pierre-Yves OUDEYER
Olivier SIGAUD
Guy LE BESNERAIS

To my parents!

**Abstract**

In this document, I first analyze some of the reasons why real-world environment perception is still strongly inferior to human perception in overall accuracy and reliability. In particular, I focus on the task of object detection in traffic scenes and present an argument why this task is in fact a good model task for other, related perception problems (e.g. in robotics or surveillance). Enumerating the difficulties encountered in this model task (and therefore, by inference, in many other detection tasks as well), I come to the conclusion that problems in object detection can in fact be, to a significant extent, traced back to problems of the learning algorithms that are used in various forms when performing object detection. Namely, the lack of a probabilistic interpretation, the lack of incremental learning capacity, the lack of training samples and the inherent ambiguity of local pattern analysis are identified and used to justify a road map for research efforts aimed at overcoming these problems. I present several of my works concerning real-world applications of machine learning in perception, where the stated problems become very apparent. Subsequently, I describe in detail my recent research contributions and their significance in the context of the proposed road map: context-based object detection, generative and multi-modal learning as well as an original method for incremental learning. The document is concluded by an outlook that addresses further work to complete the road map, and the possibilities that are offered by such an endeavour in the field of machine perception.

**Contents**

## 1. Introduction

This document gives an overview of my scientific work after obtaining my PhD degree in 2005 from the university of Bochum (Germany). This work was performed partly at the Honda Research Institute Europe Gmbh in Offenbach (Germany) from 2005 to 2010, but for the most part at the "Ecole Nationale Supérieure de Techniques Avancées" (ENSTA ParisTech) at Palaiseau (France) where I am an associate professor ("enseignant-chercheur") since 2011. It is, by its very nature, a pedagogical rather than a scientific document, intended to give an overview but not an in-depth account of my research activities. Therefore, no precise details (formulas, parameter values, experimental results) will be shown here except for pedagogical purposes. For these details, the interested reader is referred to the original publications. This is not a self-contained document either, but assumes a certain familiarity with computer vision, pattern recognition and machine learning problems.


Figure 1: Example scenes from the chosen model scenario of object detection in road traffic. This scenario covers a great variety of object classes to be detected: vehicles (A,B), pedestrians (C), road markings (D), traffic signs (E) and traffic lights (F). At the same time, the variation in scene types is considerable as well, which can mainly be distinguished by complexity: relatively simple highway traffic (B) and complex inner-city traffic (A,C,D,E,F).

### *1.1. Scientific context and model scenario*

The context of the research described here is the achievement of human-like intelligence in present-day computers using tools and insights from computer science, mathematics, biology and psychology. A particular focus of my work is intelligent **visual** environment perception, a task the difficulty of which is often underestimated as it is performed by each of us virtually flawlessly, every day of our lives. Only when we, as computer scientists, try to duplicate some of the functions humans perform routinely, we begin to see their intrinsic difficulty: after all, the last 30 years have seen intense academic and industrial research on environment perception in economically very relevant domains such as road traffic, surveillance or humanoid robotics, without producing solutions that come even close to human perceptual performance (see [129] for a recent very exhaustive comparative evaluation on pedestrian detection to this effect).

In fact, road traffic is an excellent scenario for research in intelligent environment perception: not only is it, scientifically speaking, very evident that some kind of fundamental breakthrough is required to close the gap between the current state of the art and human performance, but, at the same time, such a breakthrough would be of very high economic and societal relevance. After all, better safety systems could save hundreds of thousand of lives world-wide every year, which would otherwise be lost in traffic accidents, not to speak of a revolution in our mobility habits that could be triggered by cheap and universally available autonomous vehicles. For this reason, road traffic is often used as a model scenario in my research work, in particular road traffic as analyzed from a moving car. In such a scenario, especially in inner-city scenes, environment perception amounts to a great extent to the localization and recognition of objects, such as pedestrians, cyclists, traffic signs, road markings a.s.o., see Fig. 1. Real-time constraints usually apply, imposing upper bounds on the computational complexity of algorithms. As we are talking about a real-world scenario here, the fact of relying on physical sensors implies a deep involvement in aspects of signal- and image processing. Lastly, due to the complexity and the sheer volume of the data to be processed, machine learning must be intrinsically involved in this effort.

When we are interested in the principles of human-

Figure 2: Traffic sign detection as an example of object detection with diagnostic features. In this case, the feature in question is color, as the blue, yellow or red colors of the shown traffic signs give already a very good indication about their presence and position.

like intelligence, it is useful to limit the model scenario further by excluding the detection of all objects that are "simple" to detect, e.g., by having simple diagnostic features [168] defining them unambiguously, or almost so. An example for objects with a diagnostic feature are traffic signs which are, intentionally, characterized by a small set of forms and especially colors, see. Fig. 2, and which are therefore excluded from our considerations, as well as road markings. We are thus left with the detection of vehicles, pedestrians, the course of the road, sidewalks, (motor)cyclists a.s.o. While these object classes have properties that might facilitate the task of detecting them, there is no single, easy-to-compute visual property that would allow an unambiguous characterization.

### *1.2. Structure of the document*

By virtue of proper restriction, object detection in road traffic can therefore be considered a very difficult (and therefore very interesting) scientific problem. As it is at the same time a problem of high industrial and societal relevance, it becomes even more interesting to search for solutions. The goal of this document is to propose a road map towards a solution, based on system-oriented approaches and machine learning methods, and to summarize the most important steps I have already taken in this direction, as well as to outline future research efforts along the proposed road map. In the following section, I will first analyze the principal challenges associated with visual environment perception (and object detection in particular) in order to motivate the proposal of a road map for overcoming those challenges. Subsequently, I will outline my recent work in this direction in Secs. 3, 4, 5 and 6. Finally, the described works are discussed in the context of the road map in Sec. 7, and an outlook of potential future research efforts is given and motivated.

Figure 3: Uncontrolled environments prevent simple hypotheses: in the shown traffic situation, the usual assumption of a flat ground plane is no longer valid. If the ground plane hypothesis is used to give a priori indications about object positions, its application would even be harmful here, as the system would suddenly detect many objects that have no link at all with the true (curved) ground plane.


Figure 6: Example of occlusion in traffic scenes. This is a very common effect that essentially destroys a significant part of visual information about and object, and, what is worse, replaces it with potentially conflicting information.

## 2. Principal challenges to real-world object detection

**Uncontrolled scenario** Object detection in road traffic scenarios is faced with many challenges, most of which are linked to the uncontrolled nature of the environment in the sense that many influences (weather, light, other traffic participants, traffic rules) are totally outside the control of the experimenter, and therefore simplifying assumptions based on controlled environment properties cannot be made. For example: in indoor robotics, a flat ground-plane assumption is often made, but this is not necessarily true in traffic scenarios as the ground can be slanted or curved even on a small spatial scale, see Fig. 3.

**Illumination and weather** Another aspect that cannot be controlled is illumination and weather (see Fig. 4): it is intuitively clear that rain, low sun, fog or snow, or the change of daytime to night-time, will have a strong impact on all algorithms based exclusively on image processing. A very disturbing property of these effects is that they are often strongly non-linear, i.e., there will be no simple image transformation that corrects them: illumination changes are often inhomogeneous across the image, camera over-exposure corrupts individual pixel values in an irreversible fashion, and fog acts as a filter removing fine details that cannot be recovered.

**Unpredictable dynamics** Traffic scenes are highly dynamic, even without considering the motion of the ego-vehicle observing them. It is again very difficult to model the behavior of other traffic participants because it is based on traffic rules, infrastructure (traffic lights, barriers), and to no small extent by each driver's personal driving style. All this movement complicates the detection of objects of interest, again because simple methods based on a static-world assumption, or on the assumption of constant ego-movement, cannot be used.

**Visual ambiguities** Another issue concerns visual ambiguities: locally, the pixel pattern of a vehicle can be very similar to the pattern of a window on a building, and similar confusions of local pixel patterns can occur for pedestrians in particular (e.g., trees, street lights, traffic lights). This is illustrated in Fig. 5. This is a fundamental problem of object detection, but exacerbated by the uncontrolled nature of road traffic scenarios as conflicting patterns cannot simply be excluded from the scenario.

**Occlusions** A last challenge that occurs in many detection scenarios, not only in road traffic, is occlusion by other objects, see Fig. 6. This does not seem to represent a major problem for the human visual system [82], but poses problems for many current detection systems as essentially a large part of the information characterizing the object is not available.

### *2.1. State of the art for object detection in road traffic*

In general, object detection is an area so vast, depending on application scenarios and constraints, that is it quite impossible to list all relevant approaches, simply because what is relevant varies as a function of the intended application scenario.


Figure 4: Examples of the effect that time of day and weather can have on a visual image, the starting point for visual perception. The ideal climatic condition, an overcast sky without direct sunlight, is shown in I. In II, it can be observed what direct, low-standing sun can do in the way of shadows which are very detrimental to detection algorithms. In III, we observe that rain completely changes the appearance of a visual scene, with a reflective road surface that may even lead to "ghost detections", i.e., the detection of vehicles that are reflected on a wet road. In IV, the effect of night-time is illustrated. Contrary to intuition, the effect is not really dramatic as artificial illumination is usually homogeneous and diffuse and thus does not cause strong shadows. In V, the effects of fallen snow can be inspected. This case does not present extreme challenges to visual processing, but LIDAR signals are strongly affected due to the high reflectance of snow and ice.

Figure 5: Ambiguities in visual scene analysis. The left image shows a typical inner-city traffic situation with three modifications. The right image shows the nature of modifications: Three patterns were copied to other positions within the image. The fact that these changes go unnoticed is due to the similarity of these patterns to traffic objects. Please verify for yourself whether you could have found the changes!

When restricting this discussion to the model scenario outlined above, however, a halfway representative overview can be given, focusing on purely visual approaches, i.e., not considering the use of LIDAR or other sensors, on methods that are (in principle) real-time capable or close to it, and methods for detecting vehicles or pedestrians (see Fig. 1). In general, object detection methods can be grouped into "detection-by-recognition" and direct approaches. The latter implement algorithms for directly localizing objects of the desired class, whereas the former reduces detection to a classification or recognition problem by analyzing all potential object locations by a binary classifier that distinguishes objects (of the desired class) from non-objects ("background"), as sketched in Fig. 7. Virtually any object detection algorithm makes use of multi-scale processing, analyzing the visual image at multiple spatial scales. This is required,

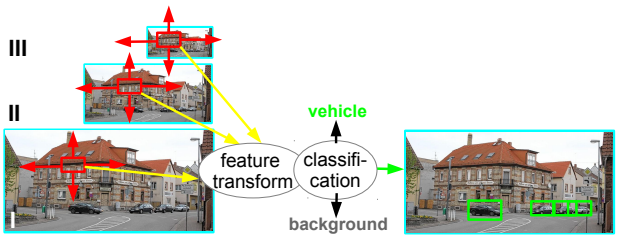


Figure 7: How a state-of-the-art sliding-window detector works, based on a vehicle detection scenario in road traffic. The original image (bottom left) is analyzed on $N = 3$ spatial scales denoted **I, II, III**. Analysis is carried out by a constant-size sliding window that assumes every position within each of the $N$ images. Window content at each position/scale is analyzed and each window is assigned an identity (vehicle or background). On the right-hand side, the results of this operation are schematically drawn into the image, where "detections" of the background class are not shown. Please note that, by having a window of constant size but different image sizes at each scale, this method effectively detects vehicles of different sizes.

since objects of the desired class need to be recognized regardless of their distance to the observer, which is usually strongly variable, and a multi-scale treatment transfers some of the signal variability thus introduced away from detection/classification models.

Object detection is an intrinsically ill-posed problem, as the space of possible objects, and above all backgrounds, is extremely large and cannot be described or even sampled completely. Especially for visual object detection, the approach for coping with this problem is twofold:

- **feature transform:** reduction of variability by image transformations. Often, image variability has a physical cause. For example, global illumination changes strongly affect image pixel values, just as rotations or translations of objects (or the observer) do. The goal of feature design is to determine image transformations that are *invariant* to some of these causes, thus significantly reducing signal variability.

- **machine learning:** the remaining variability is usually dealt with by machine learning techniques, which attempt to map the output of the feature transform step to statements about object class membership. Results depend strongly on the model complexity of the used algorithms.

Many visual objects seem to be strongly characterized by gradients: edges, corners and lines. At the same time, gradient information is insensitive to global illumination changes, or, if properly normalized, to local ones as well. In this way, a huge source of variability, namely variability due to illumination changes, is removed when computing image gradients. Therefore all relevant feature computation techniques are based on the calculation of image gradients, all the more so as gradients do not require significant computational resources. Notable feature computation techniques are histograms of oriented gradients (HOG) [20], local binary patterns (LBP) [123], aggregated channel features (ACF) [25], Haar wavelets based on integral images [176] and Canny-edges [23].

When discussing learning methods, the model scenario imposes constraints on these as well. Most feature computation techniques transform the image, or patches of it, into a representation of high dimensionality (outright or simply because an image has already high dimensionality). This requires learning algorithms capable of dealing with high sample dimensionality. Furthermore, as the problem is ill-posed, an enormous amount of samples will be necessary to achieve satisfactory performance, which privileges algorithms that have a benign scaling behavior w.r.t. the number of samples. These two constraints alone eliminate otherwise very powerful methods such as, e.g., Gaussian Processes [139] as they cannot deal with high data dimensions. Commonly used methods are therefore neural networks, boosting methods [155] and support vector machines (SVM) [9]. SVMs are in principle not very well suited for real-world detection because of their unfavorable scaling behavior for large sample numbers. However, as their generalization capability is excellent, the resulting very long training times are in practice often accepted or circumvented by parallelizing the training process using GPUs [15].

There are several systems, composed of feature transform and an associated learning algorithm, which are very frequently used for object detection in applied scenarios, including road traffic. Concern-

ing detection-by-recognition, there is the HOG+SVM approach [20] which performs a sliding window search using a linear support vector machine which is particularly efficient, operating on a HOG representation of the image. Another popular approach is the boosted cascade approach described in [176], which uses the extremely efficient and integer-based Haar wavelet feature transform coupled to a cascade of simple detectors that are arranged in a chain for performing detection-by-recognition. Object detection with higher spatial precision, and a certain tolerance to occlusions, is performed by the latent SVM method [33] that selects characteristic parts of objects and detects those objects through a detection of the parts. The most "extreme" detection-by-recognition approaches are convolutional neural networks [158], which skip the feature transform step, replacing it by a set of non-linear transformations that are learned from training data. There are direct approaches as well that have achieved a certain attention, most notably the implicit shape model (ISM) [103] which detects generic key points in the image that are subsequently classified. Then, each key point "votes" for a certain object identity, position and scale, and maxima in the space of votes are associated with object positions in a manner analogous to the Hough transform[81]. These systems constitute, more or less, the state of the art for object detection in road traffic, and any new approach needs to be compared to at least a subset of them in order to show its merit.

### *2.2. Problems with learning approaches*

While it seems that feature computation methods are both efficient and beneficial (in fact there is little fundamental dissension about the choice of features), we find several fundamental problems or at least inconveniences that are linked to learning methods, both for direct detection methods or detection-by-recognition approaches.

**Local ambiguities and occlusions:** These problems were already mentioned problems of object detection in general, see Fig. 5. However, they directly impact learning algorithms because they cannot be mitigated by whatever feature transform one chooses to use: if an object is occluded, its identity will have to be inferred based on less information, and if a local pattern is inherently ambiguous, it will stay ambiguous even after a feature transform step.

**Lack of probabilistic interpretation:** all of the approaches detailed above are based on discriminative learning [8]. This is due to the need for efficient algorithms in real-world object detection, and since discriminative methods solve a simpler task they are virtually always more efficient. They essentially decide if a sample is situated left or right of a given hyperplane, or, in other words, belongs to class "object" or class "background". A measure of confidence is not foreseen, or just by using heuristics (distance-to-hyperplane for SVMs [133], output magnitude for NNs [90]) that are not (and cannot be) derived from the theoretical framework of statistical learning theory [174]. This prevents reliable outlier detection and a mathematically well-founded (e.g., Bayesian) combination of processing results with other probabilistic information.

**Lack of training data:** First of all, since all learning is statistical [174], and the problem is fundamentally ill-posed (see above), any learning algorithm requires enormous amounts of training samples. However, the creation of these samples ("annotations" or "ground-truth") cannot be fully automated for the time being, and always requires some human participation, making a large-scale annotation of video sequences difficult and, above all, expensive. After all, training samples must be consistently and correctly annotated across multiple videos while introducing as little "annotation bias" as possible, which requires the checking and re-checking of already created samples. As a consequence, even the largest available databases [49, 26, 67, 29] for vehicle and pedestrian detection contain just a few $10^5$ samples which often belong to a much smaller group of actually distinct persons or vehicles. This does not seem to be sufficient to cover all relevant cases, and indeed vehicle and pedestrian detection using these bases for training is far away from human perceptual performance [24].

**Lack of incremental and life-long learning capacity:** a common practice to improve the power of trainable detection algorithms consists in re-training them with their own incorrect detections, often in multiple iterations, see Fig. 8. As each it-

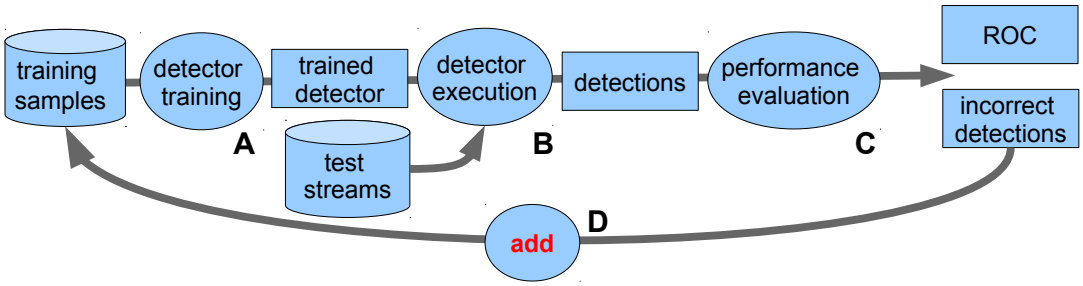


Figure 8: Illustration of detector training by repeated bootstrapping. An initial detector is trained using a database of annotated samples, and this detector is successively re-trained by its own incorrect detections on test data. In this way, very large amounts of samples are added to the original database. For non-incremental learning algorithms, this implies a complete re-training at each iteration of the bootstrapping process.

eration of this so-called *bootstrapping* process adds a non-negligible number of samples to the already huge databases of training samples, training times get progressively higher and higher, even in the case of good scaling behavior of the used learning algorithms. At some point, this makes a sensible parameter search, which has to be conducted on top of the bootstrapping process, virtually impossible because all learning algorithms considered here require complete re-training (with existing plus additional samples) if they need to refine a trained model. It would be far more efficient to re-train models with new samples only, giving a huge speed-up in practicability and efficiency and, more generally, permitting a classifier to be adapted and continuously extended by new data over long periods of time ("life-long learning").

### *2.3. Road map*

The objective of this introduction has been to elucidate that many challenges in real-world object detection, using the chosen model scenario, are in fact tied to problems of the underlying machine learning approaches. As signal processing techniques for pretreating images are well developed, it seems that no fundamental break-throughs can be achieved by enhancing those techniques. The conclusion is therefore rather simple: if we wish to improve object detection, we need to adapt the way we use machine learning techniques, or the techniques themselves. In this section, I will outline the most important steps to be implemented in new processing architectures, or in enhanced learning methods if we really wish to push visual perception and cognition to new levels of performance.

Up until this point, the discussions were strictly scientifically neutral: while I believe that the existence of the aforementioned issues in visual perception cannot be doubted since they are rigorously backed by citations of other researchers' results, their significance and their impact is of course subject to discussion. In contrast to this, the following road map is my own, very subjective view, developed during my years both in industrial and academic research institutions, on how one could go about improving visual perception in the chosen model scenario. A large part of my scientific work, as it will be presented in this document, has been conducted within the scope of this road map, making it a very appropriate means to show the intrinsic coherence of my research efforts.

**Inclusion of context information** As the existence of local ambiguities (see Fig. 5) is a property of the used signals rather than of the used learning methods, it cannot be remedied simply by improving learning methods. Instead, the proposed solution consists in harnessing additional sources of information that may provide disambiguation, and integrate this with the visual information obtained from the image. Such additional sources of information may be the geometric layout of scene elements (road, sidewalk, ground plane), information about object tra-

jectories, localized information about previous object detections, and in general anything that might conceivably resolve perceptual ambiguities. Please refer to Fig. 21 for a visualization of these concepts. In addition, it must be understood that ambiguity can arise not only from ambiguous signals but also from an insufficiently trained model. As detection problems are fundamentally ill-posed and insufficiently described by training samples (see above), the latter can occur quite often for difficult problems. Then, this is just another case of ambiguity that can be resolved by context information, which thus not only reduces local ambiguity but enhances generalization performance as well.

**Generative learning methods** In order to achieve a true probabilistic interpretation of, e.g., the output of a classifier trained using machine learning techniques, generative learning methods need to replace discriminative ones. A valid probabilistic interpretation is particularly useful for including context information (previous paragraph) or other information sources in a theoretically sound way. In particular, the class of prototype-based generative learning methods [59] is attractive here as it imposes only a weak model bias as opposed to parametric generative methods, is comparatively efficient, and can be trained efficiently in an incremental fashion (see next paragraph).

**Incremental learning** To address life-long learning, incremental learning methods are required. More precisely, we require methods that receive their training samples one by one, without knowing their number in advance, and whose performance degrades as little as possible when the statistical distribution of samples changes over time. This allows to continuously extend learned models over long time periods, and facilitates the bootstrapping process described previously (see Fig. 8).

**Multi-modal learning architectures** A simple solution for generating more training samples is to re-think the supervised training procedure that is almost exclusively used in machine learning approaches for applied perception. More to the point, a strategy similar to co-training [2] might be beneficial, trying to characterize suitable "foreground" samples by finding stable mappings to other perceptual sources. Put differently: an interesting object is a percept whose features can be stably mapped to the space of another sensor. While this characterization leaves a lot of details to be defined, it could be implemented in a purely unsupervised way, paving the ground for the generation of an unlimited number of training samples for machine learning. It is true that the quality of these samples might be less perfect than if they were created by a human, but one may argue that the greater number of samples will offset their inferior quality.

## 3. Application of machine learning techniques

I have conducted numerous studies that use machine learning techniques to solve real-world problems, mainly in the area of vehicle detection and classification [73, 54, 112, 156, 64, 51, 57, 63, 65], pedestrian detection and classification [47, 101, 66, 58], driver behavior prediction [48, 65], robotic object recognition [14, 13, 35] and human-machine interaction [88, 89, 90, 92, 94, 91, 93]. By and large, the techniques used were mainly neural networks and support vector machines, combined with problem-dependent and sometimes tailor-made feature transforms to facilitate classification and promote invariances. This has given me a very good insight into the potential but also the problems and limitations of conventional machine learning approaches, which I will discuss for all the works described here. They constitute a representative subset of my work in this area, and I have selected them as they are most clearly suited to illustrate the point I want to make in this document: where the limits of conventional machine learning really lie *in practice*.

### *3.1. 2D/3D fusion for object recognition in mobile robots*

In this body of work [14, 13, 35], a neural network based fusion approach for real-time robotic object recognition was investigated, which integrates 2D and 3D descriptors in a flexible way. The recognition architecture is coupled to a real-time segmentation step based on 3D data, since a focus of the investigations was real-world operation on a mobile robot. As recognition must operate on imperfect segmentation results, tests of recognition performance using complex everyday objects were conducted in order to quantify the overall gain of performing 2D/3D fusion, and to discover where it is particularly useful. A main result is that the fusion approach is most powerful when generalization is required, for example to significant viewpoint changes and a large number of object categories, and that a perfect segmentation is apparently not a necessary prerequisite for successful discrimination.

In this section I will only consider the steps concerned with machine learning techniques, and omit some (quite interesting and complex) other steps, namely the real-time segmentation of the 3D point cloud into disjunct 3D object candidates.

*Motivation and context*

With the advent of cheap 3D sensing technology for indoor applications, the question immediately arose of how to use 3D information to improve the performance of environment perception, and notably object recognition. 3D sensors are not (or less) affected by perspective and occlusion/overlap effects, whereas they do not provide any non-geometric details of the objects they observe, such as color or texture. This richer information, however, is provided by normal cameras, so the idea of combining the best of both sensor technologies seems very promising. Since 3D sensors are able to measure depth at high precision and resolution, it becomes possible to adopt a two-stage approach as depicted in Fig. 9, where a generic segmentation step (i.e., independent of the identity of segmented objects) is coupled to a classification stage

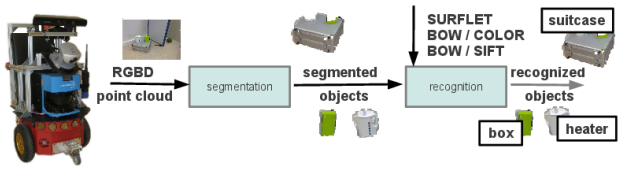


Figure 9: Implemented processing architecture for robotic 2D/3D object recognition as a block diagram.

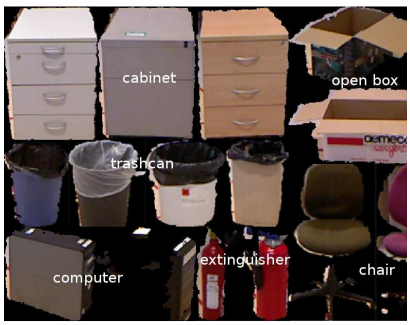


Figure 10: Example image of the objects used in the experiments on robotic 2D/3D object recognition.

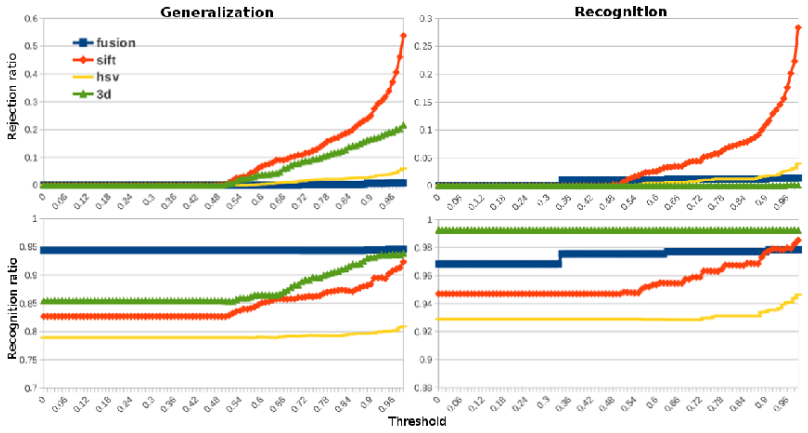


Figure 11: Results of the 3-classes experiments for robotic 2D/3D object recognition. Upper row: rejection ratio (rejected good classifications / good classifications). The rejection ratio is 1 for a threshold value of 1. Lower row: recognition ratio (accepted good classifications / accepted classifications). The recognition ratio is undefined for a threshold of 1.

that determines object identity. This is in contrast to sliding-window approaches (see Sec. 3.3) which have a priori no possibility to identify object candidates with any reasonable degree of certainty, and which therefore have to process the whole image, leading to higher computation times and error rates.

*Related work*

The most commonly used 2D object descriptors are SIFT [105] and SURF [6] which offer convenient invariance properties at a reasonable (SIFT) or optimized (SURF) computational cost. There is a large number of other approaches which will be not reviewed here as the focus is on the fusion of multimodal information.

The recognition of objects from three-dimensional data is well studied in the literature, see [12, 110] for surveys. All proposed methods expect a segmented object candidate which should be matched against templates in a database of previously registered objects. At the most fundamental level, proposed methods can be grouped into holistic and local approaches. Of the former, a prominent example is iterative closest point estimation (ICP) [190], which can match object candidates to templates if a rough alignment between the perceived object and at least one template exists. If one restricts recognition to specific object types known in advance, the Generalized Hough Transform can be a useful tool, especially for simple objects like cylinders or spheres, see, e.g., [137]. If the object class is unknown in advance, more general methods for the holistic description of objects need to be used: a prominent example is [180] where histograms of normal orientations between randomly chosen point pairs in the object candidate are computed, resulting in a holistic descriptor of object shape. Another notable holistic approach [130] attempts to find constant object signatures in views of object candidates that were taken from different directions.

In contrast to this, local approaches are rather similar to SIFT or SURF methods in RGB images, trying to find a small number of distinctive "key points" on an object candidate, whose associated descriptors can be matched against stored templates. Such methods, while potentially very powerful, are often sensitive to viewpoint changes and occlusions, so care must be taken when constructing a suitable database. A few recent and notable approaches include [165, 154, 18, 17], based on different local descriptions of object shape, for the most part based on local surface curvature or the local distribution of normals.

*System structure and methods*

Fig. 9 shows the basic architecture of the system that was implemented on a mobile indoor robot of the PIONEER type, see [13]. Once object candidates are determined, each one is analyzed based on visual information using a bag-of-word technique based on the SIFT descriptor and a HSV color space histogram for color information, as well as 3D information using a so-called surflet-pair histogram which describes holistic form. To achieve fusion for object recognition, all of this information is passed to a multilayer perception (MLP) which is trained on a large self-created database of 20 different types of everyday household objects (see Fig. 10) that are encountered by the robot, and the contribution of fusing 2D and 3D information is evaluated.

*Results and machine learning issues*

We conducted two experiments: in one we trained and evaluated recognition performance using the same viewing angles of objects. In the second experiment, we trained recognition on viewing angles that were different than those used during evaluation in order to assess generalization performance. In Fig. 11, we observe that 3D information alone is sufficient and even slightly superior to the fusion approach when training and evaluating recognition under the same conditions (viewing angle). However, when attempting to generalize to unseen viewing angles, the fusion approach performs much better than 3D information alone.

From this, we can first of all confirm the problems evoked in Sec. 2, principally the lack of training data. Generalizing to unseen views is a very common problem in robotics, and it stands to reason that a set of training samples will never contain all relevant viewing angles, with a sufficiently fine angular resolution, and in all possible appearance variations. This shows that object recognition in such an application context is an ill-defined task that is not sufficiently well specified by a training database. Is is therefore an absolute necessity to take into account additional sensor information that might compensate for such variations in appearance and viewing angle in order to obtain better generalization performance in practice. As we could see from the presented investigation, even the inclusion of very rudimentary information in a pretty simple way is already strongly beneficial w.r.t. generalization.

### 3.2. Gesture classification for human-machine interaction

Another important long-term project [88, 89, 90, 92, 94, 91, 93] was the recognition of static hand postures and dynamic hand gestures by a learning-based approach. This is opposed to methods who have a skeleton model and try to infer the skeleton configuration from observations to determine the state of a hand. The basic task is to distinguish the ten hand postures shown in Fig. 13 with high accuracy. In the course of this project, numerous aspects of this problem were investigated: the best choice of feature transform for 3D data [88], the fusion of information from two 3D sensors observing the hand from different viewpoints [89, 90], the creation of a very large benchmark database for training and validation [92], neural network architectures and optimal exploitation of all available information[94, 91], real-time implementation in a car [88] and user studies concerning the feasibility of this system for human-machine interaction [93].

*Motivation and context*

The context of this project is primarily automotive HMI, e.g., the control of an in-car infotainment system without distracting the driver (see Fig. 14). However, the subject has wider implications in the domain of human-machine-interaction in general, and particularly for robotics where natural interaction with a human is a high priority. The automotive application context imposes real-time and cost constraints, which is why a low-cost 3D sensor was chosen (see Fig. 12) and algorithms were selected based on their computational efficiency.

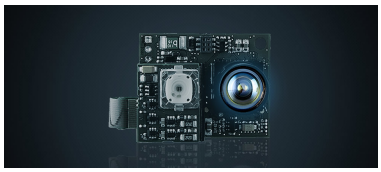

Figure 12: The used time-of-flight 3D sensor for gesture classification, the Camboard Nano. Due to its small dimensions and cheap price, it is well suited for automotive applications.

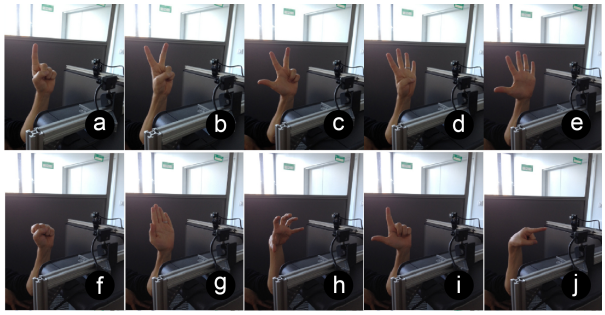


Figure 13: The hand posture database for gesture classification. From left to right, top to bottom: *ONE*, *TWO*, *THREE*, *FOUR*, *FIVE*, *FIST*, *FLAT*, *GRAB*, *PINCH*, *POINT*

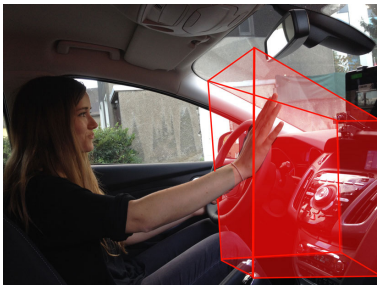

Figure 14: Demo setup of our infotainment system realized on a tablet using gesture classification. The VOI (in red) denotes the area sensible to user input.

*Related work*

Depth sensors represent an easy and robust solution for recognizing hand postures, as they can easily deal with the segmentation of the hand/arm from the body by simple thresholding as described in [125]. Moreover, it is possible to make use of depth information to distinguish ambiguous hand postures [87]. Usually a good performance is achieved with a very limited set of postures, or if designed for a specific application [164]. In [84] a single ToF-Sensor is used to detect hand postures with the Viewpoint Feature Histogram.

In general, ToF-Sensors suffer from a low resolution which makes it difficult to extract robust yet informative features. Improved results can be achieved when fusing Stereo Cameras with Depth Sensors, e.g. in [183]. Therefore, various approaches make use of the Kinect sensor's ability to extract depth and RGB

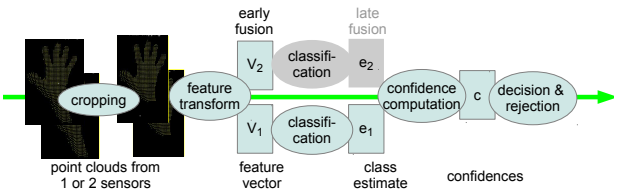


Figure 15: Overall structure of the hand gesture and posture recognition system. The choice of early or late fusion depends on application constraints: either to concatenate the feature vectors from each sensor and classify the combined vector, or to classify each feature vector individually and combine the results.

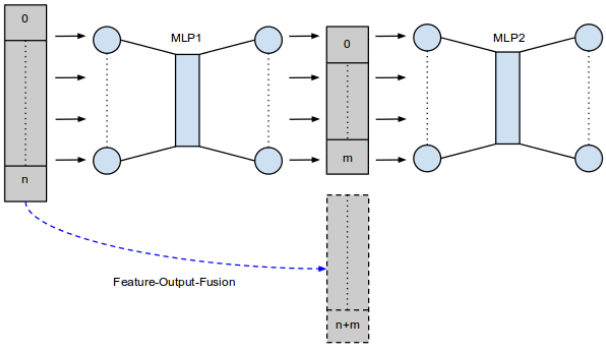


Figure 16: Illustrating the stacked MLP technique for two MLPs (MLP1 and MLP2): MLP1 is always directly trained with the samples from the training data set. A sample, represented by a feature vector of length $n$, is fed into MLP1's input layer. After MLP1 has propagated the input and calculated each neuron's activation in the output layer, MLP2 is trained on the output vector of size $m$. Optionally, the output vector is fused with the feature vector itself, forming the new input of size $m + n$.

data simultaneously [170]. although approaches using the Kinect sensor will always suffer from direct illumination by sunlight, which is not the case for ToF-sensors. The authors of [122] equally make use of the Kinect sensor's ability to acquire RGB and depth data simultaneously albeit using a hand model as a basis for hand posture recognition. In [144], a case study is made of how the Kinect sensor can be used to control E-Mail functions in a car through set of six hand gestures, although the gesture set remains small and the effect of different lighting conditions is not discussed. More comprehensive overviews are given in [141] and [179] and a good overview of automotive HMI is given in [132], with [191] describing advantages in user acceptance of in-air hand gestures in comparison to touch gestures.

*System structure and methods*

The block structure of the system implemented in the course of the project is shown in Fig. 15. Several variations are possible and have been explored in the individual contributions to this project. Processing can be based on one or two identical sensors. If there are two sensors, their contributions can be fused by simply concatenating their respective feature vectors (early fusion), or by combining the class estimates from each sensor into a single confidence measure (late fusion). Furthermore, classification can be performed by conventional multi-layer perceptrons (MLPs), by stacked or context-sensitive MLPs, or else by multi-class support vector machines (SVMs) using an one-vs-one aggregation strategy [1]. Concerning stacked MLPs, we explored a variety of strategies to provide an MLP with more useful information: basically, a second MLP was trained using the class estimates of the first, optionally enhanced by the original feature vector of a sample (see Fig. 16). Context-sensitive MLPs are similarly trained with the class estimates of another MLP, only this time the estimates come from the same MLP one time step before. This allows to take the preceding decisions into account, reflecting the assumption that a posture is something that is rather stable in time, and if a posture decision has been confidently computed, it is rather likely that it will be identical for the next sample. Again, various strategies for achieving this were explored.

*Results and machine learning issues*

Results showed that the task is a difficult one that in addition poses serious generalization issues. When performing normal N-fold cross-validation without taking into account which person (out of the $\tilde{2}0$ persons that helped to create the database) samples are coming from, it is possible to achieve approximately 90% accuracy. However, when performing

| person | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MLP1 | 83 | 49 | 69 | 58 | 79 | 62 | 57 | 68 | 84 | 70 | 89 | 75 | 90 | 74 | 90 | 80 | 73.94 |
| MLP2 | 86 | 52 | 74 | 63 | 83 | 65 | 60 | 72 | 89 | 74 | 90 | 75 | 98 | 75 | 93 | 82 | 76.94 |

Table 1: Generalization results for gesture classification (all 16 persons), comparing conventional MLP (MLP1, first row) with a stacked one (MLP2, second row). Shown are generalization result in percent for each person (represented by a single row), plus the average performance over all persons in the last row.

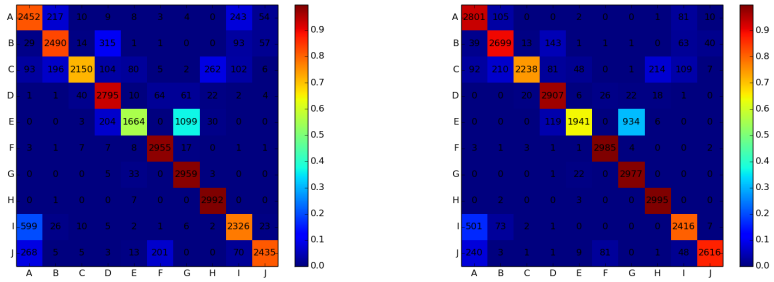

Figure 17: Comparison of normal MLP classification (left) to classification with temporal fusion (right) for the hand posture recognition task. Shown is a confusion matrix for the 10-class task that has therefore size 10x10. It demonstrates the (moderate) improvements for nearly all classes when applying the temporal fusion approach. Evaluation is performed on person 13, that is, classifiers are trained on all persons except person 13, and then tested on person 13.

a kind of leave-one-out training strategy based on persons, training on all persons save one and testing on the remaining one, we obtain (averaged over all persons) a strongly inferior performance of about 72%, see Tab. 3.2, fusion approaches pushing this result to 76%, see Fig. 17. To my mind, this approximates the true generalization performance much better since samples coming from the same person are bound to be correlated. Thus, only a measure that is based on persons will measure generalization to unknown persons, which is after all that which is desired in practice. Although using a second sensor proved to be very helpful combined with simple late fusion strategies that are computationally very efficient, it was found that the second sensor imposed too strong a computational burden for real-time operation, mainly due to the additional point cloud processing and feature transform involved. Very much in line with the reasoning in Sec. 2, I encountered the following points stemming from the chosen approach to machine learning:

- **Lack of training data** even a very large database of approximately 600.000 samples, coming from 20 persons and containing ten posture classes, does not define the problem uniquely: generalization to unseen persons is still less-than-perfect. This has very probably nothing to do with MLPs but with the fact that a purely statistical approach to learning that requires many samples.
- **Context information improves generalization** The integration of additional (or contextual) information, be it temporal (preceding detected posture) or contextual (signal from a second sensor) strongly increases performance and generalization, see, e.g., Fig. 17 which shows the effect of including temporal information.
- **Lack of probabilistic interpretation is problematic** Since MLPs are not generative methods, they do not provide true class membership probabilities. Fusion with contextual or other information, as it was successfully demonstrated in [90], will therefore always rely on

heuristics that are validated only by the fact that they seem to work, but not by theoretically sound means.

- **incremental learning could be helpful** Training times could be strongly reduced if every new person's samples were added incrementally instead of a complete retraining with all samples. Even given the benign scaling behavior of NNs, training and model selection were observed to take a very long time.

### 3.3. *Real-time pedestrian detection and pose classification*

In this contribution [47], which was conducted in the context of an industrial collaboration with Honda Research Institute USA, Inc., we presented a real-time pedestrian detection and pose classification system for road traffic scenarios, which makes use of the computing power of Graphical Processing Units (GPUs). The aim of pose classification is to determine the orientation and thus the likely future movement of a pedestrian, which can increase the prediction horizon for safety applications. Evaluation focuses on pose detection performance and shows that, without resorting to complex tracking or attention mechanism, a small number of safety-relevant pedestrian poses can be reliably distinguished during live operation. Additionally, we showed that detection and pose classification can share the same visual low-level features, achieving a very high frame rate at high image resolutions using only off-the-shelf hardware.

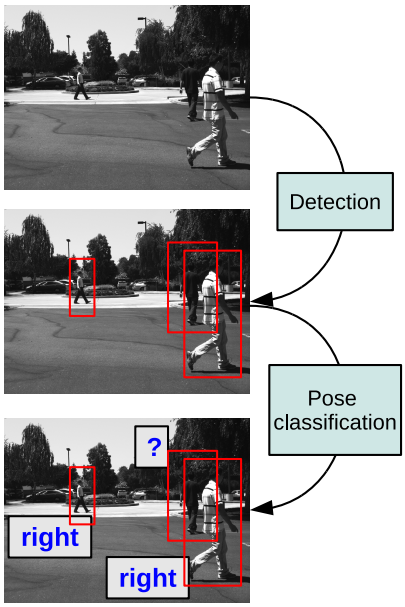


Figure 18: Block diagram of the real-time pedestrian detection/pose classification system, which is composed of a detection and a pose classification stage.

*Motivation and context*

Accidents involving pedestrians in inner-city environments are frequently fatal, even at relatively low driving speeds. Indeed, pedestrians have no protection in case of impact and are thus highly vulnerable. The goal of pedestrian detection by intelligent vehicles is, for the most part, inspired by safety considerations: if pedestrians can be detected in time, collisions might be avoided.

Inner-city scenes can be extraordinary complex, and they require the driver to focus his attention on the parts of the scene he (subconsciously) finds relevant. This prioritization has its drawbacks as the driver might simply miss something important. If the driver should fail to react, or react too late, to the appearance of a pedestrian, a Driver Assistance System could warn him about the situation, or even initiate autonomous braking. However, this requires that the system is able to robustly localize pedestrians. Pose classification takes this consideration even further as it allows, under certain conditions, to estimate a pedestrian's next actions. For this, even a small number of pose categories may be sufficient ("front view", "back view", "facing right" and "facing left"). A reliable pose classification system can be used to focus attention on a pedestrian that might cross the road even if the pedestrian is not, at the moment, in the vehicle's path.

*Related work*

The issue of pose classification has been raised by several authors (e.g. [44, 30, 102, 160, 19]), mainly in the context of road traffic and surveillance. Due to the real-time nature of our approach, we are interested in the distinction of a small number of behaviorally relevant pose *categories* (see [44, 30]) that allow a guess at a pedestrian future behavior. This is different from the determination of a precise geometric pose, i.e. the heading in a 3D space, as described in [102, 75] which is, in addition, hard to reconcile with real-time constraints. Our approach made no use of tracking as demonstrated in [189, 160], as we wanted to achieve first a sufficient performance on single-frame pose classification.

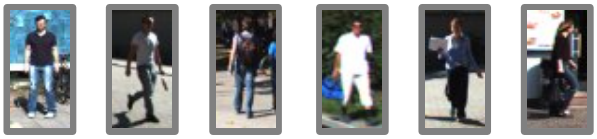

Figure 19: Some examples of the pose classes used in the pedestrian pose classification task: *front*, *back*, *left* and *right*

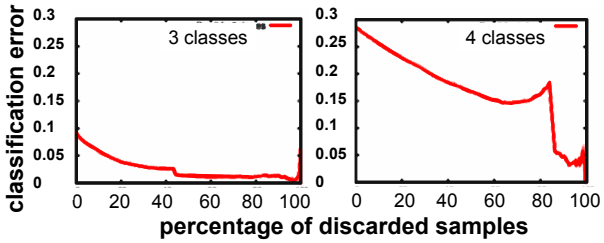


Figure 20: Results for the pedestrian pose classification using 3 (left) and 4 (right) pose categories. Shown is the overall pose classification error depending on the applied rejection threshold.

*System structure and learning methods*

Differently from the approach pursued in [47], we focus here on the machine learning aspects of the pedestrian detection and pose classification task. Several important constraints are imposed due to the system structure as depicted in Fig. 18. First of all, it must be stressed that the detection stage solves an extremely difficult and ill-defined problem: that of distinguishing pedestrian patterns from background patterns, which is subject to most of the problems evoked in Sec. 2. Please see Fig. 7 for a rough explanation of the working of this detector, which is based on the support vector machine (SVM) model of classification. Further training samples are needed for the training of the pose classification stage of Fig. 18, in this case about 20.000. Pose classification is a multi-class problem (as there are four pose categories, see Fig. 19) which is implemented using a so-called "one-against-all" aggregation strategy that combines the outputs of $N$ binary SVM classifiers.

*Results and machine learning issues*

Several issues mentioned in Sec. 2 were made very clear by the presented work on pedestrian detection and pose classification:

| | Predicted Classes | | | |
|---|---|---|---|---|
| | 60 | 10 | 23 | 7 |
| Real | 5 | 83 | 0 | 12 |
| class | 13 | 5 | 82 | 0 |
| | 10 | 37 | 0 | 53 |

Table 2: Experimental results of pedestrian pose classification using 4 pose categories. From left column to right column (or first row to last row), the categories are *right*, *front*, *left* and *back*.

| | Predicted Classes | | |
|---|---|---|---|
| | 60 | 20 | 20 |
| Real | 3 | 97 | 0 |
| class | 13 | 5 | 84 |

Table 3: Experimental results of pedestrian pose classification using 3 pose categories. From left column to right column (or first row to last row), the categories are *right*, *front or back* and *left* (the results in the last row do not add up to 100 because of rounding errors).

- **Ambiguity** As can be seen, the pose classification problem can be solved with different success depending on the organization of categories: if one chooses to distinguish 4 categories (see Fig. 19), then the problem is difficult (70% accuracy) due to the high similarity between the *back* and *front* poses, which is shown by the confusion matrix of Tab. 2. This is again a manifestation of the ambiguity of local visual patterns stated in Sec. 2, and is therefore unlikely to be solved by analyzing local visual patterns only. If one chooses to group the two offending classes into a single one, the problem becomes markedly easier (91% accuracy), see Tab. 3.

- **Lack of probabilistic interpretation** Another issue I encountered was the fact that SVMs do not provide truly probabilistic class membership probabilities. There are post-processing techniques such as Platt scaling[133] which normalize SVM scores (distances to the separating hyperplane) to look like probabilities, but they do not change the fact that the output of a discriminative classifier such as an SVM is fundamentally unrelated to probability. However, two crucial steps of the described work on pose clas-

sification rely on this assumption: the combination of several binary SVM classifiers into a single multi-class one, and the realization of a "reject" option, where the classifier decides based on normalized SVM output scores that a sample cannot be confidently assigned any of the classes in question. Again, as in Sec. 3.2, the only justification for these ad hoc steps is that they seem to work for the moment (for the efficiency of the reject option, see Fig. 20), which is dangerous especially in safety-critical functions.

- **Scalability and lack of incremental learning capacity** Lastly, one finds that the initial problem of pedestrian detection is a very difficult one, requiring extremely many training samples to achieve decent performance, in this case about 200.000, whose acquisition is extremely time-consuming, and the internal consistency of which is nearly impossible to ensure. Due to the large number of samples, the used SVM classifiers take a very long time to train (in the order of weeks), an effort that has to be repeated every time new SVM hyper-parameters are selected, or new training samples are obtained. In this way, the bad (cubic) scaling behavior of SVMs w.r.t. sample number, together with their incapacity to perform incremental learning, renders the task of training classifiers very nearly impossible in this scenario[1].

[1] They were nevertheless chosen here for simplicity of implementation, and because it is well established that they work well together with HOG features

## 4. System-level learning for context-based object detection

In this section, I will give an overview of the wide field of context-based object detection and describe some of my own contributions to this fields [58, 64, 56, 53, 66, 14]. As outlined in Sec. 1, I will focus on the model scenario of object detection in road traffic, such as vehicle detection and pedestrian detection. Nevertheless it should be stressed that the described systems are completely generic and make no assumptions, implicit or explicit, that would forbid their applications to other object detection tasks.

### *4.1. Structure of the section*

Initially, a concise and non-technical introduction to the subject will be given in Sec. 4.2. Subsequently, the basic mathematical modelling tools, namely probabilistic computations using Bayes' rule, will be introduced in the context of object detection problems in Sec. 4.3. An overview over related scientific work will be given in Sec. 4.4. In Secs. 4.5 and 4.6, I will describe two prototypical projects [64, 58] I undertook regarding context-based object detection. As they are related quite closely, I will discuss their implication for object detection and machine learning, within the larger context of this document, in a single discussion section, Sec. 4.7.

### *4.2. What is context-based object detection?*

The basic idea behind context-based object detection is that a visual object is not simply defined by a local visual (pixel) pattern, but also by its embedding into an environment, often generally referred to as "context". The characterization of such an embedding is of a probabilistic nature and can thus serve both to infer the presence or absence of a certain object. Typical statements about object-context relations (see Fig. 21) are

- hairdryers are mostly found in bathrooms
- camels are mostly found in desert landscapes but rarely in jungles
- pedestrians are rarely seen on highways but often in city traffic
- vehicles are always found close to the ground

Figure 21: Examples for the strong relations between objects and context: Left, middle: a camel and a hairdryer where we expect to see them: in the desert and in the bathroom. Right: an out-of-context example of a car that is not close to the ground. This is immediately noticed by our visual system and considered as uncanny.

These probabilistic example statements suggest that such relationships can be learned, and it stands to reason that this is what human beings actually do. Such a type of learning is rather different from conventional machine learning tasks that are working with raw or weakly processed signals, at least in most applied domains such as robotics or road traffic. Here, by contrast, the basic quantities are symbolic (classes of objects, types of rooms) or at least of a much more abstract nature than the original signals. For this type of learning, I have consistently used the term "system-level learning" (SLL) to underscore that it, on the one hand, operates on a level that is far removed from the level of signals, and, on the other hand, that it links (sub)symbolic information coming from very different modal parts of a complex processing system. The models that are learned on the system-level are termed "context models".

In object detection, context models can be used for disambiguation purposes which occur due to inherently ambiguous signals or insufficiently trained detectors (see Sec. 2). A good example for this is the detection of vehicles by night (or when it is raining heavily): usually, detectors are not exhaustively trained with such samples, and it is assumed that the feature transform stage will remove variations w.r.t. to the training data. This is, however, never fully the case, and thus one can only hope that a trained detector will generalize sufficiently to handle this problem. A better solution would be a suitable combination of

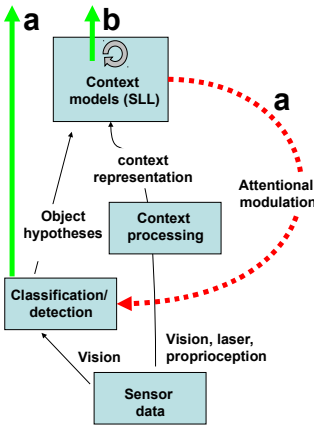


Figure 22: Two different ways to use context models: attentional modulation of low-level processing (case **a**) or purely high-level hypothesis selection (case **b**). In case **a** (attentional modulation), context models act directly on the object detection module, which is why resulting object hypotheses are directly read out from there, representing the combination of detector and context model. In case **b**, context models act as an additional classifier (or filter) and eliminate incorrect hypotheses provided by the detector, which is now operating without influence from context models.

detector and context models such that uncertainty in one component can be compensated for by the other component. This is a viable strategy as it implies the combination of (approximately) independent information, so the likelihood of simultaneous ambiguities is rather low. To stay with the example of generalizing vehicle detection between daytime and night-time, a context model expressing the fact that vehicles are close to the ground plane, which is usually detected using stereo or LIDAR data, will be completely unaffected by day/night changes. Thus a low-confidence detection that is however **close to the road plane** could be sufficiently boosted to be accepted. We thus see that context models address the issue of visual ambiguities but also, to a significant extent, the issue of generalizing to situations not (precisely) covered by training data.

### *4.3. Methods and approaches*

As for the question of how to integrate context models with a visual detector, there are fundamentally two possibilities, see Fig. 22. One way is to use context models as a post-processing filter by letting them eliminate those detection results which are incompatible with current scene context, which was shown in [56]. This option has the advantage of simplicity, but suffers from the fact that it can only reject detections. In other words: anything **not** detected by the detector is lost for good, like an unusual-looking car at night, receiving a detection score that is just too small to be considered. As it is on the road and close to the ground, context models could infer that it is a car but they do not get the chance.

Therefore the second option is to use context models for directly influencing the detection stage [64]. This approach is usually termed "attentional modulation" or "visual attention", and is strongly inspired by findings in psychology and neurobiology indicating that mammalian brains rely quite heavily on this option. Here, not only does there exist the possibility to reject detections that are inconsistent with context (or rather, prevent them before they are made), but one may also create additional detections if they are very consistent with context models. This is strongly reminiscent of probabilistic computations using Bayes' law, where a "likelihood" (like a detection result) is combined with a "prior probability" (e.g., derived from context models) to form a posterior probability distribution that can be used for decision making [171]. This deserves to be treated a bit more in-depth because it is a central cornerstone for many arguments in this and the following section:

Following [171], we suppose that an object $\vec{O}$ has, e.g., a certain size $\sigma$, a certain position $\vec{x}$, a speed $\vec{s}$ and an identity $o$: $\vec{O} = \{\vec{x}, \sigma, o, \vec{s}, \dots\}$. Assuming that the entirety of pixels in an image, or features extracted from those pixels, is contained in the vector $\vec{v}$, the goal of an object detector is to obtain $\vec{O}$ from $\vec{v}$, i.e., to infer the distribution $P(\vec{O}|\vec{v})$. Virtually all common detection schemes operate on *local* patterns $\vec{v}_L$ only and ignore the remaining non-local (contextual) information $\vec{v}_C$:

$$P(\vec{O}|\vec{v}) = P(\vec{O}|\vec{v}_L) \approx P(\vec{v}_L|\vec{O})P(\vec{O}) \qquad (1)$$

If we wish to take into account the contextual information $\vec{v}_C$, the calculation has to be performed in a

slightly more complex way:

$$\begin{aligned} P(\vec{O}|\vec{v}) &= P(\vec{O}|\vec{v}_L\vec{v}_C) = \\ &= \frac{P(\vec{O}\vec{v}_L\vec{v}_C)}{P(\vec{v}_L\vec{v}_C)} = \\ &= \frac{P(\vec{O}\vec{v}_L\vec{v}_C)}{P(\vec{v}_L|\vec{v}_C)P(\vec{v}_C)}\frac{P(\vec{O}\vec{v}_C)}{P(\vec{O}\vec{v}_C)} = \\ &= \frac{P(\vec{v}_L|\vec{O}\vec{v}_C)}{P(\vec{v}_L|\vec{v}_C)}P(\vec{O}|\vec{v}_C) \end{aligned} \quad (2)$$

Here, $\vec{v}_C$ may contain any information complementary to the local visual pattern: image pixels not in the vicinity of an object, past images, past trajectory measurements, the time of day, a.s.o. By supposing that local features do not depend on non-local ones, this can usually be simplified to

$$\begin{aligned} P(\vec{O}|\vec{v}) &= \frac{P(\vec{v}_L|\vec{O}\vec{v}_C)}{P(\vec{v}_L|\vec{v}_C)}p(\vec{O}|\vec{v}_C) \quad = \\ &= \frac{P(\vec{v}_L|\vec{O})}{P(\vec{v}_L)}p(\vec{O}|\vec{v}_C) \quad \approx \\ &\approx P(\vec{v}_L|\vec{O})P(\vec{O}|\vec{v}_C) \end{aligned} \quad (3)$$

since the term in the denominator $P(\vec{v}_L)$ does not depend upon $\vec{O}$ and can thus be disregarded for maximum determination. Eqn.(3) is an important simplification because it amounts computationally just to the multiplication of two probability distributions, the "likelihood" $P(\vec{v}_L|\vec{O}))$ coming from the detector and the "context prior" $P(\vec{O}|\vec{v}_C)$ coming from context models. The only issue with this equation is that detectors are usually formulated to approximate the probability $P(\vec{O}|\vec{v}_L)$, and as most detection algorithms are discriminative models, the required likelihood is impossible to obtain. If we had a generative algorithm at our disposal for detection, we could work directly with eqn.(3), since generative algorithms can provide a probability distribution over (local) patterns from which sampling is possible. In the absence of generative detection methods, one works often with more crude approximations of eqn.(3), or with a much more radically simplifying starting point: if we assume that context models and detection methods can be treated as independent "classifiers", we can simply model the probability of detecting an object as

$$P(\vec{O}|\vec{v}) = \underbrace{P(\vec{O}|\vec{v_L})}_{\text{detector}} \underbrace{P(\vec{O}|\vec{v}_C)}_{\text{context models}}$$

It is this equation, which can by appropriate assumptions be obtained from eqn. (3), which inspires all of my modelling work described in this section, to a large extent because it is very compatible with insights into biological processes of visual attention. Another important reason is that it is very efficient to implement because it just uses quantities that are already computed even when discriminative detectors are used, which means that the inclusion of context information into visual object detection can be extremely efficient if detection and context models are efficient themselves.

The only downside (which inspired my works described in Sec. 5) is that the quantities one obtains both from discriminative detectors and context models are usually not real probabilities (most prominently: they are not normalized), and thus neither Bayes' rule nor other laws of probability are, strictly speaking, applicable. However, a major insight is that this does not seem to matter a great deal, and very significant performance increases can be obtained by applying eqn. (3) even in a naive fashion.

### *4.4. Related work*

*Computational modelling of static visual attention.* Recent work treats visual attention as a kind of Bayesian inference process where the "attention prior" is combined with a likelihood term arising from a detection module [171, 145, 131]. Whereas a spatial attention prior can be easily expressed in a probabilistic form, the detection scores coming from a real-world object detector generally need to be "converted" to probabilities which is not always straightforward and involves a complicated calibration process [131, 133]. A strictly feature-based attention model was proposed by [80]. It focuses on feedforward processing and lateral competition, either in the form of center-surround filtering or explicit competition mechanisms. This model was applied in numerous real-world scenarios, e.g., [78], for goal-driven scene analysis [119] or fast object detection

and recognition[181]. While the work described in [119] employs high-level semantic models of object-to-object or object-to-goal relations to guide visual attention to behaviorally important locations, these models are specified by a designer and not acquired through learning. The work of [181] couples an exhaustive object detection mechanism to signal-driven saliency with beneficial results. The work of [111] focuses on car detection in road traffic scenarios, whereas the similar VOCUS model [41] targets mobile robotics applications. Both approaches use an offline optimization procedure to generate feature-based object search templates based on small numbers of training samples. These templates are fused with a bottom-up attention signal similar to [80] such that both visual saliency as well as proximity to the search template may trigger object detection. The coupling of object detection and contextual information mediated by low-level modulation is demonstrated in [115] where context information about the "gist", i.e., a low-dimensional description of a scene, is used to infer the locations of relevant objects in images by statistical models constructed from training examples. The concept of gist is taken further in [76] where a generic probabilistic model of 3D scene layout is proposed that can be queried for likely image locations of, e.g., cars or pedestrians in order to inform an exhaustive local object detector. Object detection may not only be guided by global scene properties, but also by other objects in the scene: in [22], a discriminative model of local object-to-object interaction is proposed that formalizes cooperation and competition between local detections of multiple object classes and gives a probabilistic interpretation of this process. Lastly, object detection may also be regarded as an active process in which the performed gaze actions (i.e., object detections) should maximize information acquisition. Based on the saliency map approach of [80], a POMDP formalism is used in [177, 178] to optimize gaze target selection based on the detections arising from previous gaze targets, visual saliency and global scene priors.

*Dynamic visual attention.* Most of the previously described approaches to visual attention are guided by static local image properties [79], sometimes by static spatial context [171]. Even if non-static image features, such as local motion, are used [41, 113], such attention mechanisms are always reactive in the sense that they guide attention towards the detected features but do not anticipate future events. Recent psychophysical work [70] however reveals that humans learn highly precise dynamic models *predicting* the movement of objects, and that such predictions are used to guide eye movements to the predicted locations ahead of time. This predictive mechanism is shown to permit the visual pursuit of highly dynamic objects, such as squash balls, with the very limited amount of fixations per second that can be realized by the human visual system. The closest corresponding function in technical systems is *tracking*, i.e., the trajectory analysis and pursuit of moving objects. This is a crucial functionality in robotics, or in road traffic, where the performance of single-image-based detectors must be augmented by taking into account the temporal evolution of the trajectory. Especially in pedestrian detection applications [28, 30, 45, 128, 20, 25, 34, 7, 43], this is a very necessary component mostly realized by Kalman filtering [107, 117, 157, 162, 10, 188, 46, 187]. The issue with tracking algorithms, as they are commonly used in the cited works, is that they perform a kind of "late fusion" as a post-processing step of detection. In this way, they interpolate between the present detection and a prediction derived from a pre-specified motion model. However, when the motion model is violated, e.g., because an object changes direction, the fusion with the incorrect motion model causes "ghost" detections that deviate considerably from the true object position until the motion model is updated. In case of noisy detection scenarios, such an update can take several seconds, resulting in a considerable time interval in which potential safety applications (e.g., emergency braking) receive incorrect data. There is no work I know of that makes use of object-centered dynamic attention mechanisms, as described here, to influence its own detections and thus avoiding the problem of ghost detections, except potentially [46].

*4.5. Static scene context for vehicle detection*

The work described here was born out of a long-term project in vehicle detection [156], to which I contributed the integration of (static) context information in various forms, see [64, 56, 53]. Here, I present this large-scale hierarchical system while paying particular attention to the fusion of "bottom-up" detection results with "top-down" attentional modulation obtained from context models. A particular focus is the learning of context models, and their "translation" into an image-wide attentional modulation signal, to be combined with the dense confidence map provided by a detector that analyzes local visual patterns.

*System structure*

The system structure is roughly given by Fig. 22 and follows the variant **a)** indicated there for including context models. Context models are situated at the highest hierarchy level (termed *system level*), taking information from an intermediate level of generating and processing object hypotheses (*hypothesis level*) which in turn receives low-level signals in the form of dense retinotopic maps from the *preprocessing level*. In the following, several modules of the system are described in more detail:

*The detector.* The appearance-based detector [186] generates object hypotheses in two successive steps. As a first step, it generates scale-specific retinotopic confidence maps as described in [185] by a sliding-window approach (see Fig. 7), thus making it a detection-by-recognition approach, see Sec. 2. Each

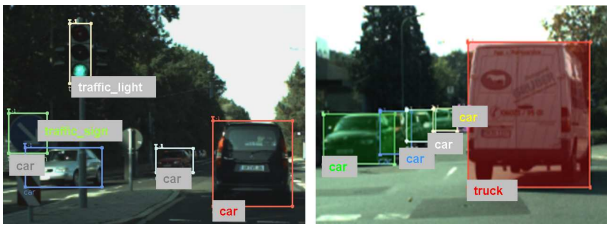


Figure 23: Examples of recorded streams and annotated information for benchmarking context-based object detection. Each annotation consists of a rectangular area, an identity and an occlusion value (not shown).

Figure 24: Example of how the appearance-based classifier works in the context-based object detection architecture: for each input image (**a**), it produces a retinotopic, dense map of vehicle confidences (**b**). This means that each pixel in this map represents the presence or absence of a vehicle, at a specific scale, centered on that pixel.

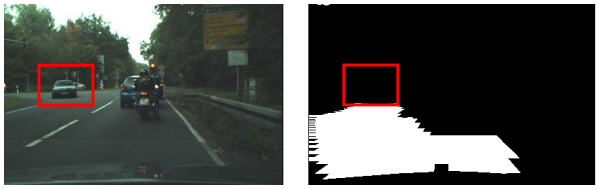

Figure 25: Performance example of free-area computation. **a)** Video image **b)** computed free area as a binary mask.

pixel of a confidence map represents the likelihood of detecting a specific view of an object (in this case: back-views of cars) at a specific position and scale in the image, see Fig. 24. In a second step, object hypotheses are generated from these confidence maps by a competitive selection process. Details about processing and detector training are given in [52].

*Free-area computation.* The *free area* is defined as the obstacle-free forward area visually similar to a road and carries significant semantic information. Since it is, by construction, bounded by all obstacles that the car might collide with, many relevant obstacles are close to the boundaries of the free area. For the purposes of the presented system, the quantity of interest is therefore the *distance* (in pixels) of an object hypothesis to the free area. Details of free-area calculation are given in [52], see also Fig. 25.

*Distance and elevation computations.* Dense stereo processing is employed for measuring the 3D position of image pixels in car-centered coordinates. For

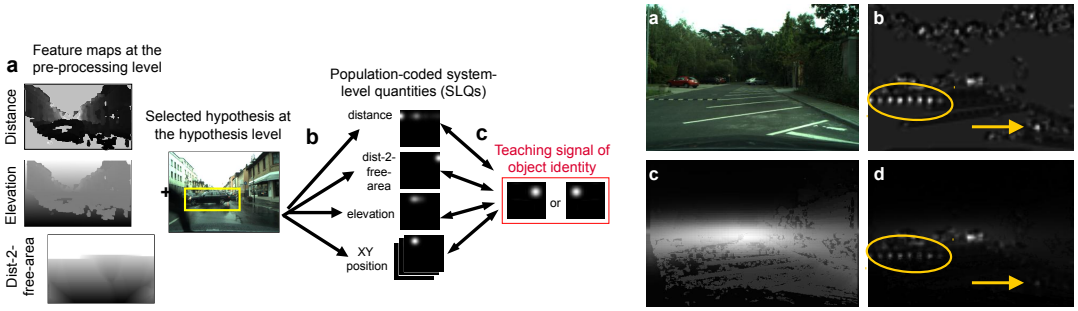


Figure 27: System-level learning of object models (left) and attentional modulation obtained thereby (right) in the context-based object detection architecture. The left-hand side diagram shows where system-level learning is situated in the complete processing system which first computes dense feature maps (preprocessing level), then selects an object hypothesis and computes system-level quantities (SLQs) in population-coded format (hypothesis level), and finally learns the mapping between object identity and SLQs (system level). A particular focus of system-level learning is the expected distribution of all SLQs given a certain object identity, which is then "translated back" into an attentional modulation map. This process is illustrated on the right-hand side: a) original image b) dense vehicle confidence map provided by detector c) attentional modulation obtained from system-level learning d) fused vehicle confidence map. It can be observed that the false detections pointed out by the arrow and the ellipse are strongly attenuated in the final vehicle confidence map.

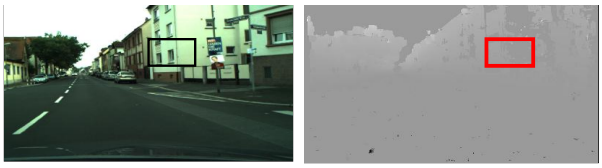

Figure 26: Examples of stereo processing for elevation and distance calculation in the context-based object detection architecture. **a)** video image **b)** dense elevation map (brighter pixels are higher over the ground plane).

obtaining hints about the identity of objects, such measurements are helpful but not optimal: it is not really the height relative to a car-centered coordinate system that carries semantic information, but rather the height over the ground plane. The ground plane is therefore estimated from stereo processing results, and a dense elevation map is created where each pixel encodes its height over the ground plane, see Fig. 26. Details are given in [52].

*Attentional modulation*

Learning of context models occurs at the system level of the architecture as depicted in Fig. 22. At this stage, all relevant quantities (termed system-level quantities - SLQs) related to an object hypothesis, such as distance, elevation, distance-to-free-area, position etc., have been computed. To facilitate the application of efficient linear learning methods, all SLQs are converted to population codes (or basis function representations [136]). I could show the computational advantage of this transformation in a related work on context-based object detection [56]. Linear regression is used to learn relations between population-coded object identity ("vehicle" or "background") and individual SLQs, see Fig. 27 (left). By learning the reverse mapping, one may obtain an *expected value distribution* for each SLQ, given that object identity is "vehicle". This distribution is then translated back into the image, mainly by means of histogram back-projection [166]. In this process, values in the dense feature map associated with an SLQ are replaced by the probability of those values in the expected value distribution. Properly fused and normalized, these modulation maps are then multiplicatively combined with detection confidences as shown in Fig. 27 (right).

*Experiments and results*

Extensive experiments were conducted to show that the inclusion of static scene context information, by means of system-level learning, can significantly improve the performance of visual vehicle detection

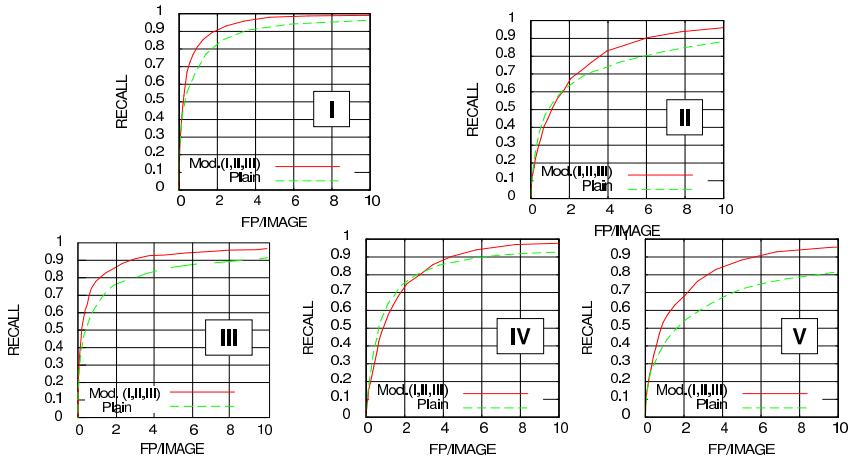


Figure 28: Performance improvement by attentional modulation for video streams I-V by ROC-like plots. Curves with more surface under them indicate better quality. The solid red/dashed green curves show performance with/without attentional modulation. System-level models were trained on parts of streams I-III not used for this evaluation. A clear overall improvement can be observed for all streams, especially in the application-relevant areas of high recall.

and, by plausible inference, visual object detection in general. To this end, we recorded five distinct video streams covering a significant range of traffic, environment and weather conditions, please see Fig. 4 for a visual impression. For the quantitative evaluation of object detection performance, we manually annotated relevant objects in the recorded video streams, please see Fig. 23 for details. Results are given by ROC analysis, where the detection threshold of the detector is varied and the two performance indicators of incorrect detections and percentage of correctly detected vehicles are plotted against each other, which is a very usual measure in object detection scenarios [129]. As the results of Fig. 28 clearly indicate, attentional modulation has a strong beneficial effect on detection performance. Improvements are strongest for unfavorable environment conditions, where the generalization ability of pure pattern-based detection reaches its limits.

### *4.6. Dynamic attention priors for pedestrian detection*

In addition to using context models derived from static quantities, I generalized this idea to include dynamic sources of context information such as an object's own position and trajectory, an idea I proposed and validated in [58]. Here, a predictive attention mechanism similar in spirit to [70] attempts to generate a probability distribution for an object's estimated position in the *next* image, which is subsequently combined with detection results from the next image. Differently from my previous work, context models are not learned here but arise from a predefined linear motion model implemented by a probabilistic *tracking algorithm*, a so-called *particle filter tracker* [4]. The distribution thus obtained I termed *dynamic attention prior* (DAP), and its contribution to performance and generalization was tested on a challenging visual pedestrian detection task, where the inclusion of DAPs works much in the same way as the inclusion of attentional modulation in 4.5. A particular point of this endeavour is to show that technical systems, in this case a state-of-the-art pedestrian detector, can profit from context models with little changes or performance overhead.

*System structure*

The processing structure of the system realizing DAPs (see Fig. 29) is very similar to that of Sec. 4.5 and implements option **a)** from Fig. 22, that is, direct modulation of detection by attentional modulation. It includes a pattern-based detector and a tracking algorithm generating predictions and thereby DAPs, which are fed back to the detector at the following time step.

*The detector.* For the pattern-based pedestrian detector, the HOG+SVM [20] sliding window detector technique (see Fig. 7) is employed, which transforms an input image into a set of scale-dependent confidence maps, see fig. 30. Each individual confidence value is considered a detection if it exceeds a so-called detection threshold. These confidence maps are modulated by the DAPs provided by the tracking algorithm, a mechanism that can "push" certain confidences, which would have otherwise been disregarded, beyond the threshold. A visualization of DAPs is given in Fig. 32.

*The tracker and DAP generation.* Tracking extrapolates a trajectory from successive object detections that is locally as linear as possible, see Fig. 31. The conceptual beauty here is that the used particle filter algorithm is intrinsically probabilistic, and thus a good choice for representing the context-based object likelihood $P(\vec{O}|\vec{v}_C)$ of eqn. (3). Each distinct object is tracked separately, and for each object a separate DAP, derived from the extrapolated trajectory, is applied to detection confidence maps as shown in Fig. 32. Given that detection confidences are not probabilistic themselves, being neither normalized nor bounded, a direct implementation of eqn. (3) is not possible and is replaced by an approximate implementation where an explicit normalization of confidences is performed based on ad hoc assumptions about their bounds.

*Experiments and results*

Evaluations are performed on 11 annotated short video sequences of single moving pedestrians, see

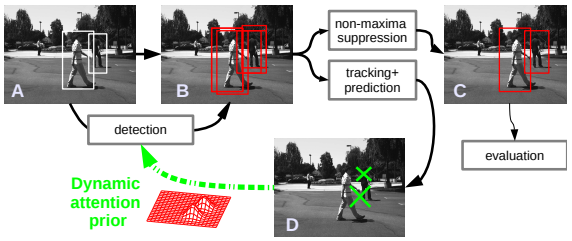


Figure 29: Block structure of the real-time pedestrian detection system enhanced by dynamic attention priors. **A** Original image, white boxes show pedestrians that are to be found. The pedestrian left of the center is too small to be detected and is thus excluded here. **B** Detections, indicated by red boxes, resulting from sliding window classification. **C** Results of non-maxima suppression (NMS) removing overlapping detections. These detections can be considered the final detection result and are passed to evaluation. **D** Predictions generated from past detections. Prediction centers and sizes are indicated by green crosses of varying size, and serve as the sources for dynamic attention priors (green dashed arrow) modulating the detection process.

Figure 30: Topologically arranged pedestrian confidence maps as computed by the HOG+SVM method from the camera image (left). Confidence maps are computed on increasingly coarse spatial scales (from left to right), which allows to detect pedestrians of different sizes.

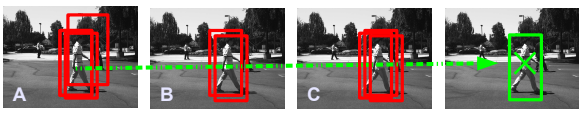


Figure 31: Illustration of the object tracking approach chosen for generating DAPs: detections of the recent past are used to compute a trajectory under a local linearity assumption. Based on this trajectory, object motion can be extrapolated to the immediate future.

Fig. 33 for a visual impression. Compared are "bottom-up" (without DAP integration) and "top-down" performance. The results, some of which are shown in Fig. 34, demonstrate that in the worst case, DAPs cause no performance degradation, and in the best case a huge performance gain. It was also investigated how DAPs affected the system when they were applied at the wrong place, which happens always when pedestrians perform abrupt direction changes (e.g., when they turn around or start walking from a standing start) so the prediction is momentarily incorrect. It was found that the effect of DAPs must always be tuned to be slight, so that they create detections only at positions where there is significant evidence from the detector. In the reverse case when DAPs are strong, "ghost" detections may be created at positions with little evidence from the detector. These detections are then fed to the tracking algorithm and influence its predictions, making them self-sustained in the absence of evidence. However, the strength of DAP feedback is governed by a single scalar parameter, and it was found that it is always possible to find appropriate values (essentially by ROC analysis varying this parameter) while

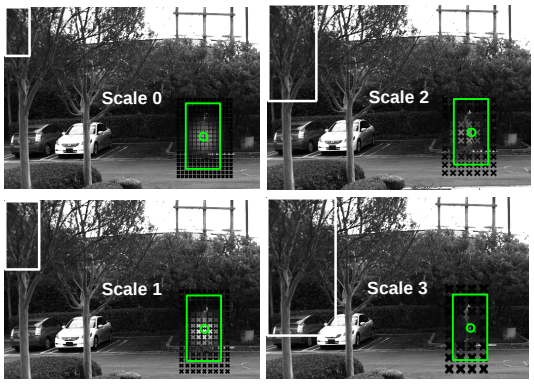


Figure 32: Multi-scale modification of detection confidences by DAPs. The green circle indicates the prediction center and the green box its associated scale (which is also predicted by tracking!). Grey crosses represent the positions of detection scores (sliding window centers) at each scale, the level of brightness indicating the strength of the boost each one receives from DAPs. The white box in the top-left corner of each image indicates sliding window size at that particular scale. As the pedestrian is predicted roughly at scale 1 (here, white and green rectangles have similar size), the scores at scale 1 get boosted more strongly than at other scales. At scale 3 no significant boost takes places any longer. Please note that only detection scores around the pedestrian are shown, in reality the whole image is densely covered at each scale.

improving detection performance significantly.

Figure 33: Example images from evaluation streams. Background and pedestrian identity and clothing vary strongly between video streams.

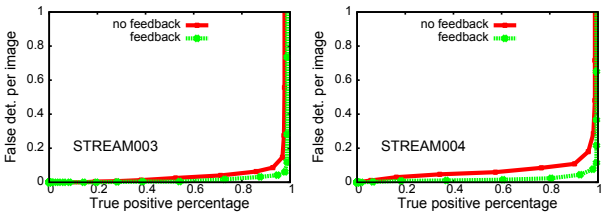


Figure 34: Showing the effects of DAPs on pedestrian detection performance by ROC-like plots. Red solid curves show the baseline detection performance without dynamic attention priors, green dashed curves show the top-down performance. Hint: A ROC-like plot is "better" than another one if it is consistently below the other.

*4.7. Discussion and machine learning implications*

- **Context information improves generalization** As the contributions described in this section have shown, the inclusion of context information is a particularly efficient way to boost the performance of object detectors, especially in terms of generalization. In practice, it was always found that the integration of a second, approximately independent information source, such as dynamic/static context information, improves generalization behavior of the machine learning models employed for object detection. Specifically, detectors became much less sensitive to changes in background, or to partial occlusions, or to changes in size which would normally degrade detection confidence in an unpredictable manner, often leading to an erratic "flickering" behavior of object detectors that is highly undesirable. These undesirable effects are due to lack of training data, or a difference between application and training scenario which comes down to the same problem: training data do not fully cover the current application scenario. The presented works have convincingly shown that the inclusion of context information effectively counteracts this problem.

- **Probabilistic interpretation** From a pure machine learning point of view, it is quite unnatural to integrate context information into discriminative detection methods as it is done in the presented works on an ad hoc basis. This cannot really be justified except by the fact that it seems to work quite well in the shown applications, which was already the case for different task presented in Sec. 3. Nevertheless, a desirable property of future context-based object detection systems is a generative detection mechanism whose outputs can be linked to probabilities (more concretely the likelihood $P(\vec{v}_L|\vec{O})$) and which can therefore be coherently combined with probability distributions $P(\vec{0}|\vec{v}_C)$ from context models.

- **System-level learning** A particular point concerns the learning of context models: it is found that learning methods are very easy to apply at this (sub-)symbolic level of a processing system, rather than at the level of signals (pixels) as it is conventionally done. At signal level, dimensionality and variation are much higher and therefore pose more of a challenge to learning techniques. In contrast, the system-level quantities (SLQs) processed by context models are often low-dimensional, allowing fast and efficient with simple methods that incur virtually no computational cost. The modulation of a detector, be it learned or imposed as in the case of DAPs, is equally computationally cheap and conceptually simple, underscoring the potential of such methods to operate in systems with real-time processing constraints.

## 5. Generative and multi-modal learning with the PROPRE architecture

Motivated by the work on context-based object detection described in Sec. 4, I grew interested in learning methods that might be intrinsically suited for this purpose. In particular, and using the notation from Sec. 4.3, I realized the necessity for learning methods that:

- are capable of outlier detection, in other words expressing the intrinsic probability of a visual pattern $P(\vec{v}_L)$
- are generative and can express the probability $P(\vec{v}_L|O)$ required, e.g., in eqn.(4)
- work for high data dimensions and sample counts, as encountered in perceptual problems
- are open-ended, or more precisely, incremental (see Sec. 6 for more precise definitions), that is, allow a continuous and long-lasting readaptation to new input statistics, in line with the road map put forward in Sec. 2.3

This severely limits the choice of learning methods: while some recent methods like deep belief networks [74] can be said to be generative methods, and can clearly handle large sample dimensionalities and numbers, they lack incremental learning capacity. Other approaches, such as Gaussian process models [139] are generative but not incremental, and in addition fail completely for sample dimensionalities typically encountered in vision problems. Parametric methods such as used in [171] for context-based object recognition also face problems for high sample dimensionalities, and it is unclear how they could be made incremental.

To address these issues, I focused my investigations on so-called *prototype-based approaches*: here, the probability distribution in data space is not expressed in parametric form but by a learned set of samples, the so-called *prototypes*. Prototype-based machine learning methods were originally motivated by prototype theory from cognitive psychology (see, e.g., [149]) which claims that semantic categories in the human mind are represented by a set of specific

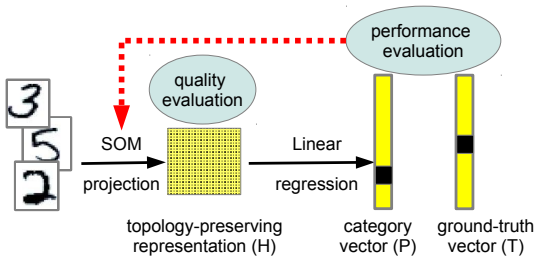


Figure 35: The generative, prototype-based PROPRE learning architecture in its most basic form.

examples (or prototypes) for these categories. Typical prototype-based approaches are the learning vector quantization (LVQ) model [83], the RBF model [114] or the self-organizing map (SOM) model [86]. A very popular prototype-based method in computer vision is particle filtering [4], where a continuous, evolving probability density function is described and updated as a set of prototypes (here denoted *particles*) whose local density represents local probability density. After some time, I decided to focus on an RBF-like model whose basic structure is shown in Fig. 35, composed of a modified SOM algorithm, coupled to a linear or logistic regression step interpreting hidden layer activities, this "read-out" being represented in the output layer $P$. This model [54, 55, 100] is termed PROPRE (short for projection-prediction) and comes in several variations concerning the precise details of the architecture, depending on the functionality that is to be achieved. In this chapter I will present some of the most interesting of these variants, but for the moment I will focus on the potential of PROPRE to be used in perceptual tasks like object recognition and detection, in the light of what has been discussed in Sec. 4.

First of all, prototype density in the internal layer H is related to probability density $P(\vec{v}_L)$ in data space [146] due to the properties of SOM learning, which is needed for outlier detection (as stated in Sec. 2). Furthermore, it can express a conditional estimate for the object identity of a pattern, $(O|\vec{v})$ in its read-out layer $P$ given a correct learning algorithm for readout. By inverse regression from $P$ to $H$, it can express the probability $P(\vec{v}_L|O)$ which is required, e.g., by context-based object detection methods, see Sec. 4. These links will be explored

more rigorously in the following text:

If we denote a prototype vector associated with unit $i$ in the internal representation $H$ by $p_i$, we can define the probability that this unit is the best-matching unit (BMU), i.e., the unit having the prototype closest to the input $\vec{v}_L$, as $P(\vec{p}_i)$. Starting from $P(\vec{v}_L)$ and using the law of total probability, we can write

$$P(\vec{v}_L) = \sum_i P(\vec{v}_L|\vec{p}_i)P(\vec{p}_i), \tag{4}$$

with $P(\vec{v}_L|\vec{p}_i)$ denoting the probability of pattern $\vec{v}_L$ given that prototype $\vec{p}_i$ is BMU. If we know, or can estimate, how patterns are distributed around prototypes, we can re-express this as

$$P(\vec{v}_L) = \sum_i P(\vec{v}_L|\vec{p}_i)P(\vec{p}_i) \equiv \sum_i a_i \nu_i, \tag{5}$$

where $a_i$ now expresses the "activity" of the unit associated with prototype $\vec{p}_i$, and $\nu_i$ is the probability that an unit becomes BMU which can easily be estimated. We observe that the probability $P(\vec{v}_L|\vec{p}_i)$ now acts as a kind of "activation function" for the unit associated to prototype $\vec{p}_i$: if it were, for example, uniform in the Voronoi volume $V_i$ of the prototype $\vec{p}_i$, we would obtain

$$a_i(\vec{v}_L) \sim \begin{cases} \frac{1}{V_i} & \text{if } \vec{v}_L \text{ in Voronoi zone of } \vec{p}_i \\ 0 & \text{else} \end{cases} \tag{6}$$

With the same assumptions and definitions, we can re-formulate the probability $P(\vec{v}_L|o_j)$, meaning the probability of observing pattern $\vec{v}_L$ given the object identity $o_j$, as:

$$P(\vec{v}_L|o_j) = \sum_i P(\vec{v}_L|\vec{p}_i)P(\vec{p}_i|o_j) \equiv \sum_i a_i \bar{w}_{ji} \tag{7}$$

where the term $\bar{w}_{ji} \equiv P(\vec{p}_i|o_j)$ can be obtained from the weights of an inverse regression, from ground-truth $G$ to internal layer $H$.

Expressions for the probability of object identity $o_j$ given a pattern $\vec{v}_L$ can be derived in a similar fashion:

$$\begin{aligned} P(o_j|\vec{v}_L) &= \sum_i P(o_j|\vec{p}_i)P(\vec{p}_i|\vec{v}_L) = \\ &\equiv \sum_i w_{ij} P(\vec{p}_i|\vec{v}_L). \end{aligned} \tag{8}$$

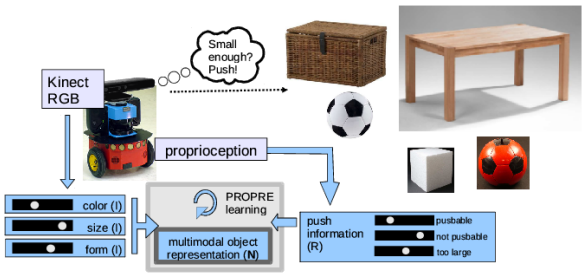


Figure 36: Illustration of the robotic model task devised for testing the bootstrapping of representations, based on the mobile robot currently in use at our lab. Based on a Kinect sensor, we obtain a rough segmentation of objects which is the basis for the subsequent determination of simple object properties. Each detected object segment is approached and an attempt to push it is made. Based on proprioceptive information, it can be judged whether the attempt is successful or not. This information, or the information that no attempt to push is currently being made, is entered into the reference representation R on the right-hand side.

When we suppose that all units are equally likely to be BMU, making the probability $P(\vec{p}_i)$ a constant, we can simplify this further to give

$$\begin{aligned} P(o_j|\vec{v}_L) &= \sum_i w_{ij} P(\vec{v}_L|\vec{p}_i) = \\ &= \sum_i w_{ij} \frac{P(\vec{p}_i|\vec{v}_L)P(\vec{p}_i)}{P(\vec{v}_L)} \equiv \\ &\equiv \sum_i w_{ij} \frac{a_i P(\vec{p}_i)}{P(\vec{v}_L)} \approx \sum_i w_{ij} a_i \end{aligned} \tag{9}$$

which again gives a very nice, simple expression in which the weights $w_{ij}$ are derived from forward regression, from internal representation $H$ to ground-truth $G$.

This treatment is concise but nevertheless shows clearly that the PROPRE architecture, being built essentially on the prototype-based SOM algorithm, gives direct access to the statistical quantities we are interested in, which are useful for a large variety of perceptual tasks. Specifically, it can address outlier detection (by evaluating $\vec{v}_L$) and compute sample likelihoods that are needed when integrating context information as in eqn. (4), and all this using extremely simple expressions that can be efficiently

computed as linear products. In light of my recent work on incremental learning (see Sec. 6), I am more actively trying to map the various entities in the PROPRE architecture to probabilistic interpretations.

### 5.1. *Developmental learning aspects and bootstrapping of representations*

What I was exploring in the early works on PROPRE [54] was the possibility of having a generative, potentially incremental model that could be guided (or modulated, to retain the biological analogy) to represent only those parts of the data space that have statistical links to another data flow. Guidance should be weak, and coming from sensory signals or quantities easily derivable from such signals. Thus, my idea was to have a self-structuring *bootstrapping* process, where the formation of a first abstract (object) representation would be modulated by very simple stimuli (e.g., pain), whereas the formed abstract representation might in turn modulate the formation of ever more refined representations.

*Architecture and model task*

To provide a starting point for experimentation, I designed a simulated robotic model task shown in Fig. 36. The idea was that a robot should develop a simple, multi-modal representation of objects that can (or cannot) be pushed, simply by interacting freely with its environment and attempting to push random objects. This task is reflected in the architecture (see Fig. 37) that is slightly adapted as compared to the "generic" architecture of Fig. 35. The fact of whether an object is pushable, not pushable or too large is entered into a dedicated *reference representation*, which in turn guides the concept formation process in the internal layer $H$ by *predictability*. Thus, the data flow from $H$ to $P$ as shown in Fig. 35 is reversed; it is now the reference representation $R$ that tries to predict activity in the internal representation

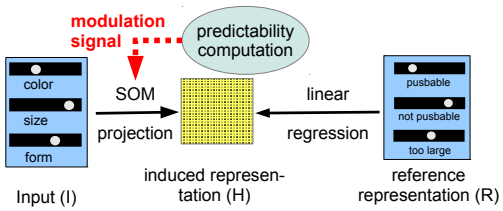


Figure 37: Block schema of the PROPRE learning architecture applied to the robotic model task.

$H$, leading to the so-called *reference-based predictions* which are used to compute a quantitative measure of predictability. Essentially, learning promotes those prototypes whose associated neural "activity" can be reliably predicted from $R$.

*Related work*

The focus on predictability as a guiding principle for learning was motivated by a conceptual work [121], arguing that symbolic quantities should be diverse on the one hand, and on the other hand be defined by their *power to predict* other quantities. My work differs from [121] in that it focuses on quantities that *can be predicted*, and in that a concrete algorithm is proposed and evaluated. Predictability is attractive because it can be naturally incorporated into local learning algorithms, whereas using predictive power necessarily involves bi-directional, non-local operations. Conceptually very close to this PROPRE variant is the *predictive coding* model originally proposed by [138] and elaborated by, e.g., [42, 118]. As in our PROPRE algorithm, predictive coding implements bi-directional learning between a receptive-field-generating process and a prediction process where receptive field generation is influenced by predictability. However there are important differences: Most basically, predictive coding was originally proposed to model observed single-neuron data, whereas PROPRE is intended for online use in robotic agents which implies a certain ease-of-use (no pre-whitening, no stability controls on learning), computational efficiency and robustness to noise. Furthermore, predictions in the predictive coding model always arise from higher hierarchy levels of the *same* processing stream, whereas PROPRE allows multimodality as it imposes no constraints on reference representations. The price for this flexibility is that, in contrast to predictive coding, PROPRE is not rigorously derived from a probabilistic model.

*Experiments and results*

Experiments are conducted with artificial input stimuli coming in the three classes "pushable", "not pushable" and "too large", as shown in Fig. 38. These stimuli are generated randomly but within well-defined parameter ranges for each class, with the

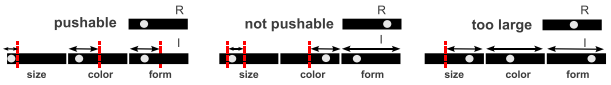


Figure 38: Input statistics used in experiments on the bootstrapping of representations. Shown are value ranges and realization examples for the artificial input representations $I$ and reference representation $R$. Left: pushable objects are defined by a certain range of colors and forms, and a small size. Middle: non-pushable objects are characterized by a slightly larger size and a certain range of colors, irrespective of form. Right: Objects too large for pushing are characterized by a large size, whereas the other two visual properties can take any value. Predictability is reduced since little can be inferred about the values of "form" and "color".

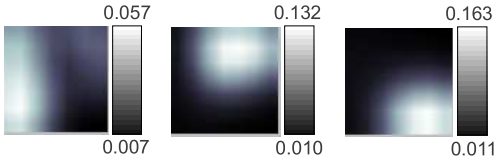


Figure 39: Examples of reference-based predictions. Roughly speaking, predictability is expressed by the difference between minimal and maximal values (please note the different value ranges indicated by the colorbars when comparing diagrams). Left: prediction derived from the "too large" signal in the reference representation $R$: in this case the predictability of the induced representation is low as compared to the case of the "not pushable" (middle) and "pushable" (right) classes.

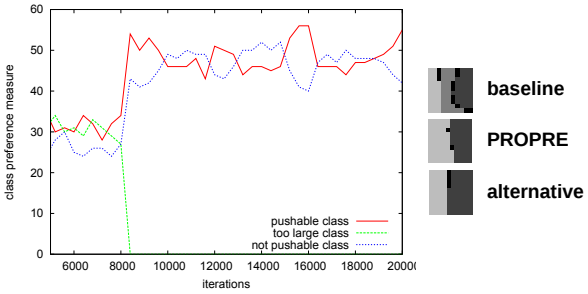


Figure 40: Development of neural selectivities in our robotic model task under the influence of PROPRE learning. Left: temporal evolution of the percentage of neurons selective to the input classes "pushable","not pushable" and "too large". Right: class selectivities in the internal layer $H$ without predictability-based modulation of learning (top), normal PROPRE modulation (middle), and alternative condition (bottom) where the definitions of the classes are slightly adjusted (not discussed here). Bright gray pixels indicate the "not pushable" class, moderately gray pixels indicate the "too large" class, and dark gray pixels the "pushable" class. Black pixels indicate no sufficient class selectivity.

goal of the classes "pushable" and "not pushable" being well predictable in contrast to the class "too large" whose predictability should be much less pronounced. As may be seen from Fig. 39, the reference-based predictions differ strongly in predictive power. In particular, the class "too large" cannot predict the internal representation as well as the other two classes. This shows that the PROPRE architecture is capable, after a learning process, to extract this information from the provided samples. As furthermore the learning of prototypes in $H$ is governed by this predictability measure, an impact on $H$ may be expected as well, and this is in fact precisely what Fig. 40 shows: a very strong reduction of neurons whose prototypes are selective for the "too large" class. The other two classes, with higher predictive power, profit from this re-organization and get allocated more neurons, which was precisely the objective of this study: to demonstrate a dynamic learning process that would distribute representational resources among the most predictable concepts and thus form a task-specific representation.

*Discussion of significance*

In the light of the road map of Sec. 2, it was shown in this section that the PROPRE architecture is well suited to fulfill the need for generative yet efficient machine learning approaches, giving access to quantities that have a rigorous probabilistic interpretation. While presenting one variant of the PROPRE architecture, Furthermore, I showed that a PROPRE variant can build up meaningful, multimodal perceptual representations by exploiting very basic quantities that are derived from behavior. In fact, PROPRE characterizes and extracts percepts that have a statistical link to behavioral quantities, and ignores other percepts that do have such a link. What is learned by PROPRE is in fact entirely dependent on the reference representation $R$; hence it

is important to state that PROPRE does not impose strong restrictions on $R$. In fact, in the following section I will show another variant of PROPRE learning where $R$ is not hand-crafted as it is done here, but derived from another learning process, leading to multi-modal learning which is another important point on the road map of Sec. 2.

### *5.2. Simultaneous concept formation and multimodal learning*

This investigation [55] was conducted in the context of developmental learning in embodied agents who have multiple data sources (sensors) at their disposal. It concerns an online learning method that simultaneously discovers "meaningful" concepts in the associated processing streams, extending methods such as PCA, SOM or sparse coding to the multimodal case. In addition to the avoidance of redundancies in the concepts derived from single modalities, the claim is that "meaningful" concepts are those who have statistical relations *across modalities*. This is a reasonable claim because measurements by different sensors often have a common cause in the external world and therefore carry correlated information. To capture such cross-modal relations while avoiding redundancy of concepts, I proposed a set of interacting generative learning processes which are modulated (in the sense of the original PROPRE architecture) by local predictability. That is to say, generative learning focuses on those concepts that can be well predicted (in a statistical sense) across modalities. To validate the fundamental applicability of the method, I conducted a plausible simulation experiment with synthetic data and found that concepts which are predictable from other modalities successively "grow", i.e., become over-represented, whereas concepts that are not predictable become systematically under-represented or even suppressed.

*Motivation and context*

The autonomous formation of representations is a very active research topic in developmental robotics [126, 85, 143, 142]. Such concepts may be formed at low abstraction levels (and are usually termed "features") or at high abstraction levels (where they tend to be termed "concepts"). While it is generally agreed that concepts derived from a single information source should be encouraged to be diverse, as it is the case in sparse coding [124], ICA [77] or competitive learning approaches [21], biological and behavioral evidence suggests a great deal of correlations between concepts derived from different sources. As individual sources are usually corrupted by (structured) noise, issues of multisensory integration become crucial for stable perception and performance, as can be seen in audio-visual facilitation[173], contour integration[5] and multisensory integration[32]. Such sources may be different sensory inputs (vision/touch, vision/audition etc.), results of divergent processing (ventral/dorsal processing in visual cortex) or even different locations on a retinotopic surface such as V1. It has furthermore been shown in various experiments that humans are able to integrate multi-sensory cues in a fashion that is close to being Bayes-optimal[32].

In this contribution, I dealt with the problem of how features or concepts may be formed that are particularly suited for performing multisensory integration. Such concepts must be statistically related across sensory modalities while being non-redundant within their own modality. For achieving this in an online learning process, a variant of the PROPRE (projection-prediction) algorithm[54] was proposed which uses predictability from another sensory modality to control learning.

*Related work*

A conceptually similar approach, which is moreover implemented in a robotic agent, is presented in [96]. This work extracts multimodal concepts from feature vectors arising from visual and haptic processing streams by concatenating them and subjecting the resulting vector to principal components analysis (PCA). By construction of PCA, the basis vectors of the resulting transformation will be those whose multimodal components are maximally correlated, which can be used to improve a multisensory classification task. The main difference to our approach is that we aim at online learning in a behaving agent, and that our approach maintains separate representations in different modalities which are however aligned to each other.

*System architecture*

What is used here is just a generalization of the PROPRE architecture described in Sec. 5.1: instead of a "god-given" reference representation that is used to modulate generative learning in a single modality, I

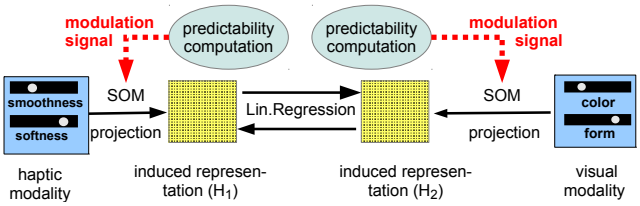


Figure 41: Generalizing the PROPRE architecture to multiple modalities (to be compared to Fig. 37).

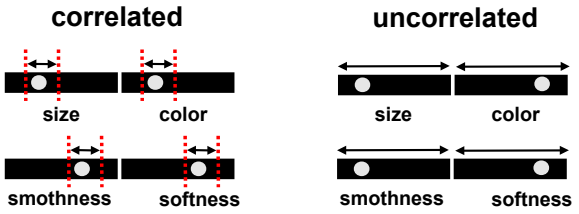


Figure 42: Input statistics for visual (upper row) and haptic (lower row) modality for multi-modal learning experiments. With a certain probability, these will be correlated (left row), or uncorrelated (right row). Values are always drawn from a uniform distribution, which is bounded in the correlated case and unbounded for the uncorrelated case.

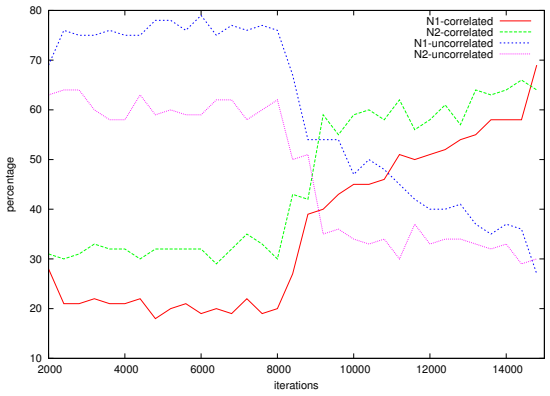


Figure 43: Percentage of neurons in both induced representations that are sensitive to the correlated class (red and green curves) and to the uncorrelated class (violet and blue curves). The deterioration of selectivities to the uncorrelated class is notable right from the start of PROPRE learning at $t = 8000$.

now use the result of another generative learning process in complementary modality. Here, one may observe an advantage of the PROPRE architecture: it is very modular, since, formally, anything may serve as reference representation $R$ because all that PROPRE does is computing a prediction from it. In particular, no analytical gradient computations need to be performed that would require supplementary knowledge about the internal workings of $R$. This generalized architecture is shown in Fig. 41.

*Experiments and results*

To assess the ability of the modified PROPRE architecture to extract correlated concepts, simulation experiments with two simulated modal data streams were conducted. As in the original PROPRE experiments, the population encoding technique was used to represent single scalar quantities, see Fig. 42. The probability distribution from which data samples were drawn in simulation were chosen to replicate the most important properties that might be encountered in real data, most importantly that correlated concepts are much less frequent than uncorrelated ones. This coincides with the reality in visual object detection where potential "background"-detections occur much more frequently than actual objects (e.g., vehicles or persons). For correlated concepts, feature values in all modalities are bounded to certain fixed intervals from which they are uniformly drawn, so knowledge of features in one modality determines to a large part feature values in another. For uncorrelated concepts, feature values are unbounded, and knowledge of one feature does not, statistically speaking, tell anything about other features. Please see Fig. 42 for a visualization of the simulated sample distributions.

Although in this generalization of the PROPRE architecture, the reference representation $R$ is represented by an induced representation fed by another modality, the principle of modulation by predictive power has been retained, the particularity of this model being that both modalities modulate each other at the same time, which can potentially lead to interesting learning dynamics. The principal interest in this work was however a proof of concept, which was achieved as shown by Fig. 43. As in the

original PROPRE experiments shown in Sec. 5.1, the lower predictability of uncorrelated samples leads to a near-suppression of such samples in the two induced representations $H_1$ and $H_2$.

*Significance of multi-modal learning*

In the context of the road map outlined in Sec. 2.3, these results are of the utmost significance, because they demonstrate that meaningful information can be extracted from multimodal data streams in a completely unsupervised manner. This approach should always work when "interesting" (useful, stable, ...) samples, characterized by correlation across modalities, are embedded into a much larger body of uninteresting samples who are not correlated across modalities. The basic assumption here is that interesting samples are identical to those that are relevant for a certain application, which is not self-evident. However, an automatic pre-selection of samples according to the needs of the application can ensure that only a subset of samples is treated for which the assumptions holds. More will be said to this effect in Sec. 7.

If this approach could be shown to work in realistic scenarios, it could mean a huge performance boost especially for very difficult problems like pedestrian or vehicle detection, which notoriously suffer from a shortage of supervised training data (see Sec. 2). With the presented method, it will be possible to "train" object -vs- background classifiers without any direct supervision on as many samples as desired, since all this would require are multi-modal recordings which are very easy and cheap to obtain nowadays.

## 6. Incremental learning

This section continues the description of my work on prototype-based generative learning methods, however now with a focus on a special functionality, the ability to learn in an incremental fashion. In this sense, the work presented here is a variant of the PROPRE architecture introduced previously. I keep the focus on vision problems who tend to exhibit a large number of high-dimensional data samples. Some of the treated problems come from the domain of road traffic, as this motivates the my entire research effort in machine learning, however I make heavy use of more conventional machine learning benchmark tasks such as MNIST [99], simply because it is easy to work with, simple in nature and yet has sufficient complexity (especially w.r.t. number of classes) to allow meaningful statements about incremental learning capacity.

### *6.1. Structure of this section*

Incremental learning being a complex and notoriously ill-defined notion, I begin with a taxonomy of this and other, related terms while also introducing a few facts about incremental learning in biological systems. The definitions I give here may not be universally adhered to, but they will help to give precise meaning to statements made in this section. After giving a survey of related work on incremental learning and related issues, I will outline my contributions to this field [61, 60, 71, 59, 72, 62], which focus on incremental learning for visual problems or problems that are similar in nature. Finally, I will discuss the relevance and significance of my particular approach to incremental learning and outline the next steps for improving them.

### *6.2. What is incremental learning?*

Incremental learning comes in various forms in the literature, and the use of the term is not always consistent. So some effort will be made here to give precise meaning to relevant terms.

*Batch -vs- online learning*

Many machine learning algorithms are trained on entire databases of samples, that is to say, they use all examples in a database at the same time, irrespective of their (temporal) order, to perform, e.g., a model optimization step. This does not preclude the repetition of this step: for example, a multilayer perceptron minimizes its cost function by iterative gradient descent, where all training samples are processed at each iteration. This approach completely ignores the temporal structure and order of samples in a database, which is of course completely acceptable when data statistics are stationary. In this case we speak of **batch learning** algorithms. When data statistics are non-stationary, it becomes interesting to take the temporal evolution of the data into account. This is realized by **online learning** approaches, which use training samples one by one, without knowing their number in advance, to optimize their internal cost function. There is a continuum of possibilities here, ranging from fully online approaches that adapt their internal model immediately upon processing of a single sample, over so-called *mini-batch* techniques that accumulate a small number of samples to perform batch learning, to the batch learning approaches described previously. As to existing machine learning models, online learning is most easily achieved by stochastic gradient descent versions of multilayer perceptrons (MLPs), but there are also extensions of the support vector machine (SVM) model ([184] for an overview) that have this capacity. Prototype-based models such as k-means [152], k-NN [152], radial basis function networks (RBF) [114], learning vector quantization [83] and self-organizing maps [86] all naturally fall into this category as well.

*Concept drift and the stability-plasticity dilemma*

When the temporal structure of data samples is taken into account, one is often faced with changes in data statistics that occur over time. Generally, such changes, denoted somewhat generally as *concept drift* [98, 172], can be gradual or abrupt. In the latter case one often uses the term *concept shift*. When data statistics do not change globally but only in a specific region of data space, sometimes the term *lo-*

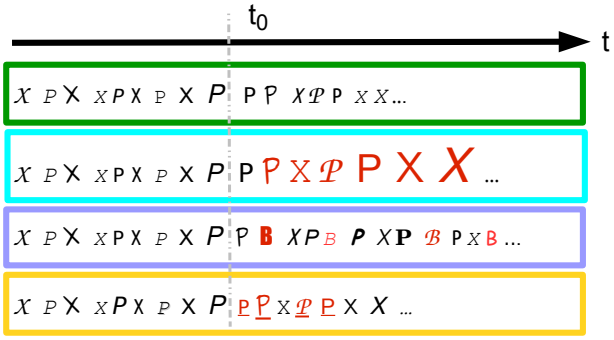


Figure 44: Examples for global and local concept drift/shift in a model classification task containing two visual classes, **P** and **X**. At $t = t_0$, concept drift sets in, with effects marked in red. Top row (green): original sequence of data samples without concept drift. Second row (cyan): gradual global concept drift. All characters increase slowly in size. Third row (violet): local concept shift with conflict. A third class **B** appears suddenly in the data stream. This is local because class **X** is far away from the new class in data space, however **P** and **B** are visually very close which leads to a conflict when trying to distinguish them. Fourth row: local concept shift without conflict. All instances of class **P** become suddenly underlined. Again this is local because class **X** is not at all affected, and there is no conflict because class assignments remain constant

*cal concept drift* is used [172]. A prominent example is the addition of a new, visually dissimilar object class to a classification problem. This in particular is an important use case for incremental learning algorithms as there is, a priori, no reason why new statistics in localized regions should disrupt learned models elsewhere. Another and much more problematic case is local concept drift/shift with conflict, for example when a new but visually similar class appears in the data: this will in any event have an impact on classification performance until the model can be locally re-adapted to separate the old from the new class. Please see Fig. 44 for a visualization of some prominent special cases of concept drift. Recognizing concept drift at execution time constitutes a challenging task, see [98, 172] and references therein. Coarsely speaking, an algorithm needs to decide whether a deviation is just due to noise, or due to a real change in data statistics. This implies a model of the data, in the simplest form a time scale on which "real" concept drift can occur. If concept drift can be reliably detected, it is possible to adapt to it, although this adaptation raises another difficulty: when old and new data statistics are in conflict, how quickly should models be updated? They can be updated quickly but in this case, old information will be forgotten equally quickly. On the other hand, adaptation can be performed slowly, in which case old information is retained longer: it really depends on the application one has in mind to correctly set these parameters. This complexity of online learning might seem intimidating w.r.t. batch learning algorithms. However, the conceptual simplicity of the latter stems from the extreme simplification that is made in discarding all temporal information in a set of data samples.

*Definition of incremental learning*

The most obvious consequence of taking into account the temporal evolution of data statistics is that machine learning algorithms, instead of being trained exactly once, need to track these changes and react accordingly. From the foregone discussion, it is clear that any incremental learning algorithm must necessarily be an online technique whereas the reverse it not true, the most famous counterexample being MLPs which exhibit a so-called *catastrophic forgetting* behavior [109, 140, 37, 36, 106] even when the new data statistics do not invalidate the old ones. Stated in a clear and concise manner, a learning algorithm that shall be called "incremental" should have the following properties:

- it must be an online learning method
- concept drift is recognized autonomously and the internal model is adapted accordingly
- previously learned knowledge is retained when facing concept drift. In case of local conflicts, old knowledge is replaced only locally.

### *6.3. Existing approaches to incremental learning*

There are a number of approaches for incremental learning with support vector machines (see [184] for an overview). Some rely on heuristics, like retraining a model with all support vectors plus a new "incremental" batch of data [27, 167], but this is without theoretical guarantees and not what would be considered fully online approaches as it makes little sense in

this context to present examples one by one. Other ideas are a modification of the cost function to be optimized by SVM training, see, e.g., [153] in order to facilitate incrementality. But in the light of the given definitions, these approaches are closer to online learning and will run into trouble under concept drift. Furthermore, it has been proposed to perform SVM training *adiabatically*, that is, presenting one example at a time while maintaining the relevant optimality conditions on all previously seen examples. However in our terminology this is neither online nor incremental, as all previously seen samples need to be stored, although the approach can considerably simplify SVM training and has numerous useful consequences in practice. Lastly, there are ensemble learning algorithms [134, 184] that achieve incremental learning simply by training new classifiers for new batches of data, and combining all existing classifiers for decision making. While this indeed achieves incremental learning under some conditions, it makes the implicit hypothesis that concept drift coincides with new data batches, whereas a *detection* of concept drift is not addressed at all.

As the problem of catastrophic forgetting was first remarked for multilayer perceptron (MLP) models [109, 140], it is hardly surprising that there was significant work on the subject of how catastrophic forgetting could be avoided. There was an initial consensus that catastrophic forgetting in MLPs arises from the completely *distributed* nature of internal representations, coupled to back-propagation-type learning (see [40] and references therein). This can maybe best be understood by considering the opposite case, so-called *localist* representations where each internal unit responds only to a very small sub-volume in input space. Learning in such networks would thus only adapt the sensitivities of internal units that are "closest" to the input, which would eliminate the catastrophic forgetting problem at the price of poor generalization performance [159]. It was concluded that *semi-distributed representations* were necessary which would not be strictly localist but not completely distributed either, thus achieving a compromise between generalization and catastrophic forgetting. A number of modifications of the MLP model were proposed with the goal of reducing the *representational overlap* and achieving semi-distributed internal representations. These relied to a significant extent on purely algorithmic approaches, namely sparsification [38], orthogonalization of internal node weights [39, 116] and the adaption of the backpropagation algorithm to reduce representational overlap [95]. These were more or less successful in mitigating but no eliminating the effect of catastrophic forgetting, although the success tends to be strongly problem-dependent and generalization capability is reduced [159]. This is not surprising as representational resources are less efficiently exploited if they are constrained to be sparse or orthogonal. In fact this was a first step towards semi-distributed or localist prototype-based representations although a vastly increased number of units is required for these approaches to generalize well. A rather recent proposition is made in [68] where a specific regularization scheme is supposed to reduce but in no way eliminate catastrophic forgetting effects.

Furthermore, there were attempts to modify the general architecture of MLPs [97, 163] which are more in the line of generative learning in that they attempt to detect newness and use different representational resources for new samples. This was taken further by connectionist models with a more elaborate organization, featuring different memory subsystems for long-term and short-term learning [151, 3], as well as models that performed explicit replay and re-learning of previous samples to alleviate forgetting [147]. It can safely be said that all of these approaches managed to reduce the problem at the price of being vastly more complex than conventional connectionist models. Contrarily to modern approaches, inspiration was taken primarily from biology and thus a solid mathematical foundation was not intended, preventing a thorough understanding of the algorithms.

To perform explicit incremental learning in the sense of Sec. 6.2, most modern approaches perform a local partitioning of the input space and train a separate classification/regression model for each partition [175, 120, 161, 11, 16]. The manner of performing this partitioning is very diverse, ranging from kd-trees [16] to genetic algorithms [11] and adaptive Gaussian receptive fields [175]. Equally, the choice of local models varies between linear models [175],

Gaussian mixture regression [16] or Gaussian Processes [120]. Since this article is concerned with high-dimensional perceptual problems, it can be stated for all cited approaches that it is really the partitioning of the input space that is costly in terms of memory. Most notably, covariance matrices used in [175] are quadratic in the number of input dimensions which makes their use prohibitive.

### *6.4. Insights into biological incremental learning*

As biological incremental learning has reached a high degree of perfection, we explicitly investigated the biological literature for hints as to how this might be achieved. Basing ourselves on observations from the basic sensory cortices, we noted that sensory representations seem to be prototype-based, where prototype-sensitive neurons are topologically arranged by similarity [169, 104, 150, 31]. Learning seems to act on these representations in a task-specific way, where more prototypes are allocated to sensory regions where finer discrimination is necessary [135], i.e., where more errors occur during learning. Learning is conceivably enhanced through acetylcholine release in case of task failures [182, 69], leading to higher "prototype density" in difficult regions of the sensory space. In particular, learning seems to respect and even generate topological layout of prototypes by changing only a small subset of neural selectivities [148] at each learning event, namely around those neurons that best matched the presented stimulus [31].

When going beyond the single-neuron level and looking at architectural issues, there is a large body of literature investigating the roles of the hippocampal and neocortical areas of the brain in learning. Generally speaking, the hippocampus employs a rapid learning rate with separated representations whereas the neocortex learns slowly, building overlapping representations of the learned task [127]. A well-established model of the interplay between the hippocampus and the neocortex suggests that recent memories are first stored in the hippocampal system and they are played back to the neocortex over time [108]. This accommodates the execution of new tasks that have not been recently performed as well as the transfer of new task representations from the hippocampus (short-term memory) to the neocortical areas (long-term memory) through slow synaptic changes.

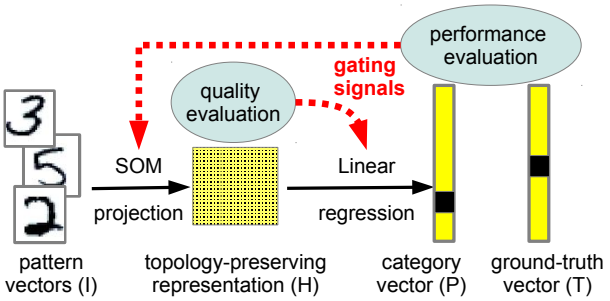


Figure 45: Adaptation of the PROPRE learning architecture for incremental learning. The most fundamental change made to the architecture is that learning in $H$ is now modulated by the ability of $H$ to predict $P$, instead of being predictable from $P$.

Figure 46: Illustration of how incremental learning is made possible through a topologically ordered prototype representation. Due to topological ordering, neighbouring prototypes are almost always situated in nearby regions of input space. Therefore, local updates of prototypes will almost always be local in input space as well, thus effectively enabling efficient incremental learning. This is shown here for a subset of prototypes trained on the MNIST database, the best-matching unit (BMU) for a "5" input being indicated by a small red circle. It is obvious that the local 2D update region, indicated by a larger red circle, is indeed local in the input space. The yellow circle indicates a region where this property does not hold (structural defect) but the reader can convince himself that this occurs but rarely.

### *6.5. Flat incremental learning architecture*

The biological findings of Sec. 6.4 were modelled by an architecture for incremental supervised learn-

ing proposed in [60, 61] which is another variant of the basic PROPRE architecture introduced in Sec. 5, the most important difference being that learning of prototypes in the hidden layer $H$ is now driven by the ability to predict $P$, as opposed to predictability from $P$. This architecture combines generative, prototype-based learning of an internal representation with discriminative learning of classification or regression outputs, see Fig. 45. From the latter, a task-related error signal is derived which adapts the internal representation in case of mismatch or classification ambiguity. This ensures that prototype density increases in regions of the input space that are difficult to classify, or in which concept drift is occurring. Prototype adaptation is stably self-terminating when no more errors are made, or when concept drift subsides.

The internal representation is topologically organized, and prototype adaptation modifies weights only locally, as observed in biology (see Sec. 6.4). It is above all this property that allows for incremental learning of prototypes: adaptation of a single prototype changes just its neighbours, which are close in data space as ensured by the topological organization of selectivities, see Fig. 46.

A read-out mechanism between hidden and output layer maps local input space regions (i.e., sets of prototypes) to class memberships using simple linear regression learning.

The mapping from hidden to output layer is adapted only when there is sufficient prototype activation in the internal representation. If this is not the case, e.g., when concept drift is occurring, adaptation is suspended, because random weak activations due to unknown inputs can disrupt already existing read-out weights.

The presented architecture is prototype-based in its hidden layer and covers data space by hyperspheres of adaptive size around prototypes. This strongly simplifies the definition of local regions which, in other algorithms, is very costly in spaces of high dimensionality $I$ (see Sec. 6.3). The quality of this approximation can be controlled by controlling the overall number of prototypes. As such a prototype-based representation approximates the distribution of data points in input space as a whole,

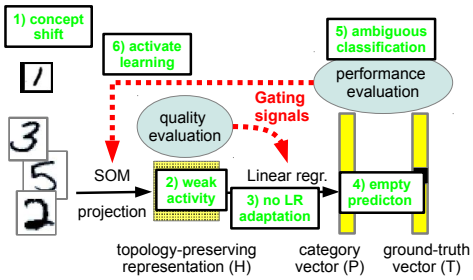


Figure 47: Sequence of events as a reaction to concept shift in the form of a new visual class, to be regarded together with Fig. 50.

Figure 48: Typical samples from the MNIST handwritten digit database used for the experiments on incremental learning. Each sample has a size of 28x28 pixels which gives a dimensionality of 784. All MNIST samples are treated "as is", that is to say, without intermediate feature transform.

it is a generative model [8] as it could be used for sampling purposes.

*Experimental validation*

The performance of this architecture, both incremental and non-incremental, was measured on the MNIST handwritten digit dataset [99] containing 70000 samples of dimension 784, grouped into 10 classes (please see a visualization of these classes in Fig. 48). To quantify incremental learning capacity, 10 experiments (Inc-0 through Inc-9) were conducted where the architecture was initially trained on 9 classes and subsequently presenting the remaining one exclusively for a short period, which was followed by a short period of retraining with all 10 classes. In the terminology of Sec. 6.2, these experiments represent local concept shift with conflict, as each MNIST class that is added at least partially overlaps with other classes. In general, performance depended in a monotonous fashion on the size of the internal representation, i.e., the (fixed) number of

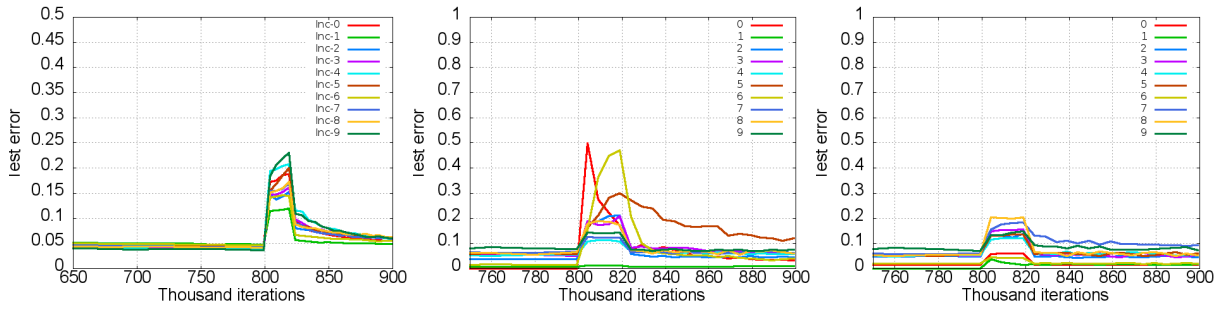


Figure 49: Results of incremental learning experiments conducted on MNIST. The missing class is presented between iterations 800.000 and 820.000, and retraining with all classes occurs between 820.000 and 840.000. Left: temporal development of overall classification error for experiments Inc-0 through Inc-9 (see text). Middle: development of class-specific errors in Inc-0. right: development of class-specific errors in Inc-1.

prototypes. Non-incremental (no concept shift, using all 10 classes) performance was around 3% for a reasonable number of prototypes (900), which is close to state-of-the-art results of around 1.5%. As can be seen from some exemplary results shown in Fig. 49, the architecture managed to predict the added class with high precision after a short period of model readaptation with transiently higher classification errors.

### *6.6. Flat incremental learning architecture with short-term memory*

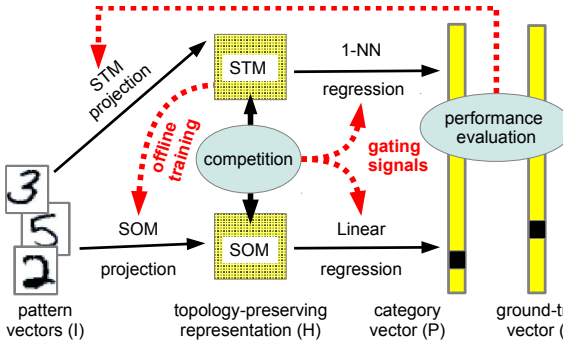


Figure 50: Extension of the original "flat" model by a short-term memory (STM) module. This module stores misclassified or ambiguous samples in order to be able to react quickly to these samples. At regular intervals, the content of the STM is replayed to the internal SOM layer of the architecture and subsequently emptied.

In [60], I also investigated the effects of interacting short-and long-term memories on the proposed model, inspired by biological findings (see Sec. 6.4). The reason for adding a short-term memory is that the used statistical learning algorithm for adapting the prototypes is slow to adapt even when learning rates are high. In contrast, real-world situations require a rapid reaction to concept drift once it is detected as such (in our case, concept drift is detected by analyzing prototype activation in the internal representation). In order to achieve this, the basic model described in the previous section has been complemented by a second memory system, a short-term memory (STM). This memory is exemplar-based, that is, it stores incoming samples "as is" until it reaches its predefined capacity limit. Its function is three-fold:

- store samples along with their class in case of wrong or ambiguous classification
- predict class from samples stored in STM in case no prototype in the internal SOM representation responds sufficiently
- periodically train internal SOM representation using samples stored in STM (which is cleared afterwards)

The prediction from short-term memory is therefore only a temporary solution, until new patterns have been transferred to long-term storage. In the STM,

generalization ability is traded for reaction speed until a transfer to long-term memory is achieved whose generalization ability is higher. The extended architecture thus has a rather complex reaction to concept drift. When concept drift occurs, e.g., by adding a new class to a problem, the different steps until complete adaptation are as follows (see Fig. 51 for a visualization):

- Newness is detected by lack of activity in the internal SOM representation. This implies an extremely ambiguous classification as all class predictions will be very low, which leads to the storage of the current sample in STM.
- Once sufficiently many samples are stored in STM, it will react with sufficient activity to the new class, and will therefore give correct predictions. Therefore, storage of samples to STM will nearly cease.
- After a predetermined interval, samples in STM are replayed to the internal SOM representation, which incorporates them into its set of prototypes, whereupon the STM is reset.

Now, the SOM prototypes will respond strongly to the new class (thus inhibiting STM use), which allows the re-adaptation of linear regression weights that control readout, at which point the new class can be said to be added to the model. A potential last step is a brief re-training with all classes in order to re-calibrate the readout weights in SOM regions where old and new classes overlap. As can be verified from the results shown in Fig. 52, the STM essentially fulfils the function for which it is created. When introducing concept shift with conflict in the form of a new class, performance is degraded only very slightly and transiently, since predictions are rapidly provided by the STM. On the other hand, the LTM is not adapted immediately, so it continues to perform well on already known classes until STM replay takes place, transferring samples of the newly added class into a form more suited for generalization.

*Significance and possible extensions*

Summarizing, the proposed model tries to incorporate as many facts about incremental learning in

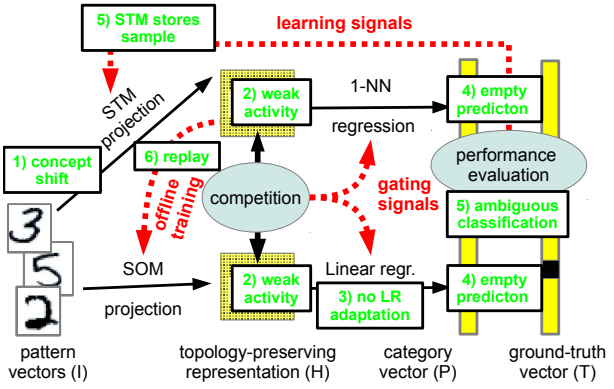


Figure 51: Sequence of events as a reaction to concept shift in the form of a new visual class when using the PROPRE-based incremental learning architecture with short-term memory, Fig. 50.

| time \ $N$ | 10x10 | 20x20 | 30x30 | 50x50 |
|---|---|---|---|---|
| GPU | 15 | 19 | 27 | 53 |
| CPU | 115 | 457 | 1037 | 2903 |

Table 4: Execution time measurements in seconds per $10^4$ iterations of the incremental learning architecture (without STM) for GPU and CPU implementations. It can be observed that GPU execution times scale (much) less than linearly in the considered range of hidden layer sizes, whereas the CPU implementation scales almost exactly in a linear fashion.

biology as possible while keeping the model as simple and efficient as possible. Modelling takes place at the architectural level, leaving aside the finer details of neural modelling such as rate/spike code or more realistic, dynamic neuron models. In the presented projects, I could show that incremental learning is both feasible and efficient for difficult real-world visual classification tasks, and that especially the high dimensionality of these tasks poses no problem at all to the proposed approach. Some experiments have been made regarding parallelization on a graphics processor (GPU), and, unsurprisingly, it turns out that a neural architecture is quite favourable to parallelization as can be seen in Tab. 4 and [61].

An obvious extension of the presented "flat" architecture is a "deep" architecture that would function

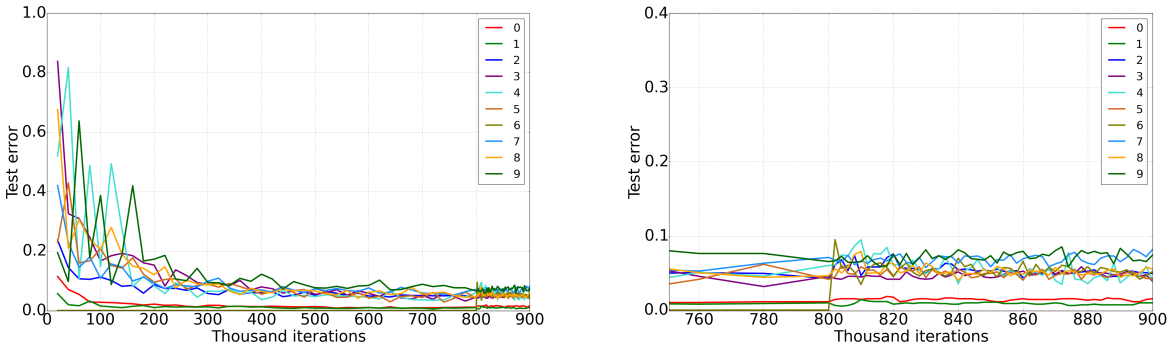


Figure 52: Results of incremental learning experiment Inc-6 with short-term memory conducted on MNIST. Left: frequency of short-term memory usage for prediction. Right: development of individual class errors for Inc-6. The increased noise exhibited by curves in the right-hand figure after 800K iterations is due to a higher sampling rate of the test error.

like a convolutional neural network (CNN), see [158] or my own work on CNNs [50]. A first step in this direction has been taken in [71], where it could be shown that the essential functions of concept drift detection can be performed in a multi-level architecture as well. Especially for prototype-based approaches such as PROPRE and its derivatives, deep architectures promise a strong reduction in resource consumption (which is already quite modest) since far less prototypes are required to describe smaller subsections ("receptive fields") of an input, especially if approximate independence properties hold between distant parts of the visual input.

## 7. Future research project

In the recent years I was able to contribute to many issues raised by the road map of Sec. 2.3, all with the long-term goal of improving environment perception in road traffic scenarios which were chosen as a template for many other perception tasks. However, most contributions were disconnected from one another, and thus the coupling aspect, which is one of the most interesting capacities of generative learning methods, was not exploited to its full potential. In the following years, I intend to create a full-blown perception system for road traffic that makes use of generative learning methods for detection, integrates context information in multiple and adaptive ways, and is capable of developing basic detection capabilities in a developmentally inspired way by exploiting multi-modality. This will require several distinct steps, some of which can be conducted in parallel:

- enhance the computational power and reduce computational cost of incremental learning as described in Sec. 6. This will mainly include creating a "deep" version of the basic PROPRE architecture, with the goal of reducing the number and the dimensionality of prototypes which can get high for difficult problems, which could limit applicability.

- fully formalize the probabilistic interpretation of PROPRE (see Sec. 5.1) for easy interfacing with other methods

- speed-up of prototype-based learning and classification methods, such as PROPRE, so they can be used for detection tasks in real time.

- experimentally investigate fully probabilistic coupling of generative object detectors and context information in the spirit of Sec. 4.

- experimentally verify that multiple learned context models, for example derived from static scene context and dynamic object trajectory analysis, can work together within a single system

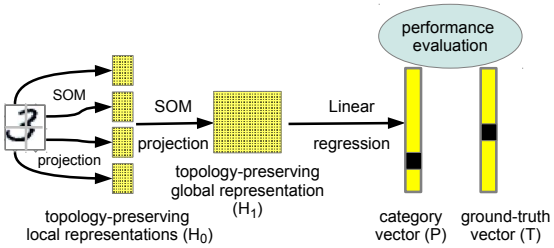


Figure 53: Envisioned extension of incremental PROPRE learning to a deep architecture, in this case with two internal layers $H_0$ and $H_1$. Additional internal layers $H_i$ may be introduced by the same mechanism of local receptive fields that was applied in $H_0$.

- demonstrate acquisition of basic perceptual skills by large-scale multi-modal learning as outlined in Sec. 5.

The fundamental building block of such an architecture will be incremental learning as outlined in Sec. 6, be it for training pattern-based object detectors or context models as in Sec. 4, and indeed considerable effort will have to go into the further development of these methods in order to make them truly applicable in such an architecture. The unsupervised training of detection skills, based on multi-modal perception, may provide a sensible starting point for a self-improvement cycle: context models trained from a roughly trained object detector, which in turn gets refined further by context models. I will discuss the most important steps a little more in-depth in the following sections:

### *7.1. Prototype-based deep learning*

One of the main drawbacks of prototype-based learning approaches such as PROPRE (see Secs. 5.1, 6) is that the number of prototypes sometimes needs to be quite high in order to satisfactorily solve a given problem. Even for simple problems such as MNIST, we found in Sec. 6 that about 1000 prototypes are needed to obtain state-of-the-art performance. This is still low in comparison to what other incremental approaches require, but if incremental learning is not a priority, then this is quite high compared to other classification methods. A simple way to reduce resource requirements is to exploit the probabilistic structure of images which we deal with in

most cases: very often, some parts of the input image are approximately independent of other parts. In this sense, it is far more computationally efficient to model them independently, again by prototypes. As the required number of prototypes can *increase* exponentially with dimensionality, in this case it should *decreases* exponentially when dimensionality is reduced by modelling only a small part of the image (a receptive field, to be coherent with deep learning terms). In its simplest form, this comes down to a four-layer architecture (see Fig. 53), where the added layer contains now prototype activities related to local descriptions of the input, which are subsequently integrated into a global representation. Preliminary experiments show that one can obtain a 10-fold decrease in free parameters (connection weights) with no loss of classification performance on MNIST when using an architecture as shown in fig. 53, but further research is needed to consolidate these results and determine the optimal way of parametrizing the architecture.

### *7.2. Speed-up of prototype-based learning for object detection*

For detection-by-recognition purposes (see Sec. 1), trained models are applied to every position and scale in an image in a sliding window fashion (see also Fig. 7). This requires a certain execution speed if real-time capability is to be maintained. At present, the prototype-based models presented here are rather costly in terms of execution time due to the input-prototype distance computation step which currently needs to be performed for all prototypes. Already, the ideas given in the preceding section can be a way out of this dilemma: by reducing the number of weights in the architecture, computational demands can be reduced by approximately the same factor. Further speed-ups may be obtained by considering a *convolutional* version of the hierarchical architecture outlined in Fig. 53. This means that the prototypes of all local SOMs in layer $H_0$ are identical, which further simplifies the architecture and reduces the number of free parameters.

Another way to speed up calculations is parallelization, be it using programmable logic (FPGA) boards or graphics processing units (GPUs). In fact, the parallelization potential of the PROPRE architecture was already established in [61] using GPUs.

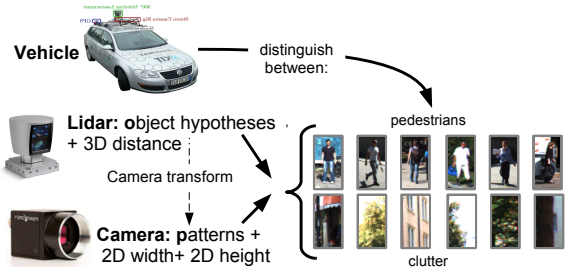


Figure 54: Approach for transferring the simulation results on multi-modal learning to object detection in a real-world domain (here, pedestrian detection/recognition is shown, but other classes such as vehicles are eligible as well): when a complementary LIDAR measurement is available for each visual object hypothesis, the multi-modal learning architecture described in Sec. 5 can be used "as is", although a preprocessing of modal data may be required.

A last idea concerns cascade-like approaches for computing input-prototype distances: making use of the topological organization of prototypes, it may be feasible to compute distances only for a subset of prototypes and then restricting further calculations to neighbourhoods of prototypes with sufficient activity.

Speeding up the *execution* (in contract to learning) of prototype-based learning is of crucial importance if these methods should ever be applied to perceptual problems in an applied setting. In contrast, learning in PROPRE is strictly local, that is, only a few prototypes around the BMU will be modified at any single time, which does not incur a high computational cost.

### *7.3. Multi-modal learning on real data*

As mentioned in Sec. 5, transferring multi-modal learning on simulated data, as described in that section, to real data coming from object detection tasks in road traffic, would mean a huge potential gain in detection performance "for free", as it is the creation of a sufficient number of supervised training samples that is very costly. First steps have been taken in that direction, using multimodal (visual and LIDAR) data coming from the KITTI dataset [49] to train pedes-

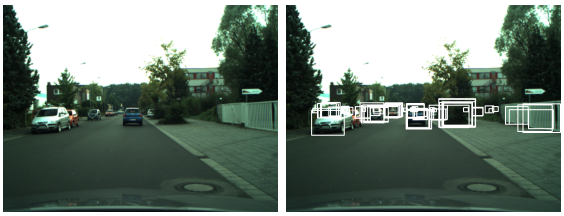

Figure 55: Results of the restricted vehicle detector presented in [73]. Left: original image. Right: vehicle detections. Please note that all vehicles are detected but many incorrect detections exist as well.

trian detectors as sketched in Fig. 54. The biggest issue in this endeavour is purely practical: how to select samples that should be subject to learning? If one selects image patches at random, the relative frequency of pedestrians will be too low for meaningful learning. On the other hand, one cannot use the pedestrian annotations in the KITTI database either (maybe for selecting pedestrians and then adding randomly selected background samples) because the goal is to learn without annotations. To solve this dilemma, a simple *restricted detector* as, e.g., proposed in [73], could be used. A restricted detector detects its class of interest by efficient but hand-coded image processing operations, e.g., finding horizontal lines for detecting lower vehicle borders as described in [73]. As can be seen from Fig. 55, this sort of detector will have a high false detection rate, which is acceptable for multi-modal learning as long as almost all object of the class of interest are localized. Such a detector can act as a filter which increases the relative frequency of the objects of interest sufficiently so that the learning algorithm can extract statistical information. When looking at the biological analogy, this can be seen as an innate attention mechanism that guides learning in the right direction. As stated at the end of Sec. 5, this kind of unsupervised multi-modal learning, together with a suitably crafted restricted detector, has the potential to boost the quality of learned models simply because the number of available training data is potentially unlimited.

### *7.4. System building*

Finally, the long-term objective of my research is to unite all of my contributions into a single perception system, with incremental learning as its basic functionality operating in different places, but at the very least in a generative prototype-based detector and, acting thereon by attentional modulation, in one of several context models (static and dynamic) as described in Sec. 4. Evidently, it will have to be investigated how learning is handled in various parts of the architecture, for preference taking inspiration from what is known about the large-scale organization of the human neocortex. Especially, the architecture of the system should contain a developmental component that would allow it to acquire a baseline performance in an unsupervised way, exploiting the principle of multi-modal learning outlined in Sec. 5.

This system will then be used to address the items on the road map outlined in Sec. 2.3, ideally leading to a detection performance that is superior to the state of the art. However, such a system may very well have a scientific merit in its own right. An example is the continuous acquisition of new training data: as the experiments on context-based object detection (see Sec. 4) suggest, many objects can be detected only with the "help" of attentional modulation derived from context models. The more context models are involved, the more certain such detections may become. Therefore, if a detection is made that would not have been recognized by the pattern-based detector alone, this should trigger learning of detection/recognition models. In case these models are "deep" networks (as evoked previously in this section), the learning of new features or new object parts can be triggered as well. Conversely, context models may be trained on object hypotheses obtained from the detector and not from ground-truth data, a possibility already investigated with some success in [53]. Ideally, the whole architecture could serve as a tool to reduce the need for ground-truth data to an absolute minimum while still achieving a detection performance beyond the state of the art.

# 8. Bibliography


[1] E. Allwein, R. Schapire, and Y. Singer. Reducing multiclass to binary:a unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

[2] Y. F. Anat Levin, Paul Viola. Unsupervised improvement of visual detectors using co-training. In *Proceedings of the ICCV*, pages 626–633, 2003.

[3] B. Ans and S. Rousset. Avoiding catastrophic forgetting by coupling two reverberating neural networks. *Academie des Sciences, Sciences de la vie*, 320, 1997.

[4] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *Signal Processing, IEEE Transactions on*, 50(2):174–188, 2002.

[5] R. Bauer and S. Heinze. Contour integration in striate cortex: Classic cell responses or cooperative selection? *Experimental Brain Research*, 2002.

[6] H. Bay, T. Tuytelaars, and L. J. V. Gool. Surf: Speeded up robust features. In *Proceedings of the 9th European Conference on Computer Vision (ECCV), Graz,Austria*, 2006.

[7] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool. Pedestrian detection at 100 frames per second. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2903–2910. IEEE, 2012.

[8] C. Bishop. *Pattern recognition and machine learning*. Springer-Verlag, New York, 2006.

[9] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[10] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Online multiperson tracking-by-detection from a single, uncalibrated camera. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(9):1820–1833, 2011.

[11] M. Butz, D. Goldberg, and P. Lanzi. Computational complexity of the xcs classifier system. *Foundations of Learning Classifier Systems*, 51, 2005.

[12] R. J. Campbell and P. J. Flynn. A survey of free-form object representation and recognition techniques. *Computer Vision and Image Understanding*, 81(2):166–210, 2001.

[13] L. Caron, Y. Song, D. Filliat, and A. Gepperth. Neural network based 2D/3D fusion for robotic object recognition. In *European Symposium on Artificial Neural Networks (ESANN)*, 2014.

[14] L.-C. Caron, D. Filliat, and A. Gepperth. Indoor RGB-D object recognition for autonomous mobile robot. In *International Conference On Computer Vision (ICCV) Workshop Paper*, 2014.

[15] B. Catanzaro, N. Sundaram, and K. Keutzer. Fast support vector machine training and classification on graphics processors. In *Proceedings of the 25th international conference on Machine learning*, pages 104–111. ACM, 2008.

[16] T. Cederborg, M. Li, A. Baranes, and P.-Y. Oudeyer. Incremental local online gaussian mixture regression for imitation learning of multiple tasks. 2010.

[17] H. Chen and B. Bhanu. 3D free-form object recognition in range images using local surface patches. *Pattern Recognition Letters*, 28(10):1252–1262, 2007.

[18] C. S. Chua and R. Jarvis. Point signatures: A new representation for 3d object recognition. *International Journal of Computer Vision*, 25(1):63–85, 1997.

[19] R. Cucchiara, C. Grana, A. Prati, and R. Vezzani. Probabilistic posture classification for human-behavior analysis. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 35(1):42–54, 2005.

[20] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[21] G. Deco and E. T. Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Res*, 44(6):621–642, Mar 2004.

[22] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *International Conference on Computer Vision (ICCV)*, 2009.

[23] L. Ding and A. Goshtasby. On the canny edge detector. *Pattern Recognition*, 34(3):721–725, 2001.

[24] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(4):743–761, 2012.

[25] P. Doll.r, S. Belongie, and P. Perona. The fastest pedestrian detector in the west. In *BMVC*, 2010.

[26] P. Doll.r, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 304–311. IEEE, 2009.

[27] C. Domeniconi and D. Gunopulos. Incremental support vector machine construction. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 589–592, 2001.

[28] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 990–997. IEEE, 2010.

[29] M. Enzweiler and D. Gavrila. Monocular pedestrian detection: Survey and experiments. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(12):2179–2195, 2009.

[30] M. Enzweiler and D. Gavrila. Integrated pedestrian classification and orientation estimation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 982–989. IEEE, 2010.

[31] C. A. Erickson, B. Jagadeesh, and R. Desimone. Clustering of perirhinal neurons with similar properties following visual experience in adult monkeys. *Nature neuroscience*, 3(11):1143–1148, 2000.

[32] M. O. Ernst and M. S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, Jan 2002.

[33] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[34] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[35] D. Filliat, E. Battesti, S. Bazeille, G. Duceux, A. Gepperth, L. Harrath, I. Jebari, R. Pereira, A. Tapus, C. Meyer, S. Ieng, R. Benosman, E. Cizeron, J.-C. Mamanna, and B. Pothier. Rgbd object recognition and visual texture classification for indoor semantic mapping. In *Proceedings of the 4th International Conference on Technologies for Practical Robot Applications (TePRA)*, 2012.

[36] R. French. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychol Rev.*, 97(2), 1990.

[37] R. French. Semi-distributed representations and catastrophic forgetting in connectionist networks. *Connect. Sci.*, 4, 1992.

[38] R. M. French. Semi-distributed representations and catastrophic forgetting in connectionist networks. *Connection Science*, 4, 1992.

[39] R. M. French. Dynamically constraining connectionist networks to produce distributed, orthogonal representations to reduce catastrophic interference. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. 1994.

[40] R. Frenhc. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4), 1999.

[41] S. Frintrop, G. Backer, and E. Rome. Goal-directed search with a top-down modulated computational attention system. In *Pattern Recognition*, Lecture Notes in Computer Science. Springer, 2005.

[42] K. Friston. Learning and inference in the brain. *Neural Networks*, 16(9):1325–1352, 2003.

[43] C. G, J. Meguro, Y. Kojima, and T. Naito. Detection of pedestrians in road context for intelligent vehicles and advanced driver assistance systems. In *IEEE International Symposium on Intelligent Vehicles (IV)*, 2013.

[44] T. Gandhi and M. Trivedi. Image based estimation of pedestrian orientation for improving path prediction. In *Intelligent Vehicles Symposium, 2008 IEEE*, pages 506–511. IEEE, 2008.

[45] T. Gandhi and M. M. Trivedi. Pedestrian protection systems: Issues, survey, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 8(3):413–430, 2007.

[46] Á. García-Martín and J. M. Martínez. On collaborative people detection and tracking in complex scenarios. *Image and Vision Computing*, 30(4):345–354, 2012.

[47] M. Garcia Ortiz, A. Gepperth, and B. Heisele. Real-time pedestrian detection and pose classification on a GPU. In *16th International IEEE Conference on Intelligent Transportation Systems(ITSC)*, 2013.

[48] M. Garcia Ortiz, F. Kummert, J. Fritsch, and A. Gepperth. Situation-specific learning for ego-vehicle behavior prediction systems. In *IEEE International Conference on Intelligent Transportation Systems(ITSC)*, 2011.

[49] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3354–3361. IEEE, 2012.

[50] A. Gepperth. Object detection and feature base learning by sparse convolutional neural networks. In F. Schwenker and S. Marinai, editors, *Proceedings of the 2nd IAPR TC3 Workshop on Artificial Neural Networks in Pattern Recognition*, Lecture notes in artificial intelligence. Springer Verlag Berlin, Heidelberg, New York, 2006.

[51] A. Gepperth. Visual object classification by sparse convolutional neural networks. In M. Verleysen, editor, *Proceedings of the 14th European Symposium on Artificial Neural Networks (ESANN), Brugge, Belgium*, pages 222–228. d-side Publications, April 2006.

[52] A. Gepperth. Implementation and evaluation details of a large-scale object detection system. Technical Report TR 10-11, Honda Research Institute Europe GmbH, 2010.

[53] A. Gepperth. Co-training of context models for real-time object detection. 2012.

[54] A. Gepperth. Efficient online bootstrapping of sensory representations. *Neural Networks*, 41:39–50, 2012.

[55] A. Gepperth. Simultaneous concept formation driven by predictability. 2012.

[56] A. Gepperth, B. Dittes, and M. Garcia Ortiz. The contribution of context information: a case study of object recognition in an intelligent car. *Neurocomputing*, 2012.

[57] A. Gepperth, J. Edelbrunner, and T. Bücher. Real-time detection of cars in video sequences. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 625–631, June 2005.

[58] A. Gepperth, M. Garcia Ortiz, E. Sattarov, and B. Heisele. Dynamic attention priors: a new and efficient concept for improving object detection. *Neurocomputing*, 197(C):14–28, 2016.

[59] A. Gepperth and B. Hammer. Incremental learning algorithms and applications. 2016.

[60] A. Gepperth and C. Karaoguz. A bio-inspired incremental learning architecture for applied perceptual problems. *Cognitive Computation*, pages 1–11, 2016.

[61] A. Gepperth and M. Lefort. Biologically inspired incremental learning for high-dimensional spaces. In *IEEE International Conference on Development and Learning (ICDL)*, 2015.

[62] A. Gepperth, M. Lefort, T. Hecht, and U. Körner. Resource-efficient incremental learning in high dimensions. In *European Symposium On Artificial Neural Networks (ESANN)*, 2015.

[63] A. Gepperth, B. Mersch, J. Fritsch, and C. Goerick. Color object recognition in real-world scenes. In J. de Sa, editor, *International Conference on Artificial Neural Networks(ICANN)*, number 4669 in Lecture Notes in Computer Science. Springer Verlag Berlin Heidelberg New York, 2007.

[64] A. Gepperth, S. Rebhan, S. Hasler, and J. Fritsch. Biased competition in visual processing hierarchies: A learning approach using multiple cues. *Cognitive Computation*, 3(1):146–166, 2011.

[65] A. Gepperth and S. Roth. Applications of multi-objective structure optimization. *Neurocomputing*, (69):701–713, 2006.

[66] A. Gepperth, E. Sattarov, and S. Rodrigues Flores. Robust visual pedestrian detection by tight coupling to tracking. In *IEEE International Conference On Intelligent Transportation Systems (ITSC)*, 2014.

[67] A. González, D. Vázquez, S. Ramos, A. M. López, and J. Amores. Spatiotemporal stacked sequential learning for pedestrian detection. In *Pattern Recognition and Image Analysis*, pages 3–12. Springer, 2015.

[68] I. J. Goodfellow, M. Mirza, X. Da, A. Courville, and Y. Bengio. An empirical investigation of catastrophic forgeting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

[69] M. E. Hasselmo. The role of acetylcholine in learning and memory. *Current opinion in neurobiology*, 16(6):710–715, 2006.

[70] M. M. Hayhoe, T. McKinney, K. Chajka, and J. B. Pelz. Predictive eye movements in natural vision. *Experimental brain research*, pages 1–12, 2012.

[71] T. Hecht and A. Gepperth. towards deep incremental learning: multilevel change detection in a hierarchical visual recognition architecture. 2016.

[72] T. Hecht, A. Gepperth, and M. Gogate. A generative learning approach to sensor fusion and change detection. *Cognitive Computation*, 2015. accepted.

[73] T. Hecht, M. Mohit, E. Sattarov, and A. Gepperth. Scene context is more than a bayesian prior: Competitive vehicle detection with restricted detectors. In *IEEE International Symposium on Intelligent Vehicles(IV)*, 2014.

[74] G. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[75] M. Hofmann and D. Gavrila. Multi-view 3d human pose estimation in complex environment. *International journal of computer vision*, 96(1):103–124, 2012.

[76] D. Hoiem, A. Efros, and M. Hebert. Putting objects into perspective. *International Journal of Computer Vision*, 80(1), 2008.

[77] A. Hyvarinen. *Independent component analysis*. John Wiley & Sons., 2001.

[78] L. Itti, C. Gold, and C. Koch. Visual attention and target detection in cluttered natural scenes. *Optical Engineering*, 40(9):1784–1793, Sep 2001.

[79] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, Mar 2001.

[80] L. Itti and C. Koch. Computational modelling of visual attention. *Nat Rev Neurosci*, 2(3):194–203, Mar 2001.

[81] B. J.hne. *Digital image processing*. Springer Verlag Berlin, Heidelberg, New York, 6 edition, 2005.

[82] J. Johnson and B. Olshausen. The recognition of partially visible natural objects in the presence and absence of their occluders. *Vision Research*, 45, 2005.

[83] B. journal=Journal of Machine Learning Research. Dynamics and generalization ability of lvq algorithms. 8, 2007.

[84] T. Kapuściński, M. Oszust, and M. Wysocki. Hand gesture recognition using time-of-flight camera and viewpoint feature histogram. In *Intelligent Systems in Technical and Medical Diagnostics*, pages 403–414. Springer, 2014.

[85] S. Kirstein, H. Wersing, and E. K.rner. Towards autonomous bootstrapping for life-long learning categorization tasks. In *IJCNN*, pages 1–8, 2010.

[86] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biol. Cybernet.*, 43:59–69, 1982.

[87] E. Kollorz, J. Penne, J. Hornegger, and A. Barke. Gesture recognition with a time-of-flight camera. *International Journal of Intelligent Systems Technologies and Applications*, 5(3):334–343, 2008.

[88] T. Kopinski, S. Geisler, L.-C. Caron, A. Gepperth, and U. Handmann. A real-time applicable 3d gesture recognition system for automobile hmi. In *IEEE International Conference On Intelligent Transportation Systems (ITSC)*, 2014.

[89] T. Kopinski, S. Geisler, A. Gepperth, and U. Handmann. Time-of-flight based multi-sensor fusion strategies for hand gesture recognition. In *IEEE International Symposium on Computational Intelligence and Informatics*, 2014.

[90] T. Kopinski, S. Geisler, U. Handmann, and A. Gepperth. Neural network based data fusion for hand pose recognition with multiple tof sensors. In *International Conference on Artificial Neural Networks (ICANN)*, 2014.

[91] T. Kopinski, A. Gepperth, and U. Handmann. A simple technique for improving multi-class classification with neural networks. In *European Symposium On Artificial Neural Networks (ESANN)*, 2015.

[92] T. Kopinski, A. Gepperth, and U. Handmann. A tof-based hand posture database for automotive applications, 2016. submitted to IEEE International Symposium on Intelligent Vehicles.

[93] T. Kopinski, S. Magand, and S. Geisler. Intuitiveness of free hand in-car gestures. In *CHI - Computer-Human Interaction*, 2015.

[94] T. Kopinski, S. Magand, U. Handmann, and A. Gepperth. A pragmatic approach to multi-class classification. In *International Joint Conference On Neural Networks (IJCNN)*, 2015.

[95] C. Kortge. Episodic memory in connectionist networks. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*. 1990.

[96] O. Kroemer, C. Lampert, and J. Peters. Learning dynamic tactile sensing with robust vision-based training. *IEEE Transactions on Robotics*, 27(3), 2011.

[97] J. Krushke. ALCOVE: An exemplar-based model of category learning. *Psychological Review*, 99, 1992.

[98] P. Kulkarni and R. Ade. Incremental learning from unbalanced data with concept class, concept drift and missing features: a review. *International Journal of Data Mining and Knowledge Management Process*, 4(6), 2014.

[99] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[100] M. Lefort and A. Gepperth. Discrimination of visual pedestrians data by combining projection and prediction learning. In *European Symposium on Articificial Neural Networks (ESANN)*, 2014.

[101] M. Lefort and A. Gepperth. PROPRE: PROjection and PREdiction for multimodal correlations learning an application to pedestrians visual data discrimination. In *International Joint Conference on Neural Networks (IJCNN)*, 2014.

[102] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool. Dynamic 3d scene analysis from a moving vehicle. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[103] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Workshop on statistical learning in computer vision, ECCV*, volume 2, page 7, 2004.

[104] D. A. Leopold, I. V. Bondar, and M. A. Giese. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, 442(7102):572–575, 2006.

[105] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60, 2004.

[106] M. M and C. NJ. Catastrophic interference in connectionist networks: the sequential learning problem. *Psychol. Learn. Motiv.*, 24, 1989.

[107] E. Maggio, E. Piccardo, C. Regazzoni, and A. Cavallaro. Particle phd filtering for multi-target visual tracking. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 1, pages I–1101–I–1104, April 2007.

[108] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102:419–457.

[109] M. McCloskey and N. Cohen. Catastrophic interference in connectionist networks: the sequential learning problem. In G. H. Bower, editor, *The psychology of learning and motivation*, volume 24. 1989.

[110] A. S. Mian, M. Bennamoun, and R. A. Owens. Automatic correspondence for 3d modeling: An extensive review. *International Journal of Shape Modeling*, 11(2), 2005.

[111] T. Michalke, J. Fritsch, and C. Goerick. A biologically-inspired vision architecture for resource-constrained intelligent vehicles. *Computer Vision and Image Understanding*, 114(5):548 – 563, 2010. Special issue on Intelligent Vision Systems.

[112] T. Michalke, A. Gepperth, M. Schneider, J. Fritsch, and C. Goerick. Towards a human-like vision system for resource-constrained intelligent cars. In *International Conference on Computer Vision Systems (ICVS) Conference Paper*, 2007.

[113] T. Michalke, A. Gepperth, M. Schneider, J. Fritsch, and C. Goerick. Towards a human-like vision system for resource-constrained intelligent cars. In *The 5th Int. Conf. on Computer Vision Systems Conference*. Universit.tsbibliothek Bielefeld, 2007.

[114] J. Moody and C. J. Darken. Fast learning in networks of locally tuned processing units. *Neural Computation*, 1, 1989.

[115] K. Murphy, A. Torralba, D. Eaton, and W. Freeman. Object detection and localization using global and local features. In J. Ponce, editor, *Toward Category-Level Object Recognition*, Lecture Notes in Computer Science. Springer, 2005.

[116] J. Murre. The effects of pattern presentation on interference in backpropagation networks. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. 1992.

[117] D. Musicki and B. La Scala. Multi-target tracking in clutter without measurement assignment. *Aerospace and Electronic Systems, IEEE Transactions on*, 44(3):877–896, 2008.

[118] M.W. and Spratling. Predictive coding as a model of biased competition in visual attention. *Vision Research*, 48(12):1391 – 1408, 2008.

[119] V. Navalpakkam and L. Itti. Modeling the influence of task on attention. *Vision Res*, 45(2):205–231, Jan 2005.

[120] D. Nguyen-Tuong and J. Peters. Local gaussian processes regression for real-time model-based robot control. In *IEEE/RSJ International Conference on Intelligent Robot Systems*, 2008.

[121] P. K. nig and N. K. ger. Symbols as self-emergent entities in an optimization process of feature extraction and predictions. *Biological Cybernetics*, 94, 2006.

[122] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, pages 1–11, 2011.

[123] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[124] B. A. Olshausen and D. J. Fieldt. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[125] S. Oprisescu, C. Rasche, and B. Su. Automatic static hand gesture recognition using tof cameras. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 2748–2751. IEEE, 2012.

[126] P.-Y. Oudeyer. Developmental robotics. In N. Seel, editor, *Encyclopedia of the Sciences of Learning*, Springer Reference Series. Springer, 2011.

[127] R. C. O'reilly. The division of labor between the neocortex and hippocampus. *Connectionist Models in Cognitive Psychology*, page 143, 2004.

[128] B. S. P. Dollar, C. Wojek and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, June 2009.

[129] B. S. P. Dollar, C. Wojek and P. Perona. Pedestrian detection: An evaluation of the state of the art. 2011.

[130] I. K. Park, M. Germann, M. D. Breitenstein, and H. Pfister. Fast and automatic object pose estimation for range images on the GPU. *Machine Vision and Applications*, pages 1–18, 2009.

[131] R. Perko and A. Leonardis. A framework for visual-context-aware object detection in still images. *Computer Vision and Image Understanding*, 114(6):700–711, 2010.

[132] C. A. Pickering, K. J. Burnham, and M. J. Richardson. A research study of hand gesture recognition technologies and applications for human vehicle interaction. In *3rd Conf. on Automotive Electronics*. Citeseer, 2007.

[133] J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[134] R. Polikar, L. Upda, S. S. Upda, and V. Honavar. Learn++: An incremental learning algorithm for supervised neural networks. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 31(4):497–508, 2001.

[135] D. B. Polley, E. E. Steinberg, and M. M. Merzenich. Perceptual learning directs auditory cortical map reorganization through top-down influences. *The journal of neuroscience*, 26(18):4970–4982, 2006.

[136] A. Pouget, P. Dayan, and R. Zemel. Information processing with population codes. *Nat Rev Neurosci*, 1(2):125–132, Nov 2000.

[137] T. Rabbani and F. V. D. Heuvel. Efficient hough transform for automatic detection of cylinders in point clouds. In *Proceedings of the 11th Annual Conference of the Advanced School for Computing and Imaging*, volume 3, pages 60–65, 2004.

[138] R. Rao and D. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *nature neuroscience*, 2(1):79, 1999.

[139] C. E. Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

[140] R. Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological Review*, 97, 1990.

[141] Z. Ren, J. Meng, and J. Yuan. Depth camera based hand gesture recognition and its applications in human-computer-interaction. In *Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on*, pages 1–5. IEEE, 2011.

[142] B. Ridge, D. S. caj, and A. Leonardis. Self-supervised cross-modal online learning of basic object affordances for developmental robotic systems. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Anchorage, USA, May 2010.

[143] B. Ridge, D. Sko˘ caj, and A. Leonardis. A system for learning basic object affordances using a self-organizing map. In T. Asfour and R. Dillmann, editors, *Proceedings of the 1st Int. Conf. on Cognitive Systems*, 2008.

[144] A. Riener, M. Rossbory, and A. Ferscha. Natural dvi based on intuitive hand gestures. In *Workshop UX in Cars, Interact*, page 5, 2011.

[145] M. Ristin, J. Gall, and L. Van Gool. Local context priors for object proposal generation. In *Computer Vision–ACCV 2012*, pages 57–70. Springer, 2013.

[146] H. Ritter and K. Schulten. On the stationary state of kohonen's self-organizing sensory mapping. *Biol. Cybern.*, 54(2):99–106, jun 1986.

[147] A. Robins. Catastrophic forgetting, rehearsal, and pseudorehearsal. *Connection Science*, 7, 1995.

[148] E. T. Rolls, G. Baylis, M. Hasselmo, and V. Nalwa. The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research*, 76(1):153–164, 1989.

[149] E. Rosch. Cognitive reference points. *Cognitive Psychology*, 7, 1975.

[150] D. A. Ross, M. Deroche, and T. J. Palmeri. Not just the norm: Exemplar-based models also predict face aftereffects. *Psychonomic bulletin & review*, 21(1):47–70, 2014.

[151] J. Rueckl. Jumpnet: A multiple-memory connectionist architecture. In *In Proceedings of the 15th Annual Conference of the Cognitive Science Society*. 1993.

[152] T. A. Runkler. *Data Analytics Models and Algorithms for Intelligent Data Analysis*. Springer Vieweg, 2012.

[153] S. Ruping. Incremental learning with support vector machines. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 641–642, 2001.

[154] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (FPFH) for 3D registration. 2009.

[155] R. E. Schapire. The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, pages 149–171. Springer, 2003.

[156] J. Schmuedderich, N. Einecke, S. Hasler, A. Gepperth, B. Bolder, R. Kastner, M. Franzius, S. Rebhan, B. Dittes, H. Wersing, J. Eggert, J. Fritsch, and C. Goerick. System approach for multi-purpose representations of traffic scene elements. In *IEEE International Conference on Intelligent Transportation Systems(ITSC)*, 2010.

[157] J. Schmuedderich, N. Einecke, S. Hasler, A. Gepperth, B. Bolder, R. Kastner, M. Franzius, S. Rebhan, B. Dittes, H. Wersing, J. Eggert, J. Fritsch, and C. Goerick. System approach for multi-purpose representations of traffic scene elements. In *International IEEE Annual Conference on Intelligent Transportation Systems*, 2010.

[158] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3626–3633, 2013.

[159] N. Sharkey and A. Sharkey. An analysis of catastrophic interference. *Connection Science*, 7(3-4), 1995.

[160] H. Shimizu and T. Poggio. Direction estimation of pedestrian from multiple still images. In *Intelligent Vehicles Symposium, 2004 IEEE*, pages 596–600. IEEE, 2004.

[161] O. Sigaud, C. Sagaün, and V. Padois. On-line regression algorithms for learning mechanical models of robots: A survey. *Robotics and Autonomous Systems*, 2011.

[162] D. Simon. *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. Wiley-Interscience, 2006.

[163] S. Sloman and D. Rumelhart. Reducing interference in distributed memories through episodic gating. In A. Healy and S. K. R. Shiffrin, editors, *Essays in Honor of W. K. Estes*. 1992.

[164] S. Soutschek, J. Penne, J. Hornegger, and J. Kornhuber. 3-d gesture-based scene navigation in medical imaging applications using time-of-flight cameras. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–6. IEEE, 2008.

[165] Y. Sun, J. Paik, A. Koschan, and M. A. Abidi. Point fingerprint: A new 3-d object representation scheme. *IEEE transaction on Systems, Man, and Cybernetics — Part B: Cybernetics*, 33:712–717, 2003.

[166] M. Swain and D. Ballard. Color indexing. *International journal of computer vision*, 7(1), 1991.

[167] A. Syed, H. Liu, and K. Sung. Incremental learning with support vector machines. 1999.

[168] D. Tanaka and L. Presnell. Color diagnosticity in object recognition. *Percept.Psychophysics*, 61, 1999.

[169] K. Tanaka. Inferotemporal cortex and object vision. *Annual review of neuroscience*, 19(1):109–139, 1996.

[170] M. Tang. Recognizing hand gestures with Microsoft's kinect. *Web Site: http://www.stanford.edu/ class/ee368/Project_ 11/ Reports/Tang_ Hand_ Gesture_ Recognition. pdf*, 2011.

[171] A. Torralba. Contextual priming for object detection. *IJCV*, 53:2003, 2003.

[172] A. Tsymbal. The problem of concept drift: definitions and related work. Technical report, Computer Science Department, Trinity College Dublin, 2004.

[173] M. Tyler and M. Spivey. Spoken language comprehension improves the efficiency of visual search. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, 2001.

[174] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2 edition, 2000.

[175] S. Vijayakumar and S. Schaal. Locally weighted projection regression: An o(n) algorithm for incremental real time learning in high-dimensional spaces. In *International Conference on Machine Learning*, 2000.

[176] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

[177] J. Vogel and O. D. Freitas. Target-directed attention: Sequential decision-making for gaze planning. In *International Conference on Robotics and Automation (ICRA)*, 2007.

[178] J. Vogel and K. Murphy. A non-myopic approach to visual search. In *Computer and Robot Vision*, volume 0, pages 227–234, Los Alamitos, CA, USA, 2007. IEEE Computer Society.

[179] J. P. Wachs, M. Kölsch, H. Stern, and Y. Edan. Vision-based hand-gesture applications. *Communications of the ACM*, 54(2):60–71, 2011.

[180] E. Wahl, U. Hillenbrand, and G. Hirzinger. Surflet-pair-relation histograms: a statistical 3d-shape representation for rapid classification. In *Proceedings of the Fourth International Conference on 3-D Digital Imaging and Modeling (3DIM)*, 2010.

[181] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional selection for object recongition - a gentle way. In *Lecture Notes in Computer Science*, volume 2525. Springer, 2002.

[182] N. M. Weinberger. The nucleus basalis and memory codes: Auditory cortical plasticity and the induction of specific, associative behavioral memory. *Neurobiology of Learning and Memory*, 80(3):268 – 284, 2003. Acetylcholine: Cognitive and Brain Functions.

[183] Y. Wen, C. Hu, G. Yu, and C. Wang. A robust method of detecting hand gestures using depth sensors. In *Haptic Audio Visual Environments and Games (HAVE), 2012 IEEE International Workshop on*, pages 72–77. IEEE, 2012.

[184] Y. M. Wen and B. L. Lu. Incremental learning of support vector machines by classifier combining. In *Proc. of 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2007)*, volume 4426 of *LNCS*, 2007.

[185] H. Wersing, S. Kirstein, B. Schneiders, U. Bauer-Wersing, and E. K.rner. Online learning for boostrapping of object recognition and localization in a biologically motivated architecture. In *Proc. Int. Conf. Computer Vision Systems ICVS. Santorini, Greece.*, pages 383–392, 2008.

[186] H. Wersing and E. K.rner. Learning optimized features for hierarchical models of invariant object recognition. *Neural Computation*, 15(7), 2003.

[187] C. Wojek, S. Walk, S. Roth, K. Schindler, and B. Schiele. Monocular visual scene understanding: Understanding multi-object traffic scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(4):882–897, 2013.

[188] J. Yu, D. Farin, and B. Schiele. Multi-target tracking in crowded scenes. In *Pattern Recognition*, pages 406–415. Springer, 2011.

[189] L. Zhang, B. Wu, and R. Nevatia. Detection and tracking of multiple humans with extensive pose articulation. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[190] Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *Int. J. Comp. Vis*, 7(3):119–152, 1994.

[191] A. Znagui Hassani, B. van Dijk, G. Ludden, and H. Eertink. Touch versus in-air hand gestures: evaluating the acceptance by seniors of human-robot interaction. *Ambient Intelligence*, pages 309–313, 2011.

# Curriculum vitae

## Personal information

| | |
|---|---|
| Name/age | Dr. Alexander Gepperth, born June 14th, 1976 in Munich |
| Familienstand | married, two children |
| Nationalität | German |

## International academic experience

| | |
|---|---|
| 5/2011 – present | Professor (tenured) at ENSTA ParisTech (France)<br>Responsibilities: research, teaching, acquisition of funds, specialization „transportation systems“, big data<br><br>Research areas: incremental and scalable machine learning, perception in robots and vehicles, human-machine-interaction |

## Industrial experience

| | |
|---|---|
| 02/2008 – 5/2011 | Premature tenure at Honda Research Institute Europe GmbH, Offenbach am Main |
| 10/2006 – 2/2008 | Four-year contract as Senior Scientist at Honda Research Institute Europe GmbH, Offenbach am Main<br><br>Responsibilities: basic research on bio-inspired machine learning in vehicles, creation of demonstrations, tutoring of internships, bachelor, master and PhD students |

## Doctoral studies

| | |
|---|---|
| 11/2002 – 04/2006 | at Ruhr-Universität Bochum, institute for neural computation<br><br>Topic: „Neural learning methods for visual object detection”<br><br>Degree: Dr. rer. nat (grade: „very good“)<br><br>Responsibilities: Research, teaching, participation in collaborative projects (Honda, Bosch, DFG SFB 475) |

## Stays abroad

| | |
|---|---|
| 02/2002 – 06/2002 | Spanish language studies at „Centro de lenguas modernas“, university of Granada (Spain) |

**Tertiary education**

| | |
|---|---|
| 10/1996 – 01/2002 | Studies of physics at Ludwig-Maximilians-Universität Munich |
| | Diploma thesis: „Non-BPS states in string theory” (grade: 1,7) |
| | Degree: “Diplom-Physiker” (grade: “very good“) |

**Alternate civil service**

| | |
|---|---|
| 8/1995 - 10/1996 | Kreiskrankenhaus Pfaffenhofen/Ilm, OP division |

**Secondary education**

| | |
|---|---|
| 6/1995 | “Abitur” at Schyren-Gymnasium Pfaffenhofen/Ilm, grade: 1,8 |

**Competences**

| | |
|---|---|
| Computers | Development: C/C++, CUDA, Python, Matlab |
| | Web development: HTML, CSS and PHP |
| | Operating systems: Windows, Linux |
| | Real-time middle-ware (ROS), embedded programming (AVR), GPU programming using CUDA |
| | LaTeX, svn, git, doxygen, bash, eclipse, gnuplot, make, cmake, .. |
| | Libraries: OpenCV, Qt, numpy/scipy, matplotlib/pylab, pycuda |
| Languages | German, Czech (mother tongues)<br>English: fluent (oral and written)<br>French: fluent (oral and written<br>Spanish: advanced level<br>Japanese: basic level |

**Interests**

Tennis, volleyball, body-building, Go, violin, StarCraft I und II

# Teaching and supervision activities

**a) Teaching**

- 1999, 2000, 2001, 2002, 2 SWS: study groups at LMU München. Subjects: calculus, experimental physics, basic mathematics
- 2004, 2005, 2006, 1 SWS: robotics seminar at Ruhr-Universität Bochum
- 2011, 2012, 2013, 2014, 2015, 2 SWS: study group an ENSTA ParisTech. subject: signal pocessing (French)
- 2012, 2014, 2015 ca. 3 SWS: planning and tutoring of a Matlab project seminar. Subject: traffic light detection (French)
- 2014, 2015, 2016: robotics project week, 3 SWS. Lecture and supervision of implementation work. Subject: system engineering using robots (French)
- 2014, 2015, 3 SWS: lecture and study group „Programming in Python“ (French)
- 2014, 2015, ½ SWS: lecture „image processing for robotics“ (French)
- 2015, ca. ½ SWS: lecture „signal theory“ (French)
- 2015, ca. 2 SWS: lecture series „Intelligent and Autonomous vehicles“ (Englisch)

**b) Supervision of bachelor, master and diploma theses**

- 2006: Thomas Kopinski, Ruhr-Universität Bochum (diploma thesis)
- 2007: Imran Bhatti, TU Darmstadt (master's thesis)
- 2008: Yongkie Wiyogo, TU Darmstadt (bachelor thesis)
- 2008: Vrushali Jedhe, TU Darmstadt (master's thesis)
- 2009: Michael Garcia Ortiz, MINATEC Grenoble (master's thesis)
- 2011: Nezha El Fakraoui, UPMC Paris (master's thesis)
- 2011: Mouloud Belounis, UPMC Paris (master's thesis)
- 2011: André Foessel, ENSTA ParisTech (master's thesis)
- 2011: Yacine Mokhtari, UPMC Paris (master's thesis)
- 2011: Brahim Belaoucha, UPMC Paris (master's thesis)
- 2013: Yang Song, ENSTA ParisTech („projet de recherche“ / bachelor thesis)
- 2013: Xiao Hu, université Paris-Sud (master's thesis)
- 2013: Evgeny Zuenko, ENSTA ParisTech („projet de recherche“ / bachelor thesis)

- 2013: Egor Sattarov, ENSTA ParisTech („projet de fin d'études“ / master's thesis)
- 2014: Yang Song, ENSTA ParisTech („projet de fin d'études“ / master's thesis)
- 2014: Tong Zhu, ENSTA ParisTech („projet de fin d'études“ / master's thesis)
- 2015: Maxime Bucher, université Paris-Sud (master's thesis)
- 2015: Mandar Gogate, Birla University, Indien (master's thesis)

**c) Supervision of PhD theses**

- since 2009: Michael Garcia Ortiz, Universität Bielefeld,successfully defended 7/2013
- 11/2012-11/2015: Louis-Charles Caron, ENSTA ParisTech, successfully defended 11/2015
- 11/2012-2/2016: Thomas Kopinski, Hochschule Ruhr-West, successfully defended 2/2016
- since 2013: Thomas Hecht, ENSTA ParisTech, defense planned for 11/2016
- since 2014: Egor Sattarov, ENSTA ParisTech/université Paris-Sud, defense planned for 1/2017

**d) Supervision of Post-Docs**

- 10/2012-10/2013: Michael Garcia Ortiz, „Real-time pedestrian detection“, funded by Honda Research Institute USA, Inc. **Permanent research position at Aldebaran Robotics (Paris) since 1/2014.**
- 10/2013-10/2015: Mathieu Lefort, „Neural algorithms for developmental learning“, funded by INRIA and the Digiteo consortium. **Permanent academic position since 10/2015.**
- 6/2015-present: Cem Karaoguz, „Incremental learning for object detection“, funded by MBDA Systems.

# List of publications and patents

## *Patents*

[P1A] Gepperth, A / Fritsch, J. Adaptive driver assistance system with robust estimation of object properties. EP Patent 2,065,842, 2012.
[P1B] Gepperth, A / Fritsch, J. Adaptive driver assistance systems with robust estimation of object properties. US Patent 8,175,782, 2012.
[P2A] Gepperth. A. ARTIFICIAL COGNITIVE SYSTEM WITH AMARI-TYPE DYNAMICS OF A NEURAL FIELD. EP Patent 2,215,588, 2012.
[P2B] Gepperth. A. ARTIFICIAL COGNITIVE SYSTEM WITH AMARI-TYPE DYNAMICS OF A NEURAL FIELD. US Patent App. 12/738,937, 2008.

## *Publications*

[62] Alexander Gepperth and Barbara Hammer. Incremental learning algorithms and applications. European Symposium on Artificial Neural Networks (ESANN), 2016. Accepted.

[61] Thomas Hecht and Alexander Gepperth. Towards deep incremental learning : automous change detection in a hierarchial generative architecture. European Symposium on Artificial Neural Networks (ESANN), 2016. Accepted.

[60] A Gepperth, M Garcia Oritz, E Sattarov and B Heisele. Dynamic attention priors: a new and efficient concept for improving object detection. Neurocomputing, 2016 (accepted).

[59] T Hecht, A Gepperth and M Gogate. A generative learning approach to sensor fusion and change detection. Cognitive Computation, 2016 (accepted).

[58] A Gepperth and C Karaoguz. A bio-inspired incremental learning architecture for applied perceptual problems. Cognitive Computation, 2016 (accepted).

[57] A Gepperth and M Lefort. Biologically inspired incremental learning for high-dimensional spaces. IEEE International Conference on Development and Learning (ICDL), 2015.

[56] T Hecht and A Gepperth. A generative-discriminative learning model for noisy information fusion. IEEE International Conference on Development and Learning (ICDL), 2015.

[55] M Lefort and A Gepperth. Active learning of local predictable representations with artificial curiosity. IEEE International Conference on Development and Learning (ICDL), 2015.

[54] E Sattarov, A Gepperth, S Rodriguez and R Reynaud. Calibration-free correspondence finding between vision and LIDAR sensors. IEEE International Symposium on Intelligent Vehicles (IV), 2015.

[53] T Kopinski, S Magand, U Handmann and A Gepperth. A light-weight real-time applicable hand gesture recognition system for automotive applications. IEEE International Symposium on Intelligent Vehicles (IV), 2015.

[52] T Kopinski, S Magand, U Handmann and A Gepperth. A pragmatic approach to multi-class classification. International Joint Conference On Neural Networks (IJCNN), 2015.

[51] M Lefort and A Gepperth. Learning of local predictable representations in partially structured environments. International Joint Conference On Neural Networks (IJCNN), 2015.

[50] T Kopinski, A Gepperth and U Handmann. A simple technique for improving multi-class classification with neural networks. European Symposium On Artificial Neural Networks (ESANN), 2015.

[49] M Lefort, T Hecht and A Gepperth. Using self-organizing maps for regression: the importance of the output function. European Symposium On Artificial Neural Networks (ESANN), 2015.

[48] A Gepperth, M Lefort, T Hecht and U Körner. Resource-efficient incremental learning in high dimensions. European Symposium On Artificial Neural Networks (ESANN), 2015.

[47] T Kopinski, S Geisler, A Gepperth and U Handmann. Time-of-Flight based multi-sensor fusion strategies for hand gesture recognition. IEEE International Symposium on Computational Intelligence and Informatics, 2014.

[46] L Caron, D Filliat and A Gepperth. Indoor RGB-D Object Recognition for Autonomous Mobile Robot. International Conference On Computer Vision (ICCV) Workshop Paper, 2014.

[45] T Kopinski, S Geisler, U Handmann and A Gepperth. Neural network based data fusion for hand pose recognition with multiple ToF sensors. International Conference on Artificial Neural Networks (ICANN), 2014.

[44] A Gepperth. Latency-based probabilistic information processing in recurrent neural hierarchies. International Conference On Artificial Neural Networks (ICANN), 2014.

[43] M Lefort, T Kopinski and A Gepperth. Multimodal space representation driven by self-evaluation of predictability. IEEE International Conference on Development and Learning (ICDL), 2014.

[42] T Kopinski, S Geisler, L Caron, A Gepperth and U Handmann. A real-time applicable 3D gesture recognition system for Automobile HMI. IEEE International Conference On Intelligent Transportation Systems (ITSC), 2014.

[41] E Sattarov, S Rodriguez, A Gepperth and R Reynaud. Context-based vector fields for multi-object tracking in application to road traffic. IEEE International Conference On Intelligent Transportation Systems (ITSC), 2014.

[40] A Gepperth, E Sattarov and S Rodrigues Flores. Robust visual pedestrian detection by tight coupling to tracking. IEEE International Conference On Intelligent Transportation Systems (ITSC), 2014.

[39] M Lefort and A Gepperth. Discrimination of visual pedestrians data by combining projection and prediction learning. European Symposium on Articificial Neural Networks (ESANN), 2014.

[38] L Caron, Y Song, D Filliat and A Gepperth. Neural network based 2D/3D fusion for robotic object recognition. European Symposium on Articificial Neural Networks (ESANN), 2014.

[37] M Lefort and A Gepperth. PROPRE: PROjection and PREdiction for multimodal correlations learning An application to pedestrians visual data discrimination. International Joint Conference on Neural Networks (IJCNN), 2014.

[36] A Gepperth and M Lefort. Latency-based probabilistic information processing in recurrent neural hierarchies. International Joint Conference on Neural Networks (IJCNN), 2014.

[35] X Hu, S Rodrigues and A Gepperth. A Multi-Modal System for Road Detection and Segmentation. IEEE International Symposium on Intelligent Vehicles(IV), 2014.

[34] T Hecht, M Mohit, E Sattarov and A Gepperth. Scene Context is more than a Bayesian prior: Competitive Vehicle Detection with Restricted Detectors. IEEE International Symposium on Intelligent Vehicles(IV), 2014.

[33] M Dubois, A Gepperth and D Filliat. A Comparison of Geometric and Energy-Based Point Cloud Semantic Segmentation Methods. European Conference on Mobile Robots (ECMR), 2013.

[32] A Gepperth. Processing and transmission of confidence in recurrent neural hierarchies. Neural Processing Letters, 2013.

[31] M Garcia Ortiz, A Gepperth and B Heisele. Real-time pedestrian detection and pose classification on a GPU. 16th International IEEE Conference on Intelligent Transportation Systems(ITSC), 2013.

[30] D Filliat, E Battesti, S Bazeille, G Duceux, A Gepperth, L Harrath, I Jebari, R Pereira, A Tapus, C Meyer, S Ieng, R Benosman, E Cizeron, J Mamanna and B Pothier. RGBD object recognition and visual texture classification for indoor semantic mapping. Proceedings of the 4th International Conference on Technologies for Practical Robot Applications (TePRA), 2012.

[29] A Gepperth. Co-training of context models for real-time object detection. IEEE International Symposium on Intelligent Vehicles(IV), 2012.

[28] A Gepperth, B Dittes and M Garcia Ortiz. The contribution of context information: a case study of object recognition in an intelligent car. Neurocomputing, 2012.

[27] A Gepperth. Efficient online bootstrapping of representations. Neural Networks, 2012.

[26] A Gepperth. Simultaneous concept formation driven by predictability. IEEE International conference on development and learning(ICDL), 2012.

[25] M Garcia Ortiz, F Kummert, J Fritsch and A Gepperth. Behavior prediction at multiple time-scales in inner-city scenarios. IEEE Symposium on Intelligent Vehicles (IV), 2011.

[24] M Garcia Ortiz, F Kummert, J Fritsch and A Gepperth. Situation-specific learning for ego-vehicle behavior prediction systems. IEEE International Conference on Intelligent Transportation Systems(ITSC), 2011.

[23] A Gepperth, S Rebhan, S Hasler and J Fritsch. Biased Competition in Visual Processing Hierarchies: A Learning Approach Using Multiple Cues. Cognitive Computation, 2011.

[22] M Garcia Ortiz and A Gepperth. Autonomous generation of internal representations for associative learning. International Conference on Artificial Neural Networks (ICANN), 2010.

[21] J Schmuedderich, N Einecke, S Hasler, A Gepperth, B Bolder, R Kastner, M Franzius, S Rebhan, B Dittes, H Wersing, J Eggert, J Fritsch and C Goerick. System approach for multi-purpose representations of traffic scene elements. IEEE International Conference on Intelligent Transportation Systems(ITSC), 2010.

[20] A Gepperth. Implementation and evaluation of a large-scale object detection system. , 2010.

[19] B Dittes, M Heracles, T Michalke, R Kastner, A Gepperth, J Fritsch and C Goerick. A Hierarchical System Integration Approach with Application to Visual Scene Exploration for Driver Assistance. The 5th International Conference on Computer Vision Systems (ICVS), 2009.

[18] M Garcia Ortiz and A Gepperth. Neural Self-adaptation for large-scale system building. International Conference on Cognitive Neurodynamics(ICCN), 2009.

[17] B Mersch, A Gepperth, S Suhai and A Hotz-Wagenblatt. Automatic detection of exonic splicing enhancers (ESEs) using SVMs. BMC bioinformatics, 2008.

[16] T Michalke, R Kastner, J Adamy, S Bone, F Waibel, M Kleinehagenbrock, J Gayko, A Gepperth, J Fritsch and C Goerick. An Attention-based System Approach for Scene Analysis in Driver Assistance. at - Automatisierungstechnik, 2008.

[15] A Gepperth, J Fritsch and C Goerick. Cross-module learning as a first step towards a cognitive system concept. International Conference On Cognitive Systems, 2008.

[14] A Gepperth, J Fritsch and C Goerick. Computationally Efficient Neural Field Dynamics. European Symposium on Artificial Neural Networks(ESANN), 2008.

[13] T Michalke, A Gepperth, M Schneider, J Fritsch and C Goerick. Towards a Human-like Vision System for Resource-Constrained Intelligent Cars. International Conference on Computer Vision Systems (ICVS) Conference Paper, 2007.

[12] A Gepperth, B Mersch, J Fritsch and C Goerick. Color object recognition in real-world scenes. International Conference on Artificial NEural Networks([ICANN), 2007.

[11] A Gepperth. Neural learning methods for visual object detection. PhD thesis at the university of Bochum (Germany), 2006.

[10] A Gepperth. Visual object classification by sparse convolutional neural networks. Proceedings of the 14th European Symposium on Artificial Neural Networks (ESANN), Brugge, Belgium, 2006.

[9] A Gepperth. Object detection and feature base learning by sparse convolutional neural networks. Proceedings of the 2nd IAPR TC3 Workshop on Artificial Neural Networks in Pattern Recognition, 2006.

[8] A Gepperth and S Roth. Applications of multi-objective structure optimization. Neurocomputing, 2006.

[7] A Gepperth, J Edelbrunner and T Bücher. Videobasierte Klassifikation von Fahrzeugen in Echtzeit. Tagungsband des 3 Workshops Fahrerassistenzsysteme, Walting, 2005.

[6] A Gepperth, J Edelbrunner and T Bücher. Real-time detection of cars in video sequences. IEEE Intelligent Vehicles Symposium (IV), 2005.

[5] A Gepperth and S Roth. Applications of Multi-objective Structure Optimization.

European Symposium on Artificial Neural Networks(ESANN), 2005.

[4] S Roth and C Igel. Multi-objective structure optimization for visual object detection. Multi-objective Machine Learning, 2005.

[3] K Weinert, O Webber, A Gepperth, Y Zhang and W Theis. Time varying dynamics in BTA deep hole drilling. Intelligent Computation in Manufacturing Engineering, 2004.

[2] K Weinert, O Webber, A Gepperth, Y Zhang and W Theis. Towards a dynamical system model of the BTA deep hole drilling process. Production Engineering - Research and Development, Annals of the German Academic Society for Production Engineering, 2004.

[1] A Gepperth. Nicht-BPS-Zustände in der Stringtheorie. Diploma thesis at the Ludwig-Maximilians-Universität München, 2002.

## *The five principal publications:*

[58] A Gepperth and C Karaoguz. A bio-inspired incremental learning architecture for applied perceptual problems. Cognitive Computation, 2016 (accepted).

[59] T Hecht, A Gepperth and M Gogate. A generative learning approach to sensor fusion and change detection. Cognitive Computation, 2016 (accepted).

[26] A Gepperth. Simultaneous concept formation driven by predictability. IEEE International conference on development and learning(ICDL), 2012.

[27] A Gepperth. Efficient online bootstrapping of representations. Neural Networks, 2012.

[60] A Gepperth, M Garcia Oritz, E Sattarov and B Heisele. Dynamic attention priors: a new and efficient concept for improving object detection. Neurocomputing, 2016 (accepted).

# Projects funded by third parties 2012-2016

1/2012-11/2015: PhD project (100.000€ für 3 Jahre). Subject: „Learning in multi-modal sensorimotor loops“, sipend of the government of Canada.

3/2012-4/2013: Post-Doc project (70.000€ for one year), Subject: GPU-based pedestrian detection and tracking in real time“, funded by Honda Research Institute USA Inc.

3/2013-9/2013: master thesis project (6000€) „real-time road detection“. Stipend of „Université Paris-Sud“, collaboration with the group ACCIS at UP-Sud.

10/2013-11/2014: PostDoc project (70.000€ for one year). Subject: „Neural learning methods for developental learning“, stipend of the INRIA consortiums (France, similar to Max-Planck-Gesellschaft). Collaboration with the INRIA FLOWERS group (Bordeaux, France) led by Pièrre-Yves Oudeyer.

since 11/2013: PhD project (100.000€ over 3 years). Subject: „Bio-inspired learning methods for optimal data fusion“, grant of the „Direction générale de l'armement“ (DGA) and „Ecole Polytechnique“.

since 1/2014: PhD project (100.000€ over 3 years). Subject: Multi-modal architecture for object recognition“. Stipend of the „Digiteo“ consortiums (France/Paris region). Collaboration with the team ACCIS at „Université Paris-Sud“.

1/2014-1/2015: PhD project (ca. 80.000€ over 2 years). Subject: „Machine learning methods for human-machine interaction“, stipend of the HRW Bottrop. Collaboration with the department of computer science of HRW Bottrop.

3/2015-10/2015: PostDoc project (ca. 65.000€ for one year). Subject: „Perception-action loops“, funded by the „Digiteo“ consortium (France/Paris region). Collaboration with the group „A-O“ at „université Paris-Sud“ led by M.Sebag.

since 6/2015: PostDoc project (85.000€ over 18 months). Subject: „Incremental learning for object detection“, funded by MBDA Missile Systems France.

since 10/2015: general research support grant (150.000€ over 3 years). Subject: „Incremental machine learning“, funded by the „Direction générale de l'armement“ (DGA).

since 12/2015: Feasibility study (50.000€ over 6 months). Subject: „Multi-view vehicle detection“, funded by MOBIS Parts Europe (Frankfurt am Main).

since 12/2015: participation in H2020-project DREAM

**total: ~ 920.000€ / ~4 years = 230.000€/year**