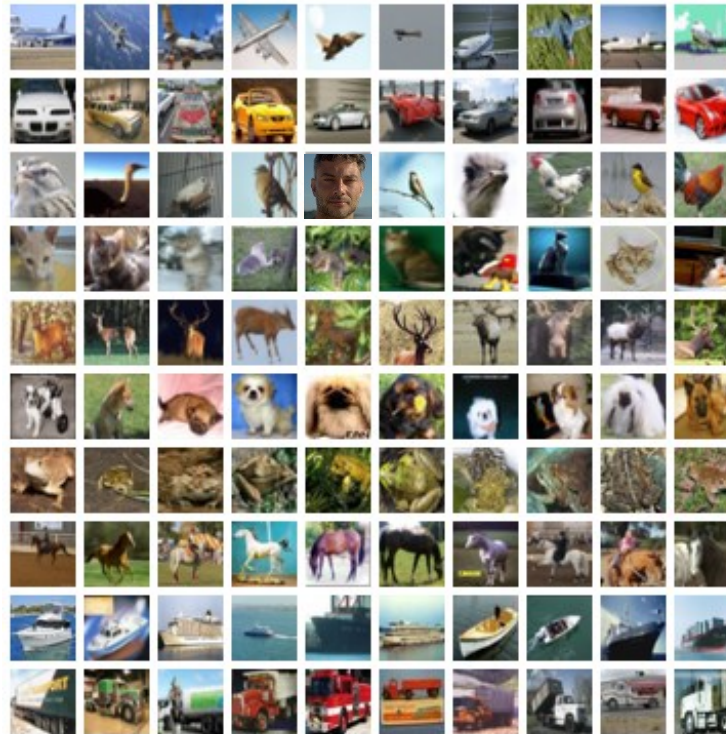


Sistemi avanzati per il Riconoscimento



Riconoscimento di oggetti e scene – metodi discriminativi

Dr. Marco Cristani



Riconoscimento di (categorie di) oggetti

- Area di ricerca primaria nella computer vision, largamente focalizzata sulle «*categorie*» anziché su oggetti specifici
 - Per esempio, bicicletta



- Esiste il problema di cosa sia una categoria, a livello semantico



Riconoscimento di (categorie di) oggetti (2)

- Soluzione: non esiste, ma esistono numerosi dataset di oggetti contenenti categorie con numerosi esempi
 - Caltech 101, Caltech 256, PASCAL (10K immagini, 10 categorie), LabelMe, Imagenet
- Ci si limita ad utilizzare tali dataset, assumendo contengano l'intera variabilità visuale di una categoria



Classificazione

- Il riconoscimento di oggetti si può suddividere principalmente in
 - **Classificazione:** presenza o assenza di almeno una istanza di categoria in un'immagine
 - Domanda: *è un'immagine che appartiene alla categoria «persone»?*
 - Domanda equivalente: *esistono delle persone nell'immagine?*



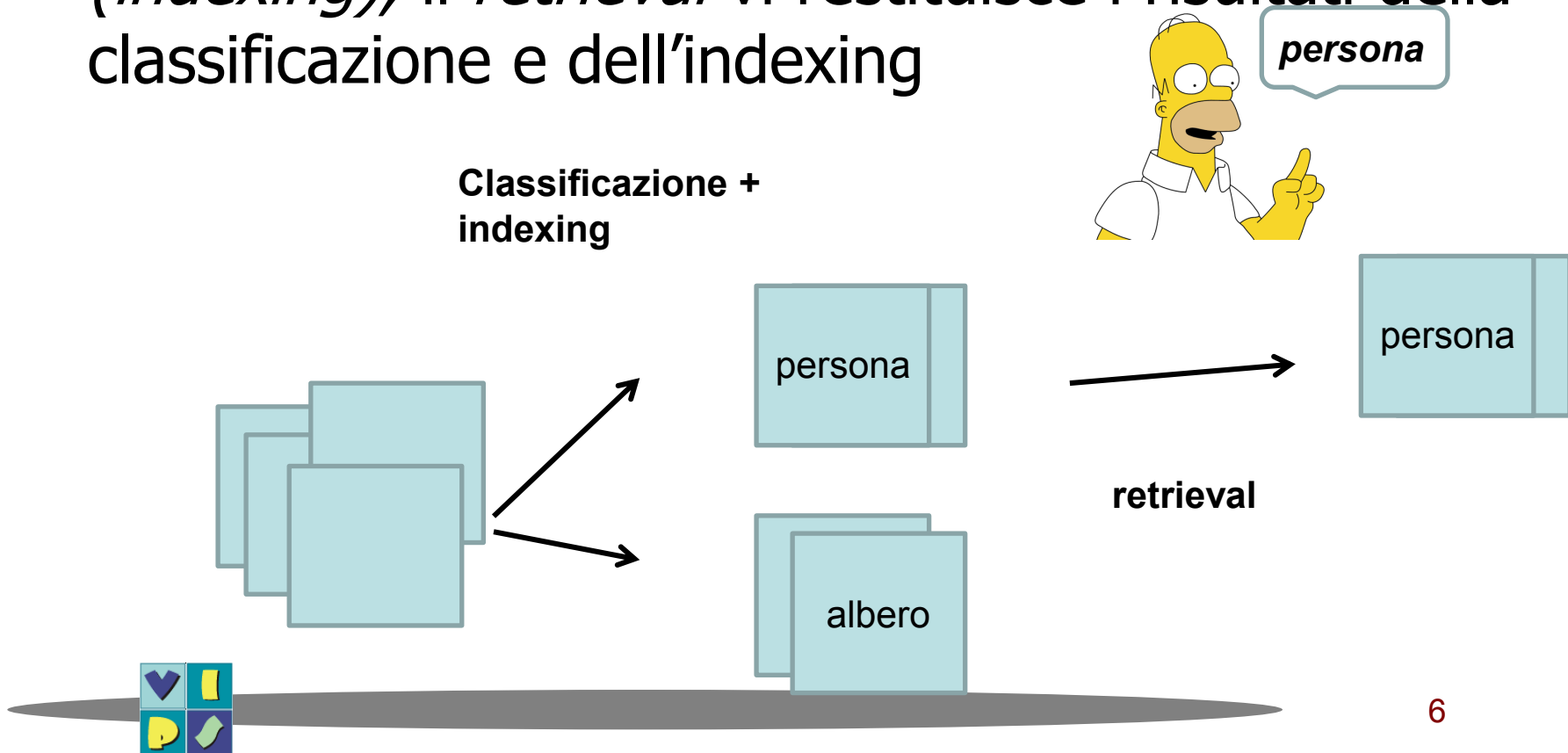
Classificazione (2)

- Questa operazione è quella a cui risponde google quando chiedete con una parola delle immagini (es.:»person«)



Classificazione (3)

- Assumendo sia stata eseguita una classificazione a monte, ed una operazione di *indicizzazione* (*indexing*), il *retrieval* vi restituisce i risultati della classificazione e dell'indexing



Localizzazione

- **Localizzazione (detection):**
trovare dove sono collocate
nelle immagini tutte le istanze
di una particolare categoria
- Domanda: *dove si trovano le
persone nell'immagine?*
- Domanda equivalente: *quante
persone si trovano
nell'immagine?*



Localizzazione (2)

- La localizzazione è necessaria solitamente nel momento in cui si deve interagire con l'ambiente
- Interfacce visuali multimodali (google glasses), sorveglianza, navigazione



Localizzazione (3)

- Più difficile della classificazione, perchè richiede in output maggiore informazione



Verifica (verification)

- Altri problemi meno affrontati sono:
 - **Verifica (verification):**
verificare se una particolare regione dell'immagine contiene un'istanza dell'oggetto di una particolare categoria
 - Domanda: *qui dentro c'è una persona?*



Identificazione (identification)

- **Identificazione (identification)**: verificare se una particolare regione dell'immagine contiene una **particolare** istanza di una particolare categoria
- Domanda: *qui dentro c'è il Potala Palace?*



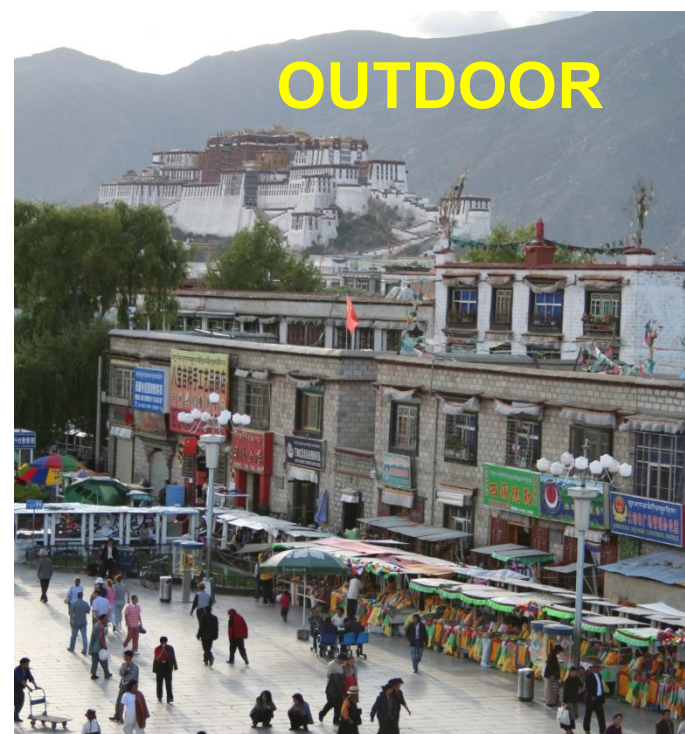
Segmentazione semantica

- **Segmentazione semantica (semantic segmentation):** segmentare l'immagine in modo che ogni segmento abbia un ben preciso contenuto semantico
- Domanda: *cosa c'è nell'immagine, e dove si trova?*



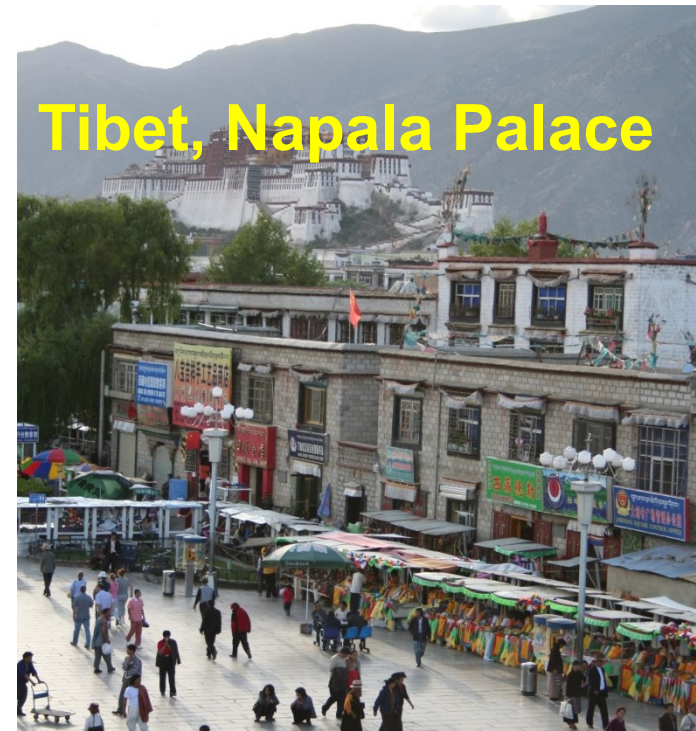
Riconoscimento della scena

- Il riconoscimento di scene si può suddividere principalmente in
 - **Categorizzazione della scena (scene categorization)** dare all'intera immagine un'etichetta
 - Domanda: *dovessi scegliere una sola parola da associare alla scena, quale le darei?*



Riconoscimento della scena

- Geolocalizzazione della scena(scene geolocalization) dare un'etichetta geografica alla scena fotografata
- Domanda: *dove mi trovo e cosa sto guardando?*



Applicazioni: Computational Photography

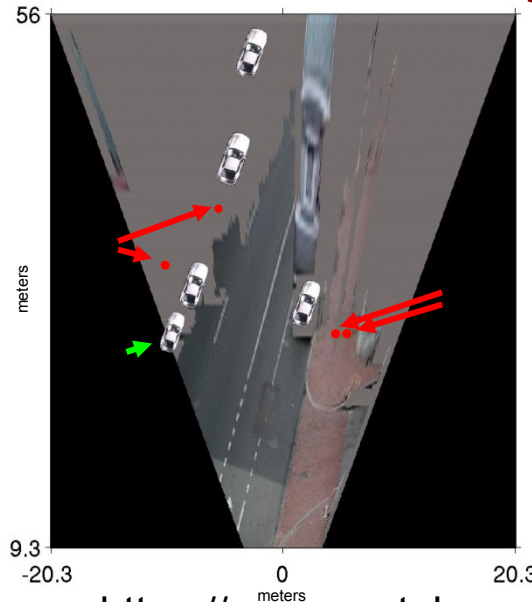


[Face priority AE] When a bright part of the face is too bright



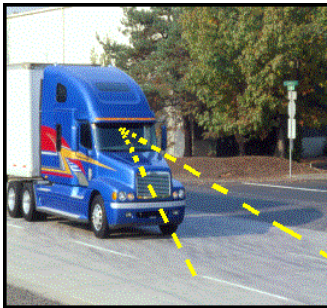
Assisted driving

Localizzazione di auto e pedoni



<https://www.youtube.com/watch?v=VuHSDpfm2AU>

Lane detection



- Collision warning systems with adaptive cruise control,
- Lane departure warning systems,
- Rear object detection systems,

<https://www.youtube.com/watch?v=yMLBKgn9Nq8>



Applicazioni: image retrieval



Places

[London](#)
[New York](#)
[Egypt](#)
[Forbidden City](#)

Celebrities

[Michael Jordan](#)
[Angelina Jolie](#)
[Halle Berry](#)
[Seth Rogan](#)
[Rihanna](#)

Art

[impressionism](#)
[Keith Haring](#)
[cubism](#)
[Salvador Dalí](#)
[pointillism](#)

Shopping

[evening gown](#)
[necklace](#)
[shoes](#)

Refine your image search with visual similarity

Similar Images allows you to search for images using pictures rather than words. Click the "[Similar images](#)" link under an image to find other images that look like it. Try a search of your own or click on an example below.

paris



[Similar images](#)



[Similar images](#)



[Similar images](#)



[Similar images](#)

temple



[Similar images](#)



[Similar images](#)



[Similar images](#)



[Similar images](#)



Challenges: variazioni di posa



Michelangelo 1475-1564



Challenges: variazioni di illuminazione



Challenges: occlusioni



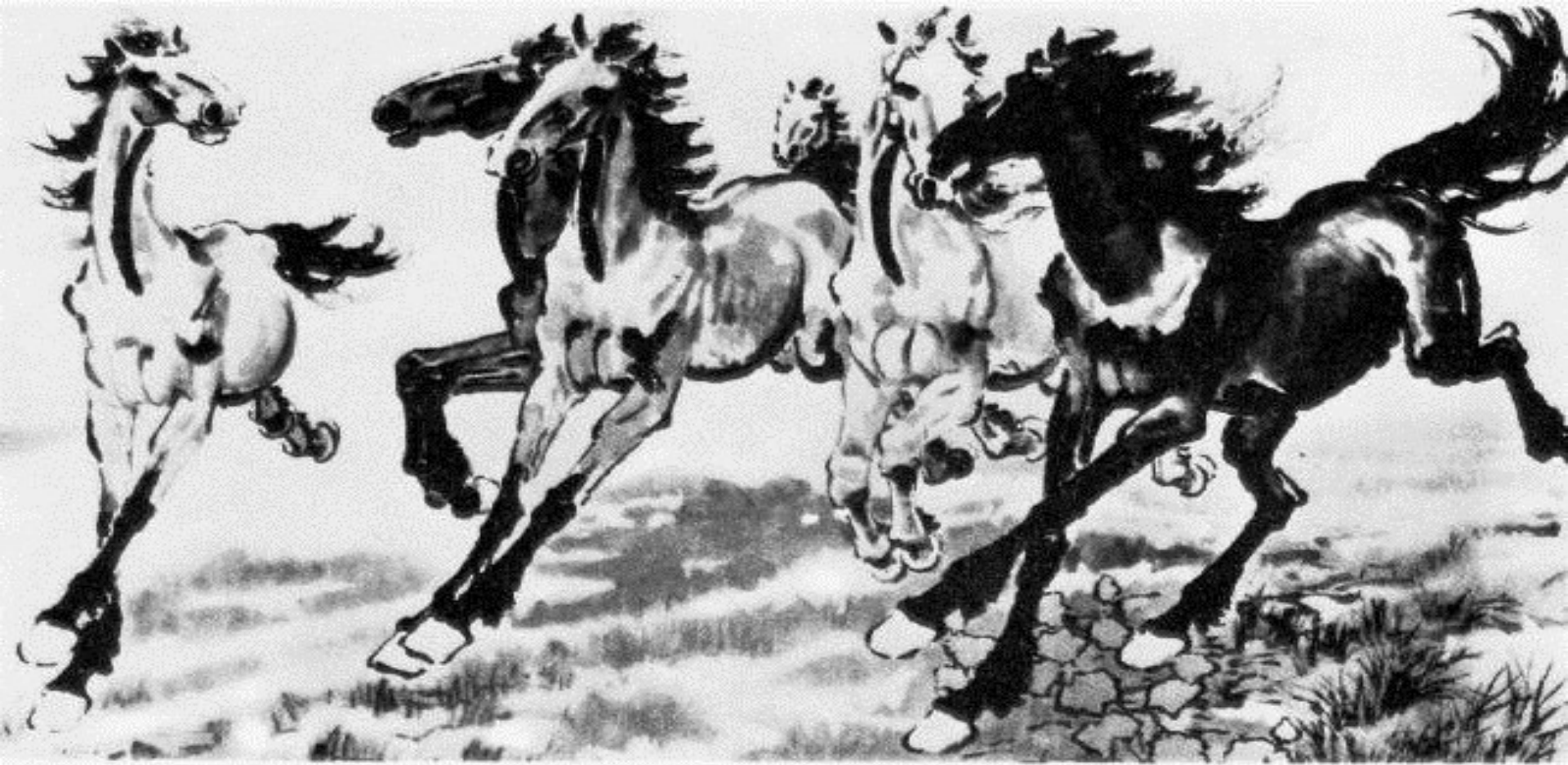
Magritte, 1957



Challenges: variazioni di scala



Challenges: deformazioni



Xu, Beihong 1943



Challenges: clutter



Klimt, 1913



Challenges: variazioni intra classe

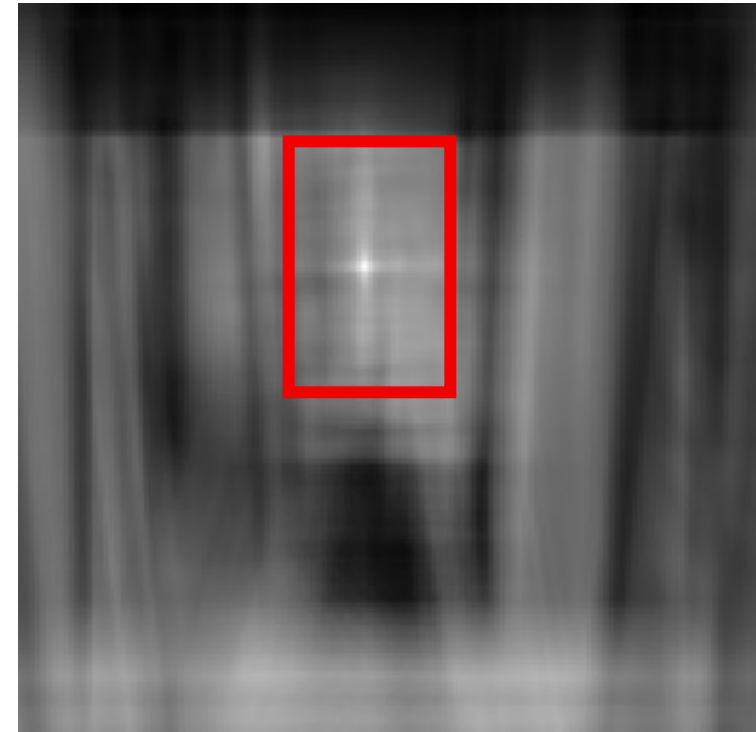
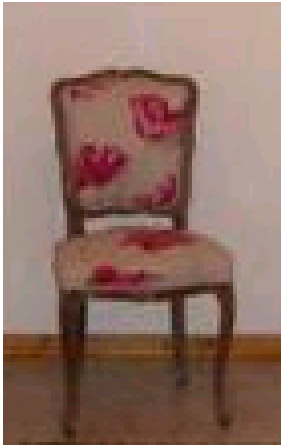


Il riconoscimento di oggetti è così difficile?

Trova la sedia nell'immagine

Faccio cross correlazione normalizzata

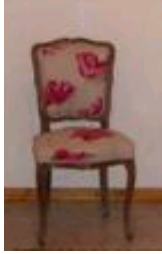
Questa è una sedia



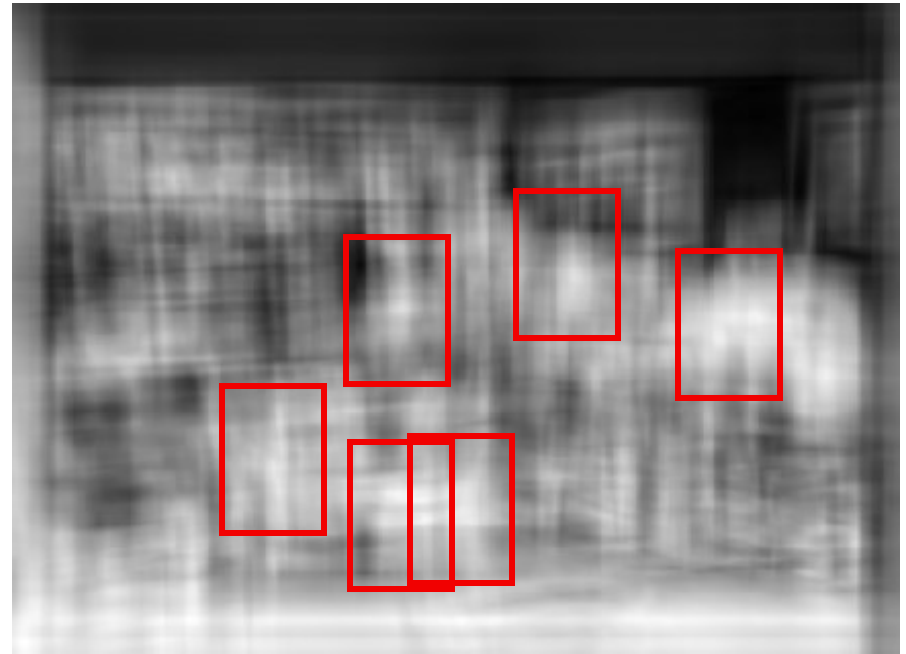
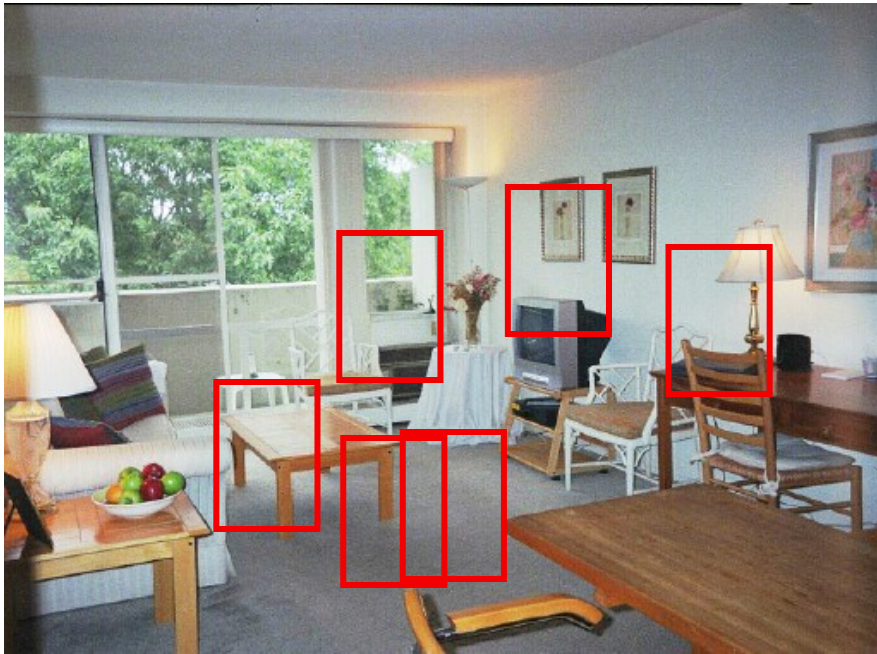
Facile!



Il riconoscimento di oggetti è così difficile?



Trova la sedia in questa immagine



Molti falsi positivi, *template matching* come tecnica a sé stante non viene piu' usata dagli anni 80



La ricerca in riconoscimento di oggetti

- La ricerca è stata negli ultimi 20 anni maggiormente focalizzata sulla classificazione anziché sulla localizzazione
- Nella comunità scientifica, i protocolli di test più conosciuti sono progettati per la classificazione
- Questo si riflette anche nei metodi a disposizione: esistono più tecniche di classificazione che di localizzazione
- Inversione di tendenza con il dataset PASCAL, e il challenge associato
- Attualmente, l'enfasi si sta spostando sulla localizzazione

<http://pascallin.ecs.soton.ac.uk/challenges/VOC/>



Approcci per classificazione di oggetti e scene

- Tassonomia: Locali VS Globali
 - *Approcci locali*: l'unità fondamentale di analisi è una *patch*, o una *regione* dell'immagine
 - Si parla anche di *feature locali*
 - Posso applicarli a parti delle immagini
 - Resistenti alle occlusioni
 - Bag of visual words (poca struttura spaziale)
 - Pictorial structures (molta struttura spaziale)



Approcci per classificazione di oggetti e scene

- Tassonomia: Locali VS Globali
 - *Approcci globali*: l'unità fondamentale di analisi è l'intera immagine
 - Non posso applicarli a parti delle immagini
 - Poco resistenti alle occlusioni
 - Si parla anche di feature globali
 - Adatti specialmente per scene categorization
 - GIST (essenzialmente, analizza lo spettro delle immagini)





Modello Bag-of-words

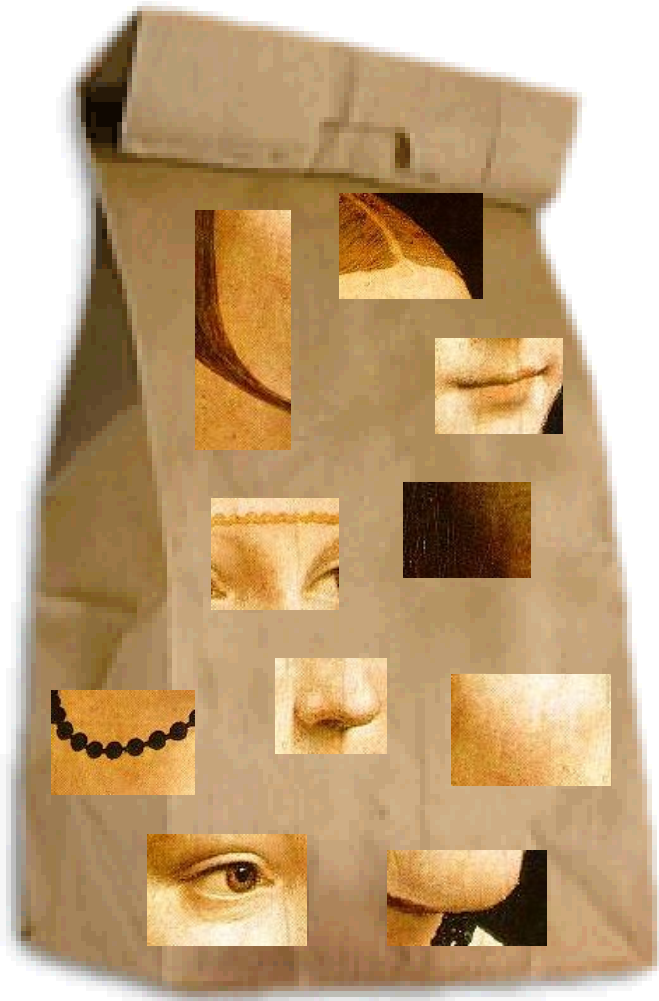
Thans to Li Fei-Fei (Oxford)



Oggetto



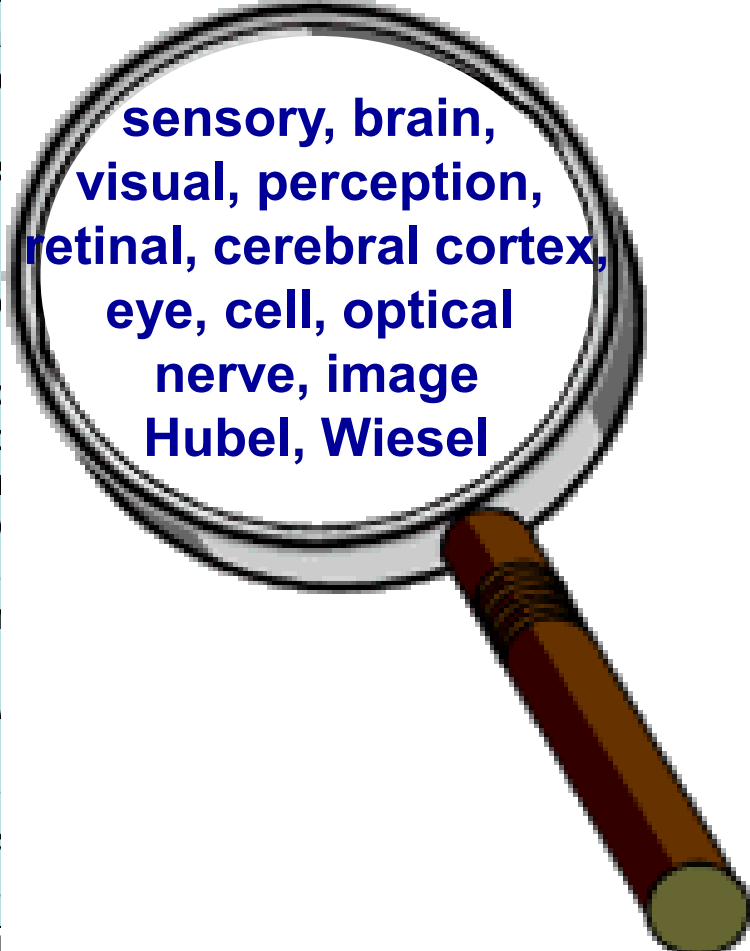
Bag of 'words'



Analogia con i documenti testuali

Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the

message. For a long time, the image centers in the brain were considered the movie image. The discovery that we know the world through perception is more complex than following the visual path. Hubel and Wiesel demonstrated that the image is not stored in the brain as a specific image.



**sensory, brain,
visual, perception,
retinal, cerebral cortex,
eye, cell, optical
nerve, image
Hubel, Wiesel**

China is forecasting a trade surplus of \$90bn (£51bn) to \$100bn this year, a threefold increase on 2004's \$32bn. The Commerce Ministry said the surplus would be created by

a predicted
compa
\$660b
annoy
China
delibe
agree
yuan i
govern
also n
dema
countr
yuan a
permit
the US
freely.
it will t
allowin

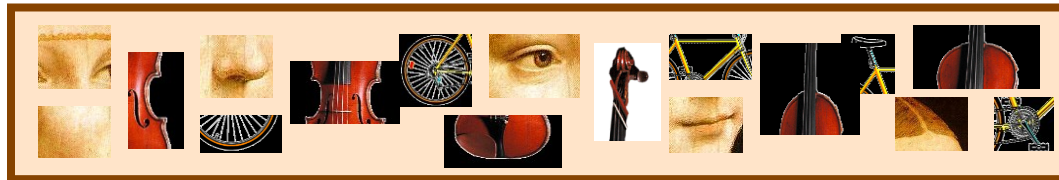
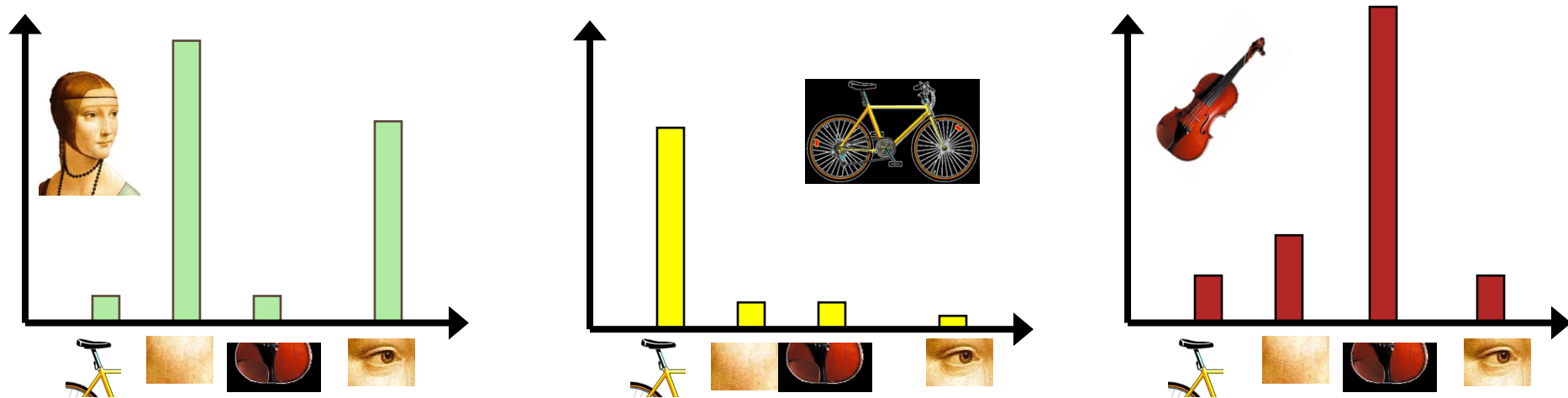


**China, trade,
surplus, commerce,
exports, imports, US,
yuan, bank, domestic,
foreign, increase,
trade, value**

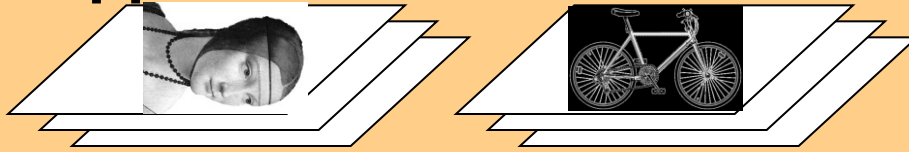


Bag of visual words: punti fondamentali

- Estraggo dall'immagine delle feature locali, che assumo essere indipendenti tra di loro
- Le rappresento attraverso un modello ad istogramma

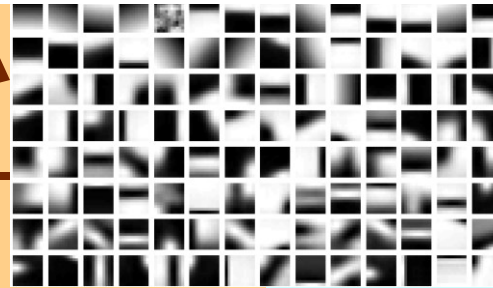


Addestramento della rappresentazione

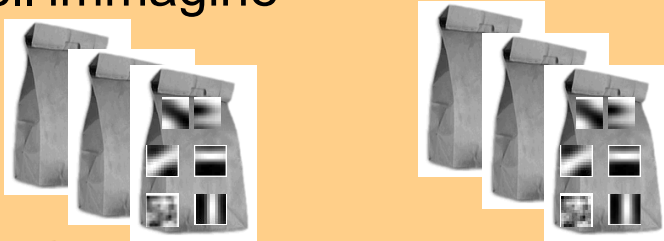


feature extraction
& rappresentazione

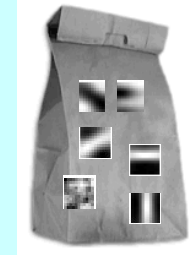
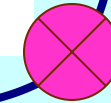
Dizionario di codewords



Rappresentazione
dell'immagine



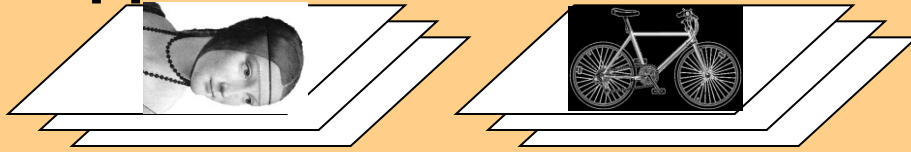
classificazione



**Decisione della
classe**

**Addestramento dei modelli
di categoria (e/o) classificatori**

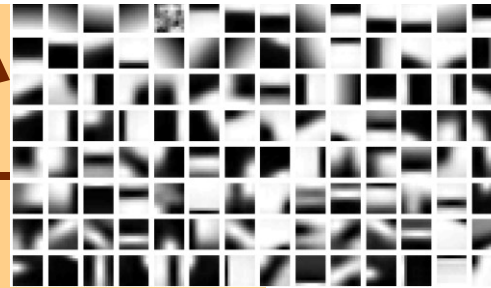
Addestramento della rappresentazione



2.

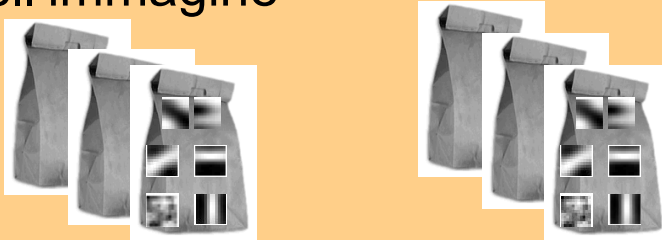
Dizionario di codewords

1. feature extraction
& rappresentazione

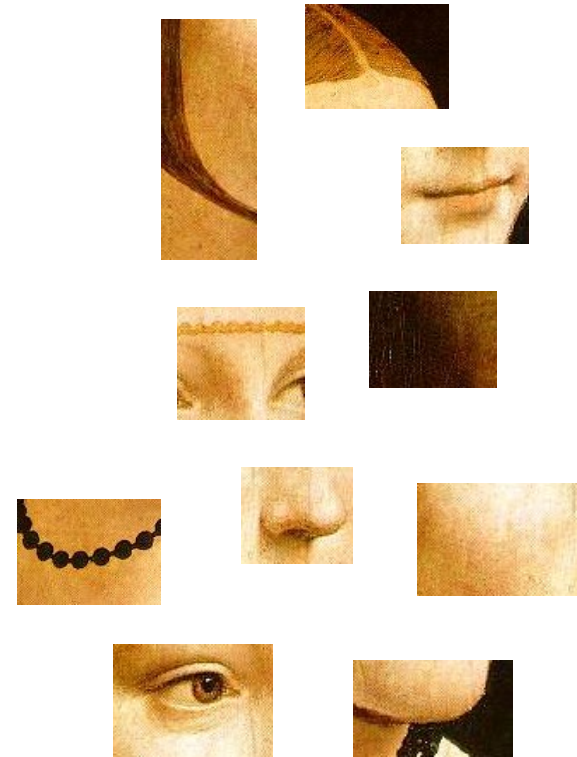


Rappresentazione
dell'immagine

3.

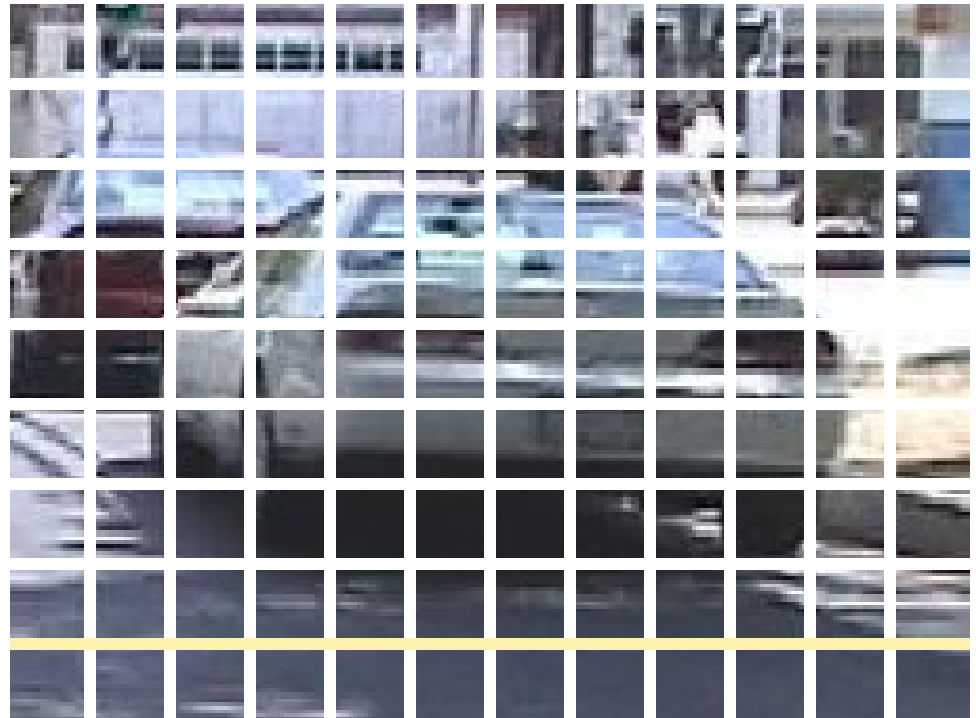


1.Feature extraction e rappresentazione



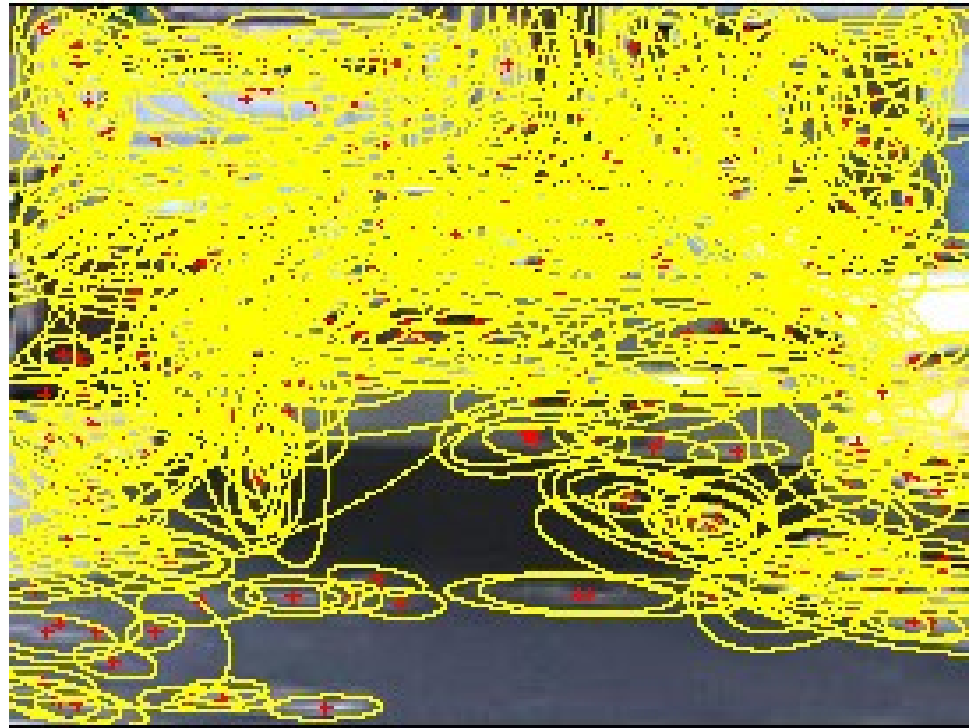
1.Feature extraction e rappresentazione

- Griglia regolare
 - Vogel & Schiele, 2003
 - Fei-Fei & Perona, 2005



1.Feature extraction e rappresentazione

- Griglia regolare
 - Vogel & Schiele, 2003
 - Fei-Fei & Perona, 2005
- Rivelamento di punti di interesse
 - Csurka, et al. 2004
 - Fei-Fei & Perona, 2005
 - Sivic, et al. 2005



1.Feature extraction e rappresentazione

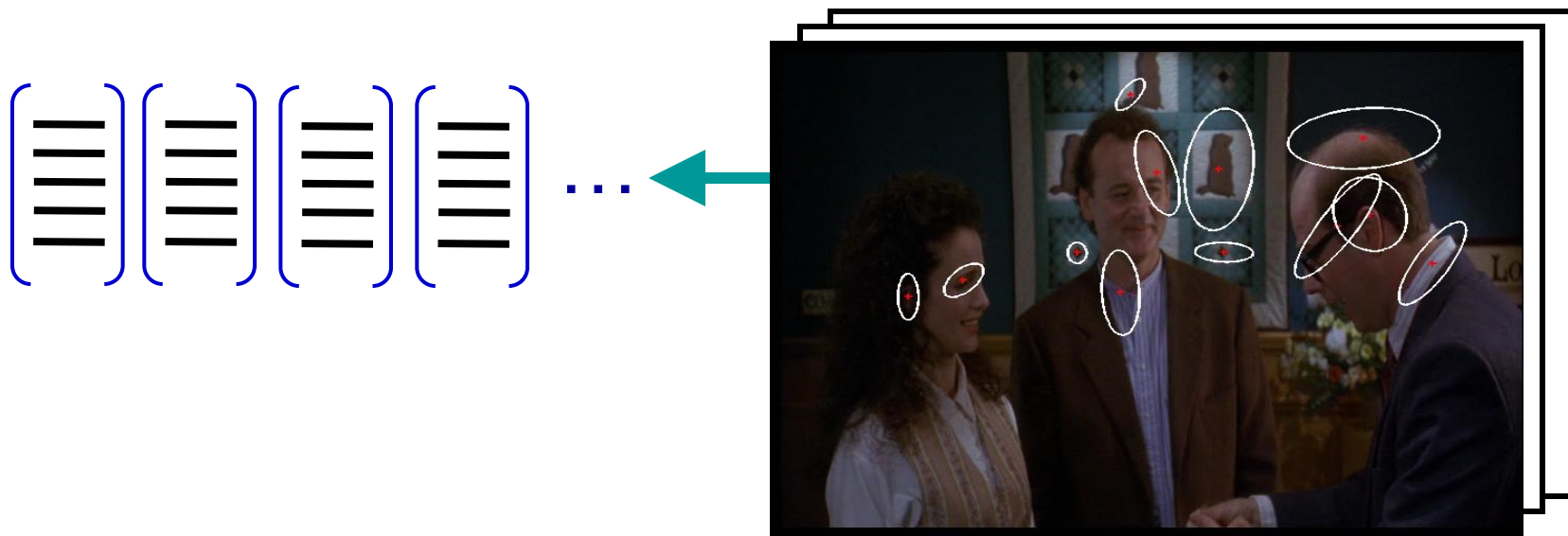
- Griglia regolare
 - Vogel & Schiele, 2003
 - Fei-Fei & Perona, 2005
- Rivelamento di punti di interesse
 - Csurka, et al. 2004
 - Fei-Fei & Perona, 2005
 - Sivic, et al. 2005
- Altri metodi
 - Random sampling (Vidal-Naquet & Ullman, 2002)
 - Patches basate su segmentazione (Barnard, Duygulu, Forsyth, de Freitas, Blei, Jordan, 2003)



1.Feature extraction e rappresentazione



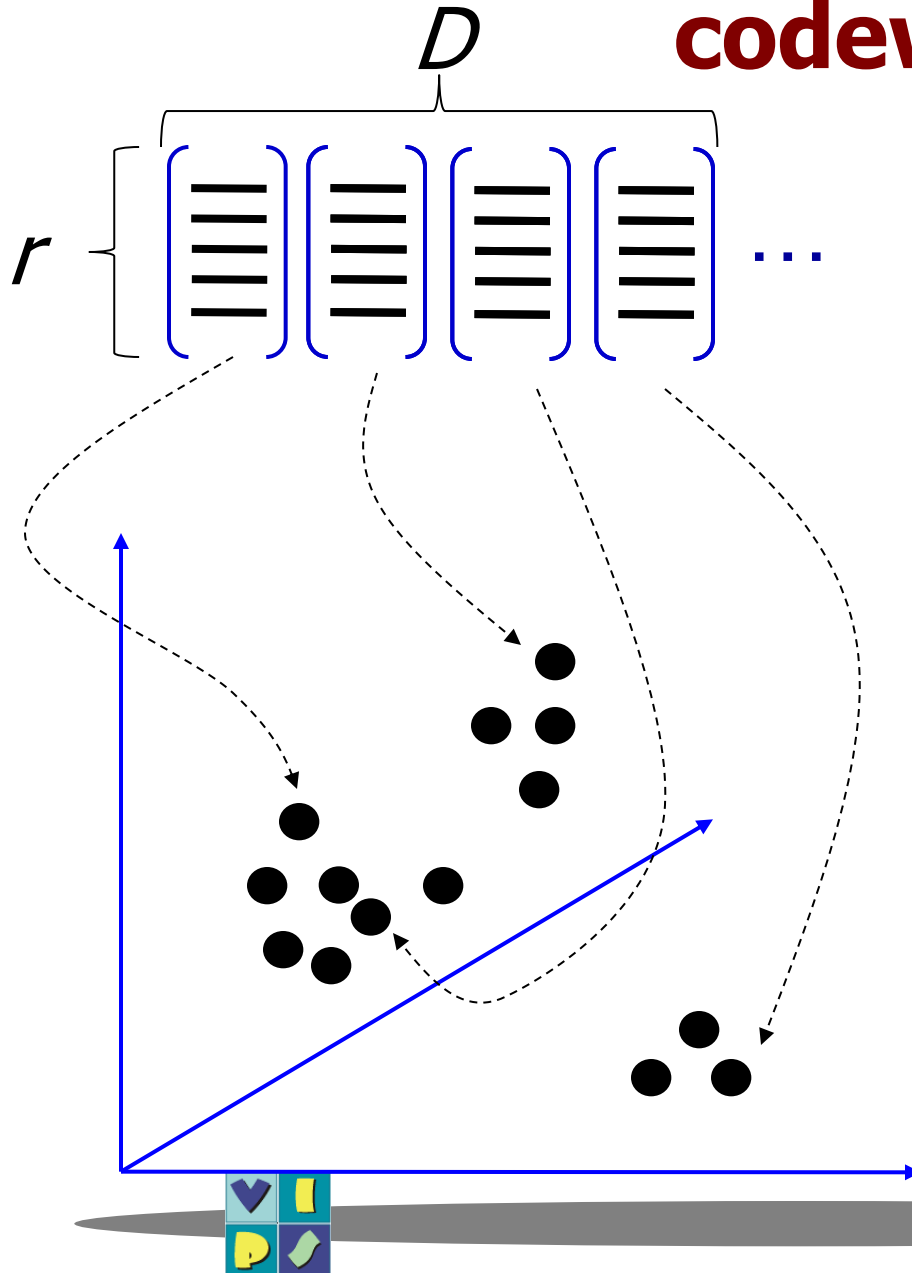
1. Feature extraction e **rappresentazione**



- In generale, per ogni feature estratta, calcolo una rappresentazione in uno spazio *r-dimensionale*

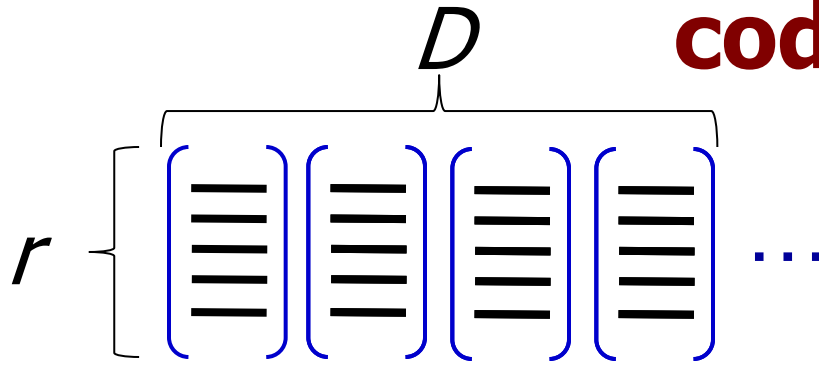


2. Formazione del dizionario di codewords

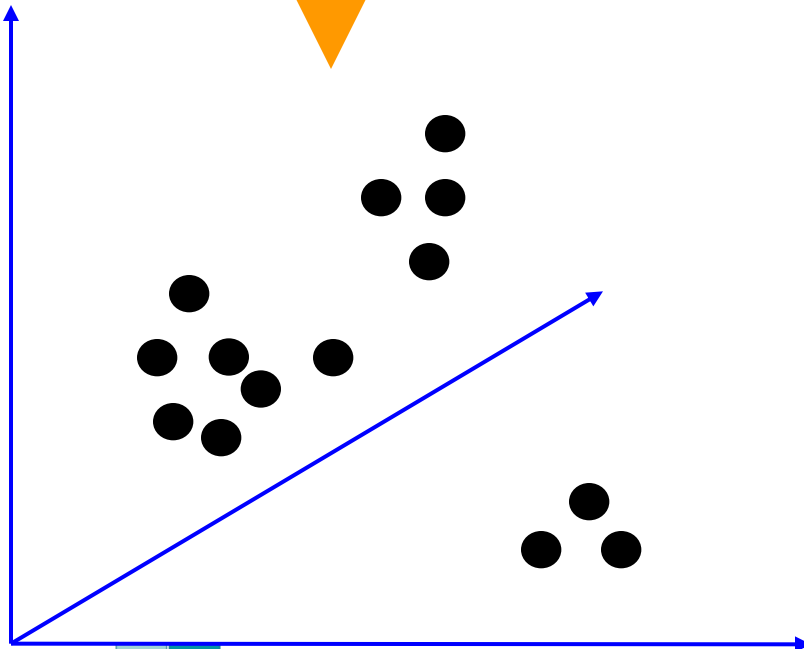
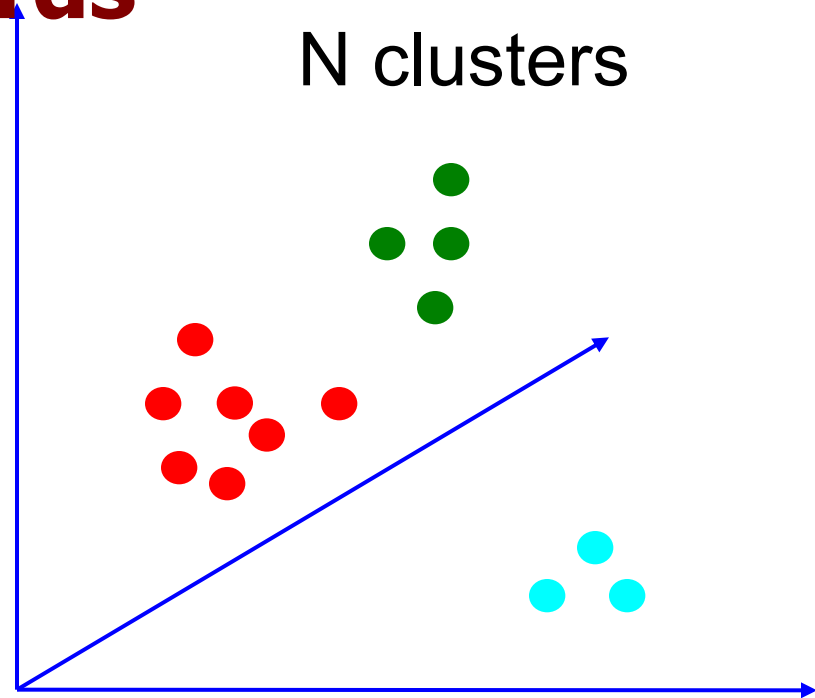


- Metto assieme tutte le D rappresentazioni nel loro spazio r -dimensionale
- In pratica, mi creo una matrice $r \times D$ con tutte le rappresentazioni

2. Formazione del dizionario di codewords



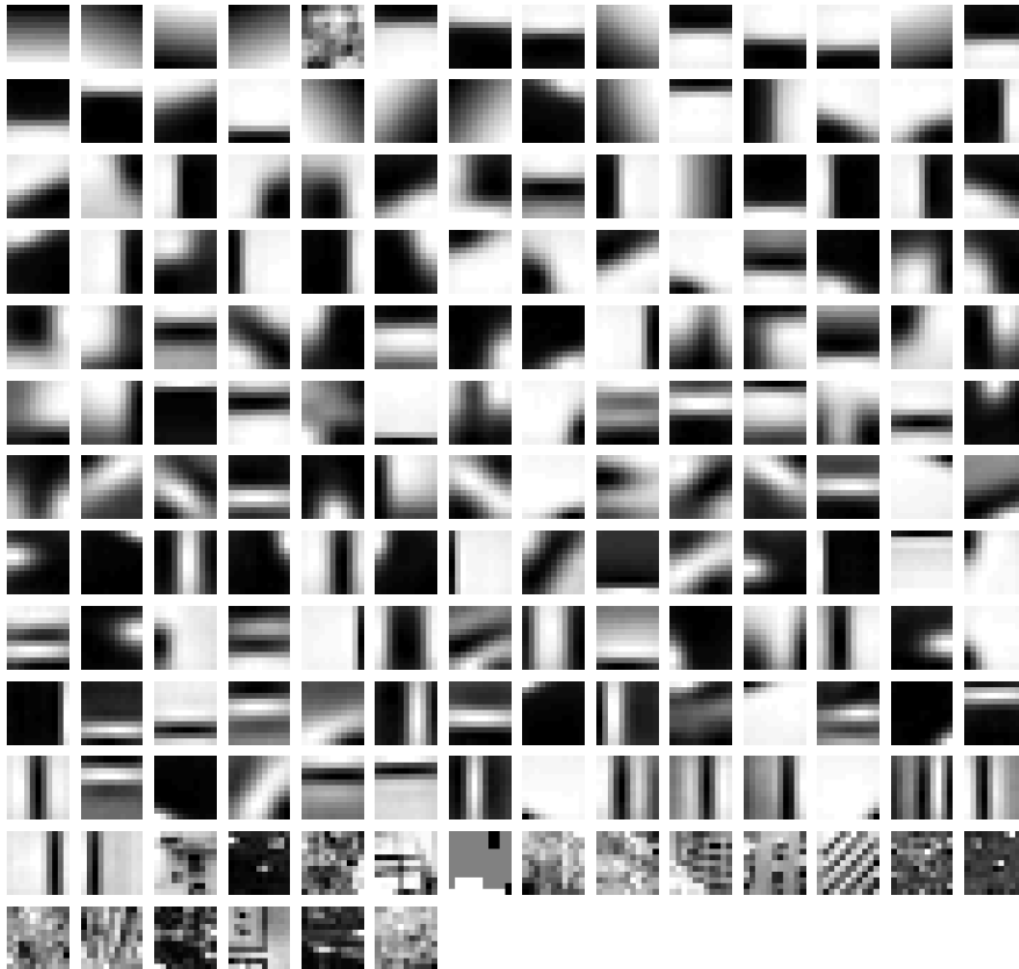
N clusters



Vector quantization
= clustering (K-means)



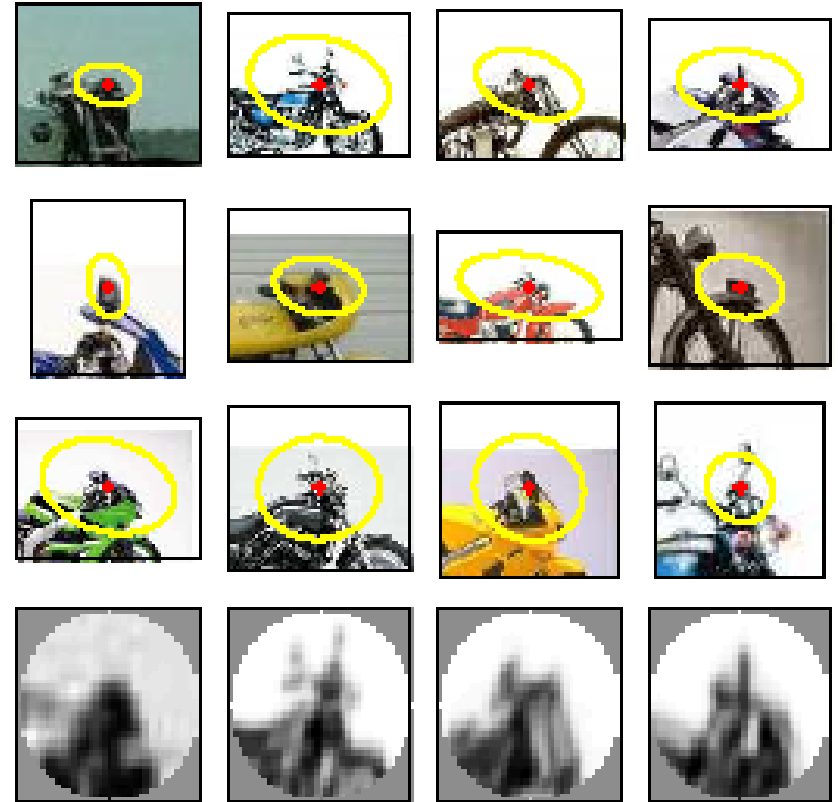
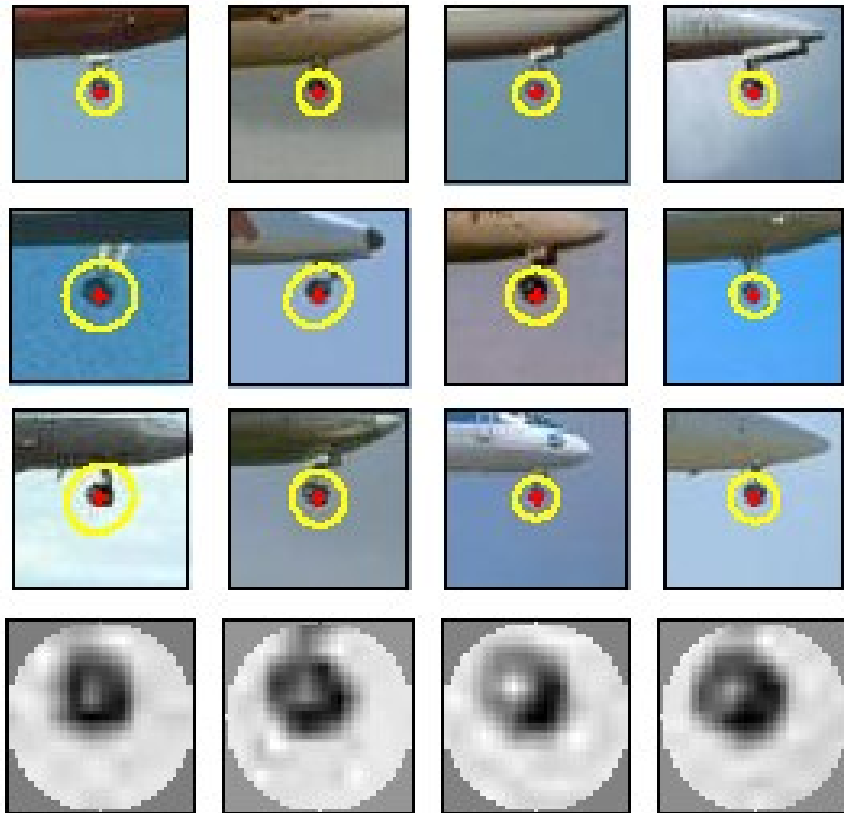
2. Formazione del dizionario di codewords



- Ognuna di queste caselle mi rappresenta, *visualmente* (ossia sull'immagine), la media delle immagini appartenenti a ciascuna delle N codewords

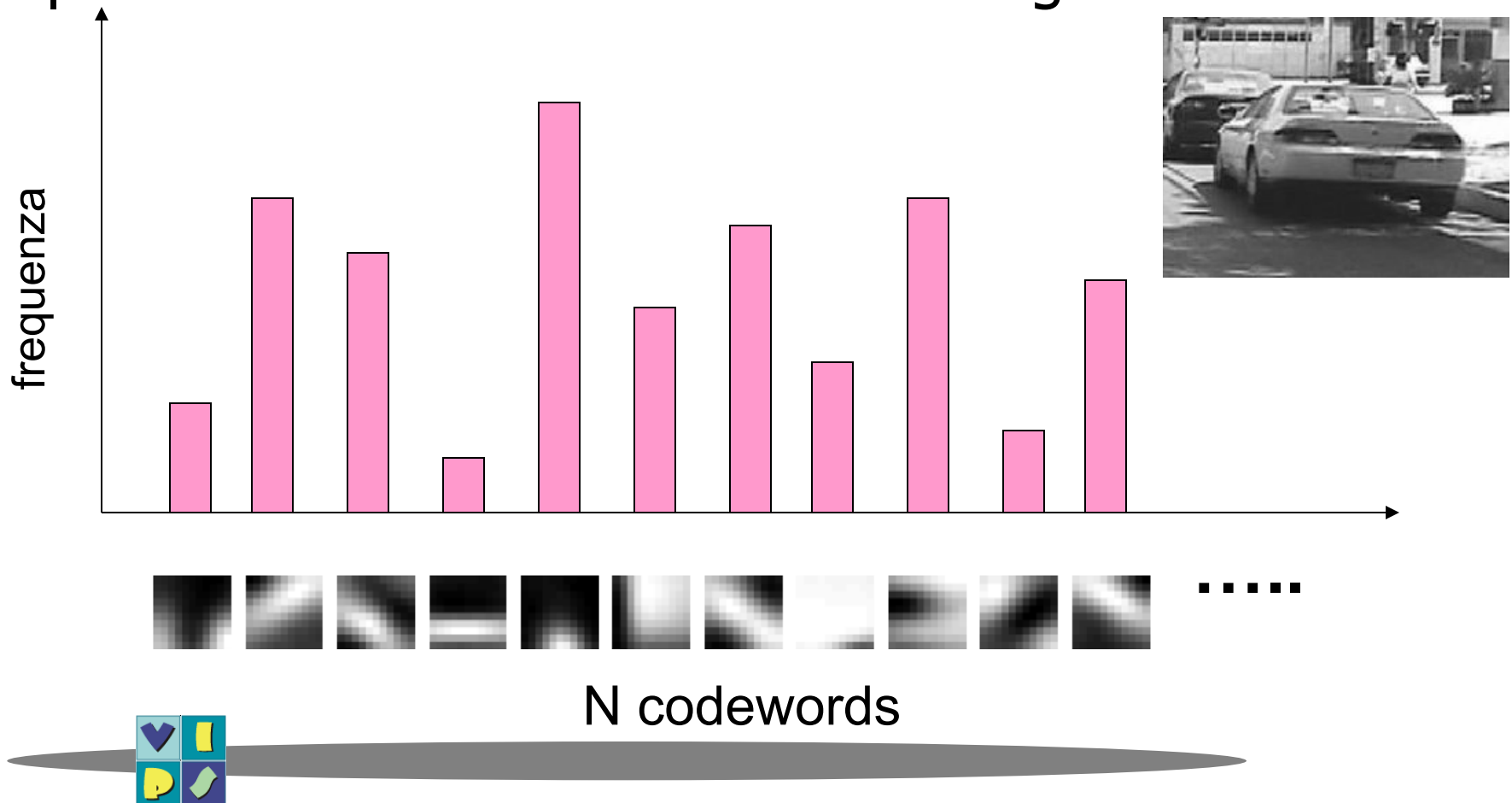


Esempi di patch assegnate alla stessa codeword

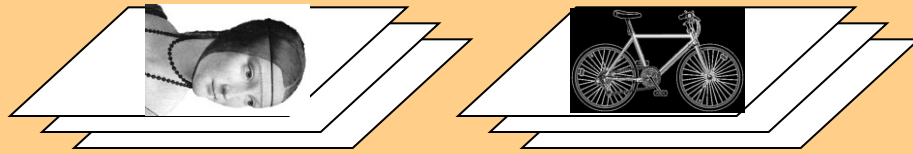


3. Rappresentazione dell'immagine

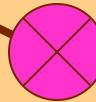
- Per ogni immagine ho un istogramma che mi conta quante istanze di ogni N parole visuale ho trovato nell'immagine



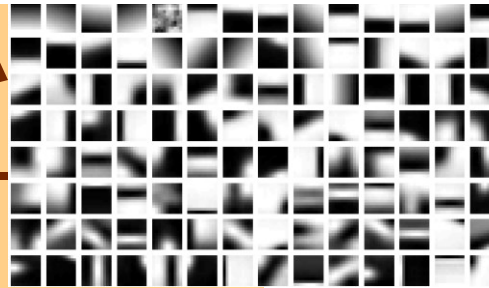
addestramento



1. feature extraction
& rappresentazione

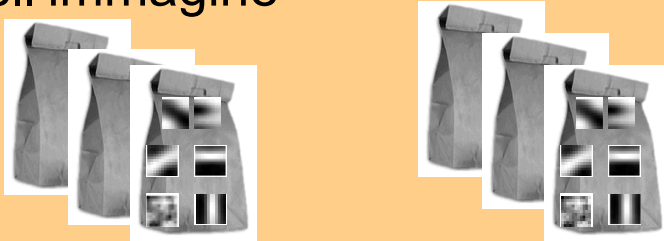


Dizionario di codewords



Rappresentazione
dell'immagine

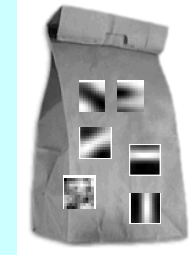
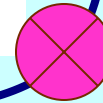
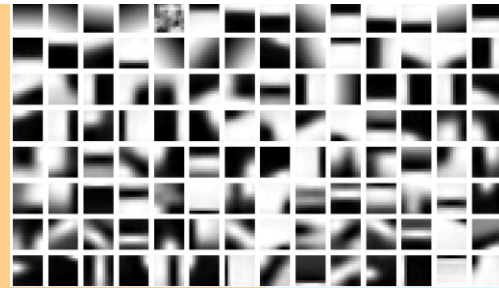
3.



classificazione



Dizionario di codewords

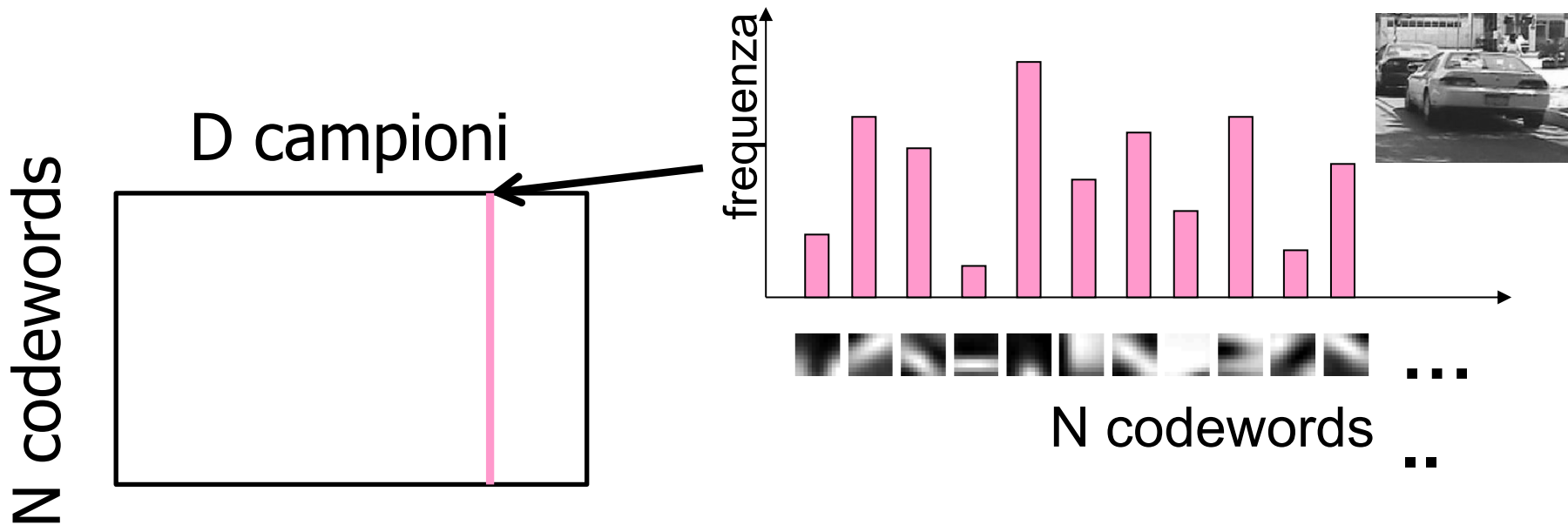


Decisione della
classe

Addestramento dei modelli
di categoria (e/o) classificatori

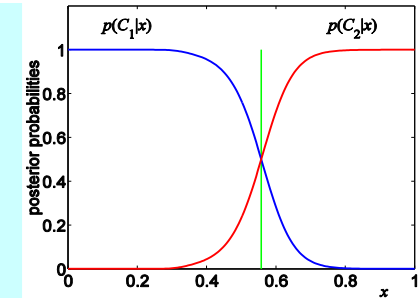
Addestramento dei modelli di categoria

- Assumendo di avere D dati di training, mi costruisco una matrice N codewords $\times D$ campioni

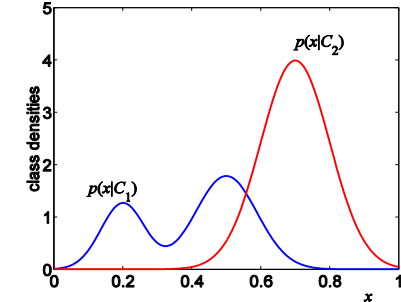


Addestramento del classificatore e classificazione

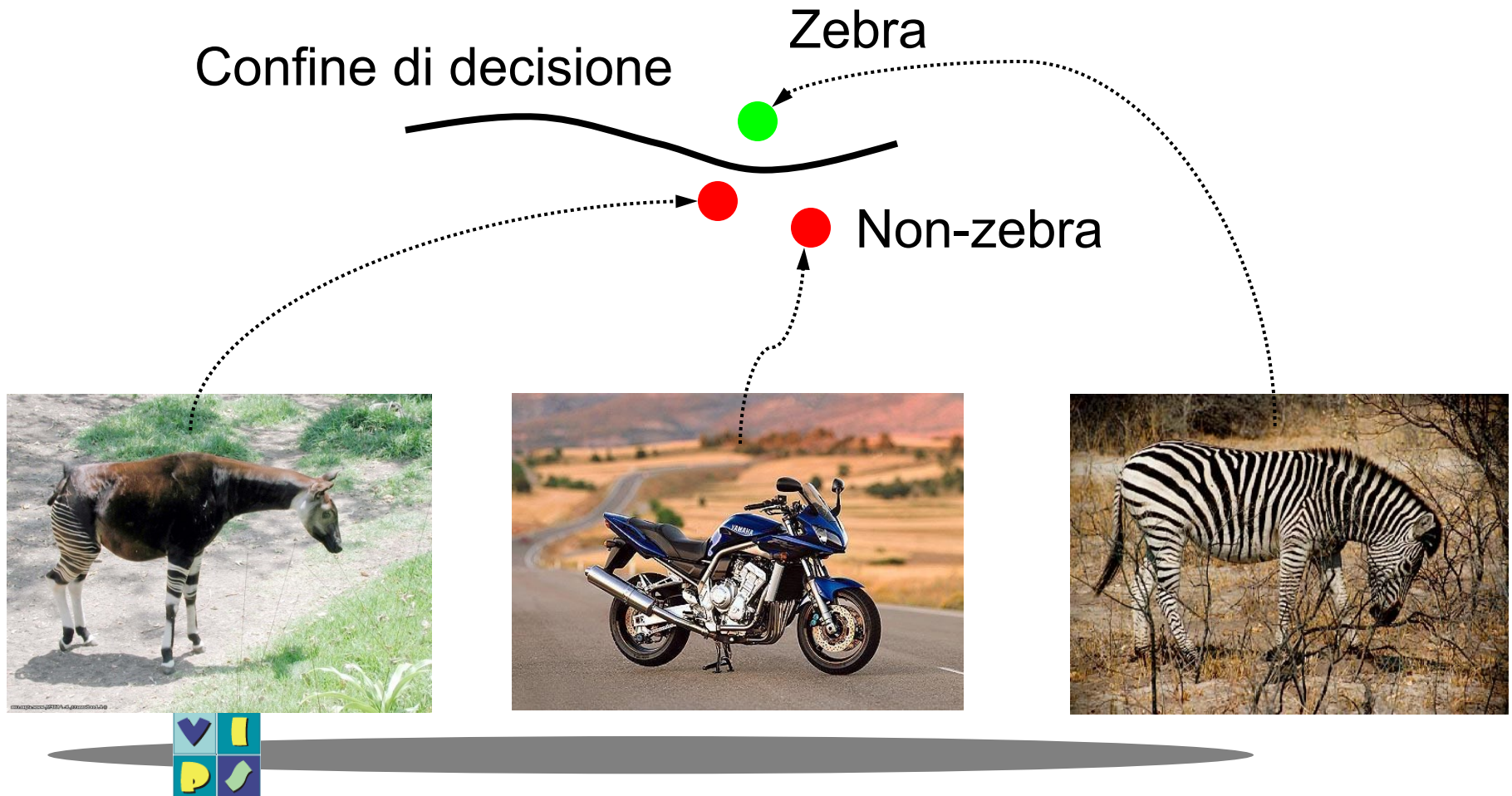
1. Metodi discriminativi:
- SVM



2. Metodi generativi:
- graphical models



Metodi discriminativi basati sulla rappresentazione BoW



Metodi discriminativi basati sulla rappresentazione BoW

- BoW + SVM
- BoW + Spatial histogram + SVM
- BoW + Pyramid matching kernel



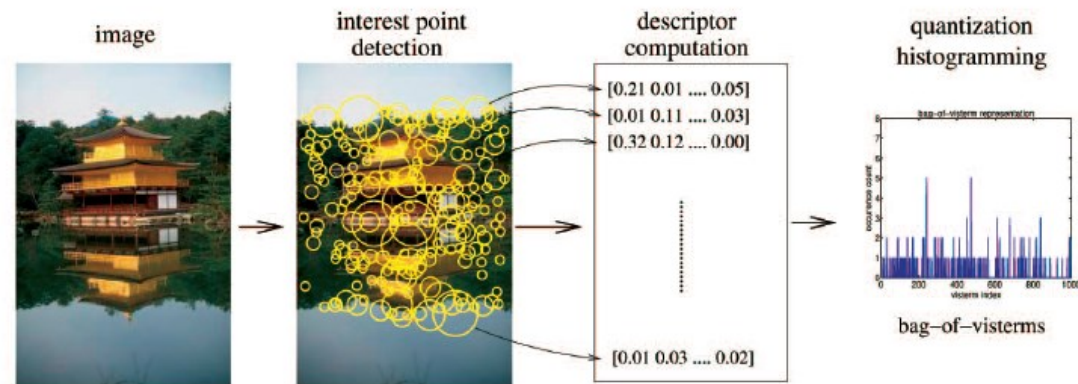
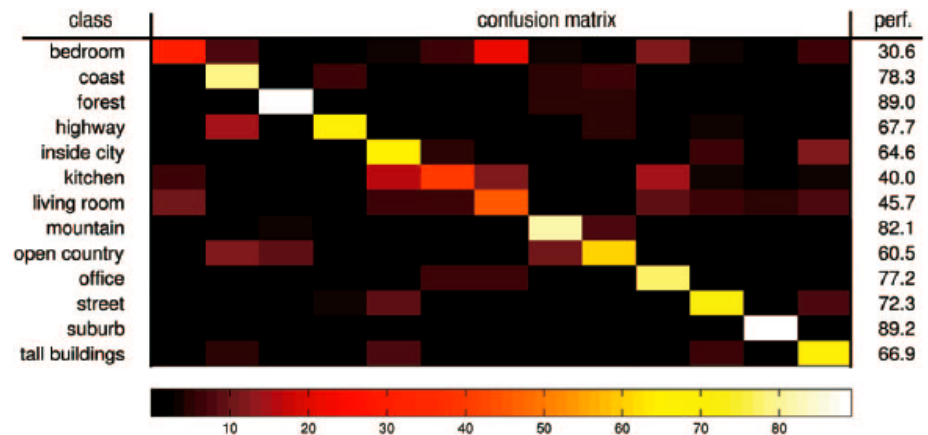
Bag of words + SVM

- Approccio molto semplice
- Assumo di avere due (C) classi di training
 1. Estraggo BoW N dimensionali
 2. Le do in pasto ad una SVM binaria (multiclasse)
 3. Classifico il test data
- Adatto particolarmente per riconoscimento di scene



Bag of words + SVM - problemi

1. Quante codewords usare (ossia, chi mi fa scegliere il migliore N)?
- Leggetevi il paper di sotto... Non esistono teorie a riguardo.

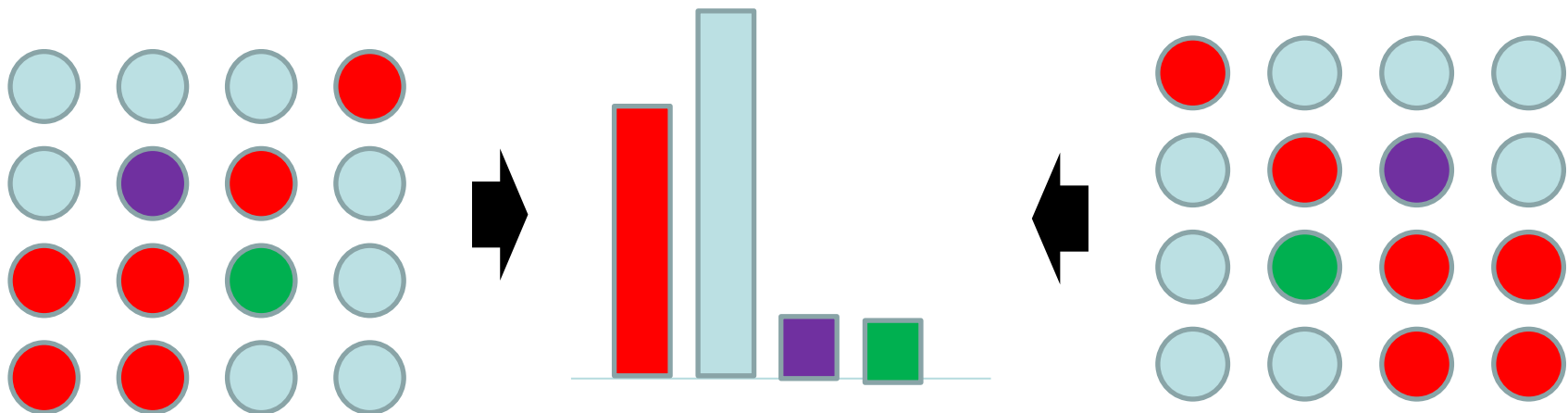


Quelhas, P., Monay, F., Odobez, J. M., Gatica-Perez, D., & Tuytelaars, T. (2007). A thousand words in a scene. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(9), 1575-1589.



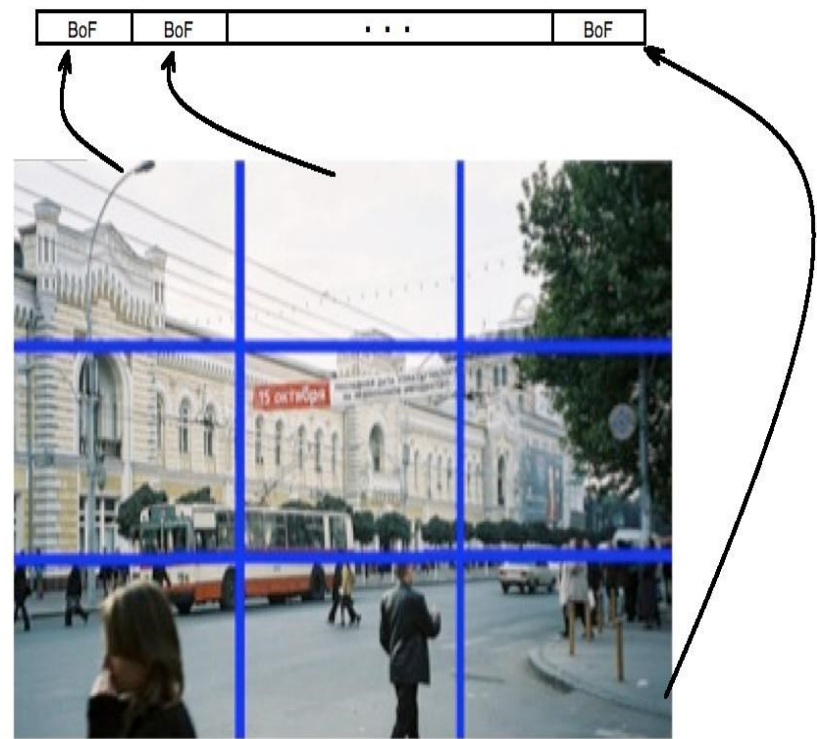
Bag of words + SVM - problemi

- Riesco a catturare la spazialità dell'immagine in questione (come sono collocati tra loro i vari oggetti presenti nell'immagine)?
- No...



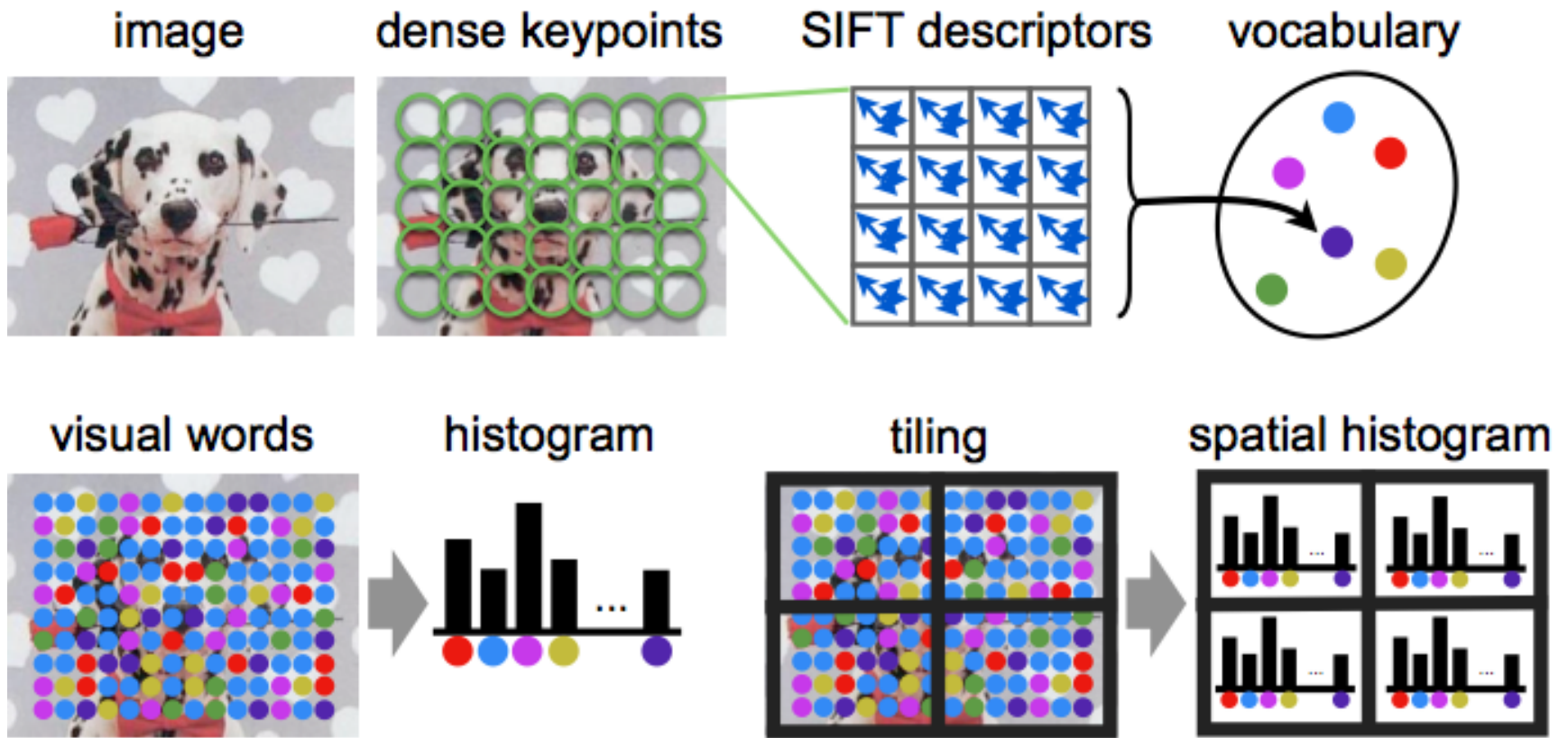
BoW+ Spatial histogram + SVM

- Uguale all'approccio precedente, ma, al posto dello step 1 del precedente approccio
 - a. divido l'immagine in piu' *settori*
 - b. Per ogni settore estraggo l'istogramma BoW
 - c. Concateno tutti gli istogrammi locali



BoW+ Spatial histogram + SVM

- Confronto

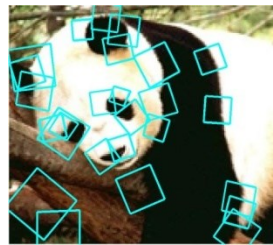
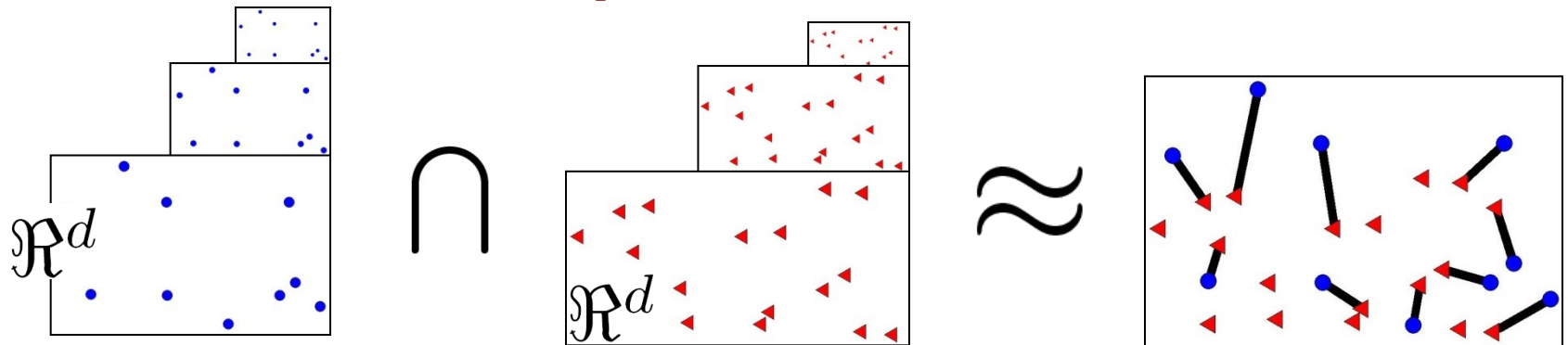


BoW + Pyramid matching kernel

- Grauman & Darrell, 2005, 2006:
 - SVM w/ Pyramid Match kernels
- Others
 - Csurka, Bray, Dance & Fan, 2004
 - Serre & Poggio, 2005



In breve: Pyramid match kernel



L'idea è di riuscire a calcolare una similarità tra le immagini che permetta di confrontare tra loro feature locali corrispondenti

Ciò avviene tramite la creazione di una funzione kernel “piramidale”

$$K_{\Delta} (\Psi(\mathbf{X}), \Psi(\mathbf{Y}))$$

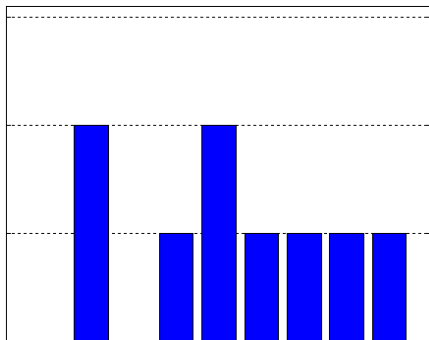
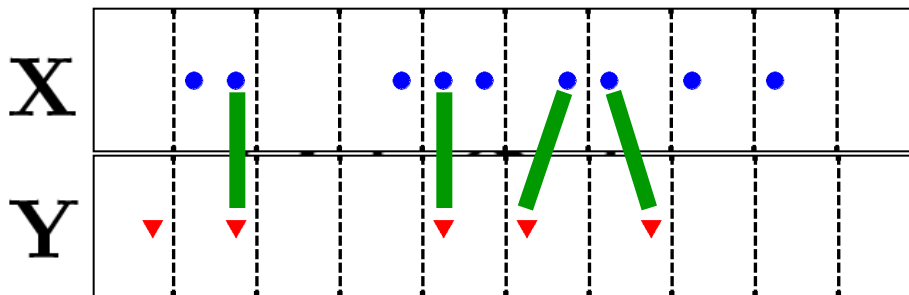


Pyramid Match (Grauman & Darrell 2005)

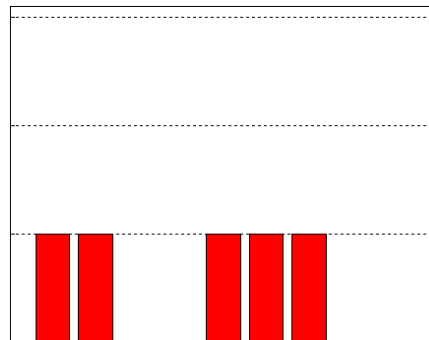
Histogram intersection

$$\mathcal{I}(H(\mathbf{X}), H(\mathbf{Y})) = \sum_{j=1}^N \min(H(\mathbf{X})_j, H(\mathbf{Y})_j)$$

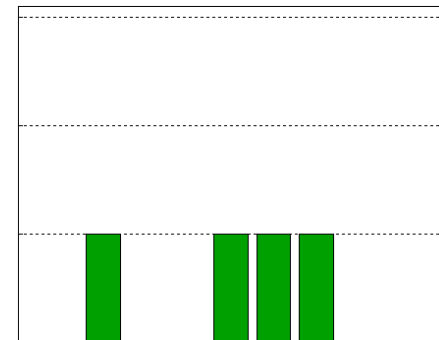
- Definisco l'operatore di intersezione tra istogrammi



$H(\mathbf{X})$



$H(\mathbf{Y})$



$$\mathcal{I}(H(\mathbf{X}), H(\mathbf{Y})) = 4$$



Pyramid Match (Grauman & Darrell 2005)

Histogram intersection

$$\mathcal{I}(H(\mathbf{X}), H(\mathbf{Y})) = \sum_{j=1}^N \min(H(\mathbf{X})_j, H(\mathbf{Y})_j)$$

$$S_i = \overbrace{\mathcal{I}(H_i(\mathbf{X}), H_i(\mathbf{Y}))}^{\text{matches at this level}} - \overbrace{\mathcal{I}(H_{i-1}(\mathbf{X}), H_{i-1}(\mathbf{Y}))}^{\text{matches at previous level}}$$

La differenza tra le intersezioni degli istogrammi a livelli consecutivi conta il numero di nuove coppie matchate tra un livello e l'altro



Pyramid match kernel

$$K_{\Delta} \left(\overbrace{\Psi(\mathbf{X}), \Psi(\mathbf{Y})}^{\text{histogram pyramids}} \right) = \sum_{i=0}^L \frac{1}{2^i} \underbrace{\left(\mathcal{I}(H_i(\mathbf{X}), H_i(\mathbf{Y})) - \mathcal{I}(H_{i-1}(\mathbf{X}), H_{i-1}(\mathbf{Y})) \right)}_{\text{Numero di nuove coppie matchcate a livello } i}$$

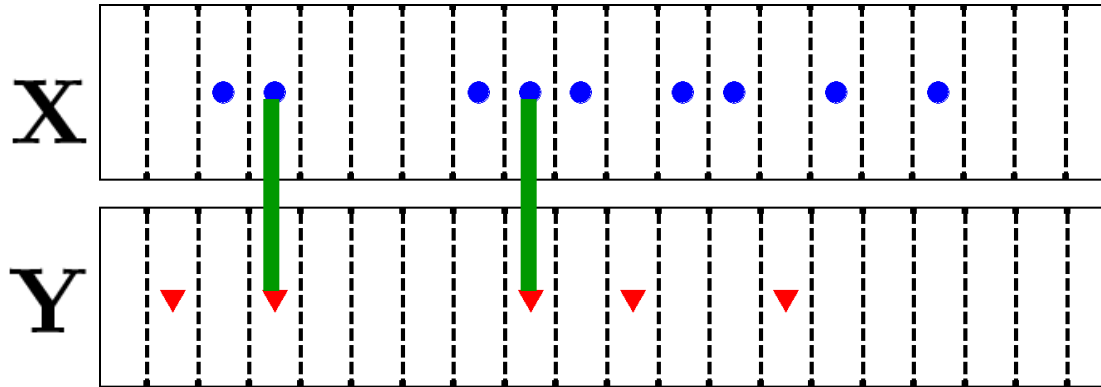
E' un peso che misura la
difficoltà di match a livello i

- I pesi sono inversamente proporzionali al bin size
- Bisogna fare attenzione alla normalizzazione delle features

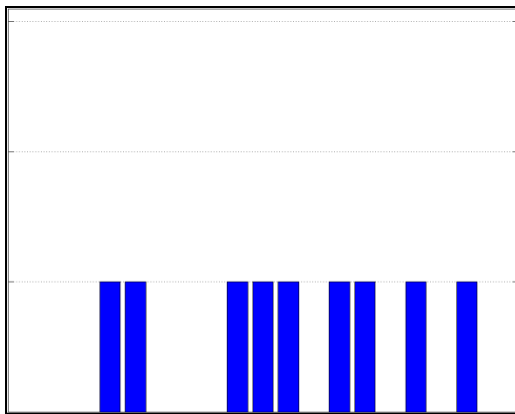


Pyramid match - esempio

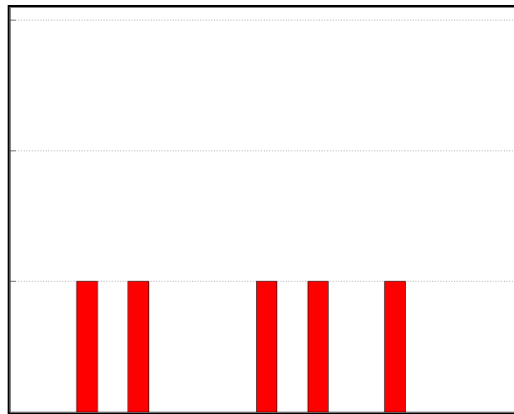
Livello 0



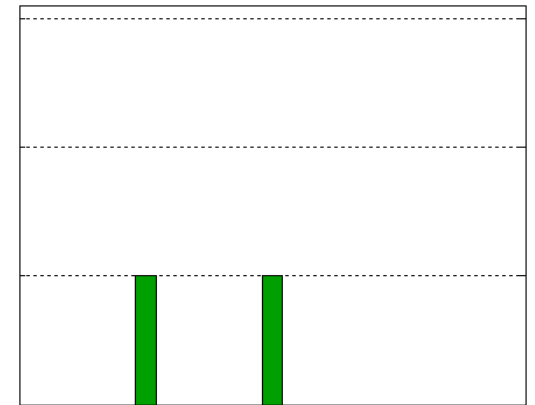
$$\begin{aligned} S_0 &= 2 \\ w_0 &= 1 \end{aligned}$$



$H_0(\mathbf{X})$



$H_0(\mathbf{Y})$

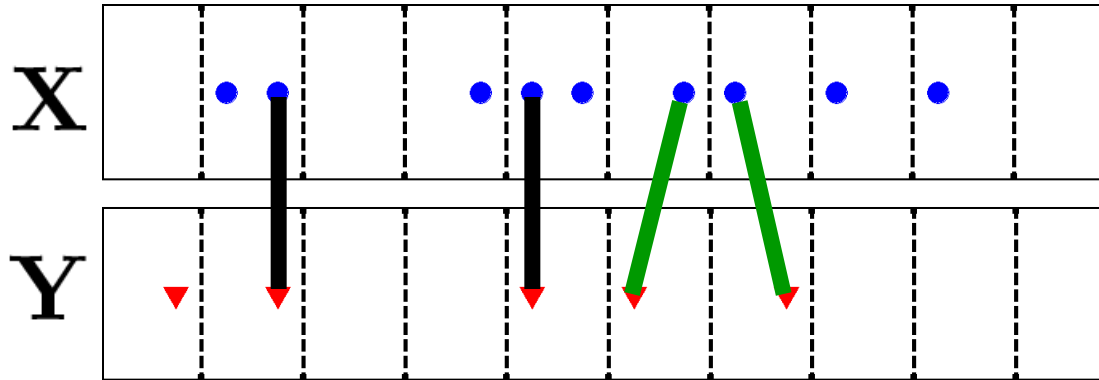


$\mathcal{I}_0 = 2$

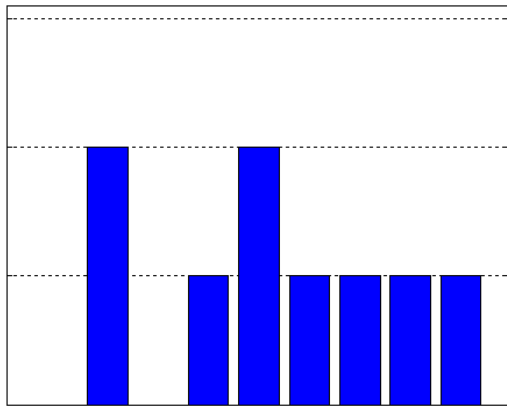


Pyramid match - esempio

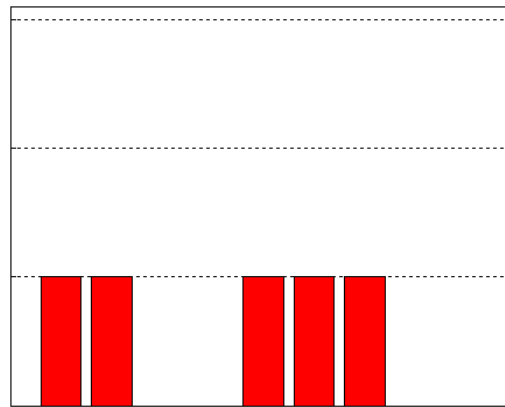
Livello 1



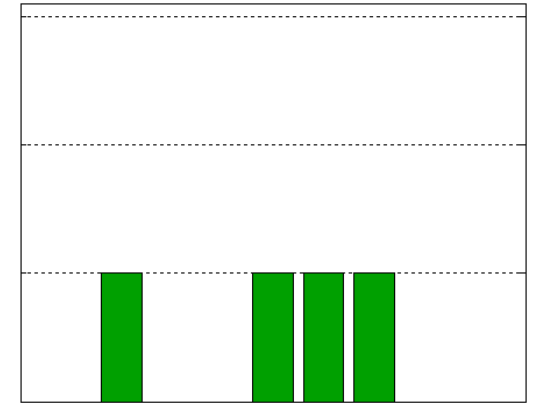
$$\rightarrow \begin{aligned} S_1 &= 4 - 2 = 2 \\ w_1 &= \frac{1}{2} \end{aligned}$$



$H_1(\mathbf{X})$



$H_1(\mathbf{Y})$

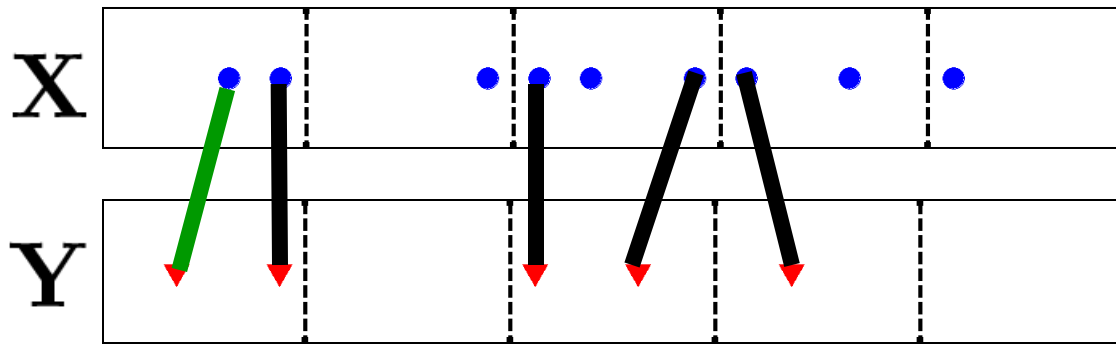


$\mathcal{I}_1 = 4$

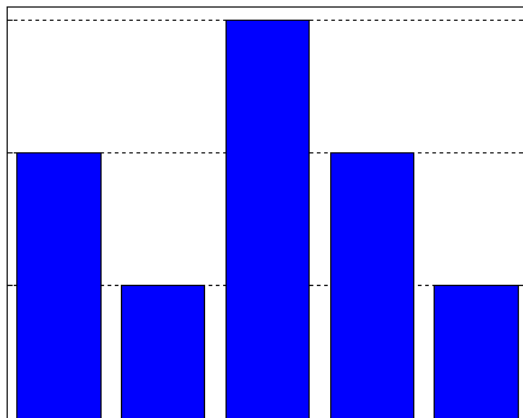


Pyramid match - esempio

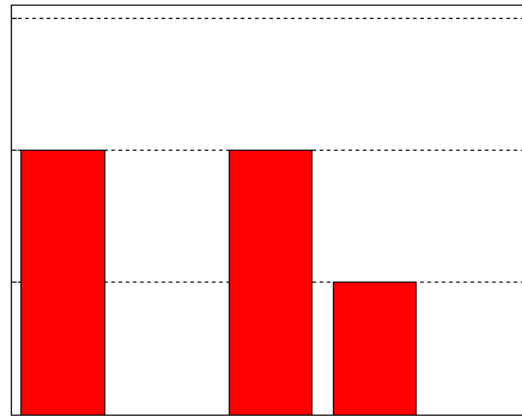
Livello 2



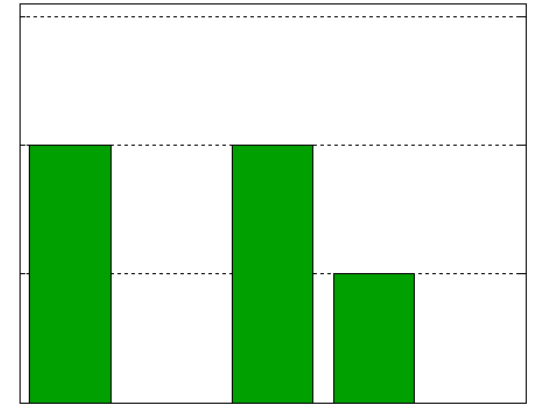
$$\rightarrow \begin{aligned} S_2 &= 5 - 4 = 1 \\ w_2 &= \frac{1}{4} \end{aligned}$$



$H_2(\mathbf{X})$



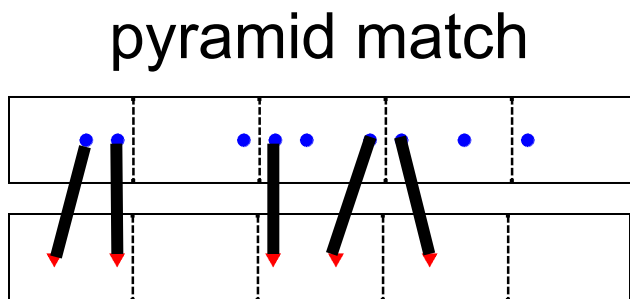
$H_2(\mathbf{Y})$



$\mathcal{I}_2 = 5$



Pyramid match - esempio



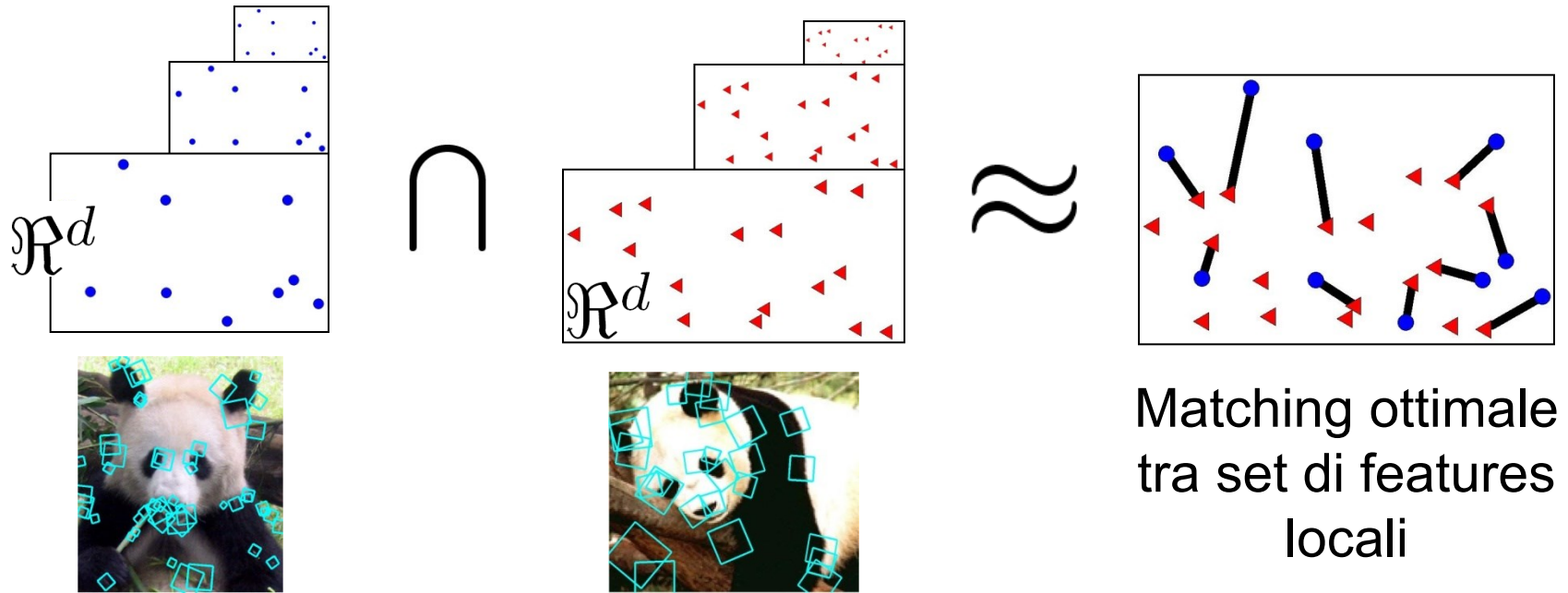
$$K_{\Delta} = \sum_{i=0}^L w_i S_i$$

$$= 1(2) + \frac{1}{2}(2) + \frac{1}{4}(1) = 3.25$$

- La definizione di questo Kernel è legittima, ossia ho una funzione semidefinita positiva



Riassunto: Pyramid match kernel



$$K_{\Delta}(\Psi(\mathbf{X}), \Psi(\mathbf{Y})) = \sum_{i=0}^L \underbrace{\frac{1}{2^i}}_{\text{Difficoltà del match a livello } i} \underbrace{\left(\mathcal{I}(H_i(\mathbf{X}), H_i(\mathbf{Y})) - \mathcal{I}(H_{i-1}(\mathbf{X}), H_{i-1}(\mathbf{Y})) \right)}_{\text{Numero di nuovi match a livello } i}$$

Difficoltà del match a livello i

Numero di nuovi match a livello i



Risultati di riconoscimento

- ETH-80 database
8 classi di oggetti
(*Eichhorn and Chapelle 2004*)
- Features:
 - Harris detector
 - PCA-SIFT descriptor, $d=10$



Kernel	Complessità	Accuratezza
Match [<i>Wallraven et al.</i>]	$O(dm^2)$	84%
Bhattacharyya affinity [<i>Kondor & Jebara</i>]	$O(dm^3)$	85%



Risultati di riconoscimento

- Caltech objects database
101 object classes
- Features:
 - SIFT detector
 - PCA-SIFT descriptor, $d=10$
- 30 training images / class
- 43% recognition rate
- 0.002 seconds per match

