

# Beyond the closed-world assumption: The importance of novelty detection and open set recognition

Joachim Denzler, Erik Rodner, Alexander Freytag, Paul Bodesheim



Computer Vision Group  
Friedrich-Schiller University of Jena

Workshop Unsolved Problems in Pattern Recognition 2013  
03.09.2013

# Big Data: A Buzzword or Real Problem?

Big data computer vision or large scale computer vision involves

- scalable image analysis algorithms
- recognition and retrieval techniques in large image datasets
- treatment of dynamic, complex, multidimensional and multi-modal data

Usually we talk about

- high dimensional features (hundreds of thousands)
- large variety of visual classes (tens of thousands)
- large number of examples (hundreds of thousands to millions)

# Big Data: A Buzzword or Real Problem?

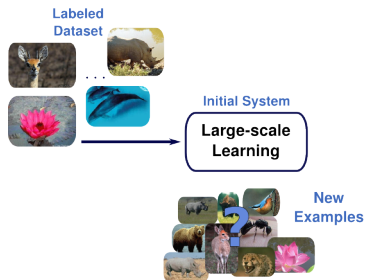
However, big data also means

- we will have **unexpected** data in our data sources
- **not every** item is labeled (either correctly or at all)
- there is the need to analyze the **unknown underlying characteristics** of a given data set
- we want to **incrementally update** our knowledge and models

Applications not only arise from computer science:

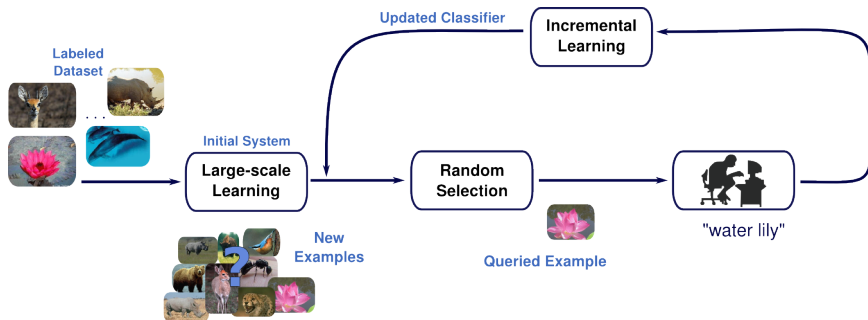
- modern imaging devices allow researchers from different disciplines to create really big data (without even looking at each image or frame)
- in some cases neither labeled data sets nor knowledge about the individual classes to be recognized are available a priori
- examples: remote sensing data, videos from confocal laser scanning microscopy, analysis of interactions between the biosphere and the atmosphere, etc.

# Our Fundamental Goal: Lifelong Learning

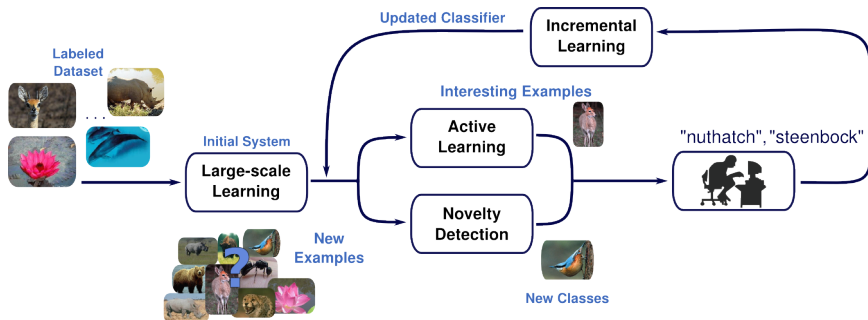




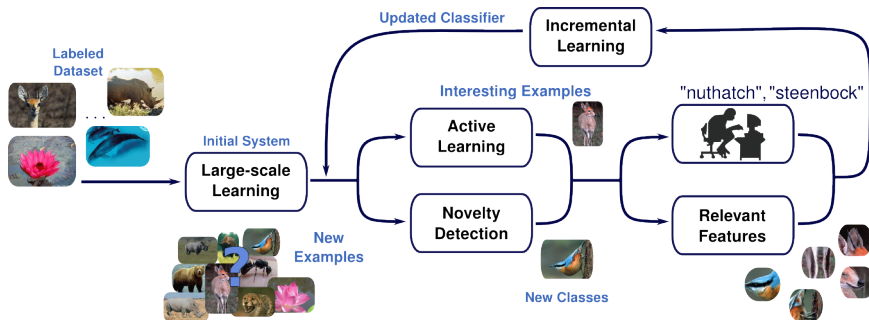
# Our Fundamental Goal: Lifelong Learning



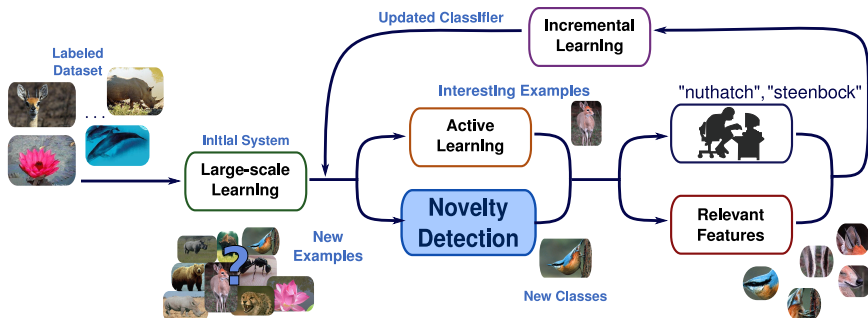
# Our Fundamental Goal: Lifelong Learning



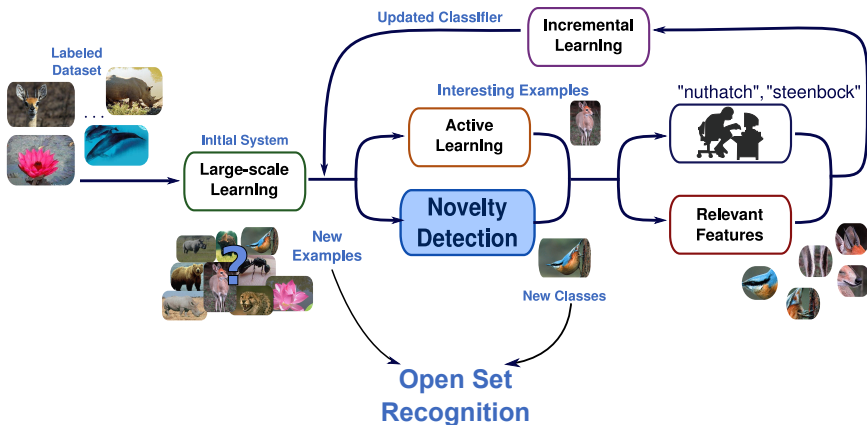
# Our Fundamental Goal: Lifelong Learning



# Our Fundamental Goal: Lifelong Learning



# Our Fundamental Goal: Lifelong Learning



# Outline

- 1 Motivation
- 2 Open Set Recognition
- 3 Novelty Detection
- 4 Multi-Class Novelty Detection: KNFST
- 5 Experiments
- 6 Conclusion

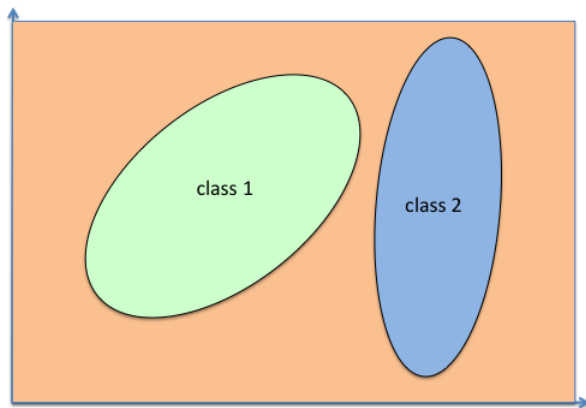
# Closed World Assumption

Standard approach in pattern recognition: closed-world assumption

- according to Niemann (1981): there is a finite set of ideal objects  $o_{\kappa,\lambda}$ ,  $\kappa = 1, \dots, k$  being the number of classes, and  $\lambda = 1, \dots, l$  being the number of ideals per class.
- the classes arise from the given problem domain
- there might be a rejection class  $o_0$  to indicate that a decision into the known classes is not possible

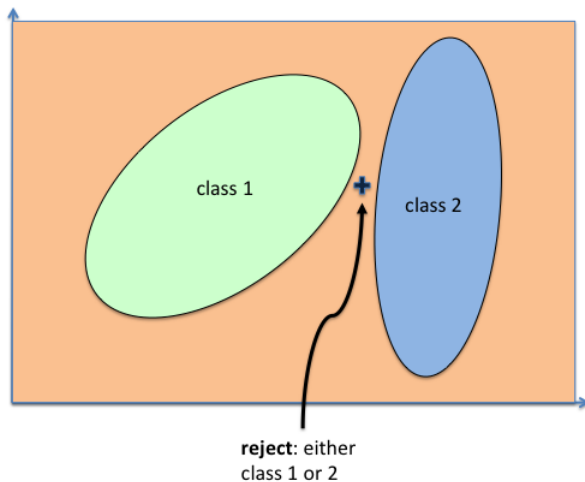
Researchers identified quite early problems with the closed-world assumption in practical applications, like speech recognition (*out-of-vocabulary problem*)

# From closed world assumption to open set recognition

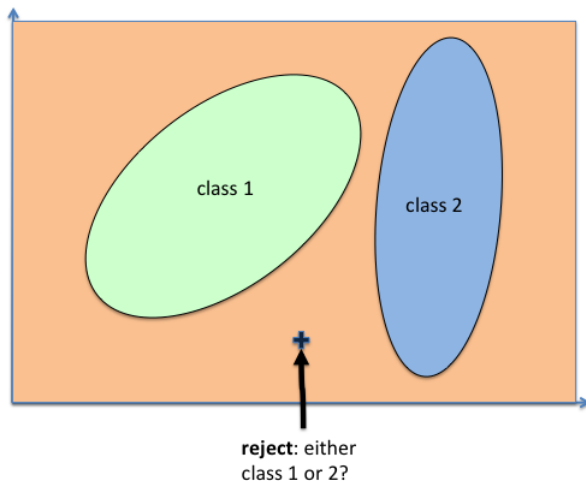




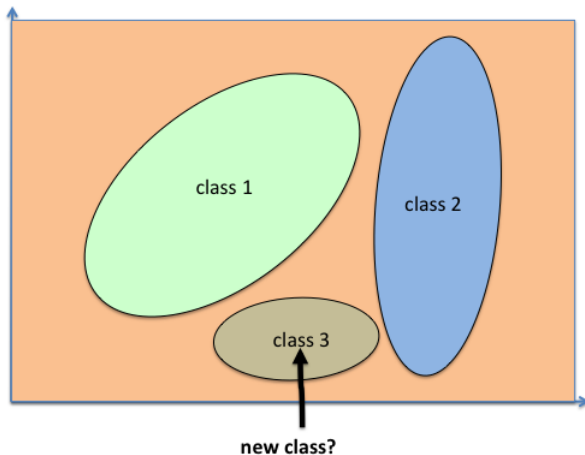
# From closed world assumption to open set recognition



# From closed world assumption to open set recognition



# From closed world assumption to open set recognition



# Definition: Open Set Recognition

Most recent work: Scheierer et al.: Towards Open Set Recognition. TPAMI, vol. 37, no. 7, July 2013

- closed set recognition: all testing classes are known at training time
- open set recognition: incomplete knowledge about the world is present at training time and unknown classes can be submitted to the algorithm during testing

We need to distinguish between the rejection option (i.e. a decision is not possible), the detection of new, novel, unknown classes (novelty detection), but at the same time being able to classify the known classes correctly

Of course several other problems arise in such scenarios, like incremental update of models, optimal feature selection, metric learning, etc.

# Practical Relevance

Closed world assumption is valid and working in practice for many applications (for example, OCR)

Big data computer vision changes the situation:

- we cannot have knowledge about the entire set of possible classes
- sometimes even under the closed world assumption, not every possible class is known a priori

Sample applications:

- service robots: new events, objects, situations will appear
- analysis of biological data: besides of known classes, sometimes the unknown are the most important ones (for examples, unknown bacteria)
- complex event detection in videos: novel events must be recognized
- quality control: impossible to define all possible defects a priori, or if so, not sufficient training data available

# Problem Specification

Open set recognition leads to new problems:

- how can we determine whether or not a **set** of (local) features is **novel** with respect to the training set?
- how can we determine the novelty of **constellations** of features?
- how to **measure** the novelty of class constellations (e.g. a car in a lake?)
- how to measure the **level of novelty** (novelty vs. abnormality vs. normality)?
- how to decide for novelty and at the same time classify objects that are not considered as novel?

# Challenges

- methods to define features to discriminate between new objects and to additionally separate them from unknown objects
- the question, whether discriminative or generative models or a mixture of both are preferable or necessary
- incremental update of features and feature space as soon as new objects or events are identified
- the question, how novelty can be measures and whether or not metrics need to be learned?
- is context information beneficial, irrelevant or even misleading?

Summary: effective methods for **multi-class novelty detection** are a preliminary to open set recognition

# Next Section

- 1 Motivation
- 2 Open Set Recognition
- 3 Novelty Detection**
- 4 Multi-Class Novelty Detection: KNFST
- 5 Experiments
- 6 Conclusion



# Novelty Detection vs. One-Class Classification

## Novelty Detection



# Novelty Detection vs. One-Class Classification

## Novelty Detection



## One-class Classification (OCC)



## Support Vector Data Description (SVDD) [Tax and Duin, 2004]

## Measuring Novelty

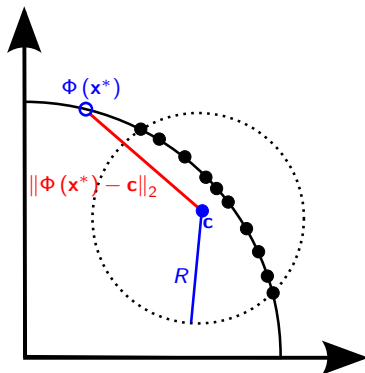
- enclose data with smallest hypersphere in reproducing kernel Hilbert space (RKHS)

$$\min_{c, R, \xi} R^2 + \frac{1}{\nu N} \sum_{i=1}^N \xi_i$$

subject to

$$\|\Phi(\mathbf{x}^{(i)}) - \mathbf{c}\|_2^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0$$

$$\text{score}(\mathbf{x}^*) = -\|\Phi(\mathbf{x}^*) - \mathbf{c}\|_2^2$$

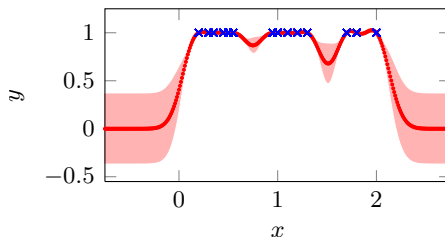


# Gaussian Process Regression (GPR) [Kemmler et al., 2010]

## Assumptions

- outputs:  $y(\mathbf{x}) = f(\mathbf{x}) + \varepsilon$
- latent fun.:  $f \sim \mathcal{GP}(\mathbf{0}, \mathbf{K})$
- noise term:  $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$

$$\Rightarrow y(\mathbf{x}^*) | \mathbf{X}, \mathbf{y}, \mathbf{x}^* \sim \mathcal{N}(\mu_*, \sigma_*^2)$$

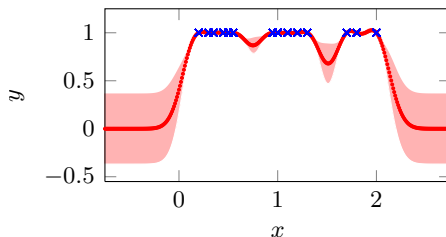


# Gaussian Process Regression (GPR) [Kemmler et al., 2010]

## Assumptions

- outputs:  $y(\mathbf{x}) = f(\mathbf{x}) + \varepsilon$
- latent fun.:  $f \sim \mathcal{GP}(\mathbf{0}, \mathbf{K})$
- noise term:  $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$

$$\Rightarrow y(\mathbf{x}^*) | \mathbf{X}, \mathbf{y}, \mathbf{x}^* \sim \mathcal{N}(\mu_*, \sigma_*^2)$$



## GPR-Mean

$$\begin{aligned} \text{score}(\mathbf{x}^*) &= \mu_* \\ &= \mathbf{k}_*^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{1} \end{aligned}$$

## GPR-Var

$$\begin{aligned} \text{score}(\mathbf{x}^*) &= -\sigma_*^2 \\ &= -\left( \mathbf{k}_{**} - \mathbf{k}_*^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{k}_* + \sigma_n^2 \right) \end{aligned}$$

# Kernel PCA [Hoffmann, 2007] and Kernel ECA [Jenssen, 2010]

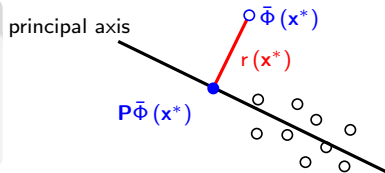
## KPCA vs. KECA

- project on principal axes in RKHS based on largest
  - variance (KPCA)
  - approx. of Renyi entropy (KECA)

## Kernel PCA [Hoffmann, 2007] and Kernel ECA [Jenssen, 2010]

## KPCA vs. KECA

- project on principal axes in RKHS based on largest
  - variance (KPCA)
  - approx. of Renyi entropy (KECA)



## Novelty Measure: Reconstruction Error in RKHS

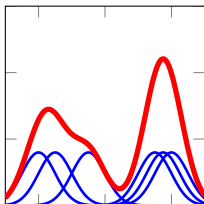
$$\text{score}(x^*) = -r(x^*)$$

$$\begin{aligned} r(x^*) &= \|\bar{\Phi}(x^*) - P\bar{\Phi}(x^*)\|^2 \\ &= \langle \bar{\Phi}(x^*), \bar{\Phi}(x^*) \rangle - \langle P\bar{\Phi}(x^*), P\bar{\Phi}(x^*) \rangle \end{aligned}$$

## Parzen Density Estimation and NN-Description [Tax and Duin, 2000]

## Parzen Density Estimation

$$\text{score}(\mathbf{x}^*) = \frac{1}{N} \sum_{i=1}^N \kappa(\mathbf{x}^*, \mathbf{x}^{(i)})$$

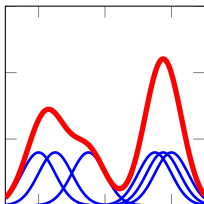




## Parzen Density Estimation and NN-Description [Tax and Duin, 2000]

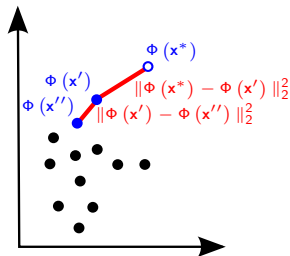
## Parzen Density Estimation

$$\text{score}(\mathbf{x}^*) = \frac{1}{N} \sum_{i=1}^N \kappa(\mathbf{x}^*, \mathbf{x}^{(i)})$$



## NN-Ratio

- $\mathbf{x}' = \text{NN}(\mathbf{x}^*)$  and  $\mathbf{x}'' = \text{NN}(\mathbf{x}')$
- $$\text{score}(\mathbf{x}^*) = - \frac{\kappa(\mathbf{x}^*, \mathbf{x}^*) - 2\kappa(\mathbf{x}^*, \mathbf{x}') + \kappa(\mathbf{x}', \mathbf{x}')}{\kappa(\mathbf{x}', \mathbf{x}') - 2\kappa(\mathbf{x}', \mathbf{x}'') + \kappa(\mathbf{x}'', \mathbf{x}'')}$$



# Next Section

- 1 Motivation
- 2 Open Set Recognition
- 3 Novelty Detection
- 4 Multi-Class Novelty Detection: KNFST**
- 5 Experiments
- 6 Conclusion

# Reminder: Novelty Detection

## Out-of-vocabulary



## One-class Classification (OCC)

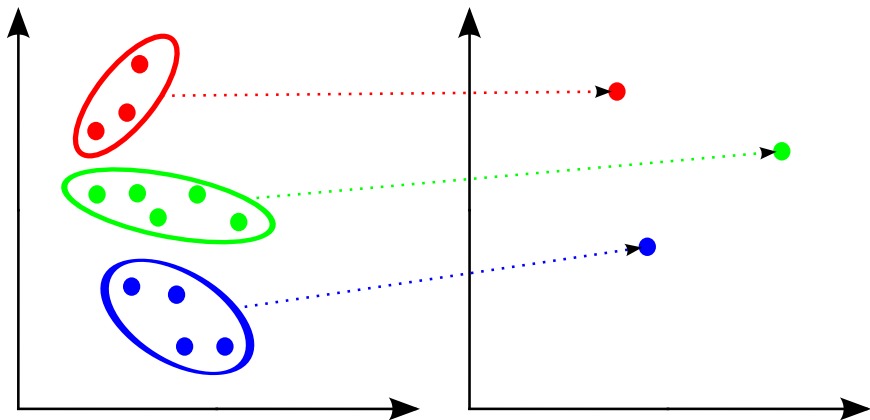


What is the difference to Multi-Class Novelty Detection?

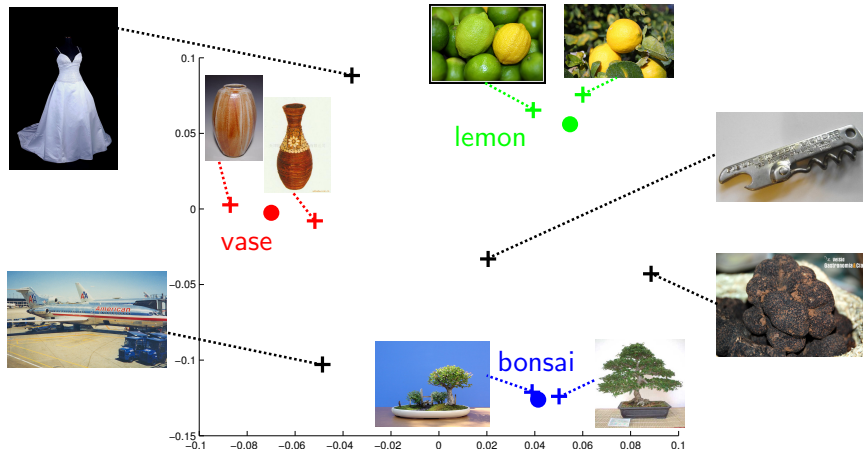
# Demands

- represent all known classes in one common space → no calibration of distance measures of one-vs-all novelty detection decisions necessary
- avoid artificial negative classes
- aim at features with zero intra-class variance
- avoid hyperparameters (outlier ratio, noise variance)

# Multi-Class Novelty Detection: Training



# Multi-Class Novelty Detection: Testing



# Null Foley-Sammon Transform (NFST) [Guo et al., 2006]

## Fisher Discriminant Criterion

- $N$  data points  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}$ ,  $c$  classes
- $\max J(\varphi) = \frac{\varphi^T \mathbf{S}_b \varphi}{\varphi^T \mathbf{S}_w \varphi}$
- $\mathbf{S}_b$ : between-class scatter matrix,  $\mathbf{S}_w$ : within-class scatter matrix

# Null Foley-Sammon Transform (NFST) [Guo et al., 2006]

## Fisher Discriminant Criterion

- $N$  data points  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}$ ,  $c$  classes
- $\max J(\varphi) = \frac{\varphi^T \mathbf{S}_b \varphi}{\varphi^T \mathbf{S}_w \varphi}$
- $\mathbf{S}_b$ : between-class scatter matrix,  $\mathbf{S}_w$ : within-class scatter matrix

## Null Space

- $J(\varphi) = \infty$  if  $\varphi$  satisfy  $\varphi^T \mathbf{S}_w \varphi = 0$  and  $\varphi^T \mathbf{S}_b \varphi > 0$
- null space spanned by null projection directions  $\varphi$
- $c - 1$  null projection directions (small sample size:  $N < D$ )



# Kernel Null Foley-Sammon Transform (KNFST)

## Kernelization

- algorithm only needs inner products of data points  
⇒ kernel trick
- using the right kernel overcomes small sample size problem

# Kernel Null Foley-Sammon Transform (KNFST)

## Kernelization

- algorithm only needs inner products of data points
  - ⇒ kernel trick
- using the right kernel overcomes small sample size problem

## KNFST and OCC

- only data of a single target class vs.  $c - 1$  null projection directions
- separating data from „minus data“
  - ⇒ one null projection direction (1D null space)
  - ⇒ one single value for data ( $v$ ) and one for „minus data“

# Next Section

- 1 Motivation
- 2 Open Set Recognition
- 3 Novelty Detection
- 4 Multi-Class Novelty Detection: KNFST
- 5 Experiments**
- 6 Conclusion

# Experimental Setup

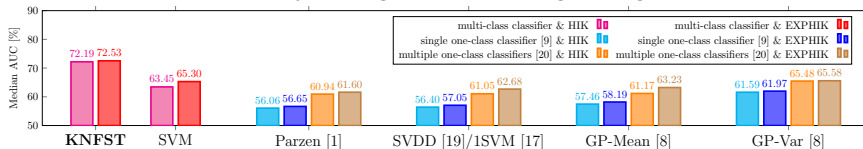
- medium scale dataset: Caltech-256 [Griffin et al., 2007]
- large-scale dataset: ImageNet [Deng et al., 2009]
  - 1,000 object classes as done for ILSVRC 2010<sup>1</sup>
  - training set: 100 samples per class
  - validation set: 50 samples per class
- evaluation and comparison: area under the ROC curve (AUC)
- features: bag-of-visual-words histograms from densely sampled SIFT features
- kernels: histogram intersection kernel (HIK) and generalized rbf-kernel

---

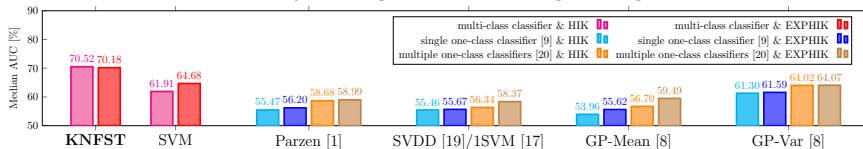
<sup>1</sup> <http://www.image-net.org/challenges/LSVRC/2010>

# Result: Caltec-256

## Five object categories known during training

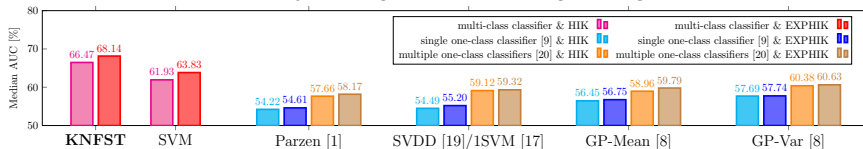


## Ten object categories known during training

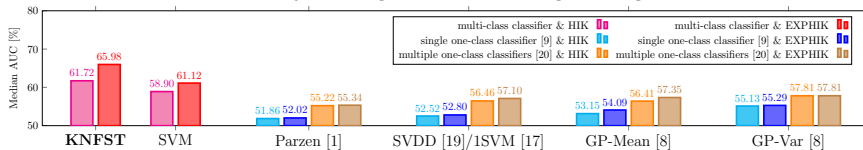


# Result: ImageNet

## Five object categories known during training



## Ten object categories known during training



# Conclusion

- definition of challenges in open set recognition
- introduction of KNFST for multi-class novelty detection
- results for two public datasets show benefits of the suggested method
- Details:
  - Bodesheim, Freytag, Rodner, Kemmler, Denzler: Kernel Null Space Methods for Novelty Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013. 3374–3381.
  - project webpage:  
[http://www.inf-cv.uni-jena.de/novelty\\_detection.html](http://www.inf-cv.uni-jena.de/novelty_detection.html)
  - code available!

# Open Questions and Future Research

- shall we treat novelty detection and classification in one framework or separate steps?
- how robust is the method, if we have outliers in the training data?
- how sensitive is the the Null-space if classes overlap?
- can we manage complexity of this method (currently result for 10 classes only)?
- are discriminative models better than generative (or the opposite), or a combination?





# Thank you for your attention!


Related publications and more detailed information can be found at


<http://www.inf-cv.uni-jena.de>


# Literature I


 Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009).  
ImageNet: A Large-Scale Hierarchical Image Database.  
In *CVPR*.


 Griffin, G., Holub, A., and Perona, P. (2007).  
Caltech-256 object category dataset.  
Technical Report UCB/CSD-04-1366, California Institute of Technology.

 Guo, Y.-F., Wu, L., Lu, H., Feng, Z., and Xue, X. (2006).  
Null foley-sammon transform.  
*Pattern Recognition*, 39(11):2248–2251.

 Hoffmann, H. (2007).  
Kernel pca for novelty detection.  
*Pattern Recognition*, 40(3):863–874.

 Jenssen, R. (2010).  
Kernel entropy component analysis.  
*PAMI*, 32(5):847–860.

 Kemmler, M., Rodner, E., and Denzler, J. (2010).  
One-class classification with gaussian processes.  
In *ACCV*, pages 489–500.

 Tax, D. M. J. and Duin, R. P. W. (2000).  
Data description in subspaces.  
In *ICPR*, pages 672–675.

# Literature II



Tax, D. M. J. and Duin, R. P. W. (2004).

Support vector data description.

*Mach. Learn.*, 54(1):45–66.