

Categorizing Nine Visual Classes with Bags of Keypoints

G. Csurka, J. Willamowski, C. Dance, L. Fan, D. Arregui



Advertisements

- ❑ **LAVS'04: Workshop on Learning for Adaptable Visual Systems at ICPR, Cambridge, August 2004**
 - Call for papers still open!
 - Collect a description if interested

- ❑ **Job Opening: Post-doctoral researcher in image processing at XRCE**
 - Currently interviewing – please email me CV + example papers, asap!
 - Collect a description if you know someone who might be interested

LAVA Project



- **Objective:** bringing learning and vision together for **visual categorization** and **event interpretation**
- IST project, 7 partners (coordinator XRCE)
- Just completed Year 2

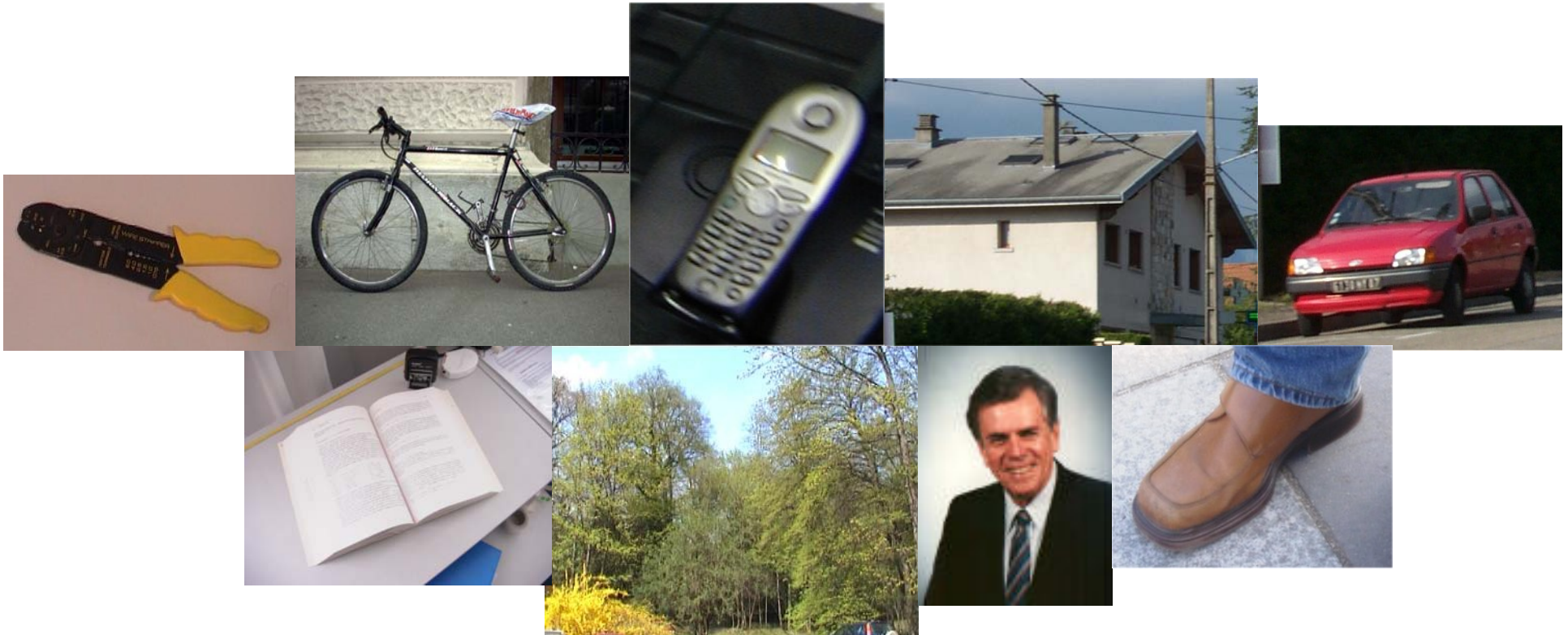
Outline

- ❑ **The problem**
- ❑ **Bag of keypoints approach**
- ❑ **Results**
- ❑ **Conclusions**

[Problem statement]

Generic **Visual** Categorization

- ❑ Common framework for many image and object categories
- ❑ Cope with lighting, view, background, occlusion variations



[Problem statement]

Generic Visual **Categorization**

- ❑ **Cope with intra-object-within-class variations and an open set of object instances**



Applications

❑ **Tagging images with content:**

- web image retrieval (combined with text information)
- images in documents
- photographic archives

❑ **Assisting other processing:**

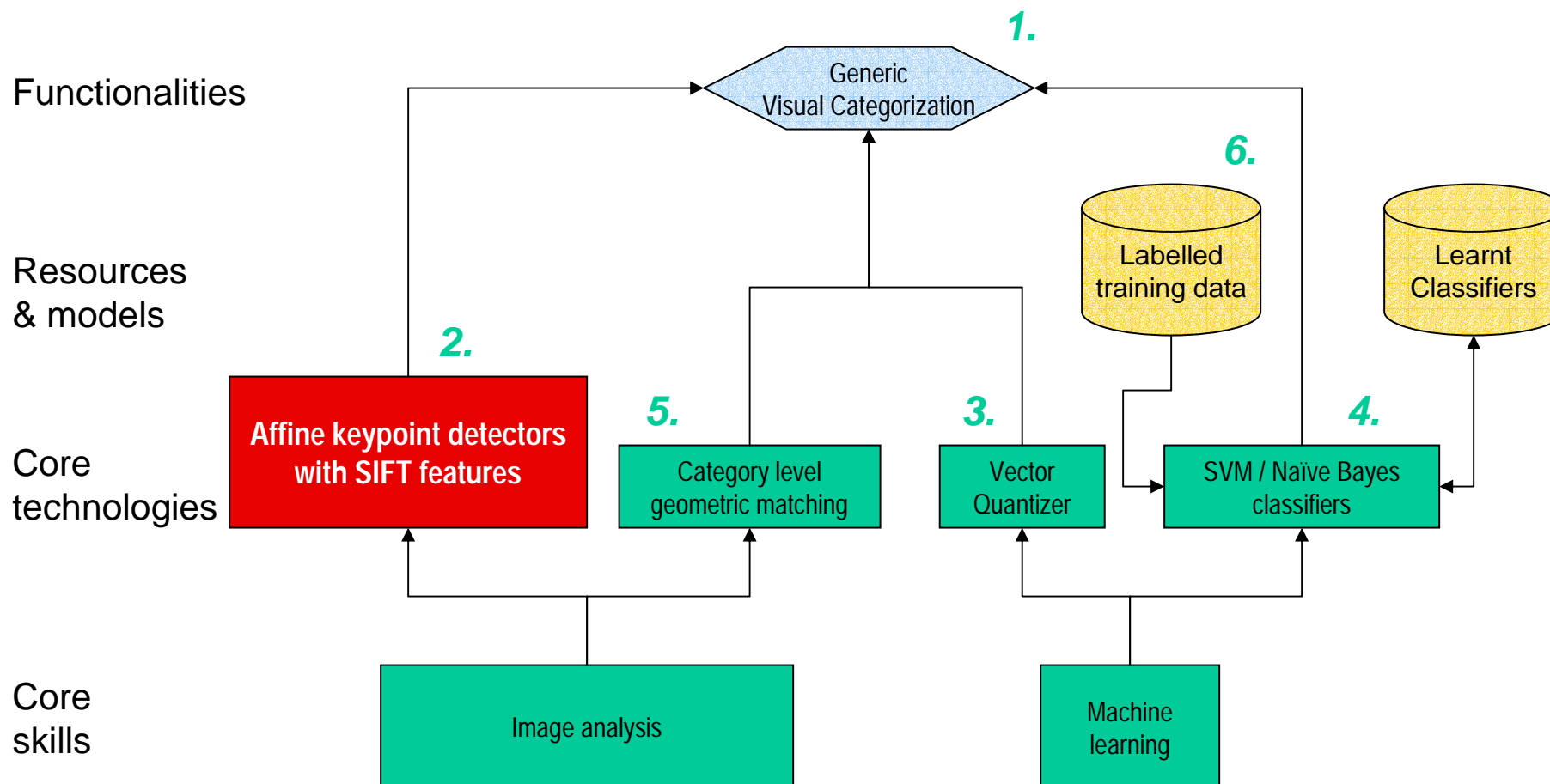
- eg image enhancement: red-eye, teeth whitening filters

[Approach] Outline

1. Get local appearance **descriptors** for the input image
2. **Vector quantize** these descriptors
3. Make a histogram of quanta = “**bag of keypoints**”
4. **Classify** histograms into visual categories

[Approach] Keypoint Detectors

Repeatably detectable "quasi-invariant" parts

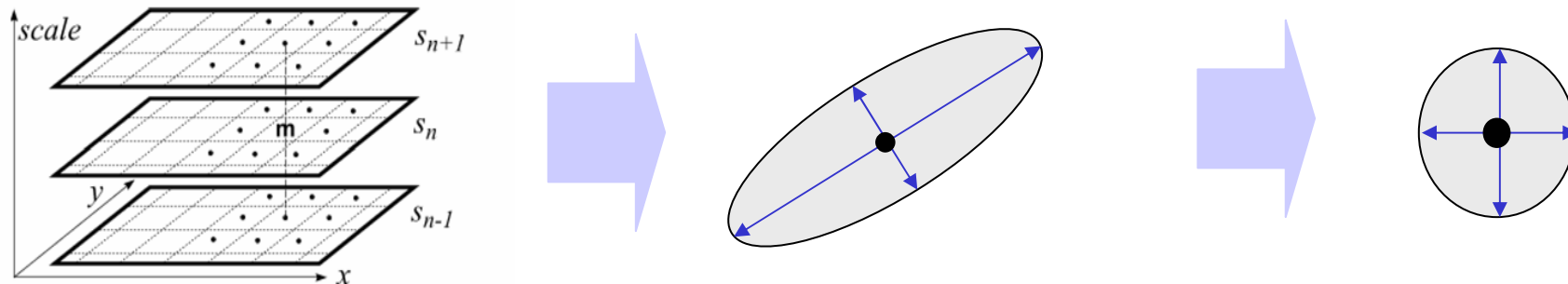


[Approach] Keypoints Sparse image description

- ❑ Local image features give robustness to occlusion and characterize multi-part objects
- ❑ Need to define: point detector and descriptor



[Approach] Harris affine detector



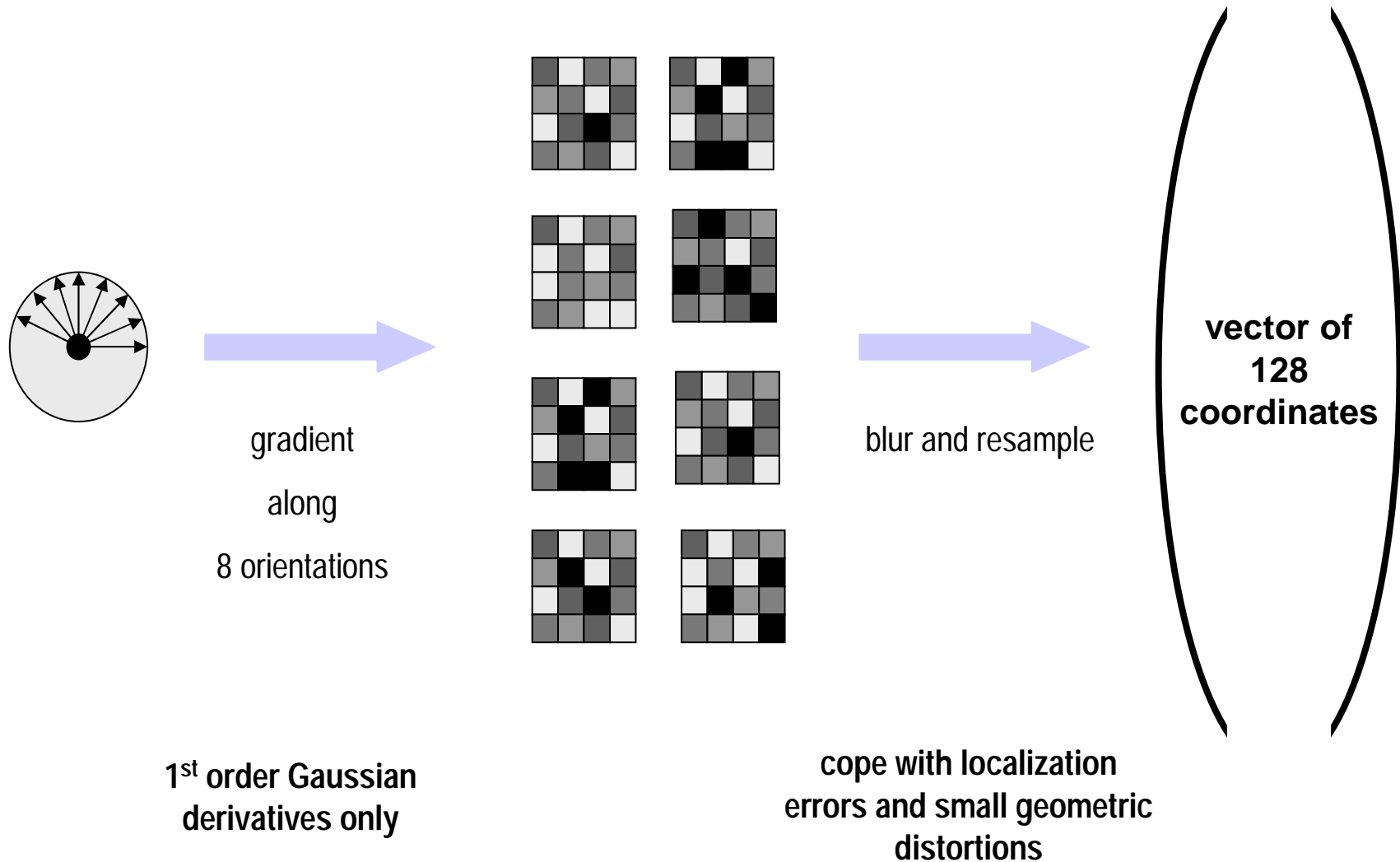
- ❑ Harris affine detector returns localizations
- ❑ Scale selection at maxima of the Laplacian

❑ Local affine invariant neighborhood is determined by a search related to a statistical bound.

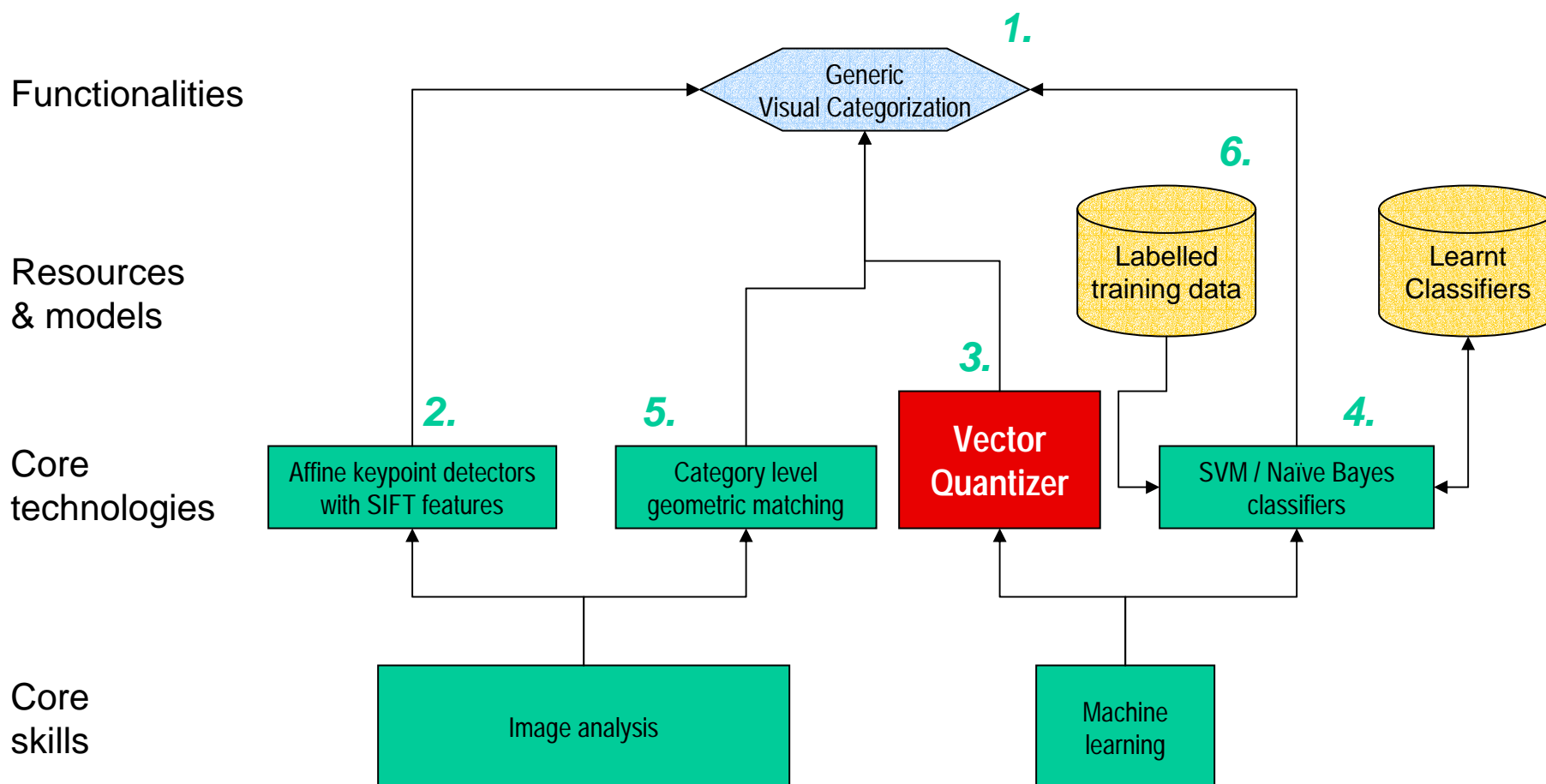
❑ Affine region is mapped to a circular region normalized according to scale, orientation and illumination

Aim: repeatable, invariant regions

[Approach] SIFT orientation maps



[Approach] Vector Quantizer Making the Visual Alphabet



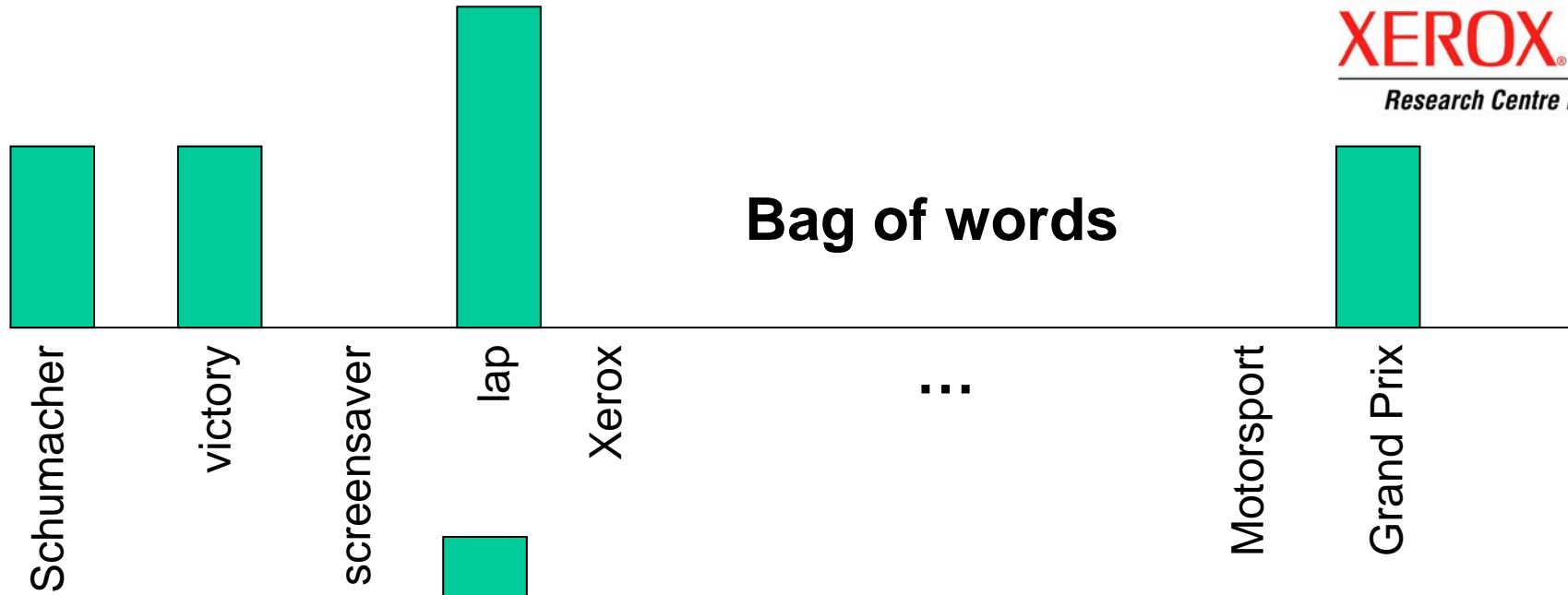
[Approach] Vector Quantization

- **Cluster a representative set of feature vectors**
 - We employ a **single clustering** for all categories
 - ◆ Hence scales to large number of categories
 - ◆ Ensures same features for each category: hence a simple classification

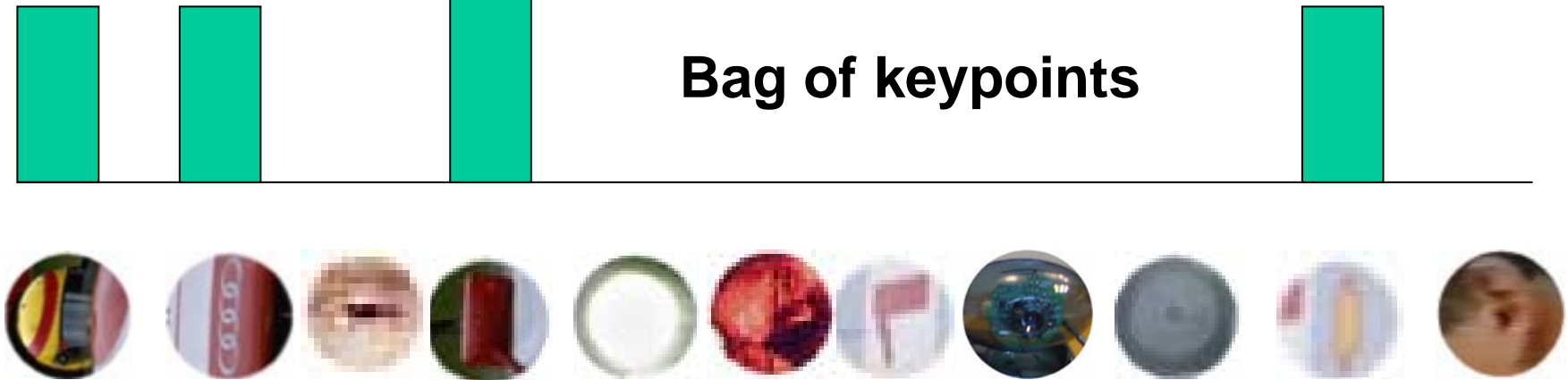
[Approach] Selected VQ Technique

- **K-Means: simple and efficient.**
- **Selection of K (number of clusters)**
 - We take an easy and well-founded approach
= **exploit classification results** to select the best
 - Results are **initialization dependent**
therefore work with many options and pick the best

Bag of words



Bag of keypoints

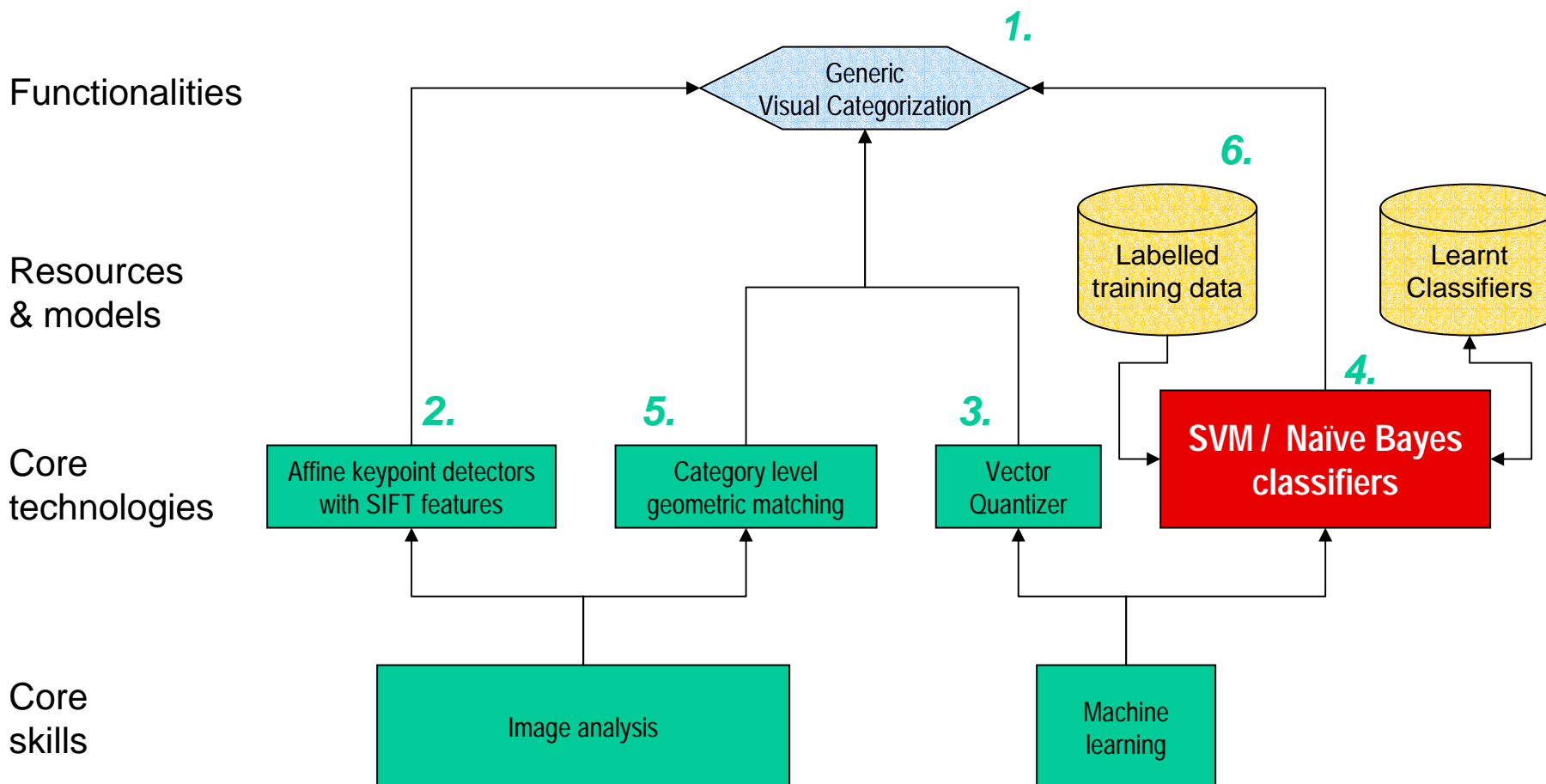


[Approach] Text-Image Analogy

- ❑ Text tricks: remove irrelevant variance by removing **stop words, stemming**
- ❑ ... but provide information about relevant variance: **part of speech, named entities** (multiple word “keywords”)
- ❑ Image approach
 - Quantized Affine Descriptors - reduces variations due to lighting, view, noise
 - Are we at the level of “words” or “characters”?
 - ◆ **Both** – keypoints exist at multiple scales eg keypoints for pupil, eye, face overlap

[Approach]

Generic Visual Categorization System



[Learning] Naïve Bayes is ...

Naïve

- Assumes features are generated independently given class
- eg key-points are independent given image class

Bayes

- Categorizes with “max a posterior probability” rule

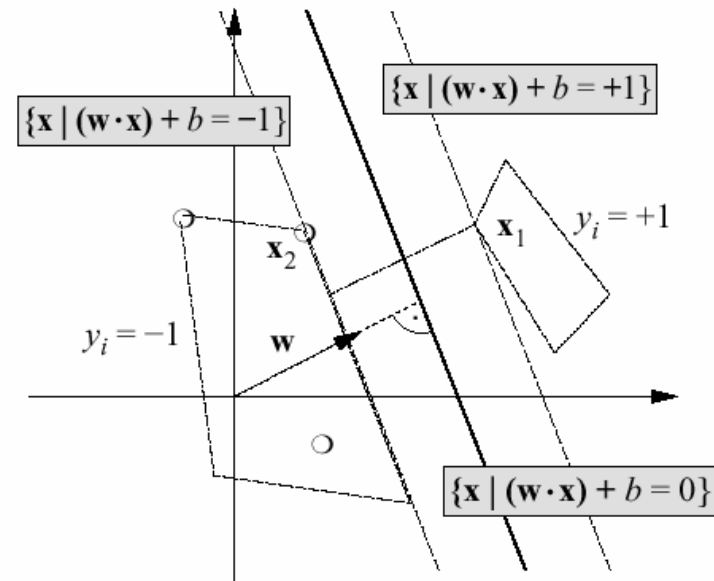
$$\begin{aligned}c^* &= \arg \max_{c_j} \Pr(c_j | d_i) \\ &= \Pr(\mathbf{c}_j) \prod_{k=1..n_i} \Pr(\mathbf{w}_k | \mathbf{c}_j)\end{aligned}$$

Prior Class-conditional feature

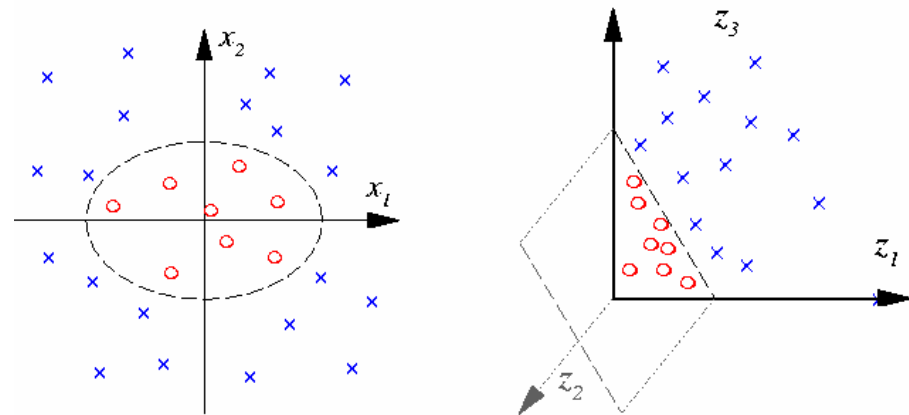
... assume **UNIFORM** densities ... just **COUNT**

[Learning] SVM

**Linear hyperplane
maximizing margins
separating data**



**Enables use of kernels for
non-linear mapping**



[Learning] Multi-Class Approach

□ Use

- "scores" for Naive Bayes
- one-against-all for SVM

[Data] Challenges

Machine learning needs lots of data

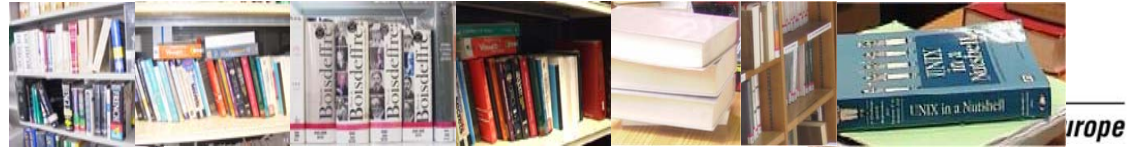
- > 1000 samples per class for handwriting & news categorization

However

- big image collections – Getty, Corbis – not usually public
- public data – usually small / only faces, cars, pedestrians
- gathering own data – must overcome legal barriers
 - ◆ for digital photos of people
 - ◆ for pictures in shops



[Data] LAVA



Acquired by Graz and XRCE

* = Thanks for written permission from Darty, the French consumer electronics shop

9 *new* classes, > 100 images per category



















Using

- Nokia 7650 Phone Camera
- Nikon Coolpix 700
- SONY Digital Video Camera DCR-230E
- Ricoh i-900 Image Capture Device

Class name	#
Shoes	396
Trees	305
Signs	254
Phones*	216
Buildings	206
Books	122
Radio_alarm*	115
Chairs	100
Tools	100

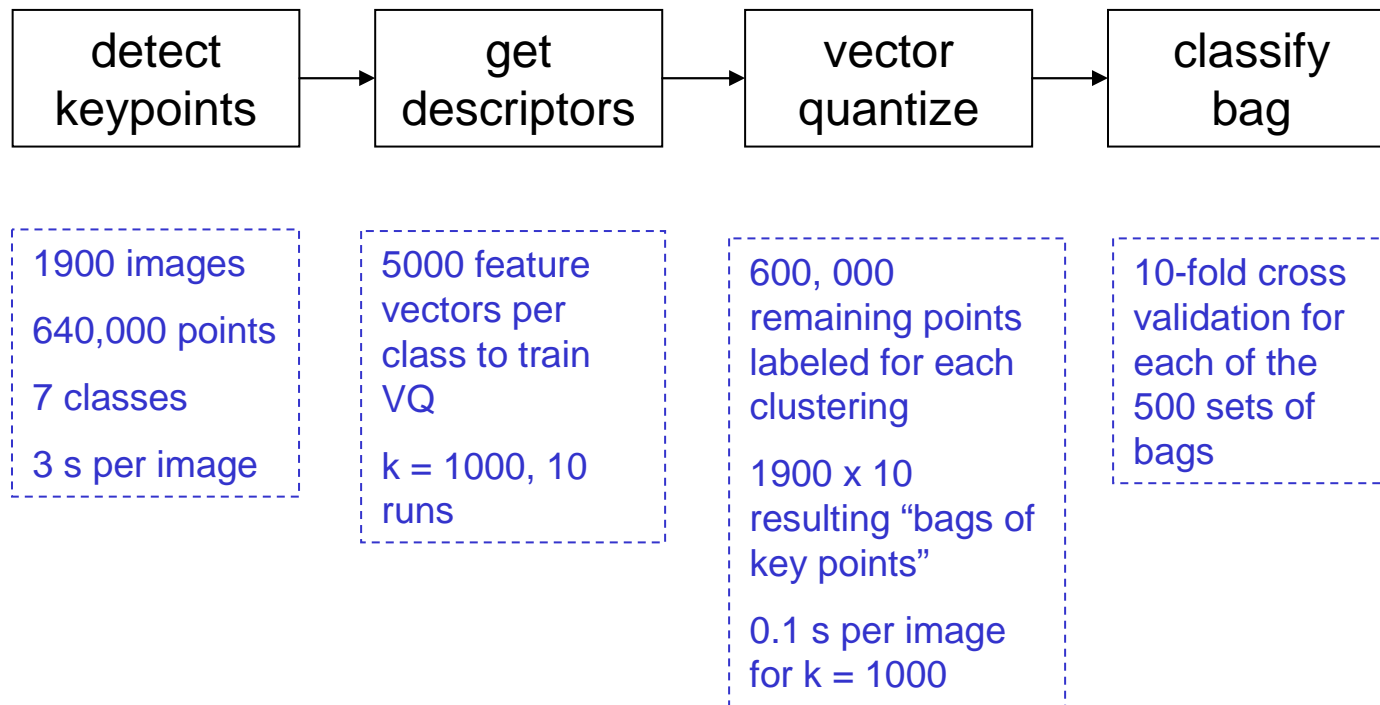


[Data] Fergus Perona Zissermann

Airplanes (side)	Cars (rear)	Cars (side)	Faces	Motorbikes (side)	Background
					
					
					
1074	651	720	450	826	451

- ❑ **NB: Both datasets have color images, but our experiments don't exploit the color information!**

[Data] Typical Numbers



Qualitative Investigation

- ❑ **Before quantitative analysis we should see if the results make sense**
 - What do the clusters look like?
 - Can we handle multiple instances?
 - What happens with partial visibility?
 - What happens when multiple object types are present?
 - What about background clutter?

[Qualitative] What the clusters look like



all keypoints



2 clusters

[Qualitative] Multiple object instances



❑ All correctly labeled

[Qualitative] Handling partial visibility



- ❑ All correctly labeled: face, car, house

[Qualitative] Handling clutter

❑ It is common to have more keypoints on the background than on the target

❑ But labels are still correct

❑ ***NB: circles just indicate the location of interest points – not their shapes which are elliptical and overlap!***






[Qualitative]

What happens in multi-label cases?

		
phones, books, cars	bikes, buildings, cars	buildings, cars, faces

- ❑ Each image was given only one training label
- ❑ The dataset is not totally clean
- ❑ However results above the margin are usually correct!

[Qualitative] Promising ... but not perfect!

		
face(1)... book(2)	trees(1)...face(5)	phones(1)... cars(2)

- ❑ **Need quantitative results to improve**

Quantitative Questions

- 1. What is the effect of 'k'?**
- 2. What is the relative performance of Naïve Bayes and SVM?**
- 3. What SVM kernel does best?**
- 4. Where do multi-class errors occur – class_i vs class_j?**
- 5. Where do detection errors occur – class_i vs background?**
- 6. How robust are the clusters?**

[Quantitative] Performance Metrics

- ❑ **Overall correct rate** $\varepsilon = \Pr(\text{output class} = \text{true class})$
- ❑ **Confusion matrix** $C_{ij} = \Pr(\text{output class is } \mathbf{i} \mid \text{true class is } \mathbf{j})$
- ❑ **Mean rank** $\rho_i = \mathbf{E}[\text{rank of class } \mathbf{i} \text{ in sorted output} \mid \text{true class is } \mathbf{i}]$

[Quantitative] n-fold cross validation

- Cut data into **n** chunks = **folders**

- **Example n = 10**

- ◆ Train on 2, 3, ..., 10: result 1 = test on 1

- ◆ Train on 1, 3, ..., 10: result 2 = test on 2

- ◆ ...

- ◆ Answer = average of result1, result2, ... result10

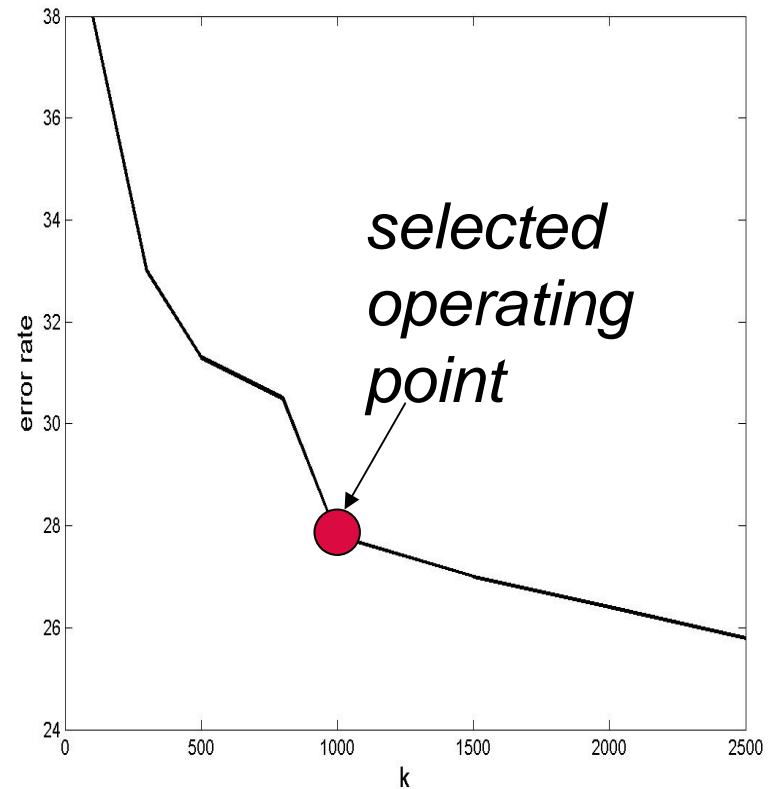
[Quantitative] Effect of k

□ Settings

- LAVA data, Naïve Bayes
- 10-fold CV

□ Result

- Error rate decreases with k
- Even for $k > 3000$
- But decrease is slow for large k



[Quantitative] LAVA Data, 10-fold CV

Confusion matrix for Naïve Bayes

	Faces	Buildings	Trees	Phones	Cars	Bikes	Books
Faces	75	4	2	2	4	3	9
Buildings	2	42	5	0	5	3	3
Trees	2	2	80	0	0	5	0
Phones	4	0	0	76	3	0	3
Cars	8	15	1	15	67	13	13
Bikes	2	14	11	0	9	73	0
Books	4	19	0	5	7	1	69
Mean error rate	22	54	19	22	28	25	28
Mean rank	1.49	1.88	1.33	1.33	1.63	1.57	1.57

Overall correct rate = 72%

[Quantitative] LAVA Data, 10-fold CV Confusion matrix for SVM: linear kernel

True classes →	<i>faces</i>	<i>bldgs</i>	<i>trees</i>	<i>cars</i>	<i>phones</i>	<i>bikes</i>	<i>books</i>
<i>faces</i>	98	14	10	10	34	0	13
<i>bldgs</i>	1	63	3	0	3	1	6
<i>trees</i>	1	10	81	1	0	6	0
<i>cars</i>	0	1	1	85	5	0	5
<i>phones</i>	0	5	4	3	55	2	3
<i>bikes</i>	0	4	1	0	1	91	0
<i>books</i>	0	3	0	1	2	0	73
Mean ranks	1.04	1.77	1.28	1.30	1.83	1.09	1.39

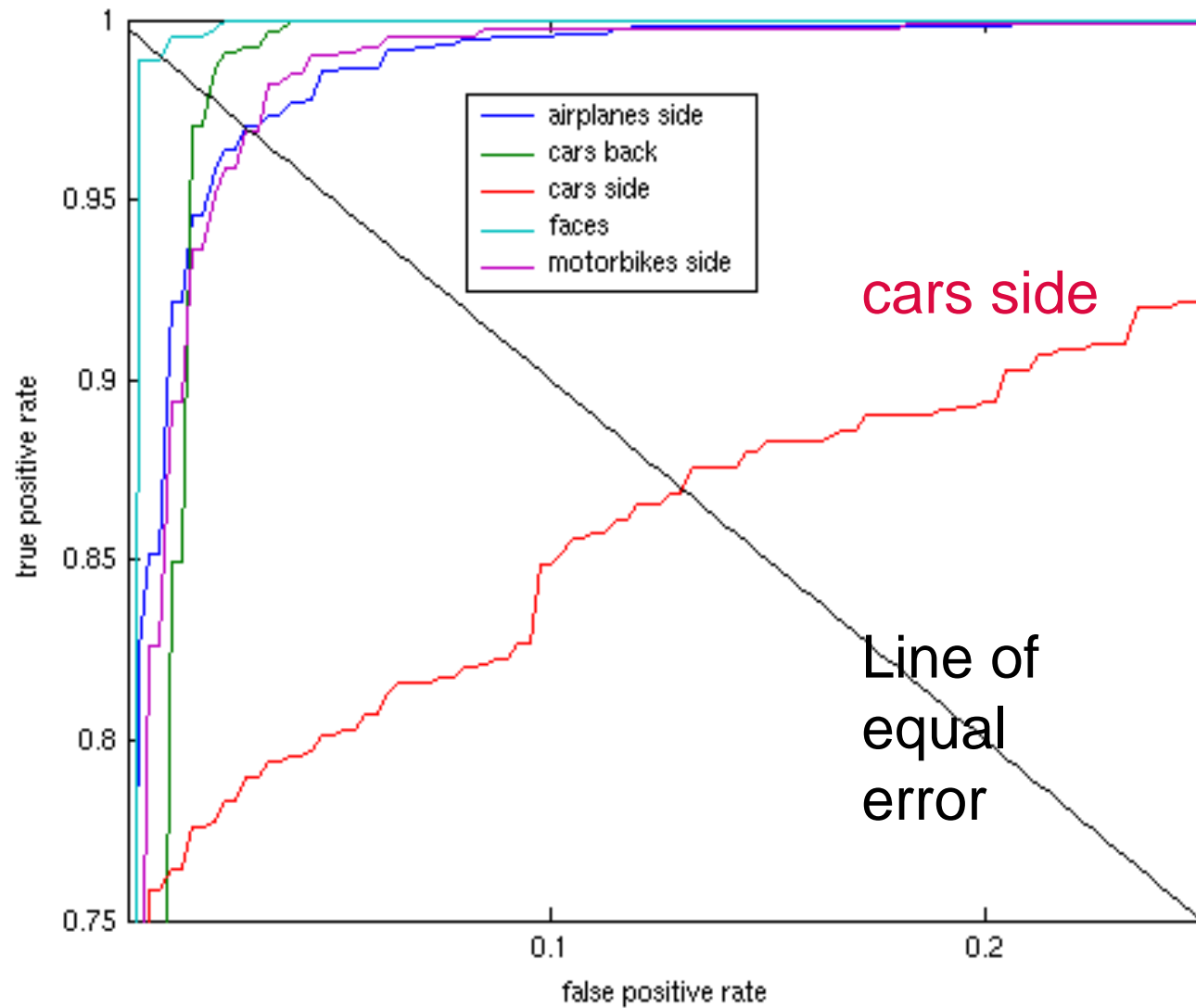
- ❑ Overall correct rate 82% >> Naïve Bayes
 - Except for phones = 76% correct rate with Naïve Bayes
- ❑ Find linear > quadratic > cubic
 - except for cars where quadratic is best

[Quantitative] Detection Performance

- Detection classifier decision
 - **class_i vs background**

- We measure the **Receiver Operating Characteristic (ROC)**.
 - As we vary the SVM output threshold, the true positive (TP) and false positive (FP) rates change
 - ROC = plot of TP against FP
 - Also measure **equal error operating point** where $FP = FN$

[Quantitative] ROC for FPZ, 2-fold CV



[Quantitative] FPZ Equal Error Points

2-fold CV	FPZ	Set 1	Set 2	Set 3
Airplanes	90.2	96.4	97.1	97.0
Cars (rear)	N/A	97.9	98.6	98.6
Cars (side)	88.5	86.1	87.3	86.7
Faces	96.4	98.9	99.3	99.1
Motorbikes	92.5	97.3	98.0	97.1

Method described in
FPZ paper

K=1000 clusters
derived from LAVA
data

Clusters trained on
FPZ data including /
without background

□ Observations

- clusters are very robust
- performance significantly better than FPZ method (less than 1/3 error rate)
- ... except cars (side)

[Quantitative] Why is cars (side) so bad?

□ Tabulate average number of keypoints detected for each class

- Simple threshold classifier based on number of keypoints has error
 - ◆ airplanes vs background = 15%
 - ◆ cars (side) vs background = 39%
- Affine interest point detector finds fewer keypoints than scale invariant detector

Airplanes	507
Cars (rear)	592
Cars (side)	46
Faces	1110
Motorbikes	736
Background	152

[Quantitative] FPZ Multiclass performance

2-fold CV	Set 1	Set 2	Set 3	10-fold CV	Set 1	Set 2	Set 3
Airplanes	94.9	96.4	95.4	Airplanes	95.4	96.7	96.7
Cars (rear)	94.6	97.1	97.2	Cars (rear)	96.2	97.5	98.2
Cars (side)	97.3	97.1	97.4	Cars (side)	97.0	97.3	97.6
Faces	89.8	92.4	91.1	Faces	92.2	94.4	94.2
Motorbikes	90.5	92.4	92.3	Motorbikes	91.9	93.5	93.4

Overall error rates for 5-class case

- ❑ **Problem is a bit harder than detection**
 - eg face detection correct rate was 99.3% previously
- ❑ **Some benefits are obtained from a larger dataset (10-fold)**
 - Particularly for faces, which were the least populous class

[Quantitative] FPZ Confusion Matrix

<i>True classes</i> →	Airplanes	Cars (rear)	Cars (side)	Faces	Motorbikes
Airplanes	96.7	0.2	0.6	2.0	3.4
Cars (rear)	0.4	98.2	1.0	1.1	2.4
Cars (side)	0.2	0.0	97.6	0.2	0.3
Faces	1.0	0.6	0.1	94.2	0.6
Motorbikes	1.8	1.1	0.8	2.4	93.4
<i>Mean ranks</i>	1.04	1.03	1.06	1.06	1.09

□ Observations

- Dataset is considerably easier than LAVA one
- (Not shown) If we add color information, get dramatic improvement

Conclusions

- ❑ **We have presented a new and efficient generic visual categorizer based on “bags of keypoints”**

- ❑ **Thorough performance evaluation demonstrates**
 - State-of-the-art performance is obtained
 - Method is robust to
 - ◆ choice of clusters, clutter, multiple objects, partial visibility

Further Work

□ Ongoing experiments indicate

- Improving performance by including geometry isn't simple
- The benefit of affine keypoints rather than scale-invariant keypoints is not clear-cut
- Recent work from Montanuniversitat Leoben suggests appropriate segmentation-based descriptors are better
- Color information helps a lot on these datasets