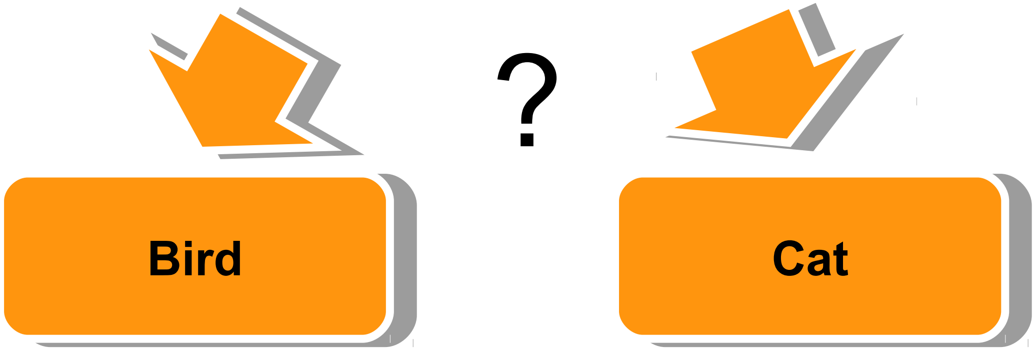
Towards Optimal Naive-Bayes Nearest Neighbor for image classification

Régis Behmo , Paul Marcombes, Arnak Dalayan*, Véronique Prinet

Institute of Automation, Chinese Academy of Sciences

(*) with Ecole Ponts-ParisTech, France

About image classification ...



Source:http://www.migueldearriba.es/index2.html

About image classification ...

Bird

About image classification ...

Cat

About image classification ...

Cat

About image classification ...

Bird ? Cat

About **supervised learning** …

Given …

Source: http://www.image-net.org/,

Some Applications

- Image ***indexing***
  - Huge quantity of data recorded daily
  - Indexed by date and location only
  - → index them by visual content

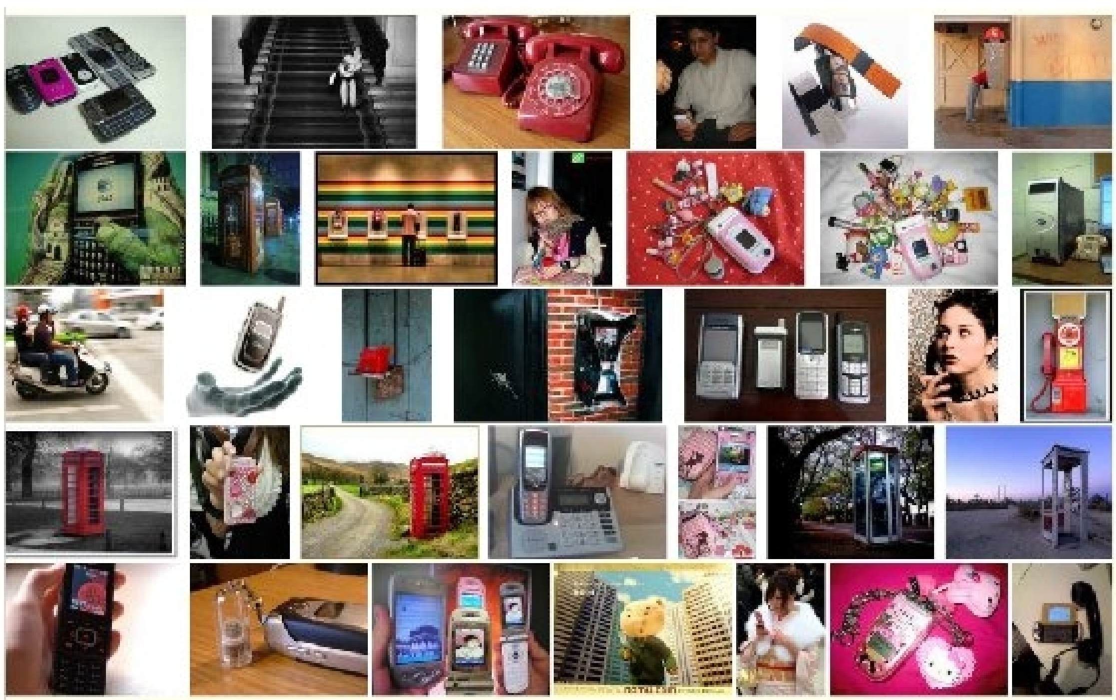
- ***Visual search***
  - Tags as defined by users, if any, do not necessary reflect the visual content of the image.

Tags by user: Milou, Sydney

Categories: dog, sea

Challenges

- Low inter-class variability
- High intra-class variability
- Semantic gap : image → concepts



P
h
o
n
e

Classification tasks

- Description & representation

  - Descriptors (local) : SIFT (Lowe99), HOG [Dalal05], Shape Context [Belongie]…

  - Representation : Bag of Feature [Schmid01], proximity distribution [Ling07],….

Classifier

  *- Generative models (e.g: MRF/CRF [Lafferty01], pLSA [*Hofmann01]*, …)*

  - Discriminative models (e.g.: SVM [Vapnik95], Boosting[Viola01], …)

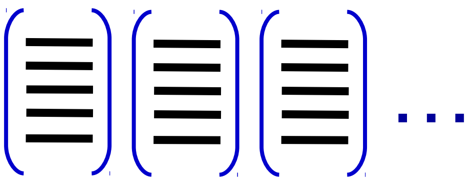
  …

- Learning

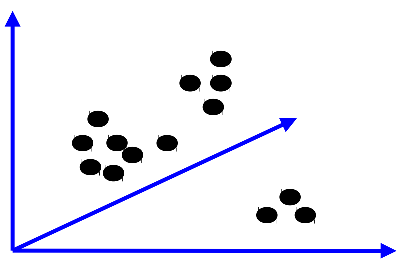
BoF + SVM pipeline

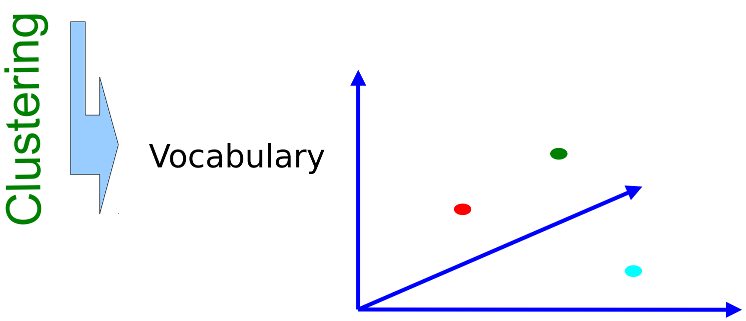
1- Learning a visual vocabulary from image (training) set

Local descriptors

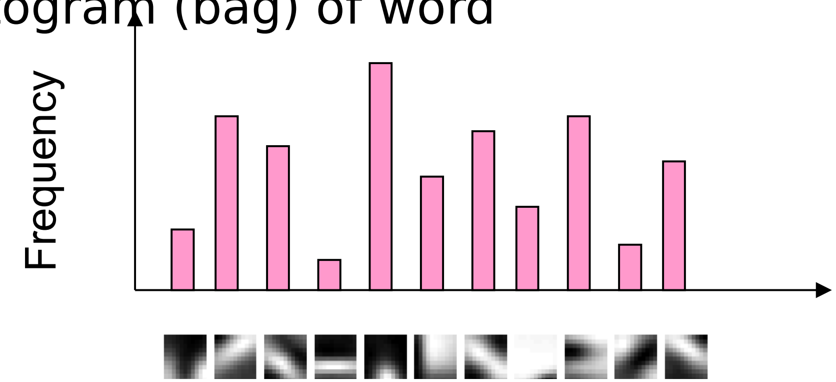


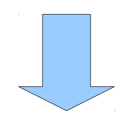
Feature space





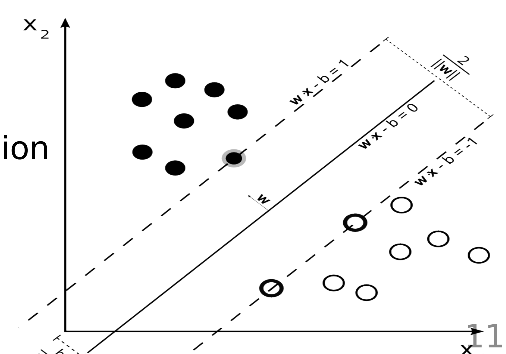
2- Representing each image by a histogram (bag) of word



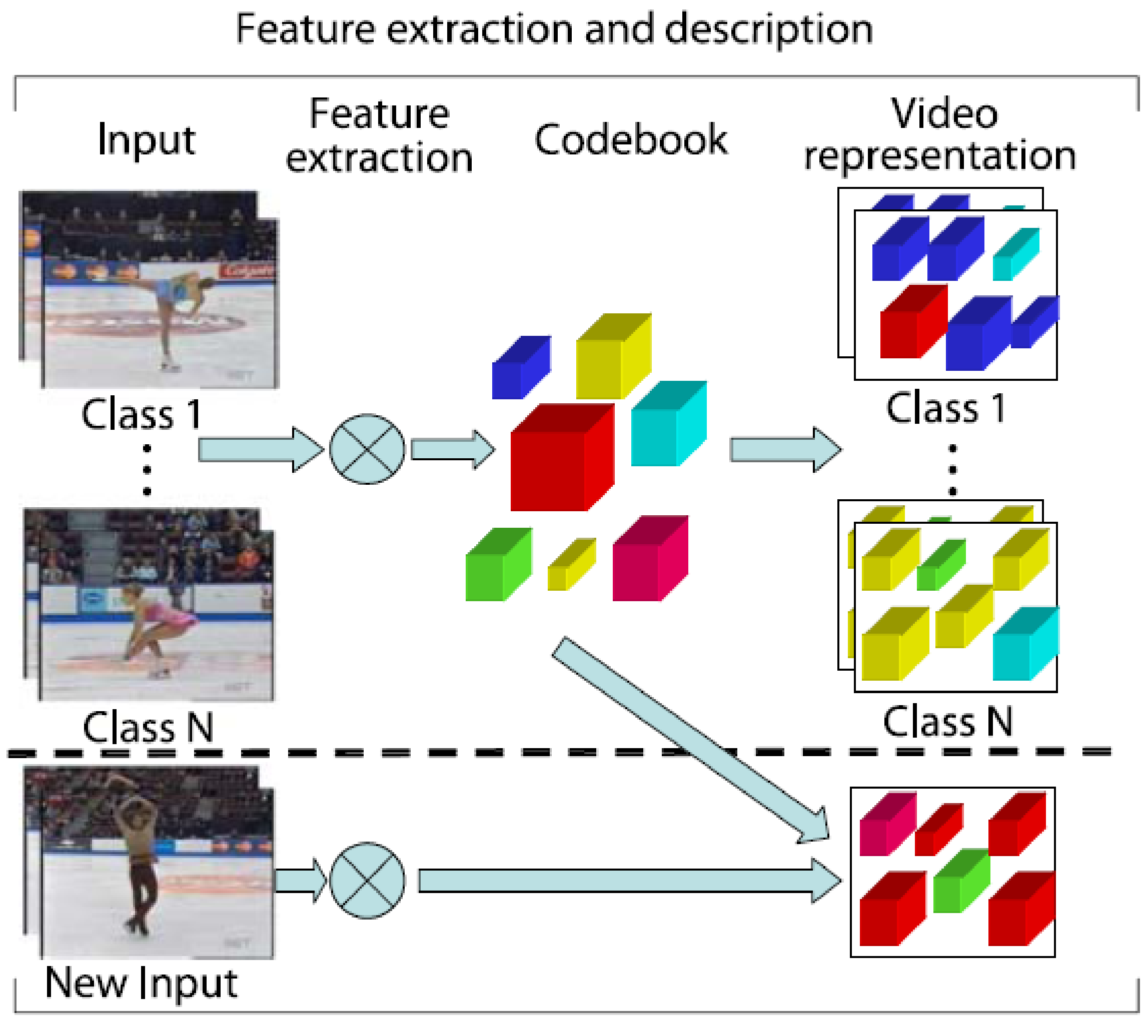


3- Classification via SVM

hyperplane separation in feature space



Bags of features for action recognition



Source : Niebles, Wang, Fei-Fei, Unsupervised learning of Human action category using spatio-temporal words, IJCV 2008.

In defense of Nearest Neighbor based image classification

(Boiman, Shechtman, Irani. CVPR 2008)

BOF + SVM : drawback

- ✗ Quantization (codebook creation)

  - descriptors discriminative power drop

  - rare but discriminative information is lost

  → *no quantization*

- ✗ Image-to-image classification (SVM)

  → *Image-to-class classification*

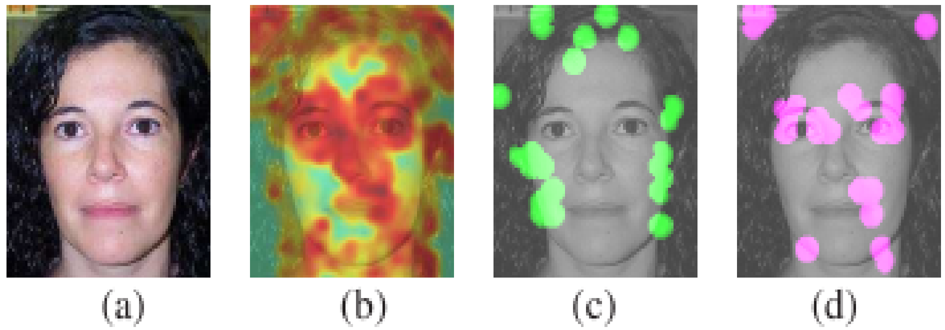
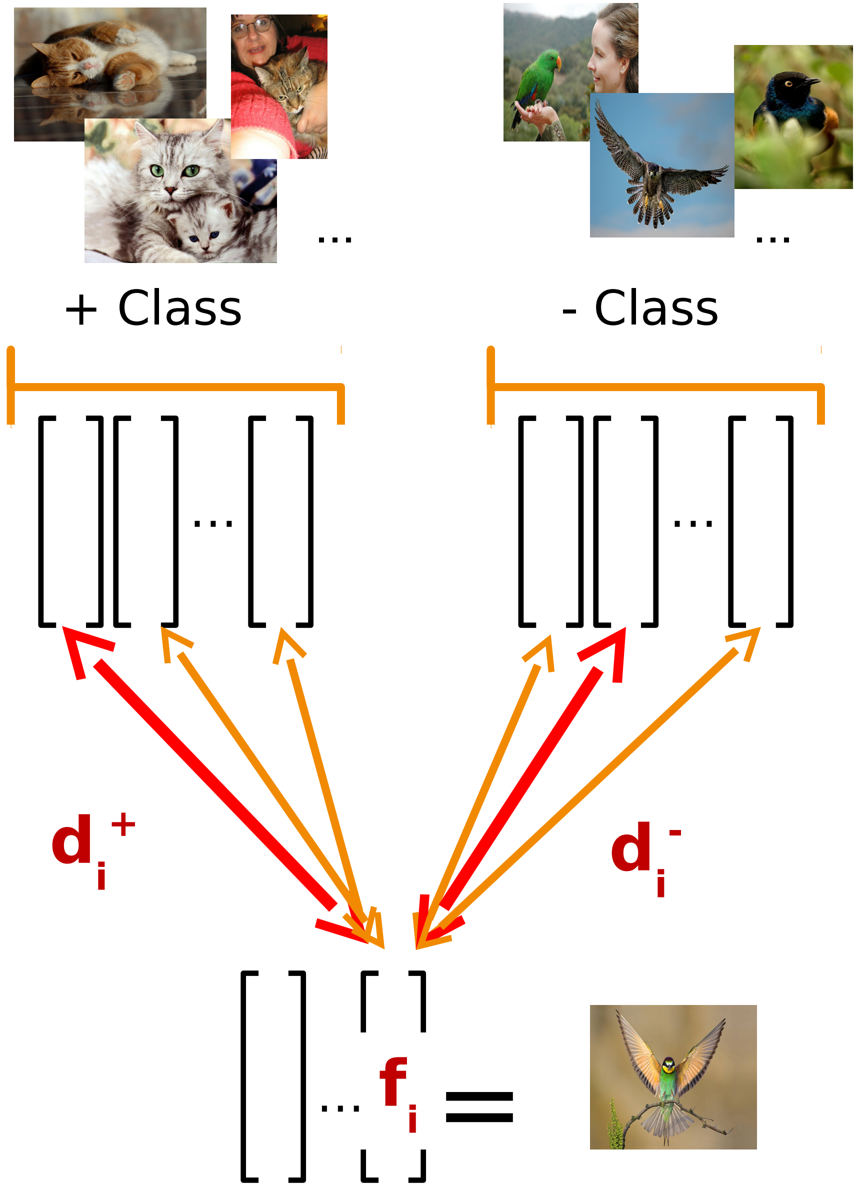


Figure 1. **Effects of descriptor quantization – Informative descriptors have low database frequency, leading to high quantization error.** *(a) An image from the Face class in Caltech101. (b) Quantization error of densely computed image descriptors (SIFT) using a large codebook (size 6,000) of Caltech-101 (generated using [14]). Red = high error; Blue = low error. The most informative descriptors (eye, nose, etc.) have the highest quantization error. (c) Green marks the 8% of the descriptors in the image that are most frequent in the database (simple edges). (d) Magenta marks the 8% of the descriptors in the image that are least frequent in the database (mostly facial features).*



$$D^{\pm}(I)=\sum_{i=1}^{N}(d_i^{\pm})^2$$

Compute features from all images from the training set ; create a "feature space" for class +, and for class - .

- Given a new image *I*, compute its *feature points* $\{f_i\}_{i=\{1\ldots N\}}$, and find nearest neighbor $NN(f_i)$ in each class pool.

- Compute *feature-to-class* distance

$$d_i^{\pm}=\|f_i-NN^{\pm}(f_i)\|$$

- Classifying rule :

$$\arg_{\pm} \min\left(\sum_{i=1}^{N}(d_i^{+})^2, \sum_{i=1}^{N}(d_i^{-})^2\right)$$

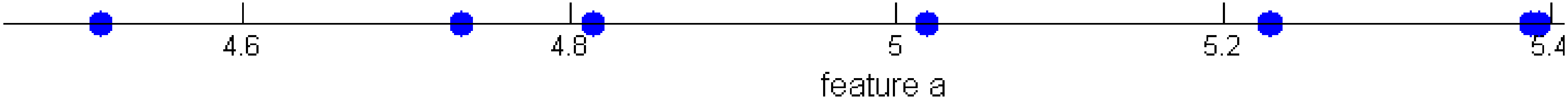
Why does it work *ok* ?

- No quantization
- “NBNN” classifier approximates the optimal MAP Naive-Bayes classifier

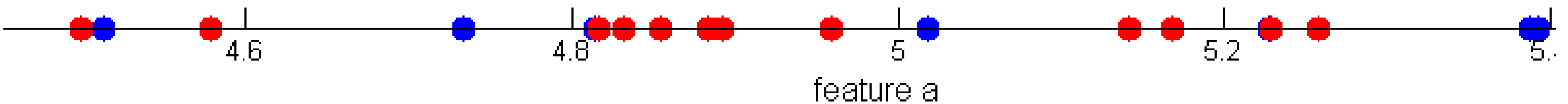
Improvement of NBNN

- New (parametric) distance
- Generalises to multi-channel (features)
- Parameters estimation by hinge loss minimisation
- Applies to object detection

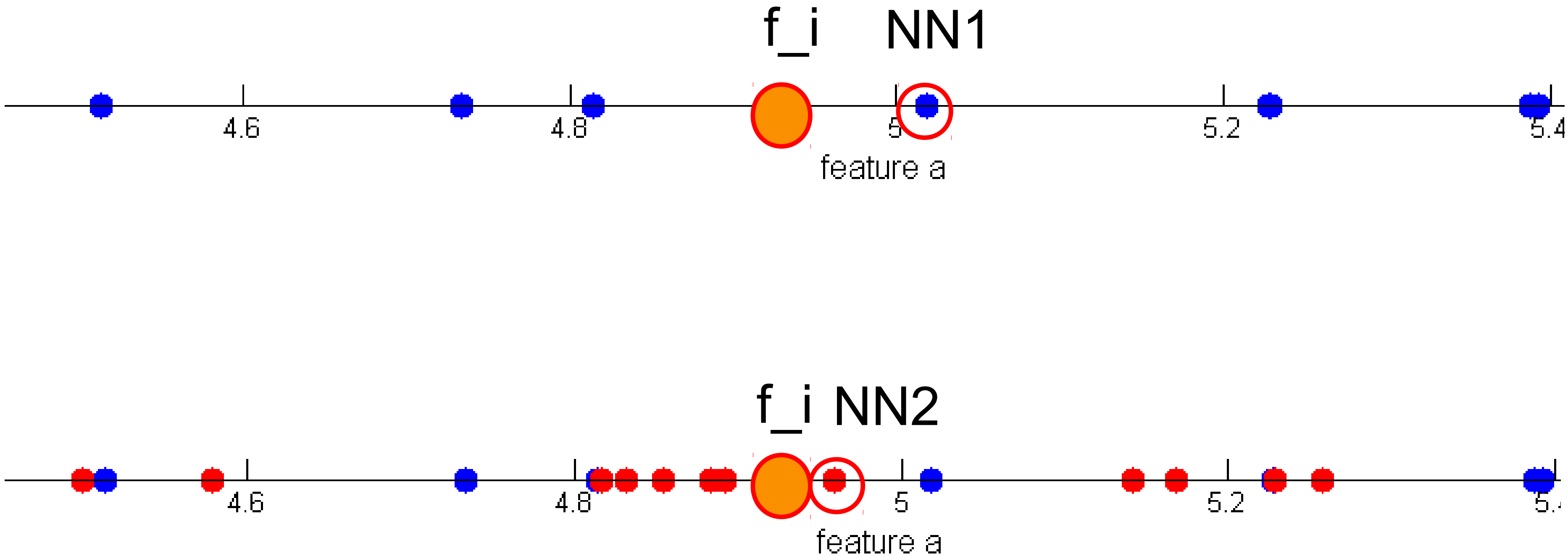
Impact of data sample size ...



Features of the N images



Features of the 2N images



Basic NN *distance* depends on the number of 'training' feature points

=> *Classification* is biased toward the most densely sampled class in the training set.

Distance correction

$$d_i^{\pm\cdot} = \| f_i - NN^{\pm\cdot}(f_i) \|$$

$$d_i^{\pm\cdot} = \alpha^{\pm\cdot} \| f_i - NN^{\pm\cdot}(f_i) \| + \beta^{\pm\cdot}$$

How do you:	Represent an image?	Classify an image?
Bag of Words (BoW)	Set of **quantised** features	Linear/Kernel SVM
Naive Bayes NN (NBNN)	Set of **unquantised** features	Linear classifier, 0 parameter
Optimal NBNN (oNBNN)	Set of **unquantised** features	Linear classifier, 2 parameters/class
Multi-channel oNBNN	Multiple sets of **unquantised** features	Linear classifier, 2N parameters/class

$$I \in \color{orange}{\blacksquare} \; ?$$

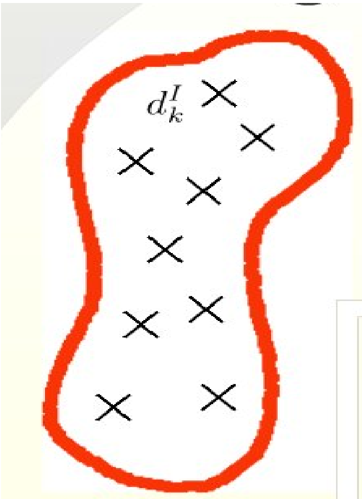
NBNN	$\sum_{x\in F(I)} (x \leftrightarrow \color{orange}{\blacksquare})$	$<$	$\sum_{x\in F(I)} (x \leftrightarrow \color{green}{\blacksquare})$
Optimal NBNN	$\sum_{x\in F(I)} \left[\alpha_{\color{orange}{\blacksquare}} (x \leftrightarrow \color{orange}{\blacksquare}) + \beta_{\color{orange}{\blacksquare}}\right]$	$<$	$\sum_{x\in F(I)} \left[\alpha_{\color{green}{\blacksquare}} (x \leftrightarrow \color{green}{\blacksquare}) + \beta_{\color{green}{\blacksquare}}\right]$
Multichannel	$+\sum_{x\in F'(I)} \left[\alpha_{\color{orange}{\blacksquare}'} (x \leftrightarrow \color{orange}{\blacksquare}') + \beta_{\color{orange}{\blacksquare}'}\right]$		$+\sum_{x\in F'(I)} \left[\alpha_{\color{green}{\blacksquare}'} (x \leftrightarrow \color{green}{\blacksquare}') + \beta_{\color{green}{\blacksquare}'}\right]$

Basic

Classification rule

$$\tilde{c}(I) = \arg\max_{c \in \gamma} p(I|c), \quad for \quad p(c) = cst$$

$$. = \arg\max_{c \in \gamma} \prod_{i=1}^{N_I} p(f_i|c)$$



Density approximation

$$p(f|c) = p_c(f) = \frac{1}{Z} \sum_{e \in \chi^c} \phi(\|f - e\|, \sigma), \qquad \forall f \in \mathbb{R}^d$$

$$\chi^c = \{ f_j^I \in \mathbb{R}^D | c = c(I), \forall I \in D^t, 1 < j < N_I \}$$

Training set

[Boiman & al. 08]

Density approximation

$$p_c(f) \approx \frac{1}{Z} \; max_{e \in \chi^c} \exp(-\|f-e\|^2 / 2(\sigma)^2)$$

- Kernel density : exp()

- SUM approximated by MAX

Define *feature-to-class* distance

$$-\log p_c(f) \approx \underbrace{min_{e \in \chi^c} \|f-e\|^2}$$

Underlying hypothesis : density parameters are class independant

... and *classification rule*

$$\tilde{c}(I) = arg\, min_c \sum_{i=1}^{N_I} \|f_i - NN^c(f_i)\|^2$$

- Naïve bayes assumption

NN distance revisited

Density approximation

$$p_c(x) \approx \frac{1}{Z^c} \; max_{e \in \chi^c} \exp(-\|x-e\|^2 / 2(\sigma^c)^2)$$

- Density parameters are class dependent !!!

Define *feature-to-class* distance

$$\tau^c(f) = -\log p_c(f). = \alpha^c \, min_{e \in \chi^c} \|f-e\|^2 + \beta^c$$

- Reparametrisation

... and *Image-to-Class* distance

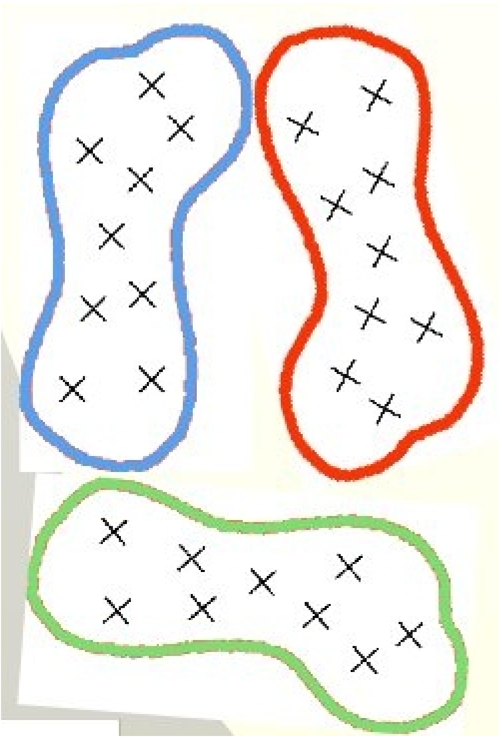
$$\tau^c(I) = \sum_{i=1}^{N_I} \alpha^c \; \|f_i - NN^c(f_i)\| \; + \; \beta^c$$

- Linear in the model parameters

Multi-channel

- Image = multiple points cloud

More descriptors → better results



$$I \sim (\chi_n(I))_n \qquad n \in \mathbb{N}$$

$\chi_n(.)$: n[th] feature channel

$$\tilde{c}(I) = arg\, min_c \sum_n \sum_{f_i \in \chi_n(I)} \underbrace{-\log p(f_i | c)}$$

$$\tau_n^c(f_i) = \alpha_n^c \; \| f_i - NN^c(f_i) \| \; + \; \beta_n^c$$

'measure' of channel discriminative power

Parameter estimation

- Binary classification : c={+,-}
- Hinge loss minimisation
- Cross-validation

$$E = \sum_{I \in D} max\left(0,\ 1 - c(I)\left(\tau^{-\cdot}(I) - \tau^{+\cdot}(I)\right) \right)$$

Training dataset

Ground truth
c \in {+,-}

Distance to negative
feature set

Distance to positive
feature set

Parameter estimation

- Constrained linear program (using off-the-shell library)
- Distance correction parameters are 'optimal'
- Overfitting if number of channels is large

Results

Experimental setting

- Dataset : Caltech101, Gratz02, SceneClass13
- Rate of good classification
  (per class or averaged)
- Locality-sensitive
  hashing
  for NN search

Single-channel classification (using SIFT)

	BoW/SVM	BoW-Chi2/SVM	NaiveBayes [Boiman08]	oNBNN
SceneClass 13	67.85 (±0.78)	76.7 (±0.60)	48.52 (±1.53)	**75.35 (±0.79)**
Graz02	68.18 (± 4.21)	77.91 (±2.43)	61.13 (± 5.61)	**78.98 (± 2.37)**
Caltech101	59.2 (± 11.89)	89.13 (±2.53)	73.07 (± 4.02)	**89.77 (± 2.31)**

Correct classification rate and associated variance

Caltech 105 (detail per class)

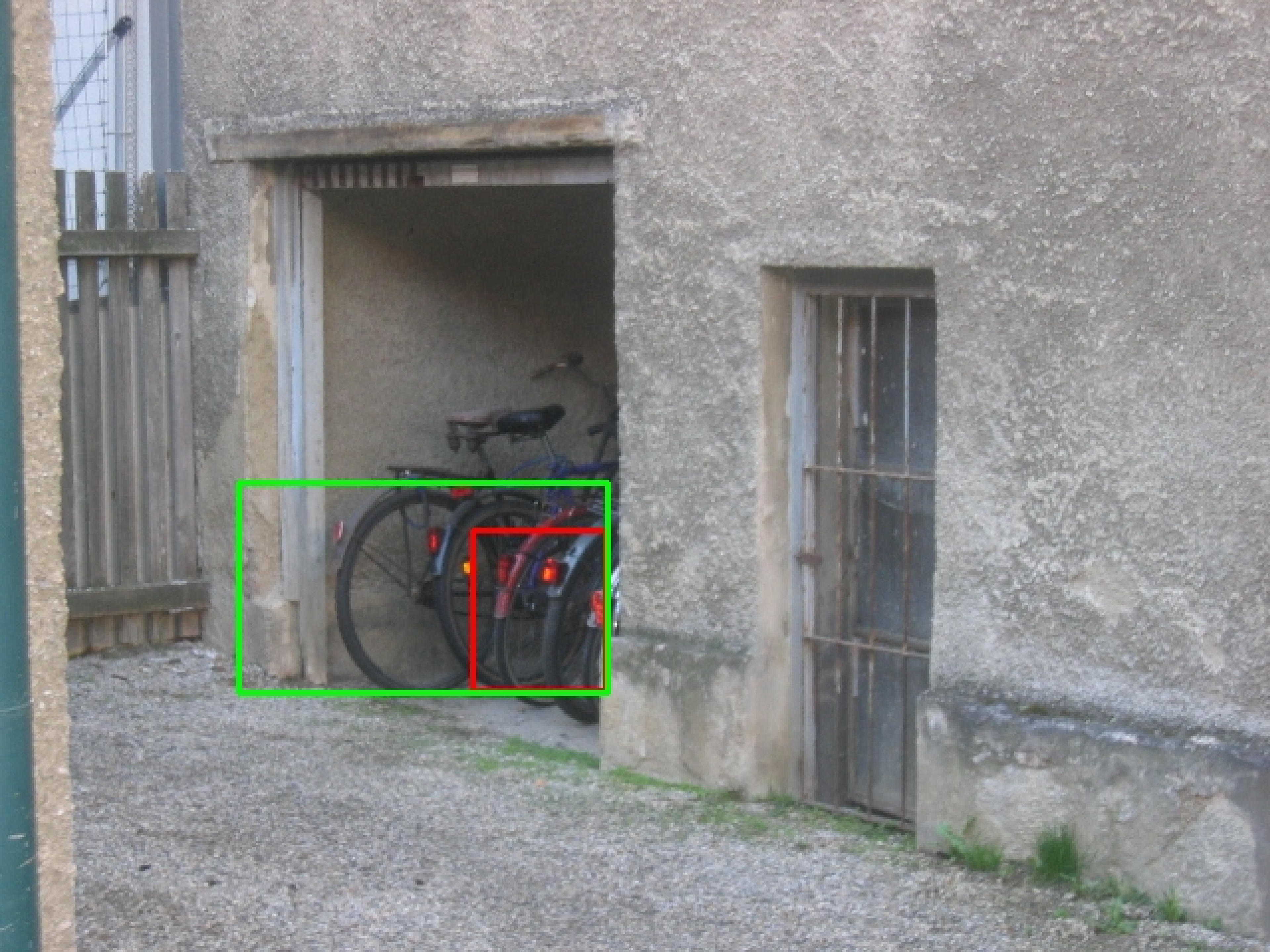
Class	BoW/χ^2-SVM	[Opelt 2004]	NBNN	Opt. NBNN
Airplanes	91.99 ±4.87	97.5	34.17±11.35	95.00 ±3.25
Car-side	96.16±3.84	100.0	97.67 ±2.38	94.00±4.29
Faces	82.67 ±9.10	100.0	85.83±9.02	89.00 ±7.16
Motorbikes	87.80±6.28	94.3	71.33±19.13	91.00 ±5.69
Background	87.50 ±6.22	-	76.33±22.08	78.93±10.67

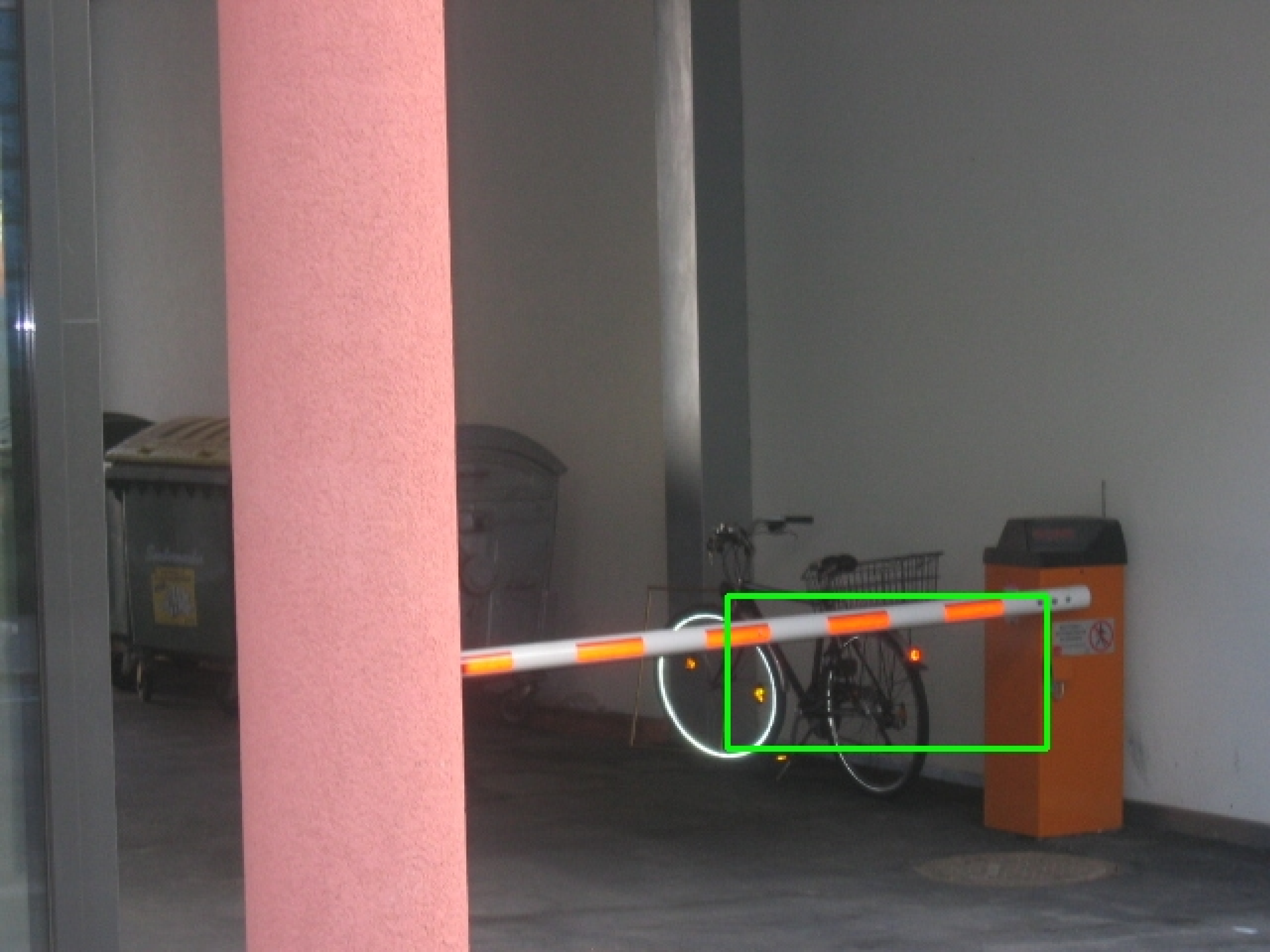
Caltech 105 (color features)

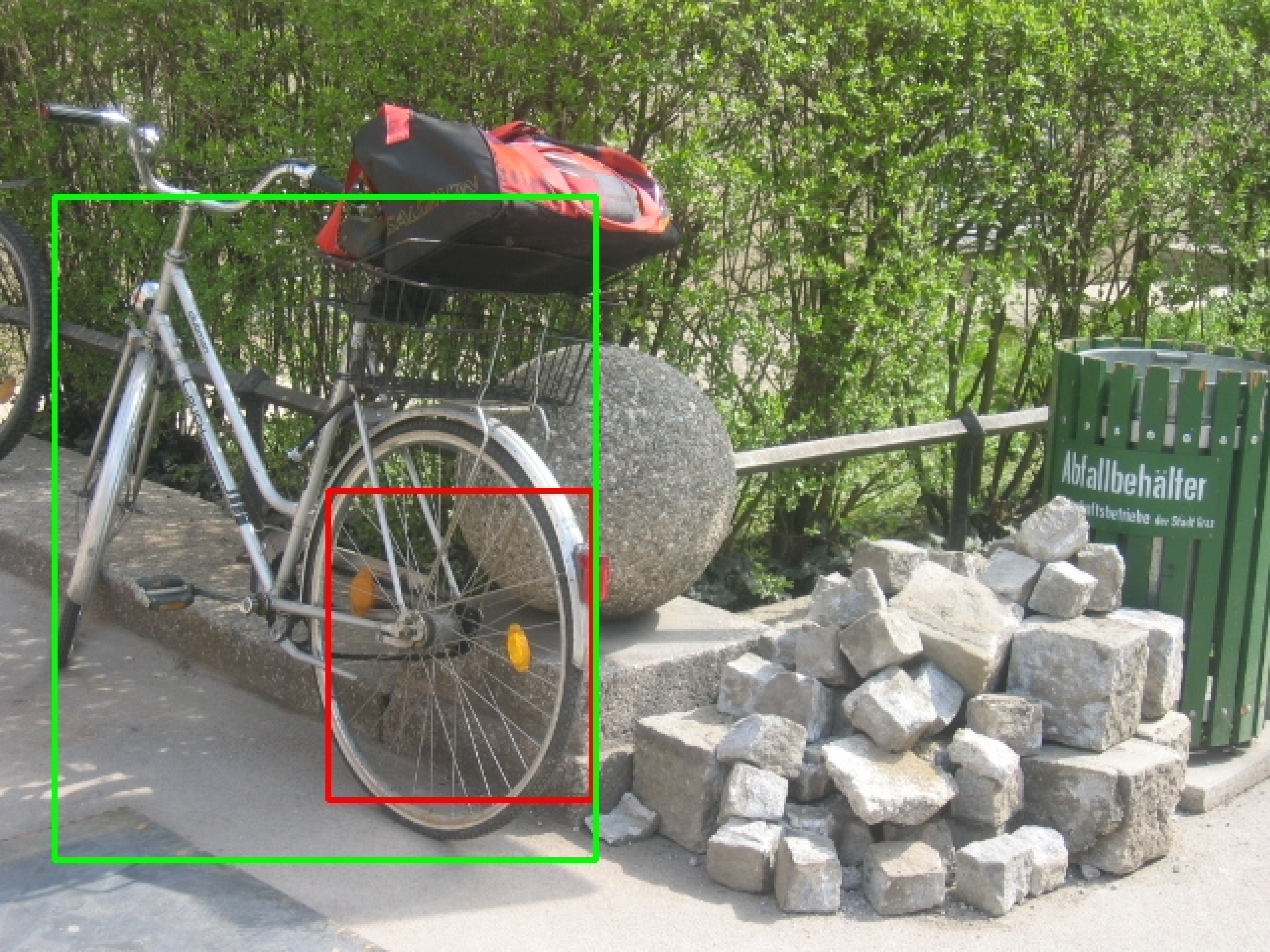
Feature	BoW/χ^2-SVM	NBNN [1]	Optimal NBNN
SIFT	88.90 ±2.59	73.07 ±4.02	89.77 ±2.31
OpponentSIFT	89.90 ±2.18	72.73 ±6.01	91.10 ±2.45
rgSIFT	86.03 ±2.63	80.17 ±3.73	85.17 ±4.86
cSIFT	86.13 ±2.76	75.43 ±3.86	86.87 ±3.23
Transf. color SIFT	89.40 ±2.48	73.03 ±5.52	90.01 ±3.03

Classification by detection (Gratz02)

Class	NBNN	Optimal NBNN	Optimal NBNN (classif. detect.)
bike	68.35±10.66	78.70±4.67	83.60
people	45.10±12.30	76.20±5.85	-
car	42.40±15.41	82.05±4.88	






Abfallbehälter

Conclusion

- ***Optimal nearest neighbor distance***
  - generalises to multi-channel classification
  - '*optimal*' parameters estimation
  - nearest neighbour search using multi-probe LSH
  - classification by detection.

- ***Limitations***
  → *model overfitting*
  → *NN search*
  → *spatial dependency between feature points*

References

In Defense of Nearest-Neighbor Based Image Classification,
Boiman , Shechtman, Irani, CVPR 2008

Learning The Discriminative Power_Invariance Trade-Off,
Varma, Debajyoti, ICCV 2007

Discriminative Subvolume Search for Efficient Action Detection,
Yuan, Liu, Ying, CVPR 2009

Beyond Sliding Windows: Object Localization by Efficient Sub-window Search
Lampert, Blaschko, Hofmann, CVPR2008

Modeling LSH performance tuning (multi-probe LHS index),
Dong & al. , CIKM 2008

GNU linear programming kit: http://www.gnu.org/software/glpk/

Unsupervised Learning by Probabilistic Latent Semantic Analysis
Hofmann, Machine Learning. 2001

Scene Classification via pLSA
Anna Bosch , Andrew Zisserman , Xavier Muoz, ECCV 2006

Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data
J. Lafferty, A. McCallum, and F. Pereira, ICML-2001

Efficient Piecewise Learning for Conditional Random Fields,
Karteek Alahari, Chris Russell, Philip H.S. Torr, CVPR 2010

Vladimir Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

Proximity Distribution Kernels for Geometric Context in Category Recognition
H. Ling and S. Soatto. In Proceedings of the International Conference on Computer Vision, 2007

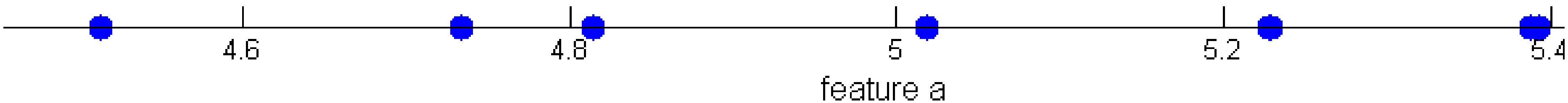
Object recognition with informative features and linear classification
D. Lowe, ICCV99

Local Features and Kernels for Classification of Texture and Object Categories: a Comprehensive Study
Zhang, Marszałek, Lazebnik, Schmid , IJCV 2001

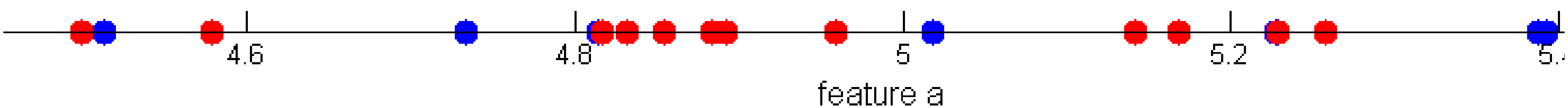
Histograms of Oriented Gradients for Human Detection
Dalal, Triggs, CVPR 2005

Viola, Jones: Robust Real-time Object Detection, IJCV 2001

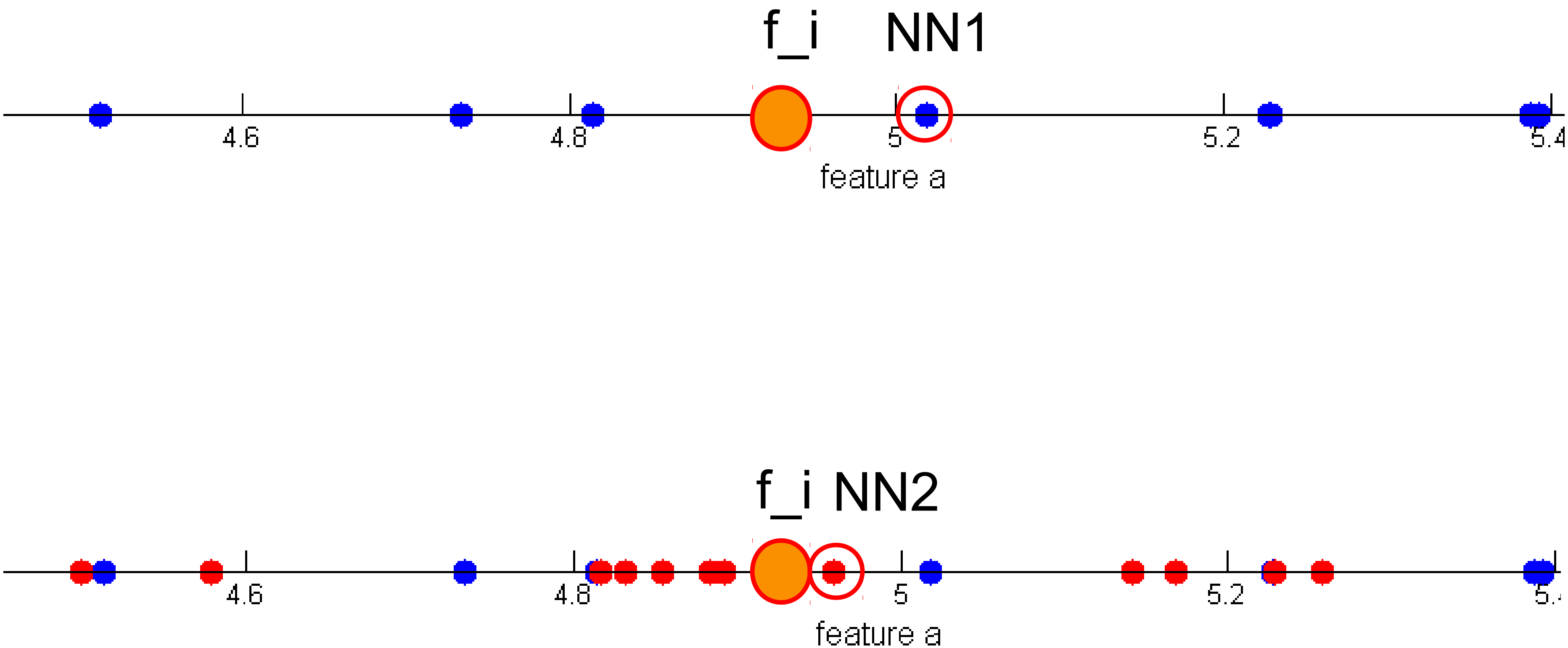
Limitations



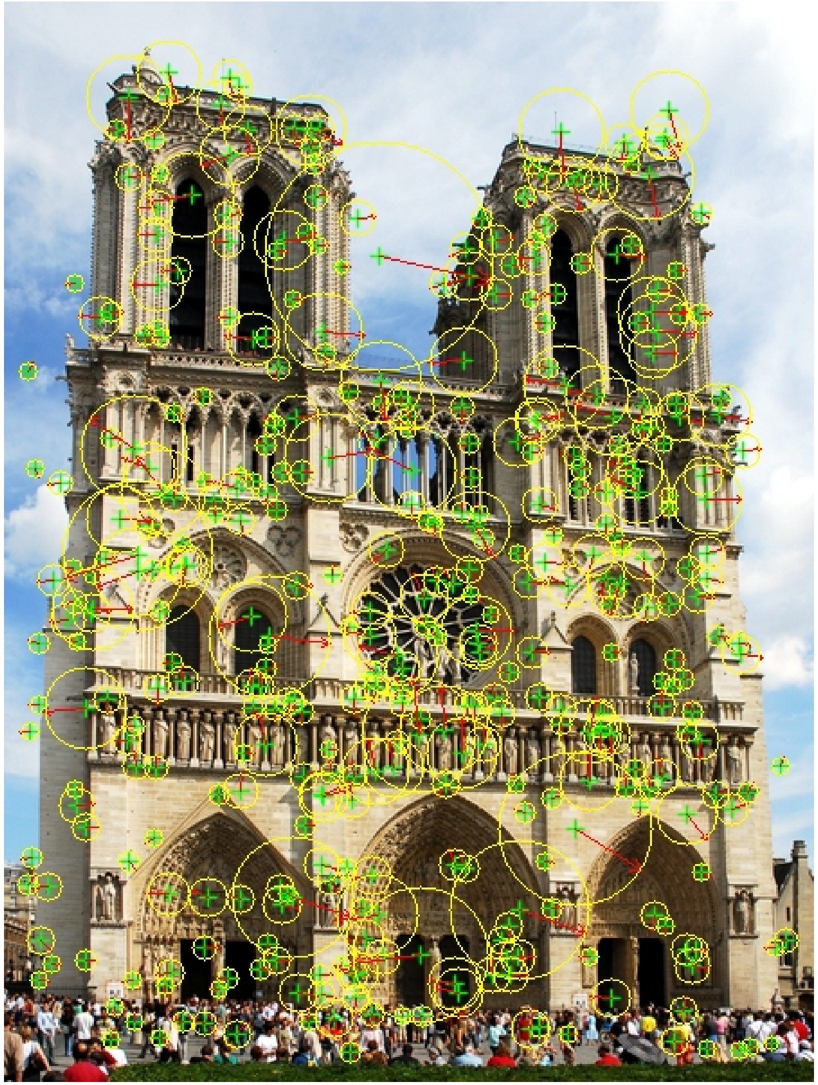
Features of the N images



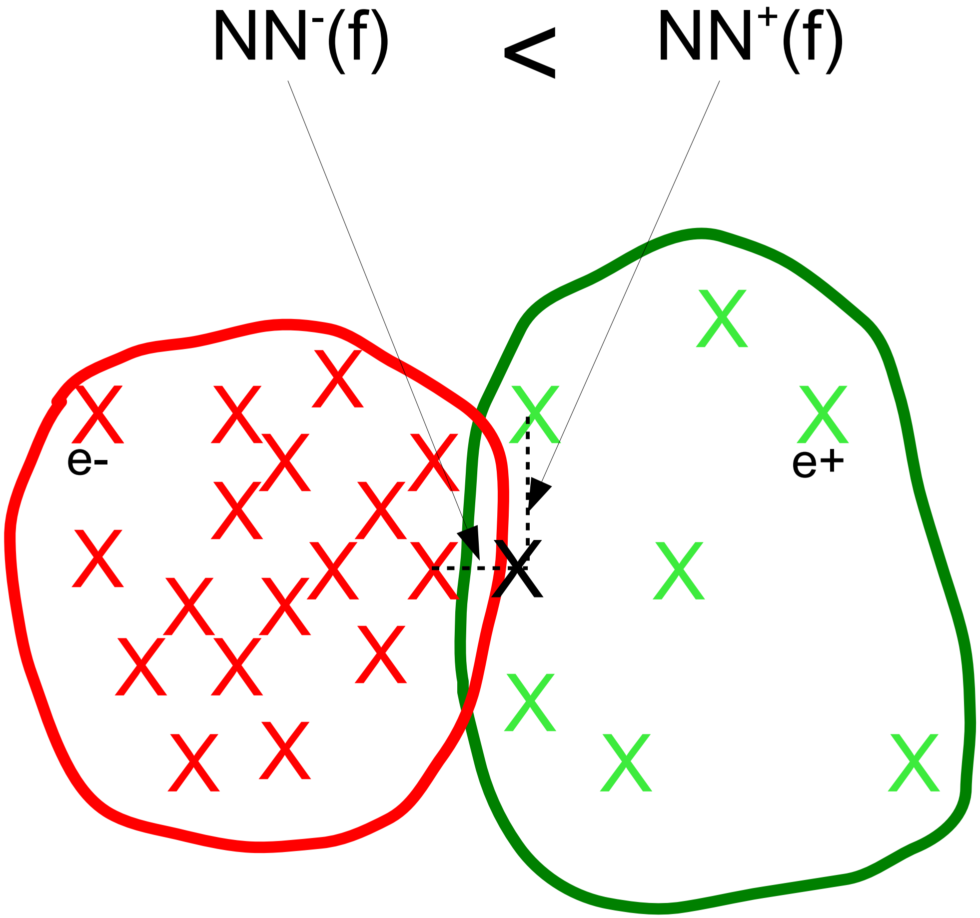
Features of the 2N images


f_i
NN1
4.6
4.8
5
5.2
feature a
f_i NN2
4.6
4.8
5
5.2
feature a

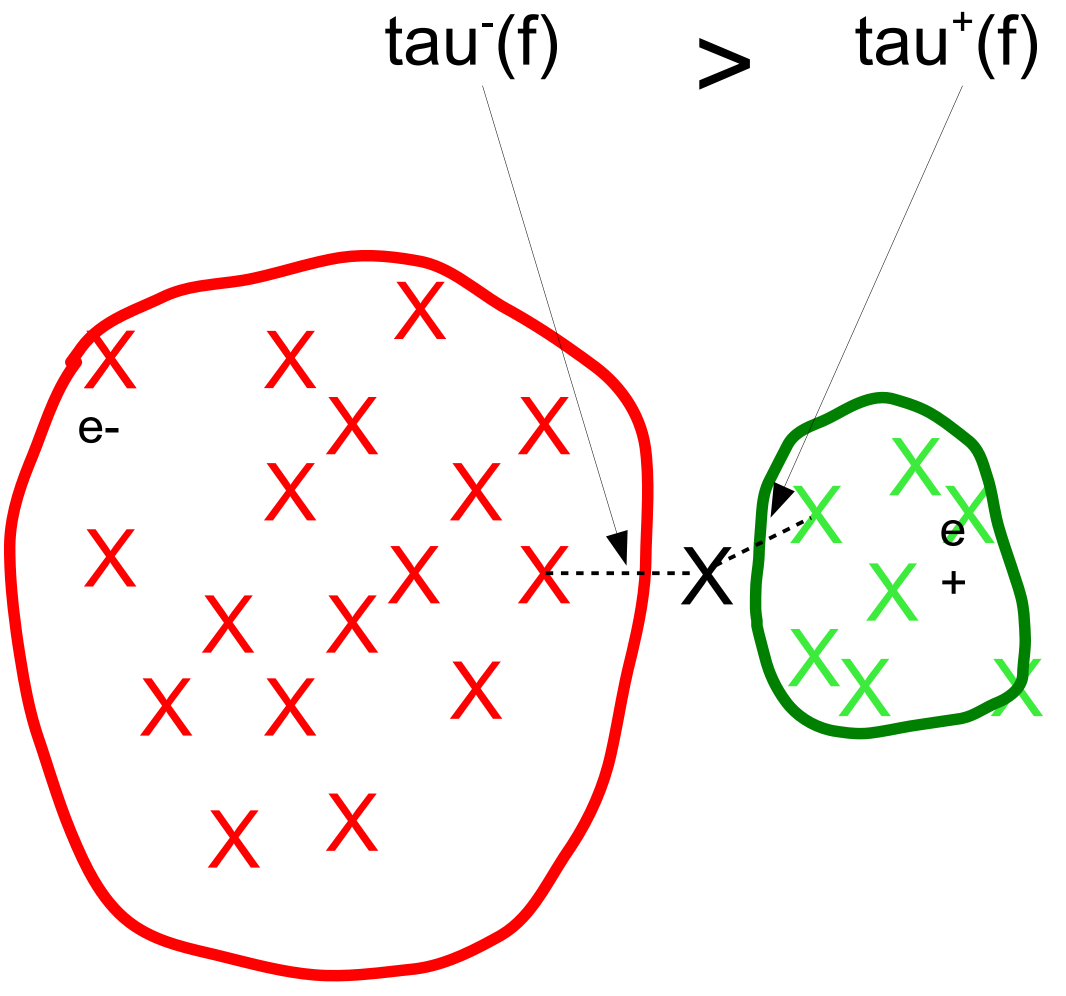
A new channel



- Oriented graph
  - Features (SIFTs) as nodes
  - Connectivity rule
-



NN⁻(f) < NN⁺(f)
e-
e+



tau-(f)
>
tau+(f)
e-
e
+

Nearest Neighbour Distance

- Multi Probe LSH Index
  (Search NN in the most probable bin)
  Dong et al., CIKM08: Modeling LSH for performance tuning

Classification by detection

- Efficient subwindow search
  (Branch and bounds)
  Lampert et al., CVPR08: *Beyond SlidingWindows: Object Localization by Efficient Subwindow Search*

Multi class

- Use 1 vs 1 classification
  for each pair of classes [*c d*]
  compute '*c*' against '*d*' classification
  chose the class with the best score

Multi-class classifier

$$\tilde{c}(I) = arg\,max_c \sum_{c' \neq c} H\left(E^{(c,c')}(I)\right)$$

Multi-class classifier
decision rule

$$E^{(c,c')} = \tau^{c}(I) - \tau^{c'}(I)$$

Binary prediction function

$$H(x) = \begin{cases} 1 & if \quad x > 1 \\ -1 & if \quad x < -1 \\ x & otherwise \end{cases}$$

Score function

Classification by detection

- Goal : finding *position* *w* and *class* *c* of an object

- Prediction rule

$$\frac{\prod_{f_i \in \omega} p(f_i | c, \omega) \prod_{f_i \in \bar{\omega}} p(f_i | \bar{c}, \omega)}{\prod_{i \in I} p(f_i | background)}$$

$\left\{ \begin{array}{ll} c & \text{object class} \\ \bar{c} & \text{'non-object' class} \end{array} \right.$