Vedere Artificială
(Computer Vision)

Bogdan Alexe

bogdan.alexe@fmi.unibuc.ro

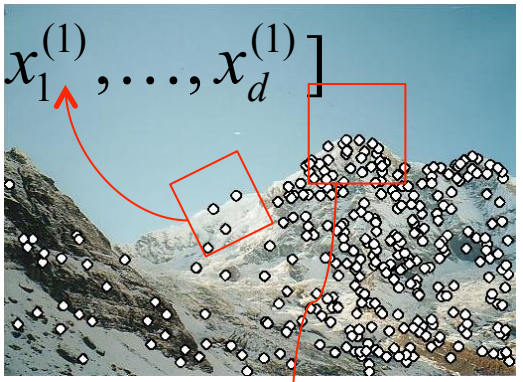
anul 2, master Informatică, semestrul I, 2019-2020

Recapitulare – cursul trecut

1) **Detectare**: detectori de puncte de interes invarianți la scală: detectorul DoG, Harris-Laplace

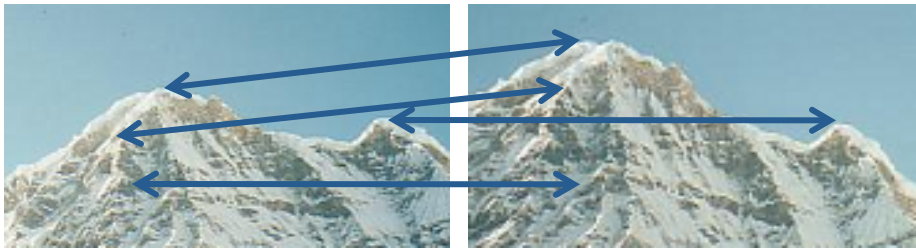
2) **Descriere**: descrie conținutul vizual din vecinătatea fiecărui punct de interes printr-un vector printr-un descriptor vizual (feature vector) - SIFT

$$\mathbf{x}_1 = [x_1^{(1)}, \ldots, x_d^{(1)}]$$



$$\mathbf{x}_2 = [x_1^{(2)}, \ldots, x_d^{(2)}]$$

3) **Matching**: determină corespondențele dintre descriptorii dintre imagini



Comparație între detectoare de puncte de interes

Table 7.1 Overview of feature detectors.

Feature Detector	Corner	Blob	Region	Rotation invariant	Scale invariant	Affine invariant	Localization Repeatability	Localization accuracy	Robustness	Efficiency
Harris	√			√			+++	+++	+++	++
Hessian		√		√			++	++	++	+
SUSAN	√			√			++	++	++	+++
Harris-Laplace	√	(√)		√	√		+++	+++	++	+
Hessian-Laplace	(√)	√		√	√		+++	+++	+++	+
DoG	(√)	√		√	√		++	++	++	++
SURF	(√)	√		√	√		++	++	++	+++
Harris-Affine	√	(√)		√	√	√	+++	+++	++	++
Hessian-Affine	(√)	√		√	√	√	+++	+++	+++	++
Salient Regions	(√)	√		√	√	(√)	+	+	++	+
Edge-based	√			√	√	√	+++	+++	+	+
MSER			√	√	√	√	+++	+++	++	+++
Intensity-based			√	√	√	√	++	++	++	++
Superpixels			√	√	(√)	(√)	+	+	+	+

Tuytelaars Mikolajczyk 2008

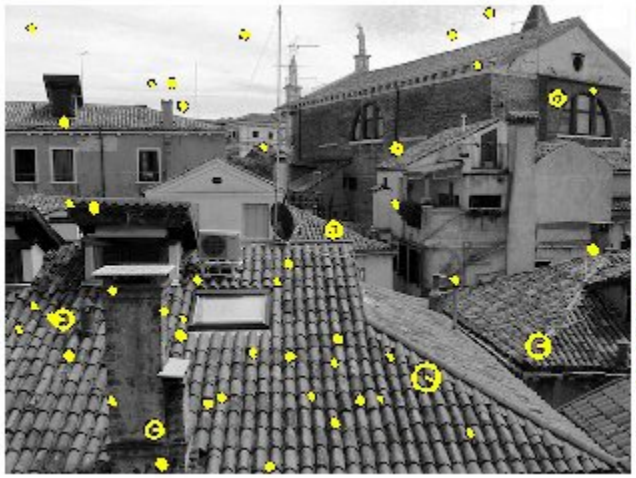
Descriptorul SIFT [Lowe 2004]

1. Rulează detectorul DoG
   - găsește maximele în poziție/scală
   - elimină punctele cu contrast mic + punctele de pe muchii

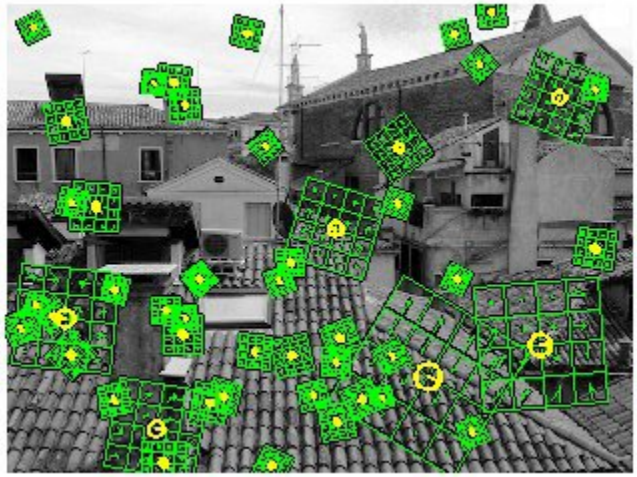
2. Găsește principalele orientări
   - grupează orientările pixelilor într-o histogramă cu 36 de intervale
     - pondere dată de magnitudinea gradientului + distanța către centrul intervalului
   - returnează orientările care se încadrează în 80% din valoarea maximă
     - în jur de 15% din puncte sunt dublate cu orientări multiple

3. Pentru fiecare (x,y,scală,orientare), calculează descriptorul:
   - împarte regiunea de 16x16 pixeli în 4x4 blocuri = 16 blocuri
   - calculează pentru fiecare bloc o histogramă de orientări ale pixelilor (intervalul $0\text{-}360^0$ împărțit în 8 intervale)
   - valoare maximă a gradienților la 0.2 și apoi re-normalizare
   - descriptorul SIFT = 16 blocuri = 16 histograme x 8 valori = dimensiune 128

Descriptorul SIFT [Lowe 2004]



Puncte de interes cu scalele și orientările asociate (selecție aleatoare de 50 de puncte)



Descriptori SIFT

http://www.vlfeat.org/overview/sift.html

Corespondențe ambigue

Imagine 1

Imagine 2

Pentru ce distanțe avem o corespondență bună?

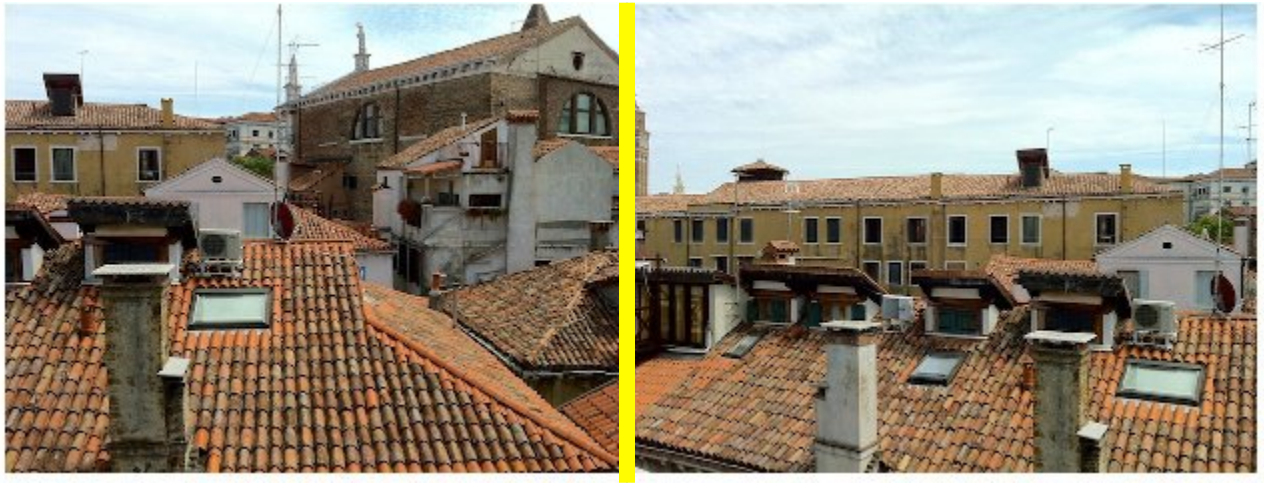
Pentru găsirea de corespondențe robuste, considerăm raportul:
distanța față de first NN / distanța față de 2nd NN

Dacă raportul e mic, totul pare ok (nu există ambiguități)

Dacă raportul e mare, nu pare ok (pot exista ambiguități)

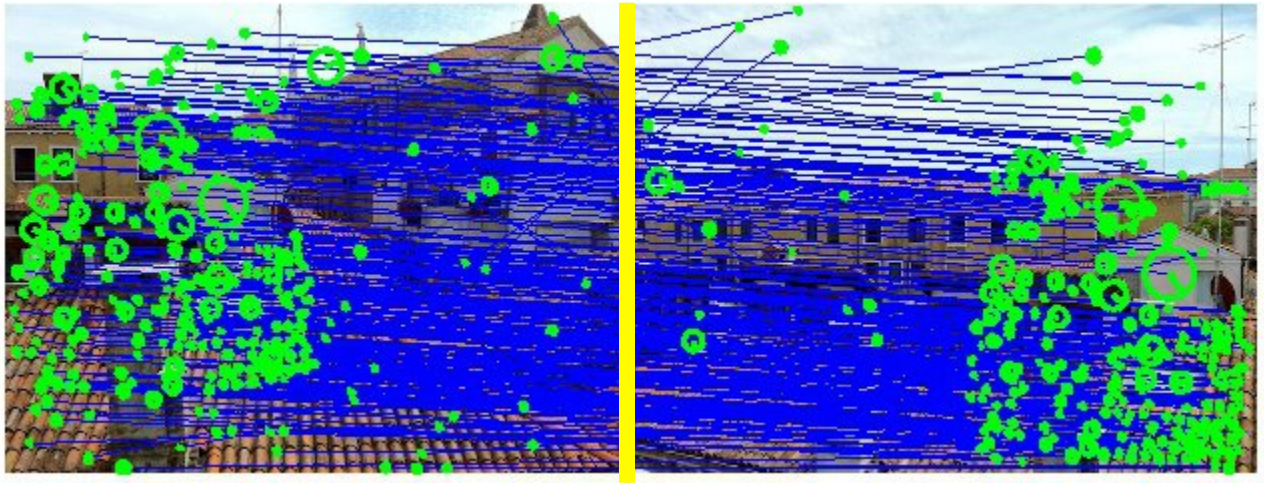
Corespondențe pe baza descriptorilor SIFT

img1



img2

img1

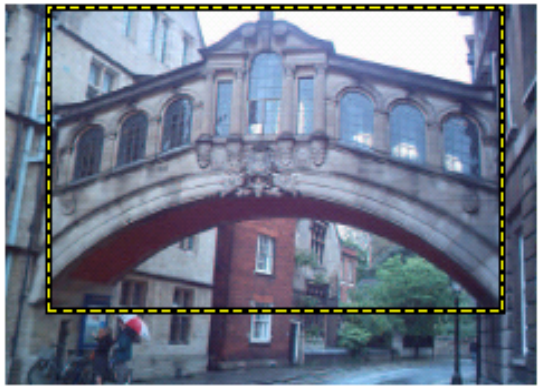
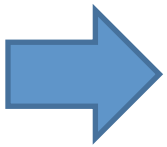
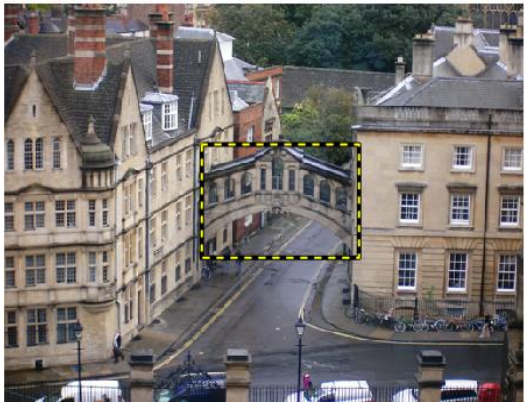
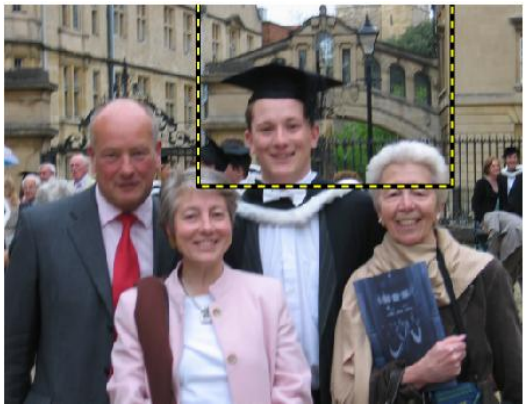


img2

http://www.vlfeat.org/overview/sift.html

Planul cursului de azi

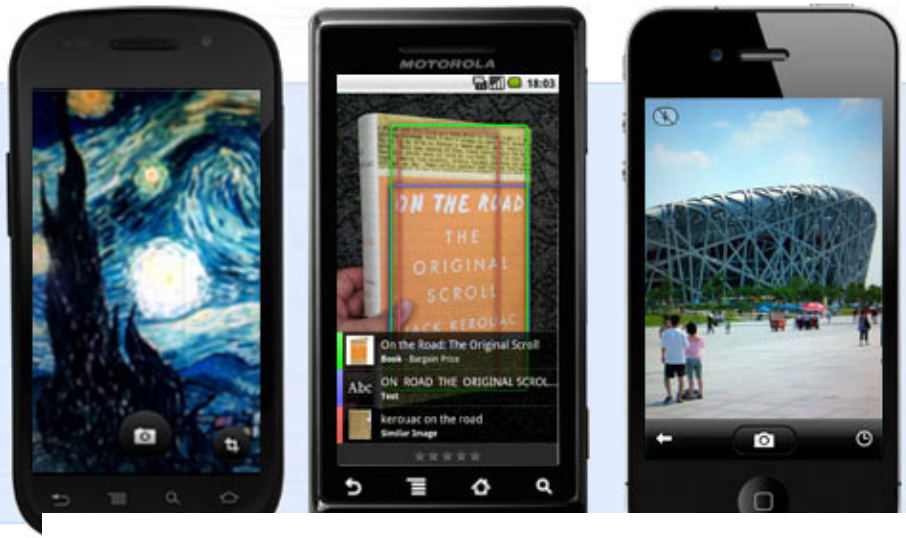
- Recunoașterea specifică de obiecte
  - indexarea eficientă a trăsăturilor locale
  - modelul Bag of Visual Words
  - verificare spațială

Google Goggles

Use pictures to search the web. ▷ Watch a video

https://www.youtube.com/watch?v=Hhgfz0zPmH4



Get Google Goggles

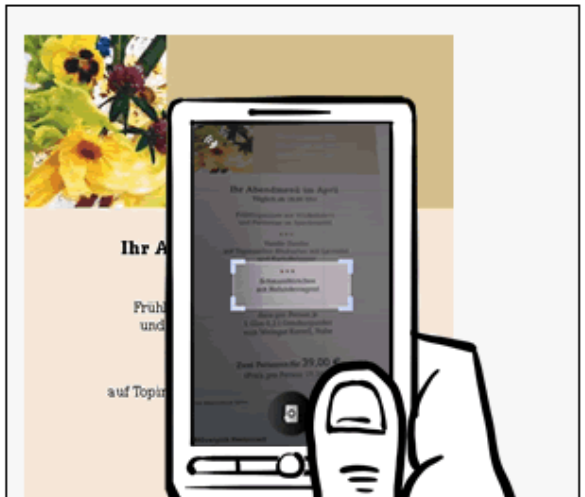
Android (1.6+ required)
Download from Android Market.

Send Goggles to Android phone

New! **iPhone (iOS 4.0 required)**
Download from the App Store.

Send Goggles to iPhone

New!
Text | Landmarks | Books | Contact Info | Artwork | Wine | Logos



Google Goggles (2009)

Google Lens (2018)

Recunoașterea sau regăsirea obiectelor specifice

Examplu I: căutare vizuală în video-uri

Interogare definită vizual

"Găsește acest ceas"

"Găsește acest loc"

"Groundhog Day" [Rammis, 1993]

Sivic & Zisserman, 2003

Video Google: A Text Retrieval Approach to Object Matching in Videos

Josef Sivic and Andrew Zisserman
Robotics Research Group, Department of Engineering Science
University of Oxford, United Kingdom

Abstract

We describe an approach to object and scene retrieval which searches for and localizes all the occurrences of a user outlined object in a video. The object is represented by a set of viewpoint invariant region descriptors so that recognition can proceed successfully despite changes in viewpoint, illumination and partial occlusion. The temporal continuity of the video within a shot is used to track the regions in order to reject unstable regions and reduce the effects of noise in the descriptors.

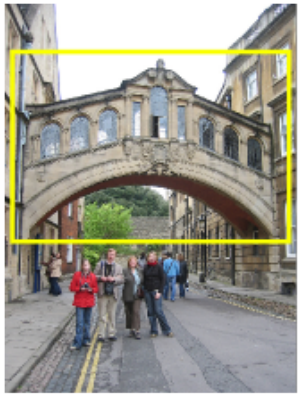
The analogy with text retrieval is in the implementation where matches on descriptors are pre-computed (using vector quantization), and inverted file systems and document rankings are used. The result is that retrieval is immediate, returning a ranked list of key frames/shots in the manner of Google.

The method is illustrated for matching on two full length feature films.

Sivic & Zisserman, 2003

Recunoașterea sau regăsirea obiectelor specifice

Examplu II: caută imagini pe web după locuri particulare



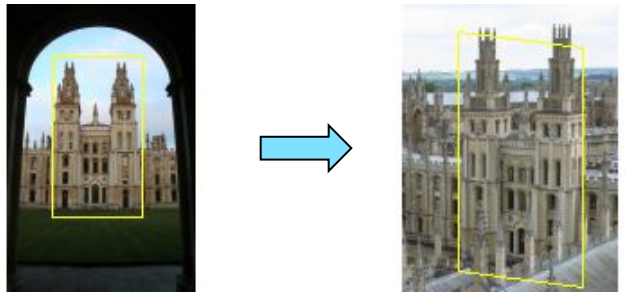
Găsește aceste locuri (ladmarks)

...într-o bază de date cu imagini mare (1M)

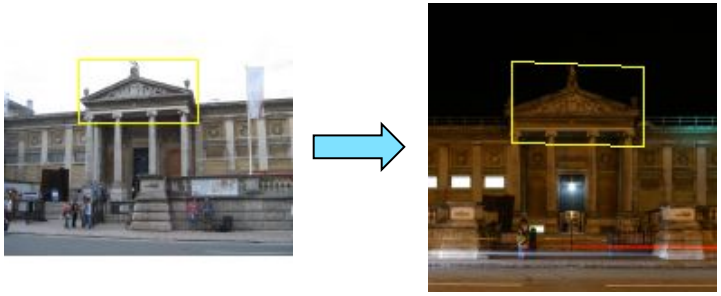
Dificultate

Vrem să găsim obiectul în ciuda unor schimbări în mărime, unghi, iluminare și ocluzii parțiale

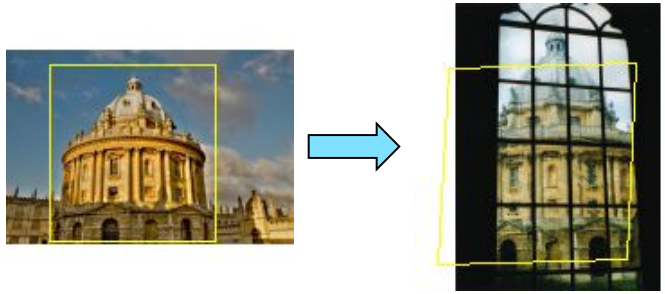
Scale



Viewpoint

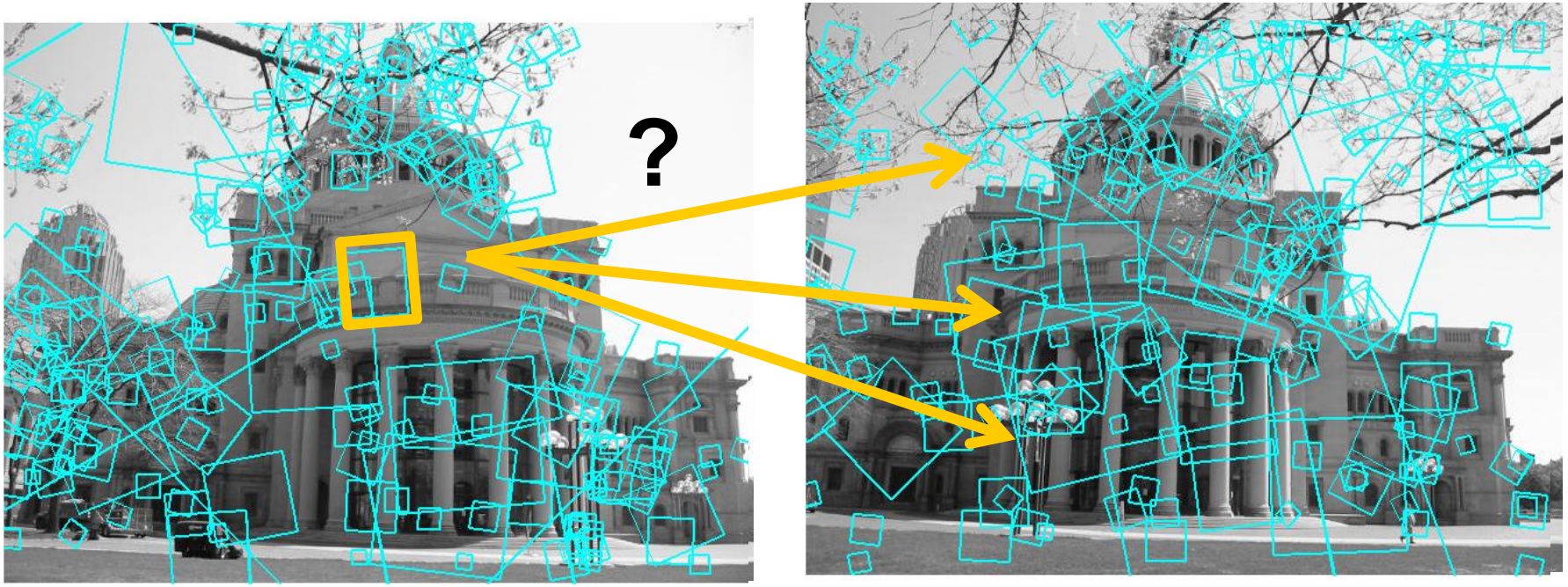


Lighting



Occlusion

Cursul trecut: Corespondențe pentru trăsături locale



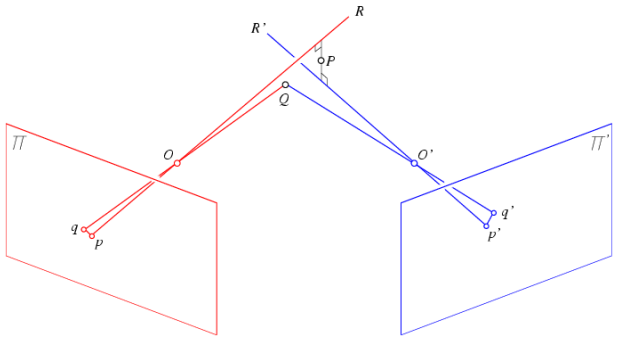
Imagine 1

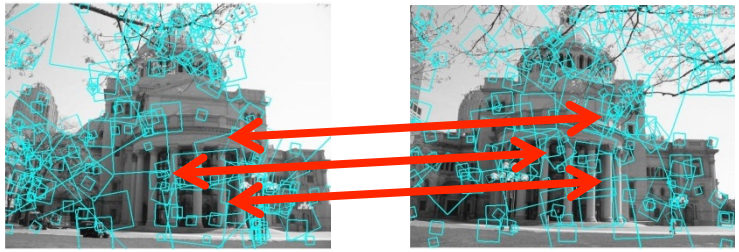
Imagine 2

Pentru găsirea de trăsături locale corespondente, găsește patch-uri similare în înfățișare – calculează o distanță

Abordare simplă: k nearest neighbors, cu o distanță maximă admisă

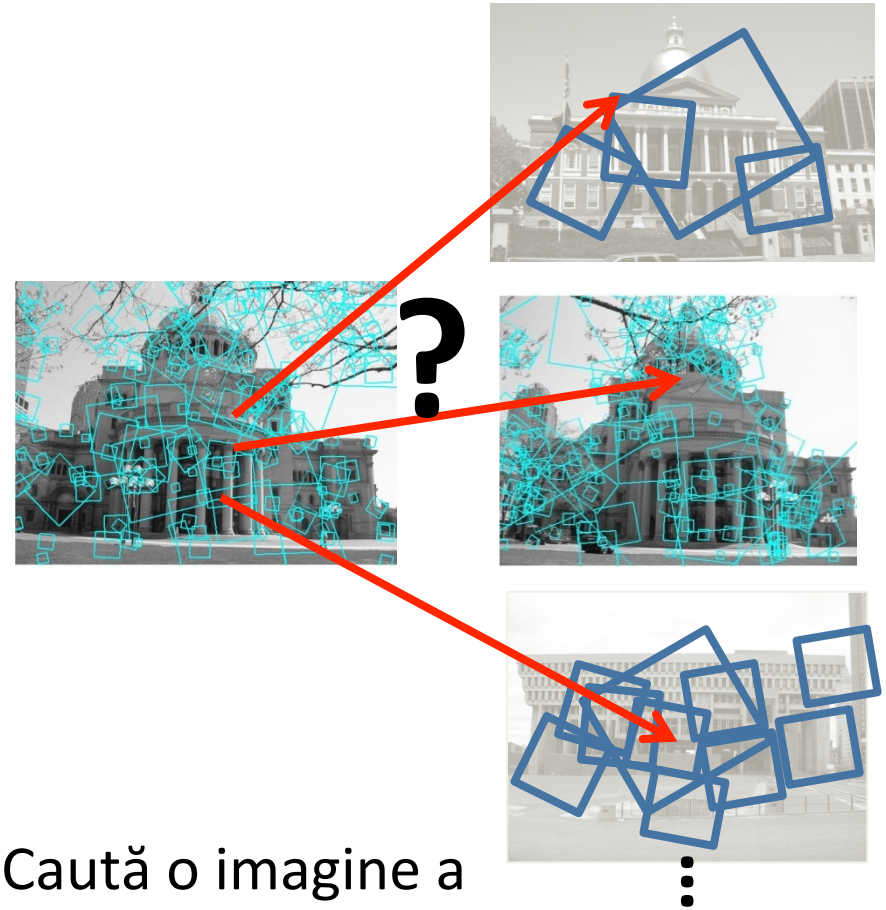
Corespondențe între imagini





Corespondențe din mai multe imagini ale aceleași scene

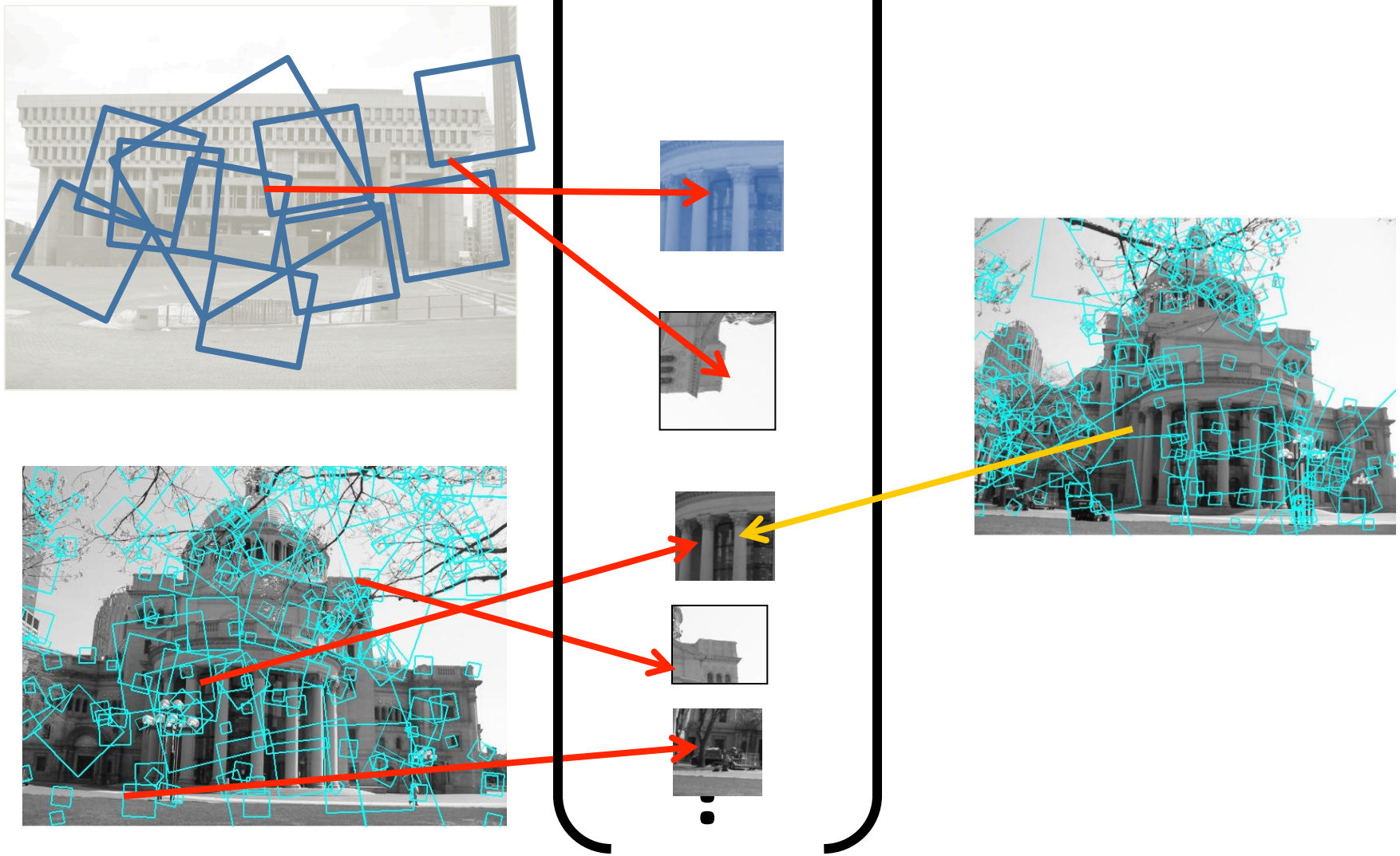
vs



Caută o imagine a aceleiași scene

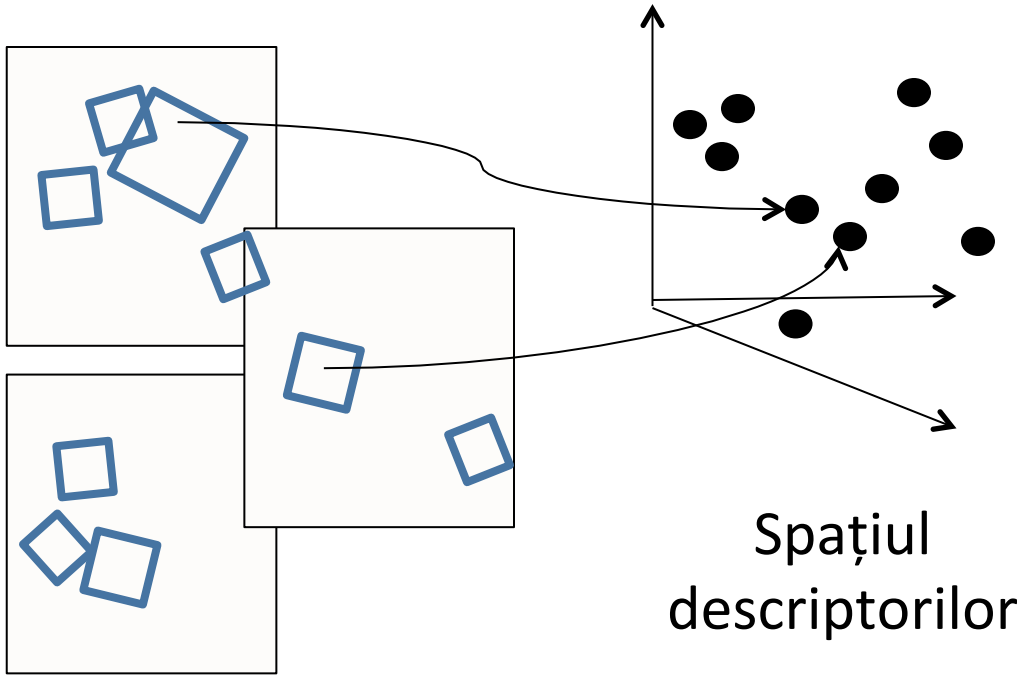
Găsirea de corespondențe între trăsături locale invariante este folositoare și pentru recunoașterea de obiecte/scene.

Indexarea trăsăturilor locale



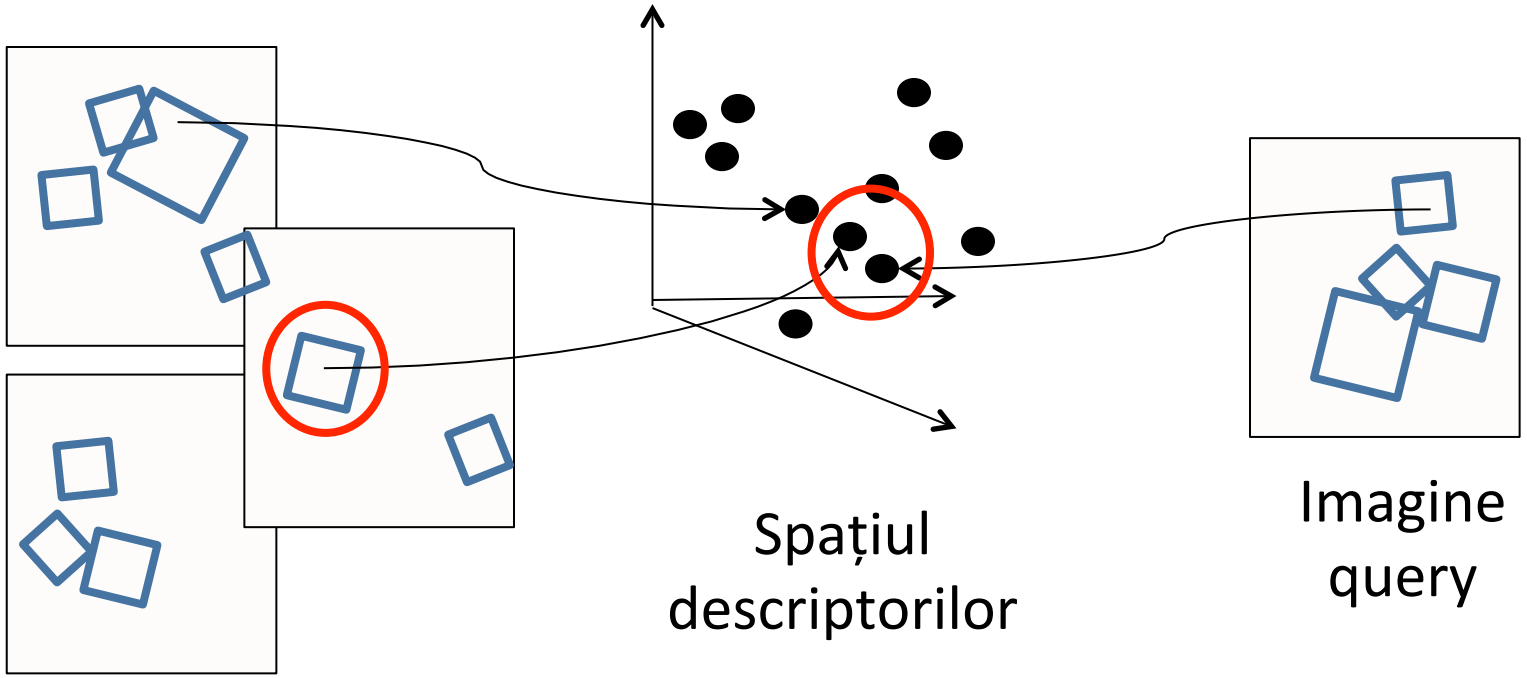
Indexarea trăsăturilor locale

- Fiecare patch/regiune este caracterizată de un descriptor, care este un punct într-un spațiu de dimensionalitate mare (spre exemplu, SIFT – are 128 dimensiuni)



Indexarea trăsăturilor locale

- Puncte apropiate în spațiul descriptorilor indică conținut vizual local similar

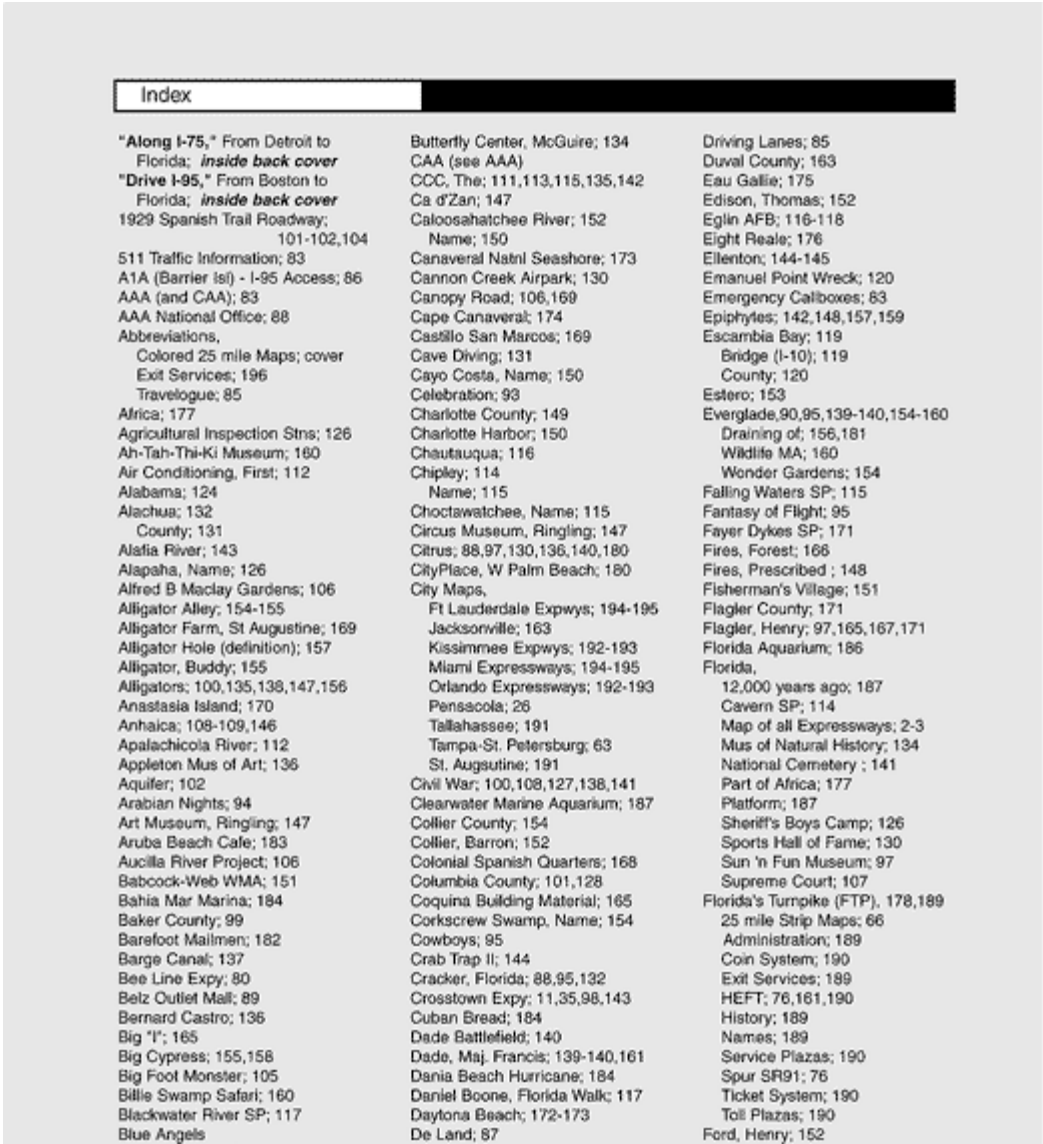


Indexarea eficientă a trăsăturilor locale

- este posibil să avem în jur de câteva mii de caracteristici (descriptori SIFT) într-o imagine și câteva sute de mii/ milioane de imagini în baza de date
- cum să găsim eficient imaginile apropiate în conținut vizual de imaginea query?

- Soluții posibile:
  - Inverted file index
  - Structuri de date pe bază de cei mai apropiați vecini
    - Kd-trees
    - Hashing

Indexarea trăsăturilor locale: inverted file index

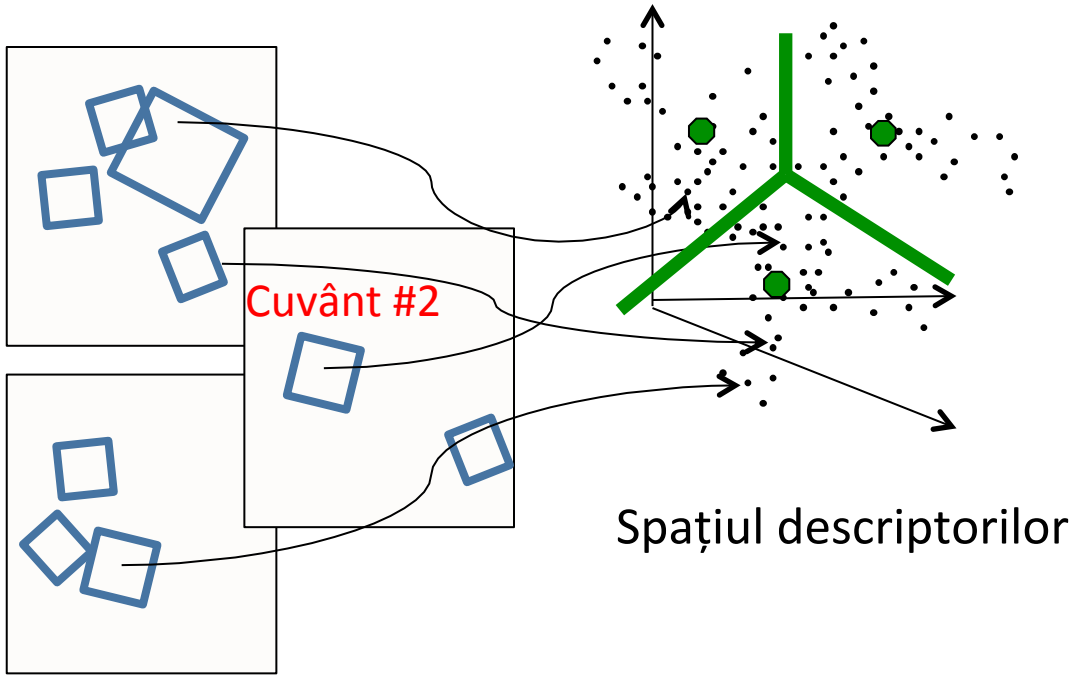


Index

- Pentru documente text, o metodă eficientă pentru a găsi toate *paginile* în care apare un *cuvânt* este de a folosi un index…
- Vrem să găsim toate *imaginile* în care apare o *trăsătură locală*.
- Pentru a pune în practică o asemenea idee, mapăm trăsăturile locale în “cuvinte vizuale”.

Cuvinte vizuale = visual words

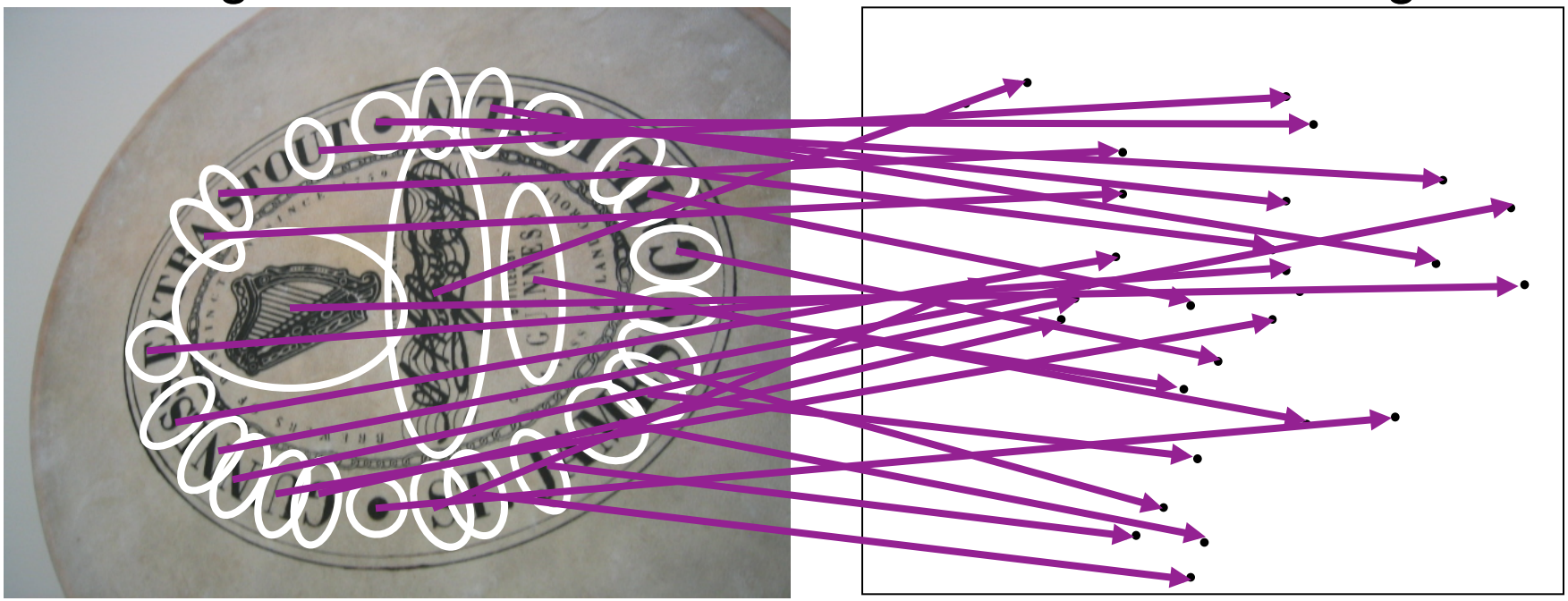
- Mapăm descriptorii în token-uri (cuvinte vizule) partiționând spațiul descriptorilor



- Partiționăm (cuantizăm) spațiul descriptorilor prin clusterizare, centri clusterilor sunt "cuvinte vizuale"
- Determinăm ce cuvânt asignăm unei regiuni noi găsind cel mai apropiat cluster

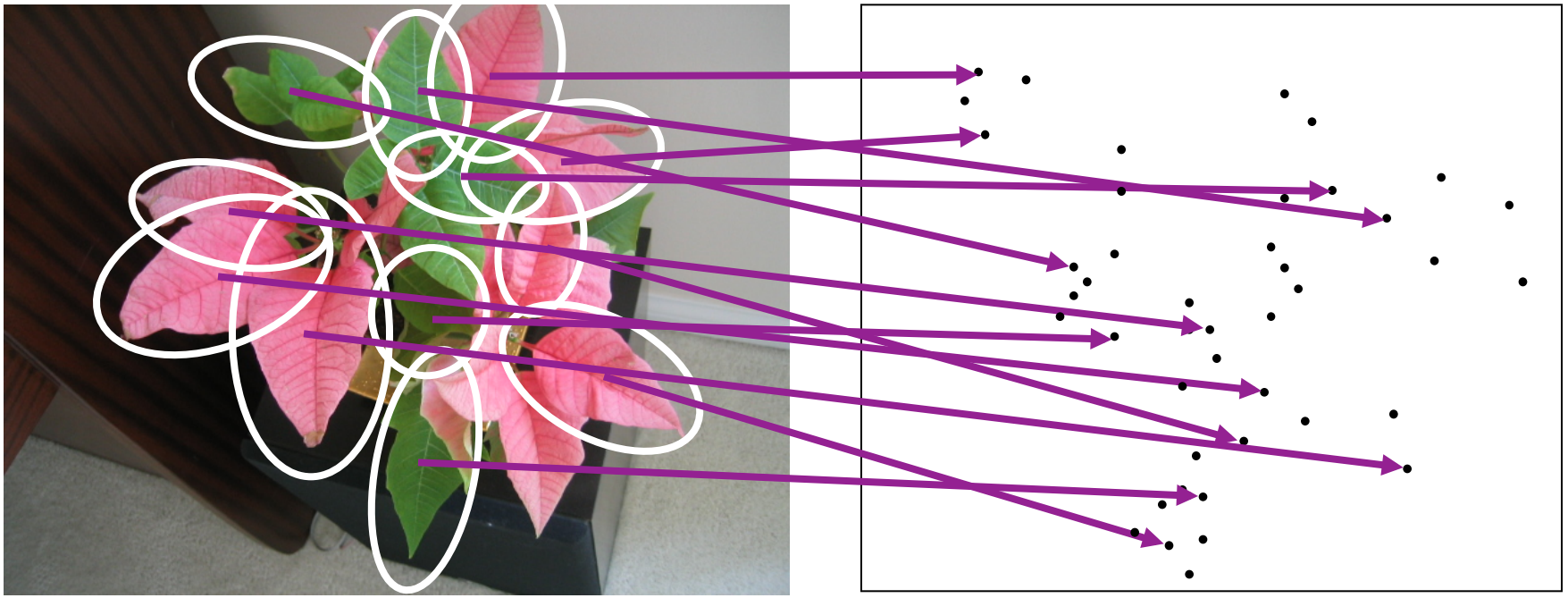
Cuvinte vizuale: ideea de bază

- Extragem caractersitici locale dintrun număr de imagini…

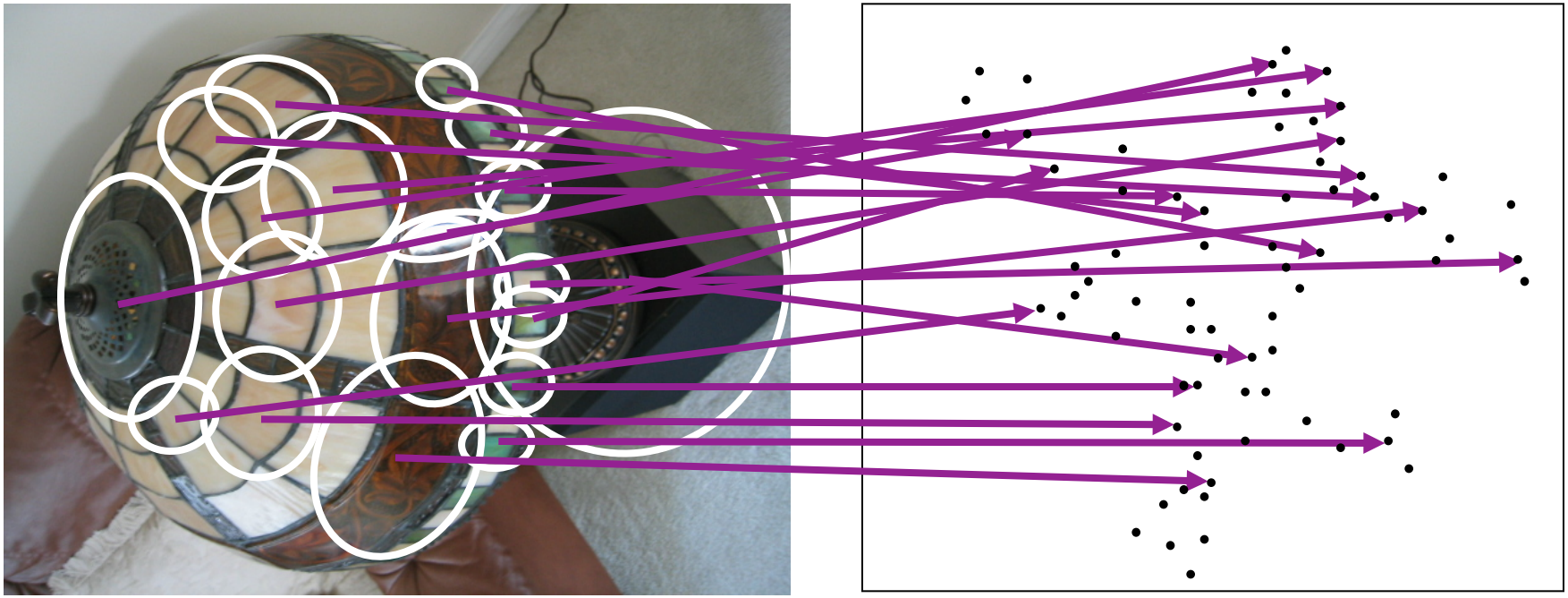


Spre exemplu, descriptori SIFT de dimensiune 128

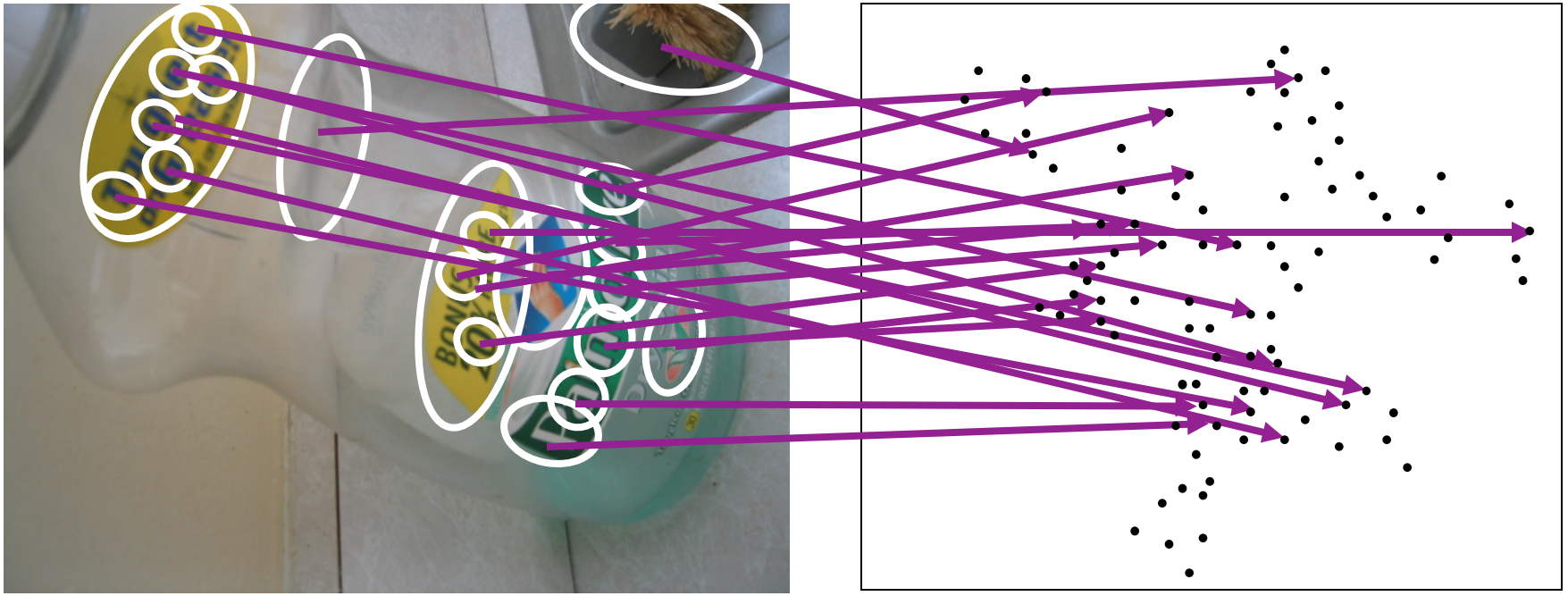
Cuvinte vizuale: ideea de bază

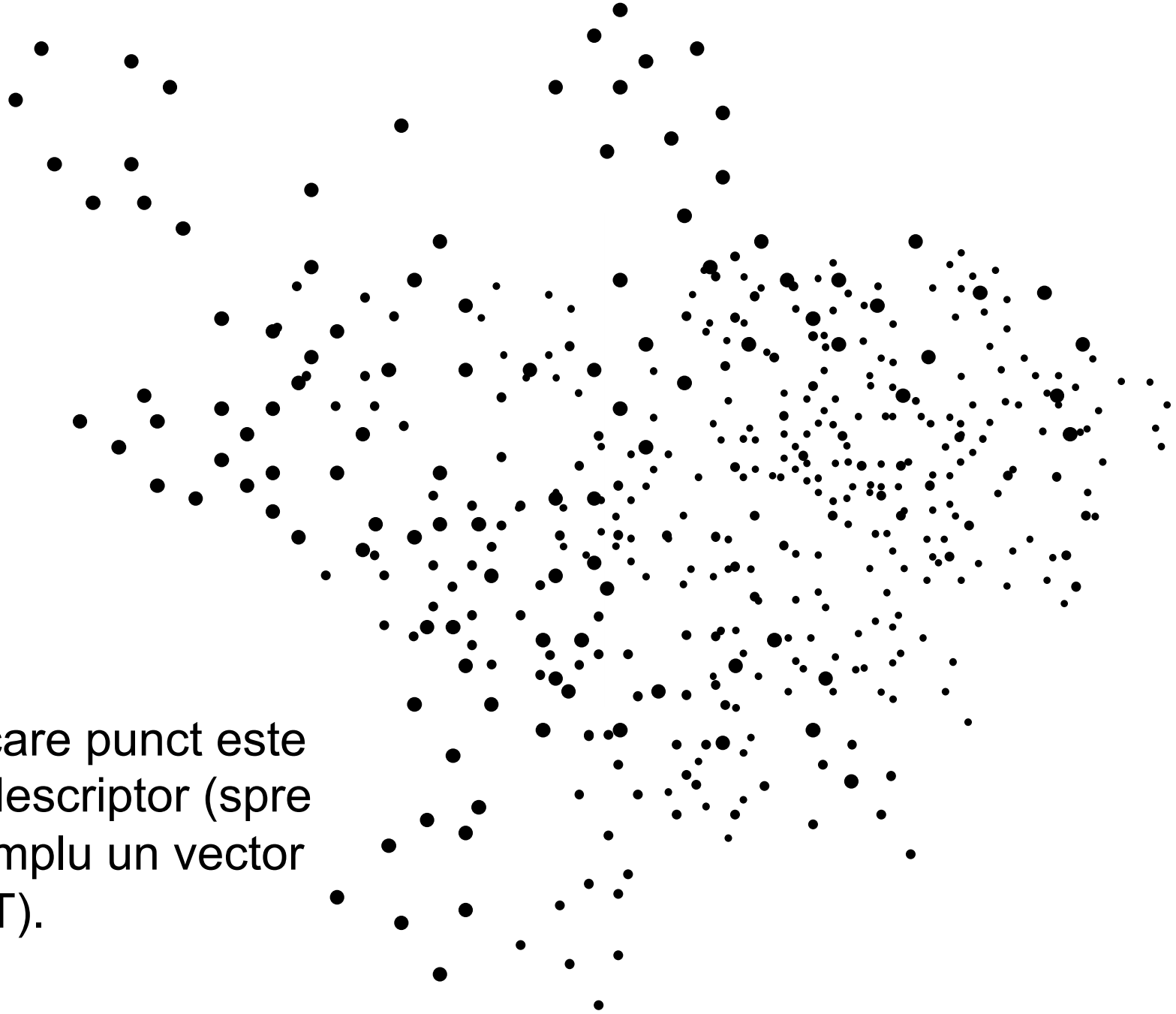


Cuvinte vizuale: ideea de bază

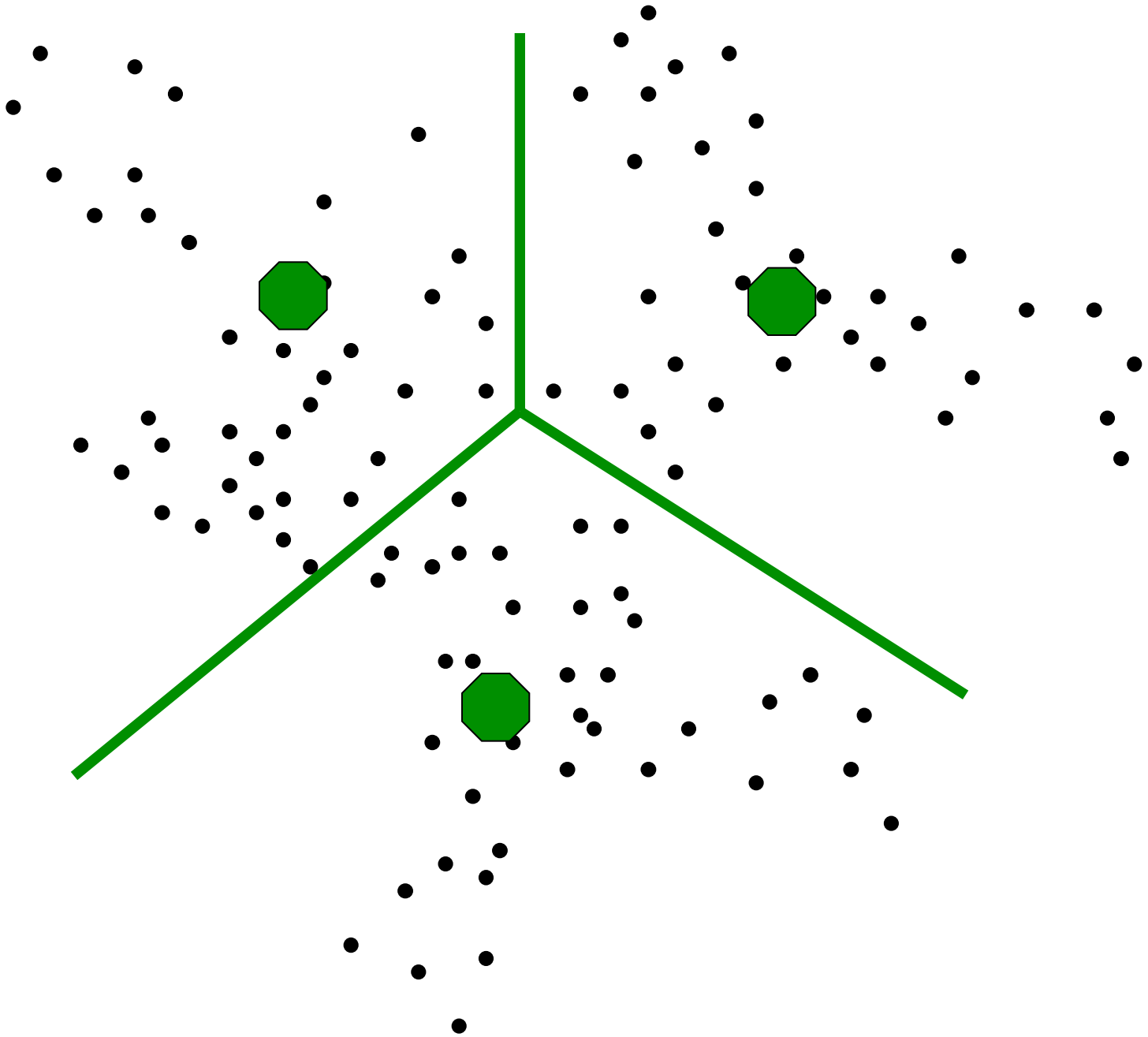


Cuvinte vizuale: ideea de bază



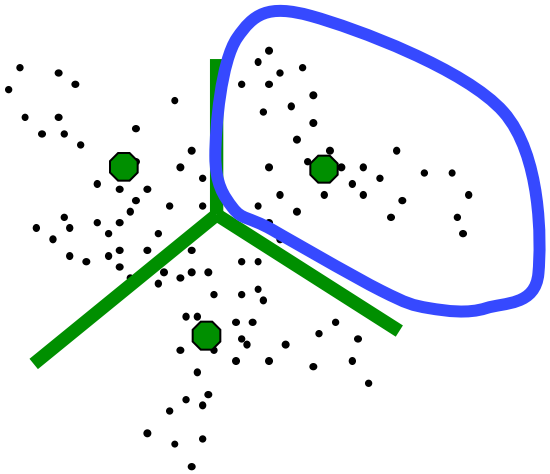


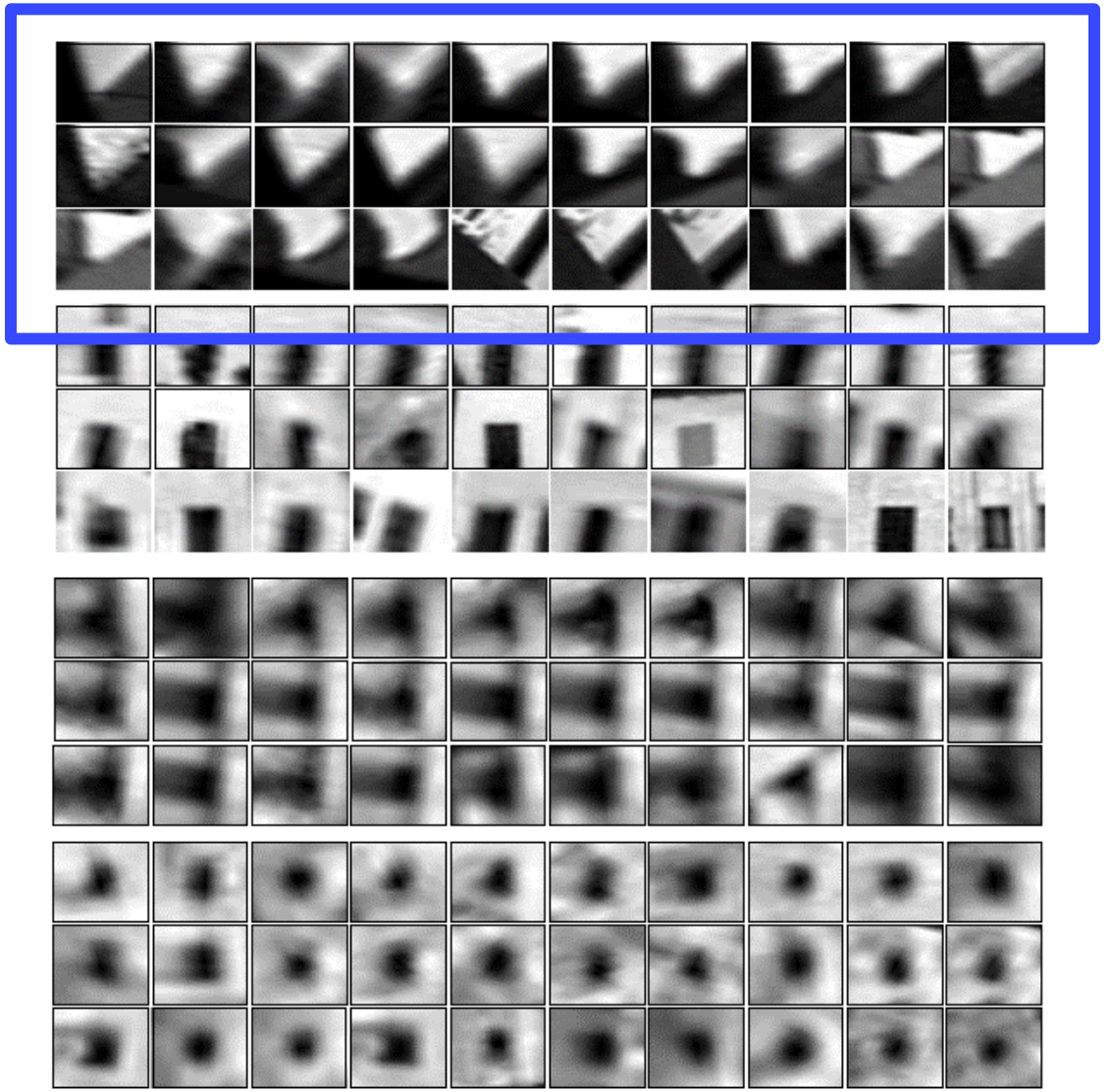
Fiecare punct este un descriptor (spre exemplu un vector SIFT).



Cuvinte vizuale

- Exemple: fiecare grup de patch-uri aparține aceluiași cuvânt vizual

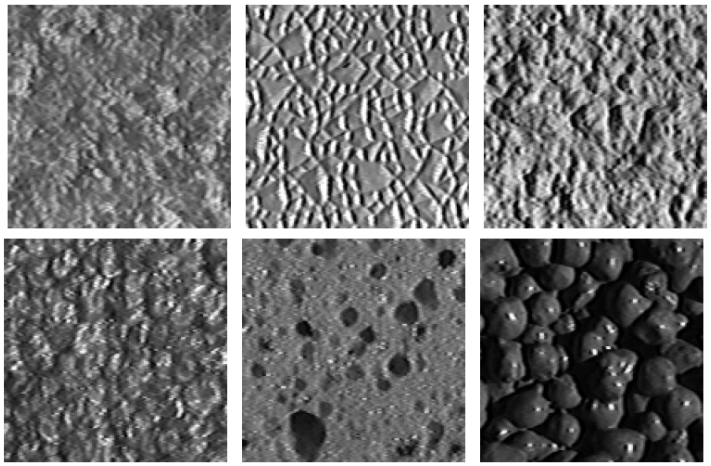


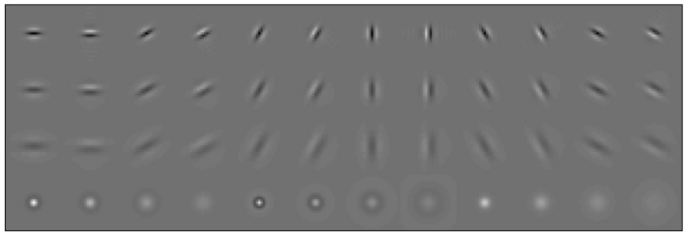


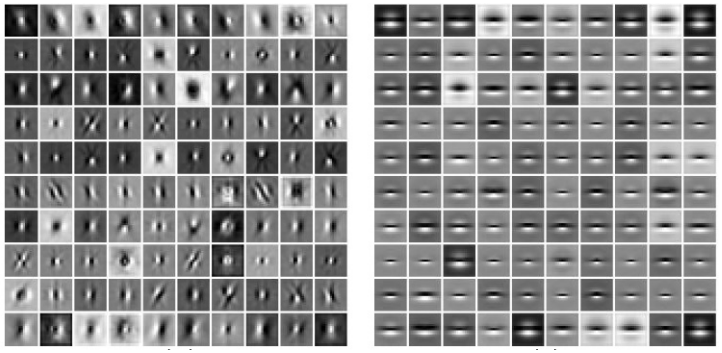
Figură din articolul Sivic & Zisserman, ICCV 2003

Cuvinte vizuale și textoni

- Cuvintele vizuale au fost explorate întâi din perspectiva reprezentării texturii și a materialelor
- *Texton* = centrul unui cluster de răspunsuri obținute pe baza filtrării unor imagini
- Descrie textura și materialele pe baza distribuției textonilor



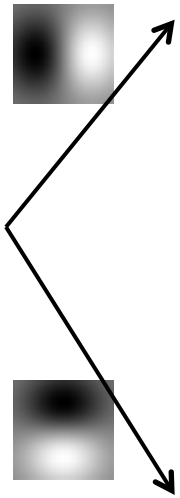


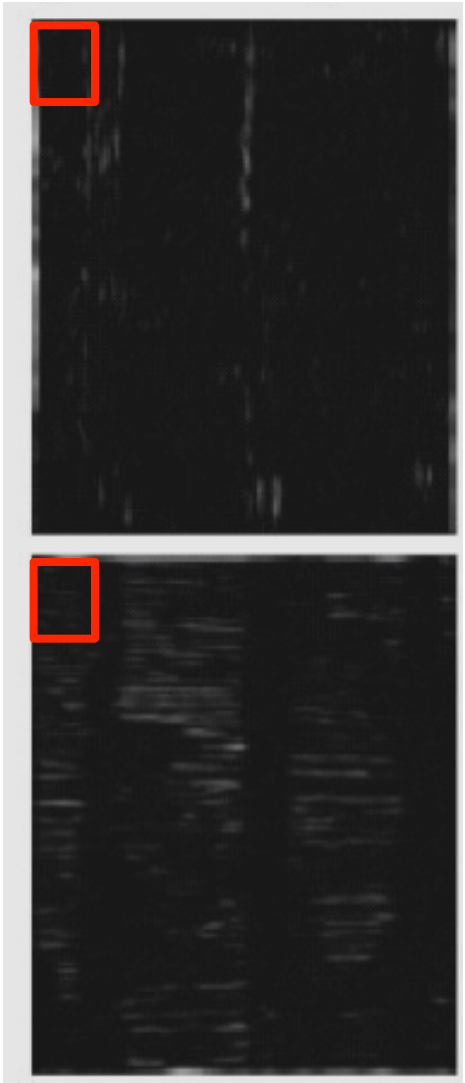


Leung & Malik 1999; Varma & Zisserman, 2002

Reprezentarea texturii: exemplu

imagine inițială





imagini filtrate (filtre pentru derivate)

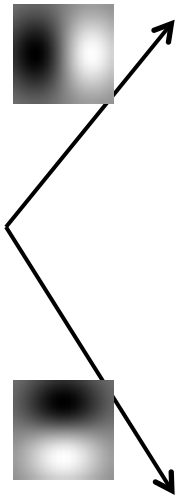
	media d/dx	media d/dy
fereastra #1	4	10

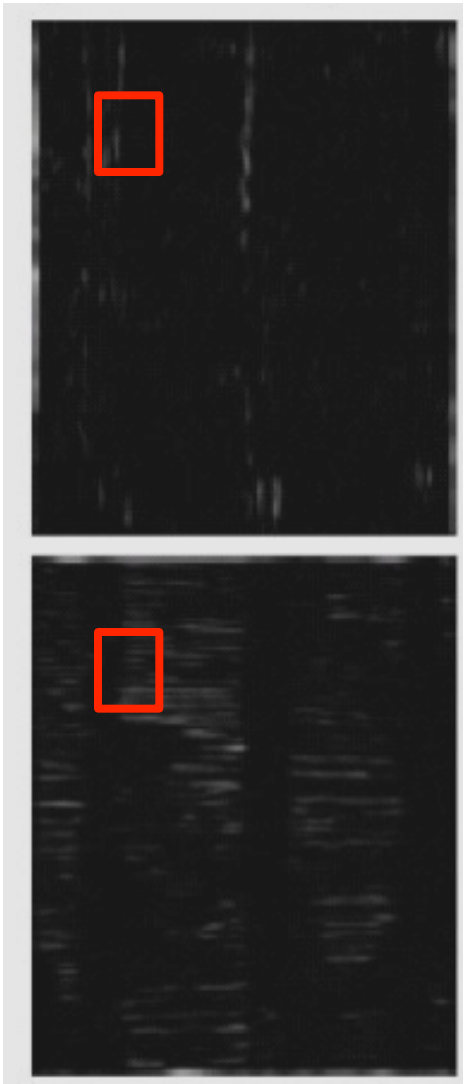
⋮

Statistici pentru a descrie pattern-urile în ferestre mici

Reprezentarea texturii: exemplu

imagine inițială





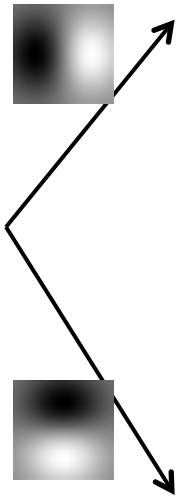
imagini filtrate (filtre pentru derivate)

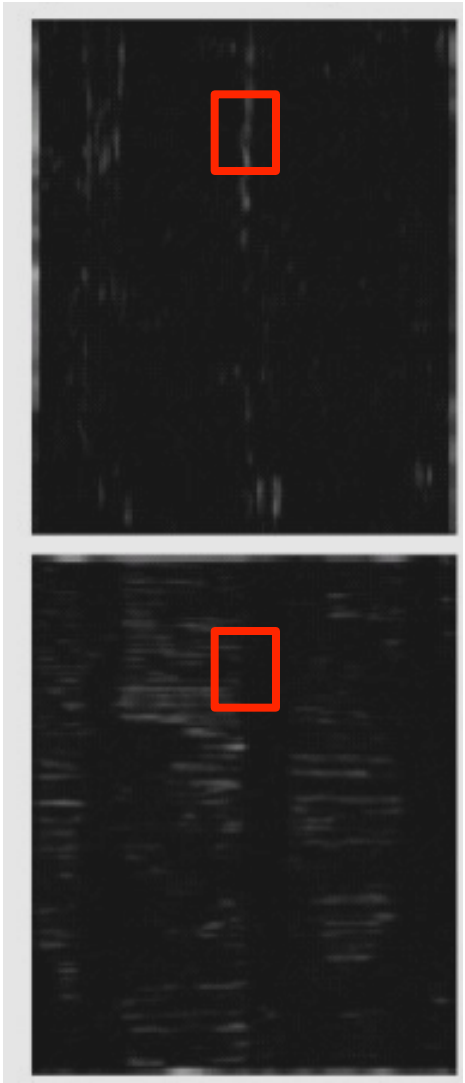
	media d/dx	media d/dy
fereastra #1	4	10
fereastra #2	25	7		

Statistici pentru a descrie pattern-urile în ferestre mici

Reprezentarea texturii: exemplu

imagine inițială





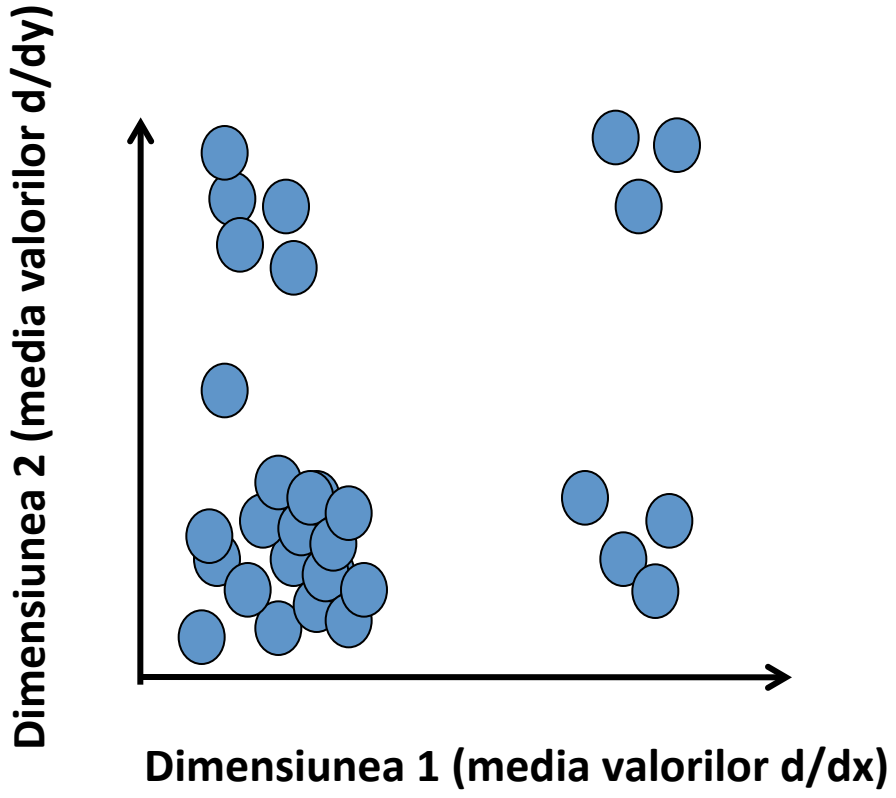
imagini filtrate (filtre pentru derivate)

	media d/dx	media d/dy
fereastra #1	4	10
fereastra #2	25	7
fereastra #3	18	20

⋮

Statistici pentru a descrie pattern-urile în ferestre mici

Reprezentarea texturii: exemplu

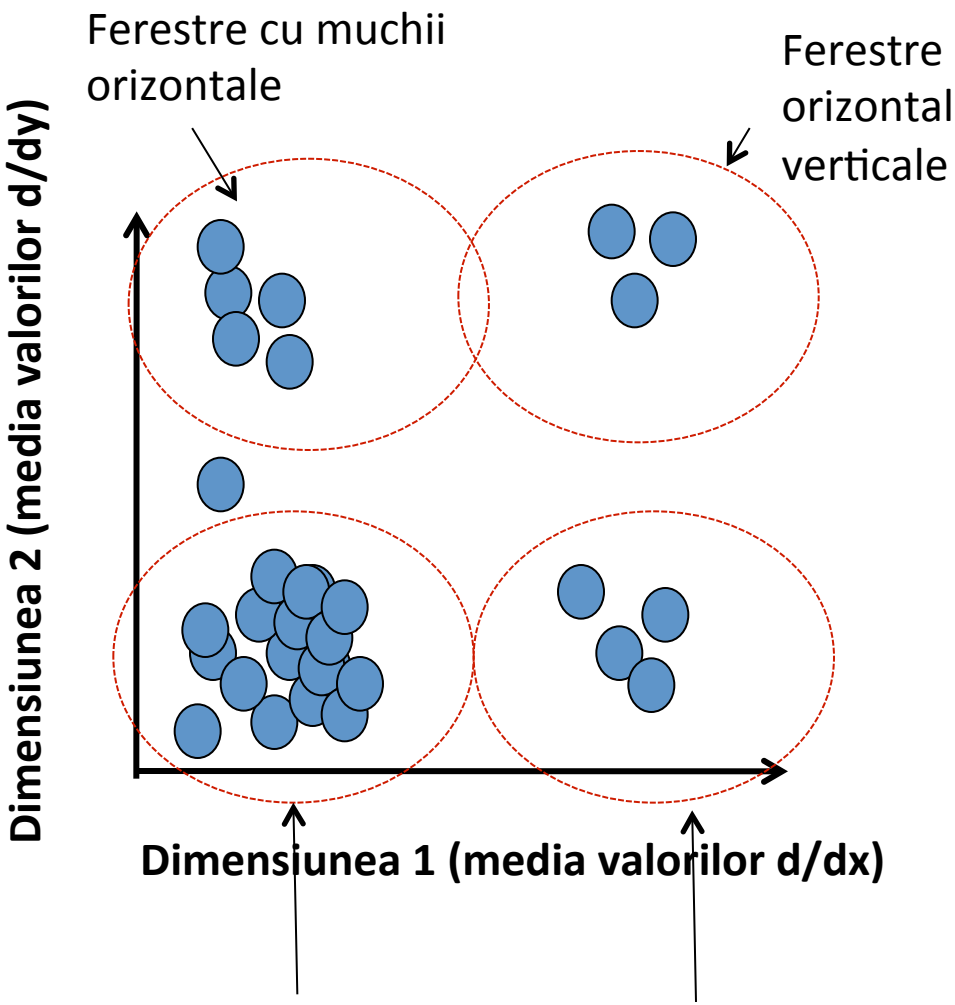


	media d/dx	media d/dy
fereastra #1	4	10
fereastra #2	25	7
fereastra #3	18	20

⋮

Statistici pentru a descrie pattern-urile în ferestre mici

Reprezentarea texturii: exemplu



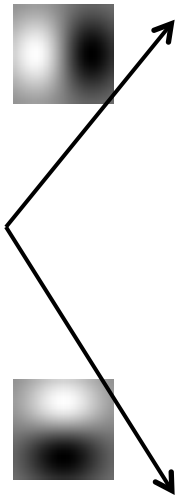
	media d/dx	media d/dy
fereastra #1	4	10
fereastra #2	25	7
fereastra #3	18	20

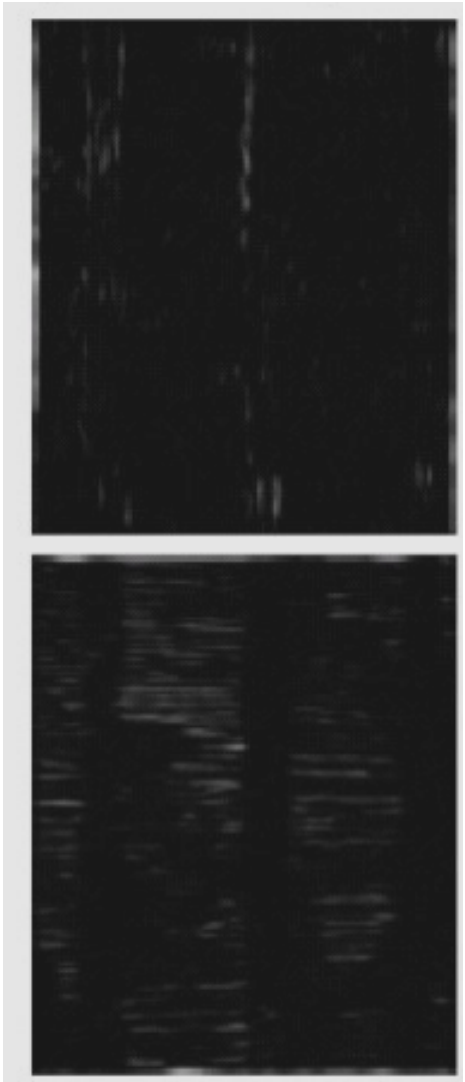
⋮

Statistici pentru a descrie pattern-urile în ferestre mici

Reprezentarea texturii: exemplu

imagine inițială

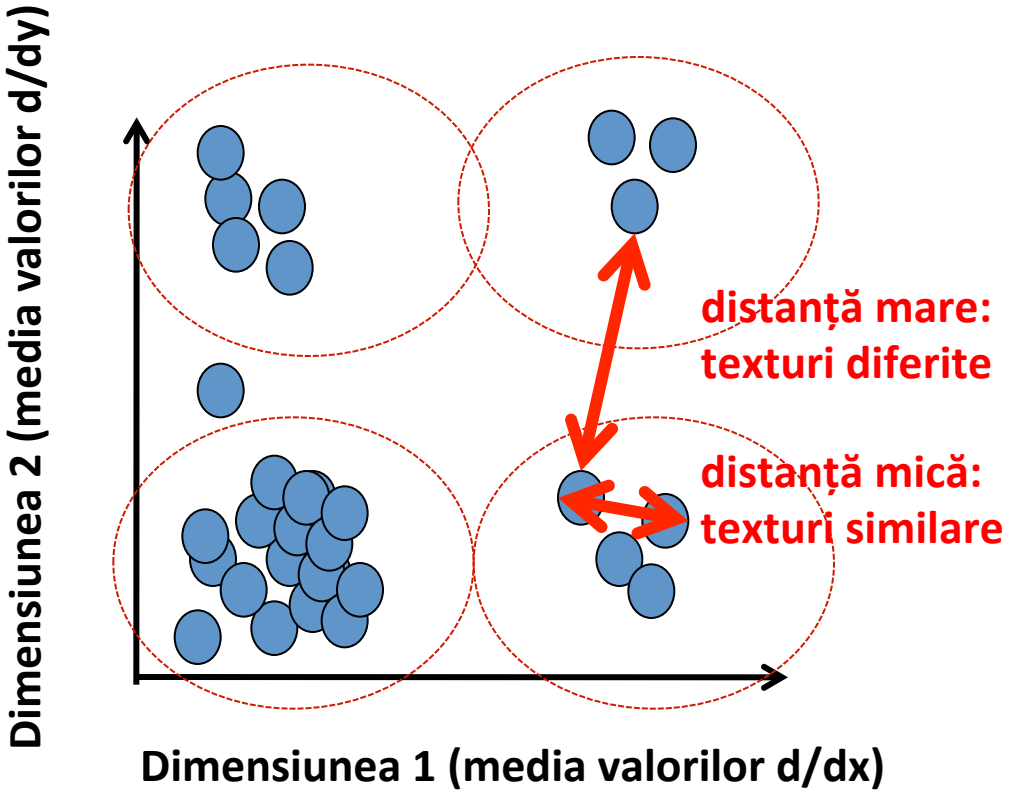




imagini filtrate
(filtre pentru derivate)

Vizualizare a tipurilor de textură.

Reprezentarea texturii: exemplu



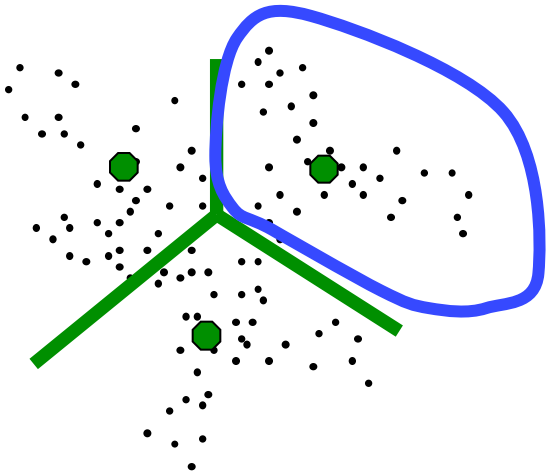
	media d/dx	media d/dy
fereastra #1	4	10
fereastra #2	25	7
fereastra #3	18	20

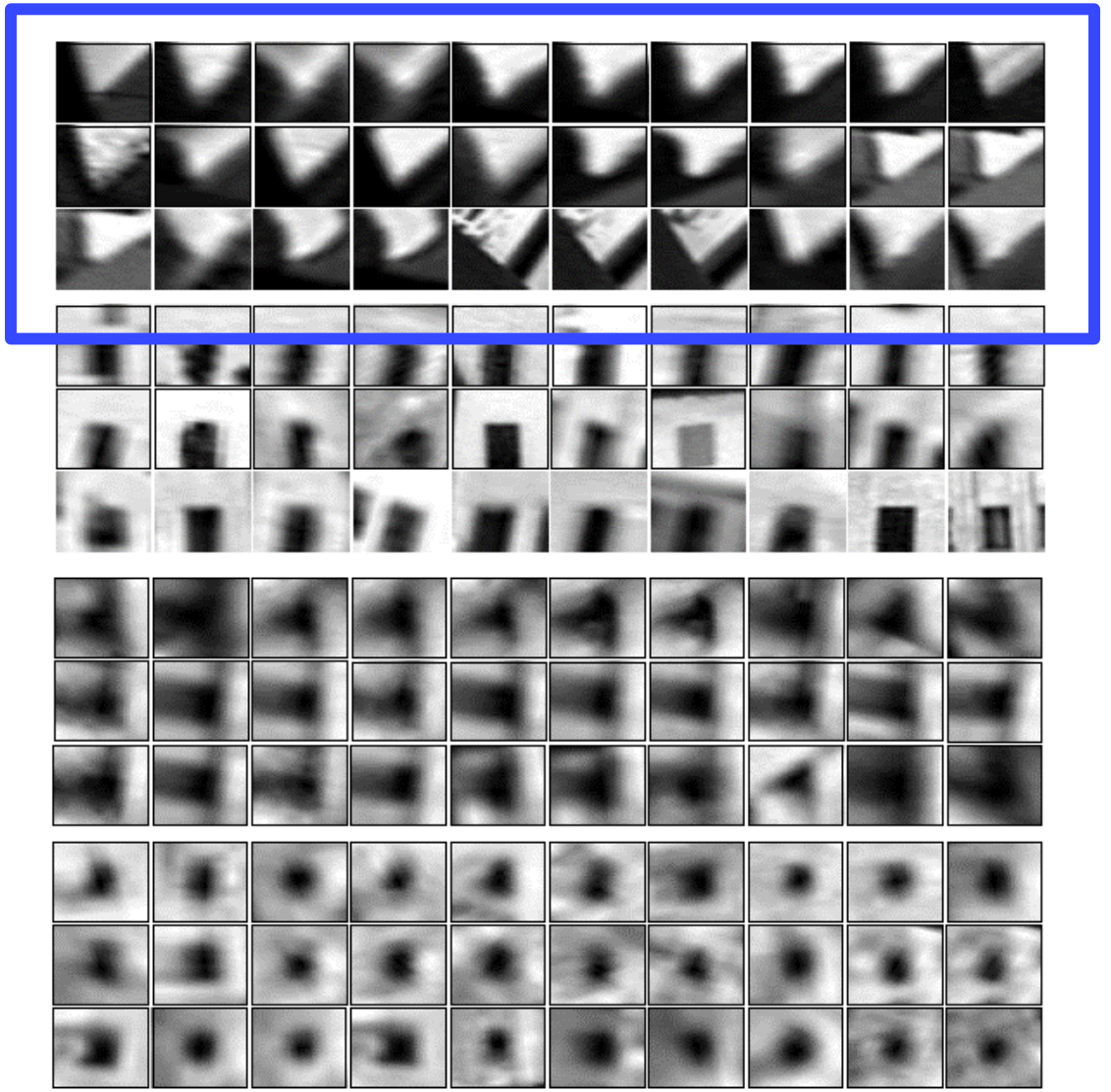
⋮

Statistici pentru a descrie pattern-urile în ferestre mici

Cuvinte vizuale

- Exemple: fiecare grup de patch-uri aparține aceluiași cuvânt vizual





Figură din articolul Sivic & Zisserman, ICCV 2003

Inverted file index

Image #1

Image #2

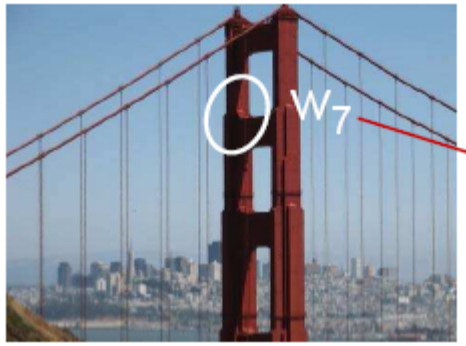
Image #3

Word #	Image #
1	3
2 …	
7	1, 2
8	3
9	
10 …	
91	2

…

- Inverted file index este o structură de date asemenea unui index de carte. Conține o intrare pentru fiecare cuvânt vizual și înregistrează toate imaginile în care acesta apare.

Inverted file index



New query image

Word #	Image #
1	3
2	
…	
7	1, 2
8	3
9	
10	
…	
91	2

…

- Pentru fiecare imagine query se găsesc cuvintele care apar în ea și i se calculează vectorul de frecvențe. Se găsește apoi imaginea cu reprezentare vectorială cea mai apropiată.

Probleme

- Cum sumarizăm conținutul vizual al unei imagini? Putem compara conținutul vizual a două imagini?
- Cât de mare ar trebui să fie vocabularul vizual? Cum ar trebui realizată eficient cuantizarea spațiului de trăsături?
- Este suficient ca două regiuni să aibă aceeași mulțime de cuvinte vizuale pentru a putea recunoaște obiecte/scene specifice? Cum verificăm spațial acest lucru?
- Cum măsurăm rezultatele regăsirii pe baza unei imagini query?

Imagine

Bag of ‘visual words’

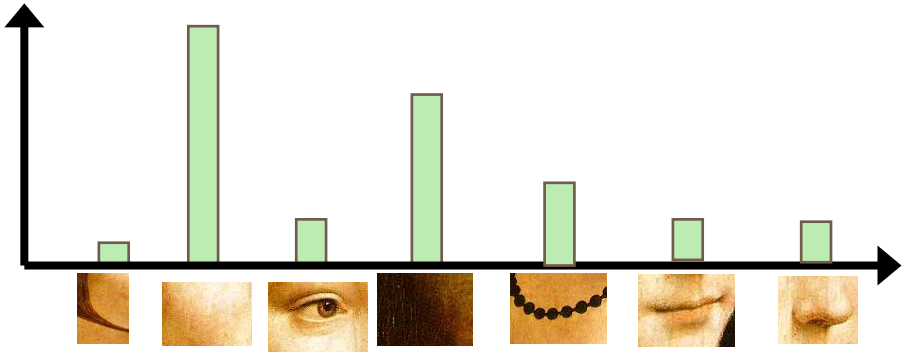
(sac de cuvinte vizuale)

Slide adaptat după L. Fei-Fei

Imagine

Bag of ‘visual words’

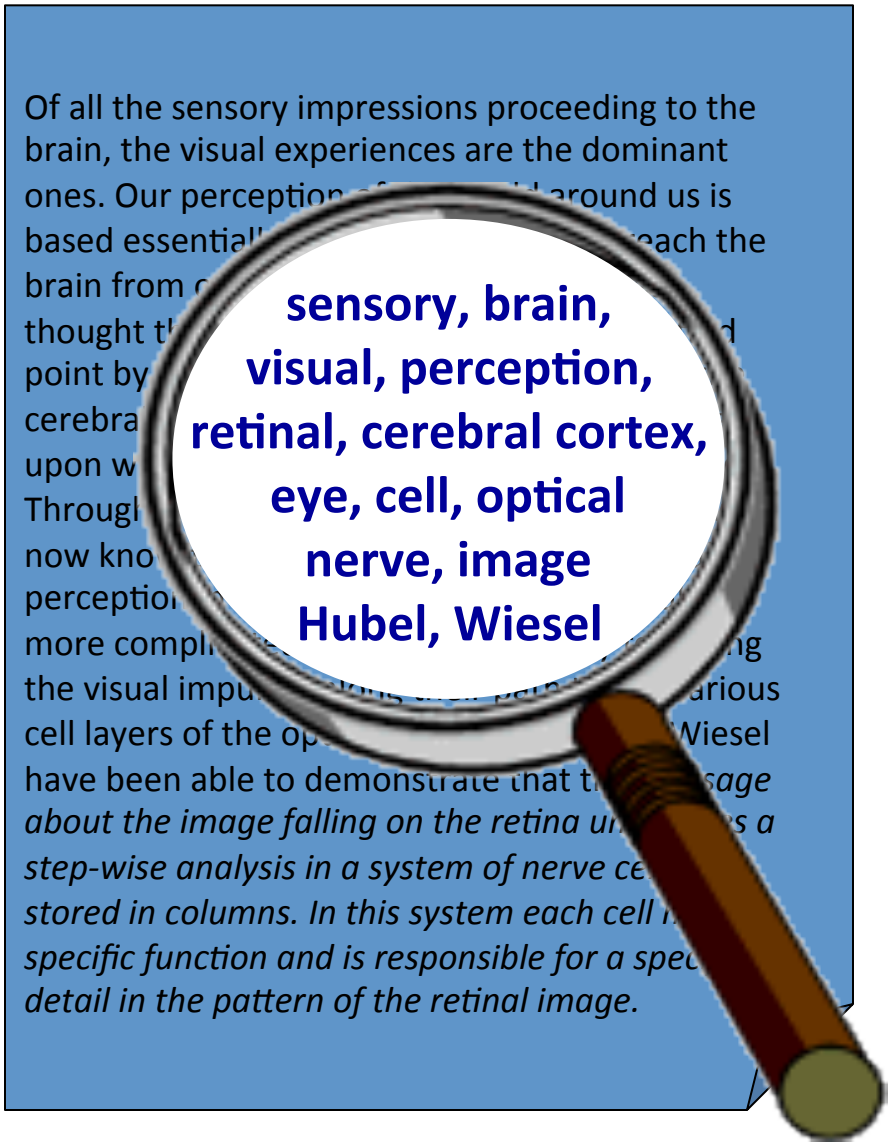
(sac de cuvinte vizuale)



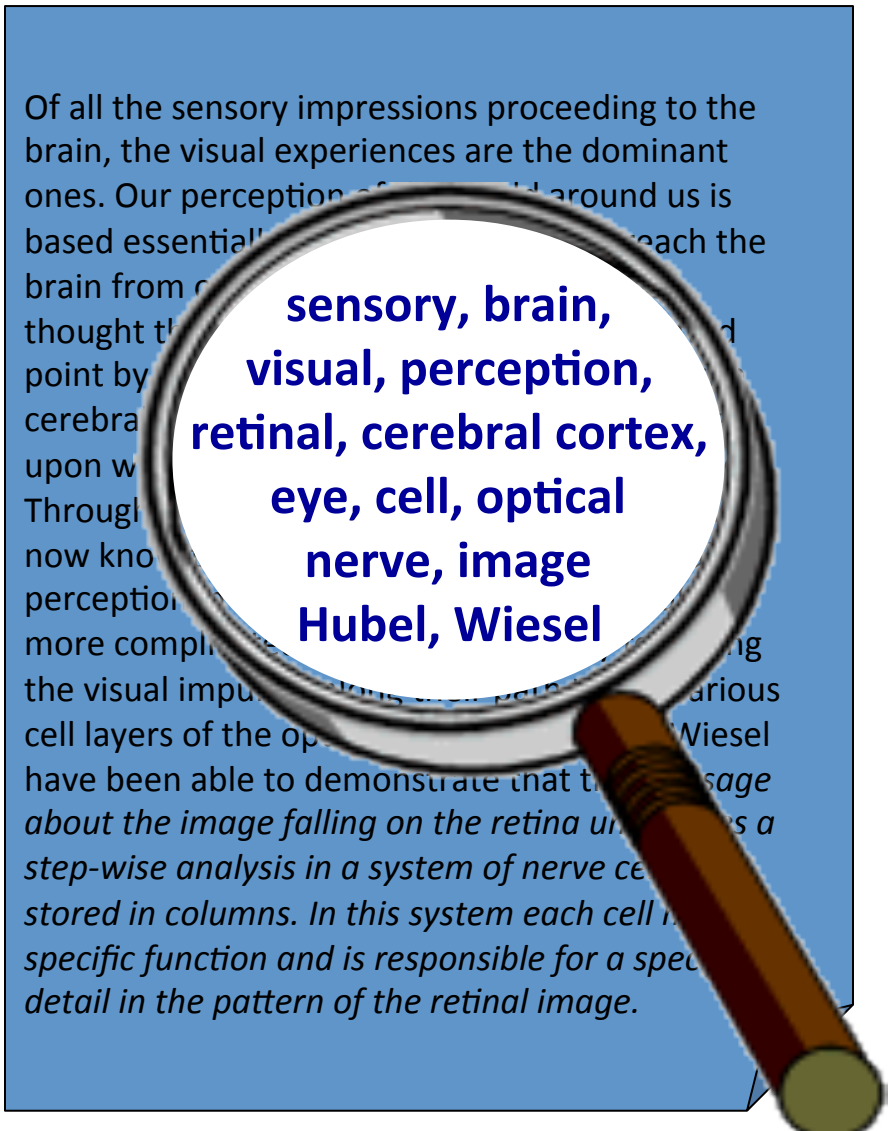
Idee de bază: reprezentăm o imagine ca o histogramă de pattern-uri prototip (cuvinte vizuale)

Slide adaptat după L. Fei-Fei

Analogia cu documentele



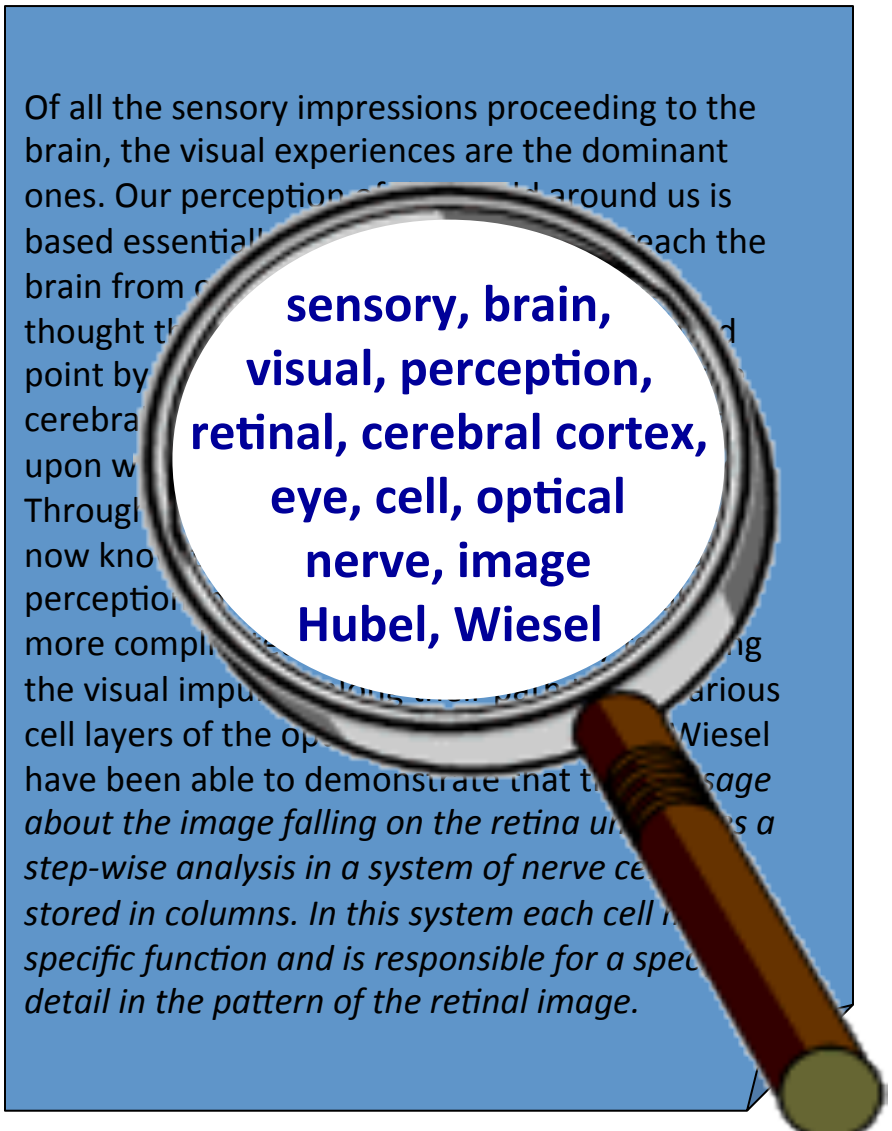
Modele de tip bag-of-words: reprezentarea documentelor



Care este domeniul (sport, politică, divertisment, medicină, economie, etc.) despre care se face referire în document?

MEDICINĂ

Modele de tip bag-of-words: reprezentarea documentelor



Modele de tip bag-of-words: reprezentarea documentelor

- Reprezentarea documentelor pe baza frecvențelor cuvintelor dintr-un dicționar

Cuvintele cu frecvența cea mai mare: Iraq, terrorists, economy, …

Salton & McGill (1983)

Modele de tip bag-of-words: reprezentarea documentelor

- Reprezentarea documentelor pe baza frecvențelor cuvintelor dintr-un dicționar

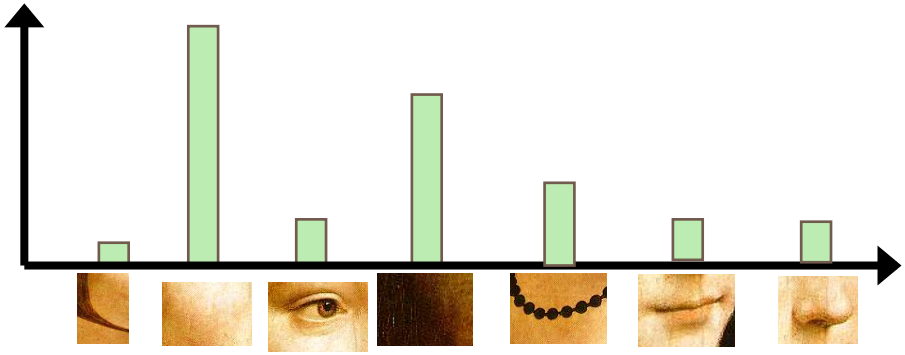
Modele de tip bag-of-words: reprezentarea documentelor

- Reprezentarea documentelor pe baza frecvențelor cuvintelor dintr-un dicționar

Imagine

Bag of ‘visual words’

(sac de cuvinte vizuale)



Idee de bază: reprezentăm o imagine ca o histogramă de pattern-uri prototip (cuvinte vizuale)

Slide adaptat după L. Fei-Fei

Imagine

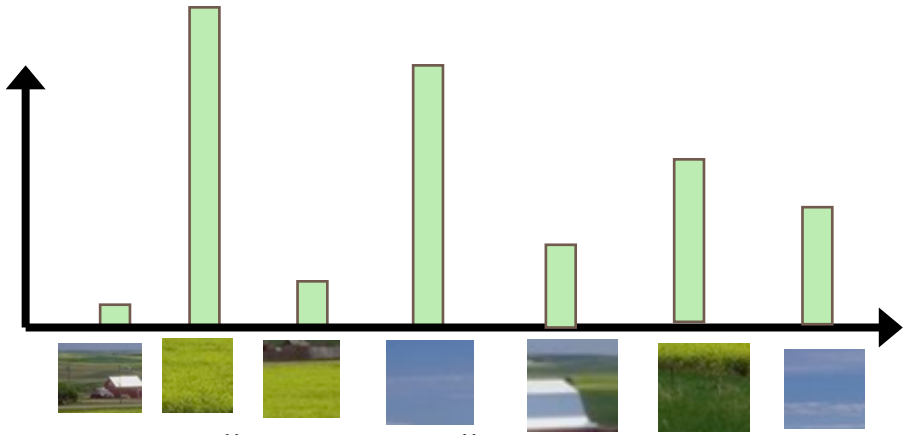
Bag of ‘visual words’

(sac de cuvinte vizuale)

Imagine

Bag of ‘visual words’

(sac de cuvinte vizuale)



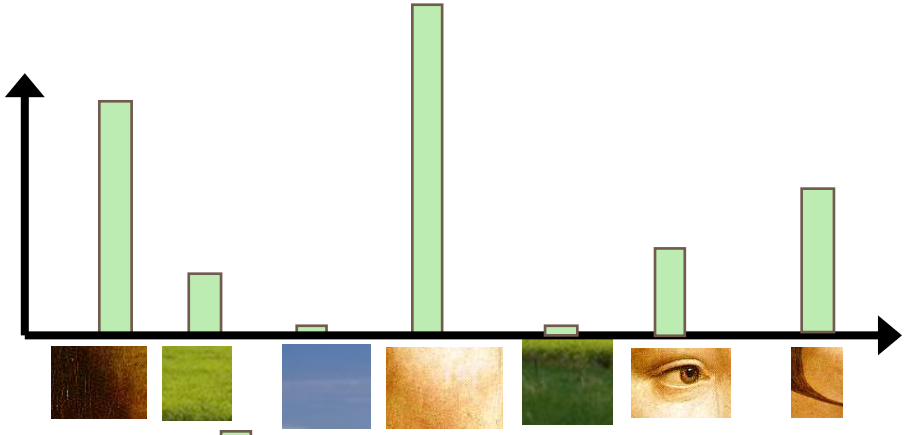
Idee de bază: reprezentăm o imagine ca o histogramă de pattern-uri prototip (cuvinte vizuale)

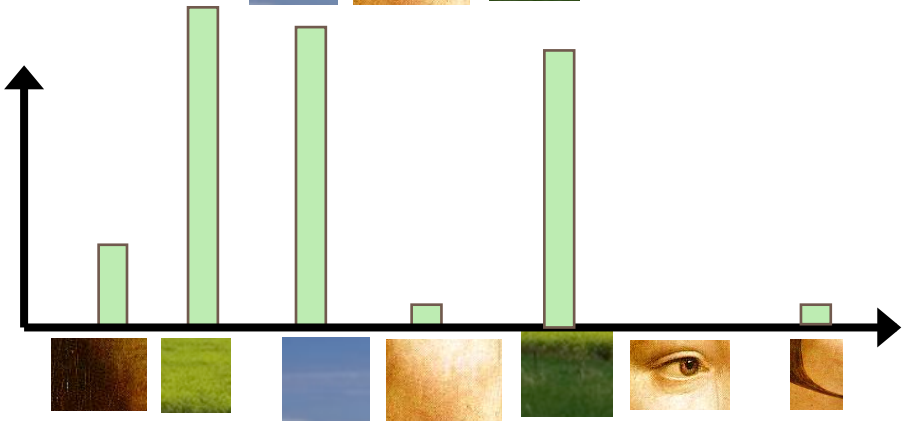
Imagine

Bag of ‘visual words’

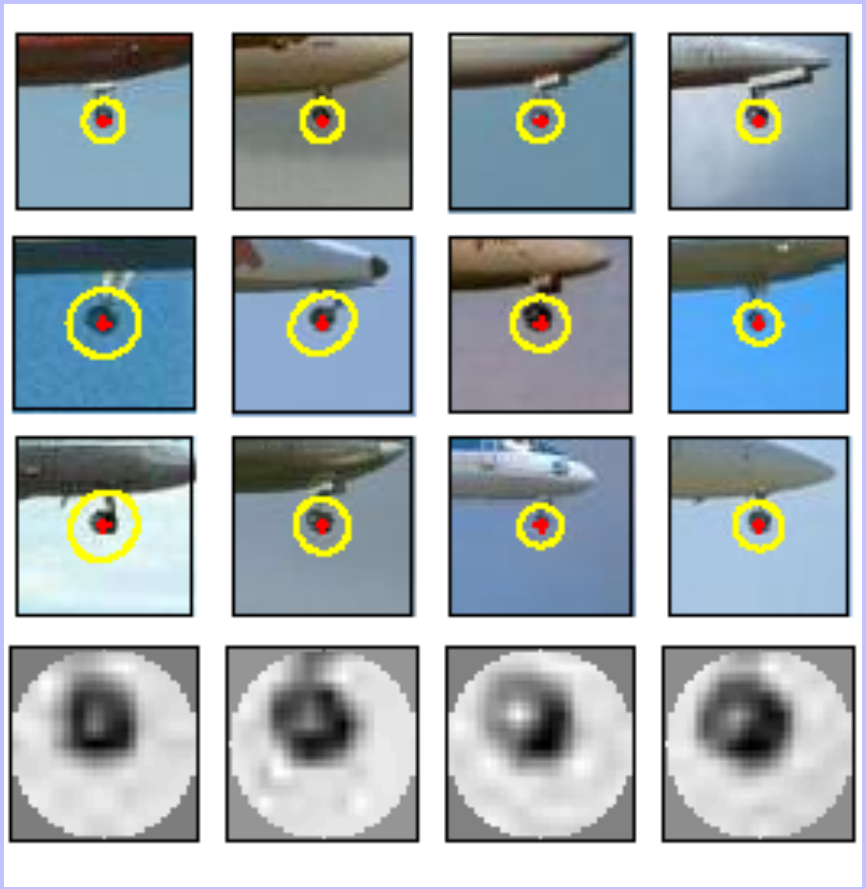
(sac de cuvinte vizuale)

Dacă vreau să compar imagini între ele nu pot să am pattern-uri prototip specifice pentru fiecare imagine. Vreau să am aceleași pattern-uri prototip pentru toate imaginile astfel încât comparația să fie posibilă.

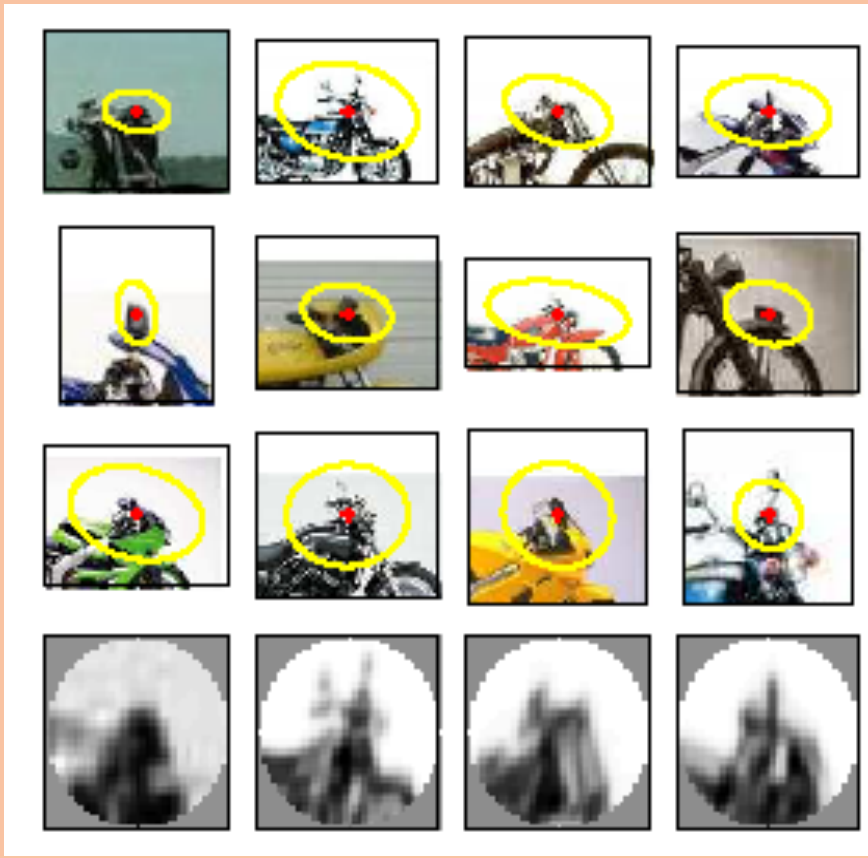




Visual words = cuvinte vizuale



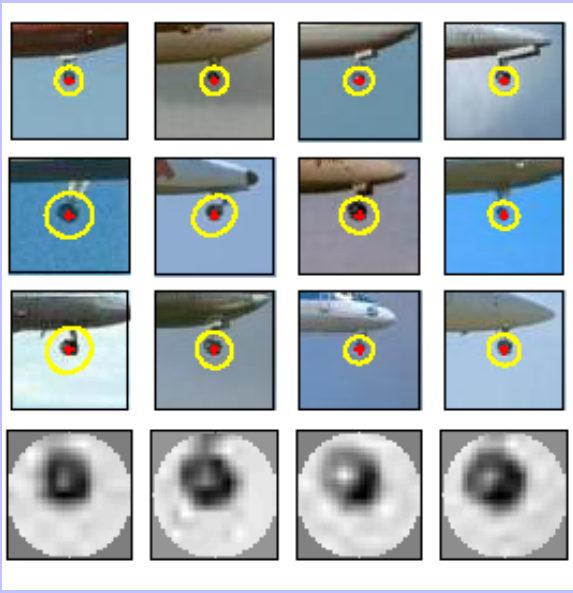
Subimaginea (patch-ul) care conține roata unui avion este diferită în fiecare imagine, dar foarte asemănătoare.

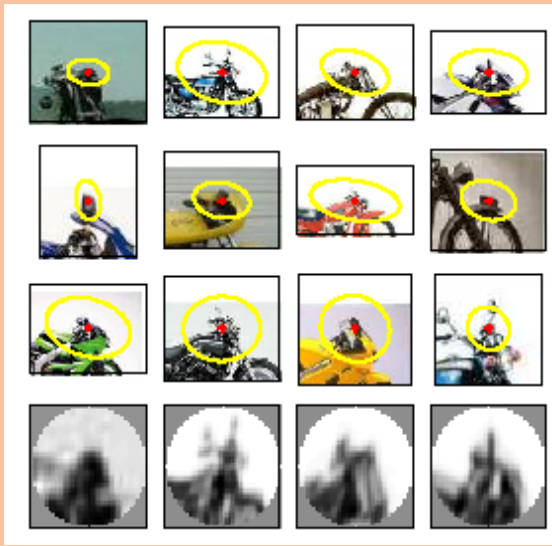


Subimaginea (patch-ul) care conține ghidonul unei motociclete este diferită în fiecare imagine, dar foarte asemănătoare.

Visual words = cuvinte vizuale

- Descriem cu ajutorul lor înfățișarea obiectelor
- Înfățisarea obiectelor variază foarte mult chiar şi pentru aceeaşi clasă de obiecte
- Înfățişarea locală a părților componente variază mai puțin
- **Ideal: cuvânt vizual = parte a unui obiect**
- **Descriem imaginile/patch-urile ca histograme de cuvinte vizuale**





Modelul ‘bag of visual words’

1. Mulțime de imagini

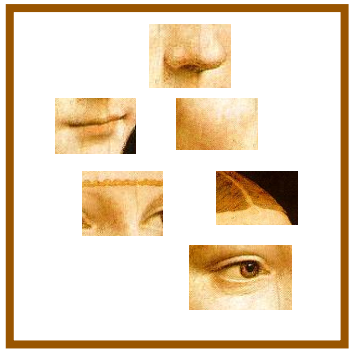


Modelul ‘bag of visual words’

1. Mulțime de imagini



2. Extragem caracteristici din fiecare imagine



Modelul ‘bag of visual words’

1. Mulțime de imagini



2. Extragem caracteristici din fiecare imagine
3. Învățăm un vocabular vizual = dicționar

Modelul ‘bag of visual words’

1. Mulțime de imagini



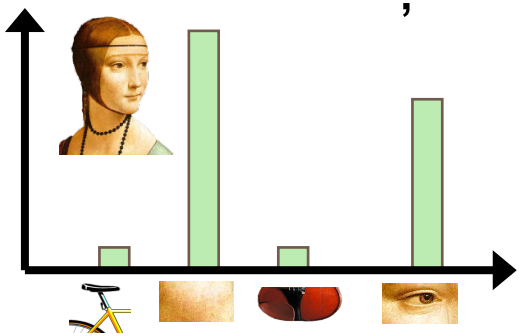
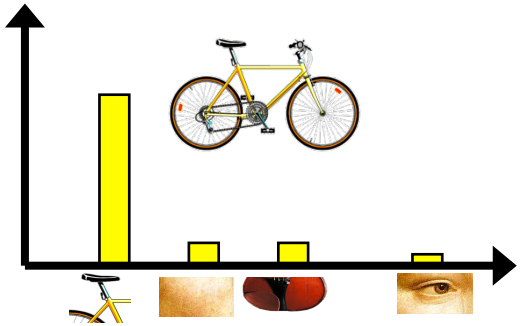
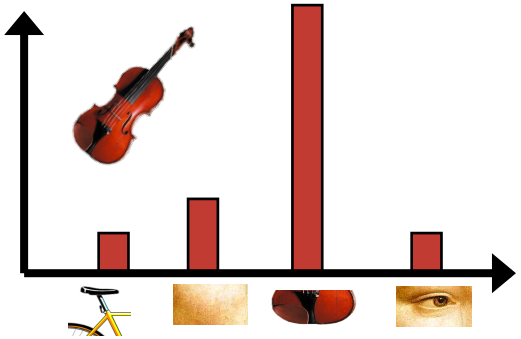
2. Extragem caracteristici din fiecare imagine
3. Învățăm un vocabular vizual = dicționar
4. Fiecare caracteristică este asignată ‘cuvântului vizual’ din dicționar cel mai apropiat

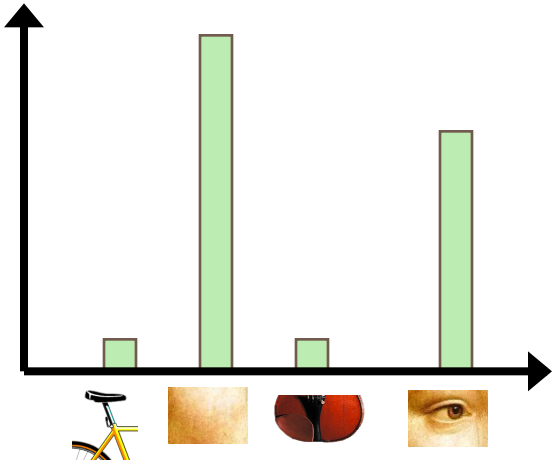
Modelul ‘bag of visual words’

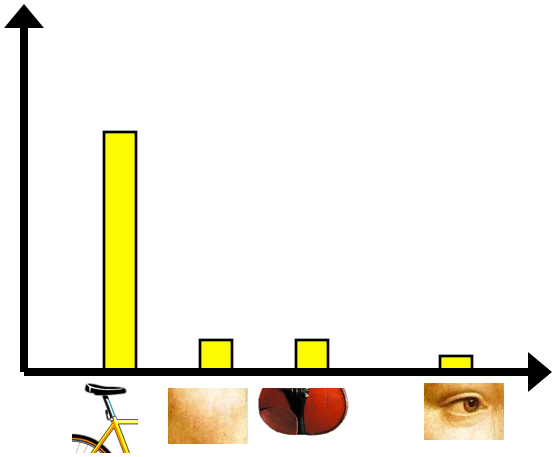
1. Mulțime de imagini

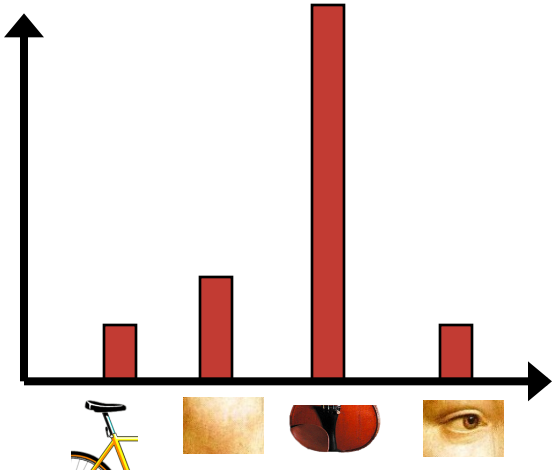


2. Extragem caracteristici din fiecare imagine
3. Învățăm un vocabular vizual = dicționar
4. Fiecare caracteristică este asignată ‘cuvântului vizual’ din dicționar cel mai apropiat
5. Reprezentăm imaginile pe baza unei histograme de frecvențe a ‘cuvintelor vizuale’



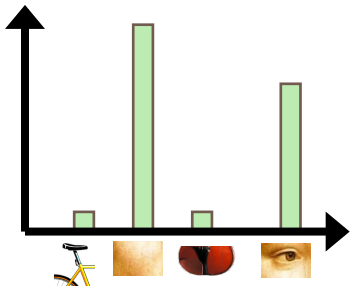




Compararea reprezentărilor BOW

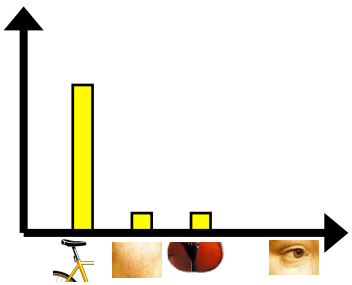
- Compararea histogramelor BOW între două imagini
- Putem folosi similaritatea cosinus

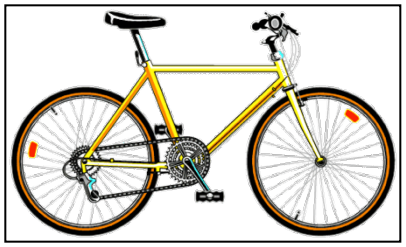
[1 8 1 4]



$\vec{d}_j$

[5 1 1 0]





$\vec{q}$

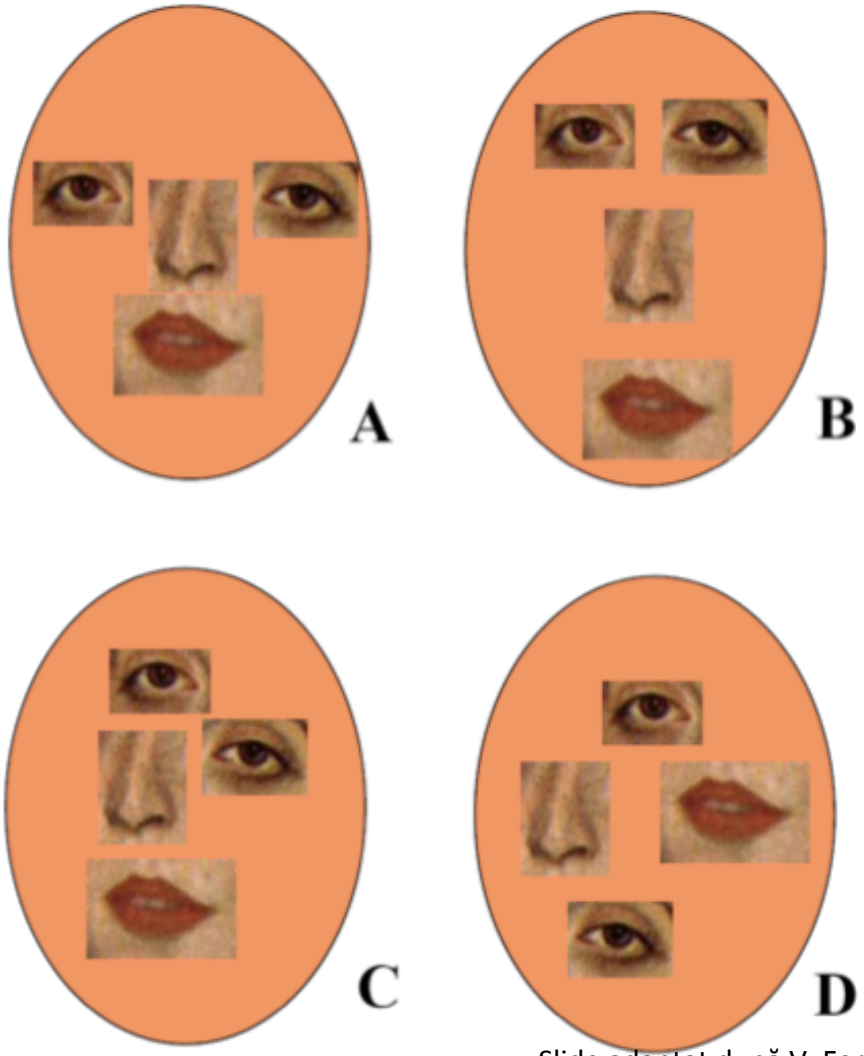
$$sim(d_j, q) = \frac{\langle d_j, q \rangle}{\|d_j\| \|q\|}$$

$$= \frac{\sum_{i=1}^{V} d_j(i) * q(i)}{\sqrt{\sum_{i=1}^{V} d_j(i)^2} * \sqrt{\sum_{i=1}^{V} q(i)^2}}$$

Pentru vocabulare cu V cuvinte

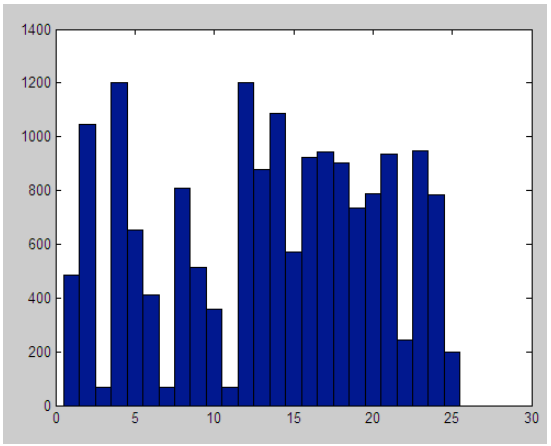
Limitări ale modelului bag of visual words

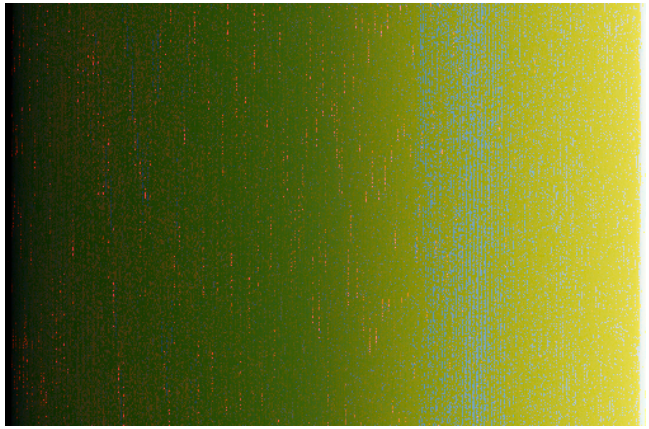
- Reprezentare care nu ține cont de poziția în imagine a cuvintelor vizuale
- Avantaje?
  - flexibil în poziționarea spațială a cuvintelor vizuale
- Dezavantaje?
  - mult prea flexibil?



Slide adaptat după V. Ferrari

Poziția în imagine a caracteristicilor



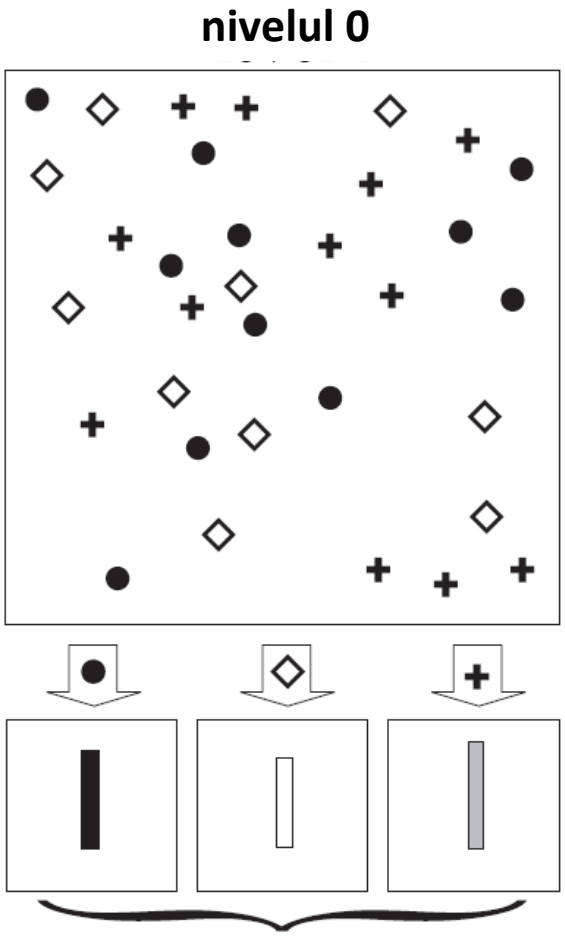


Cele trei imagini au aceeaşi histogramă de culori.

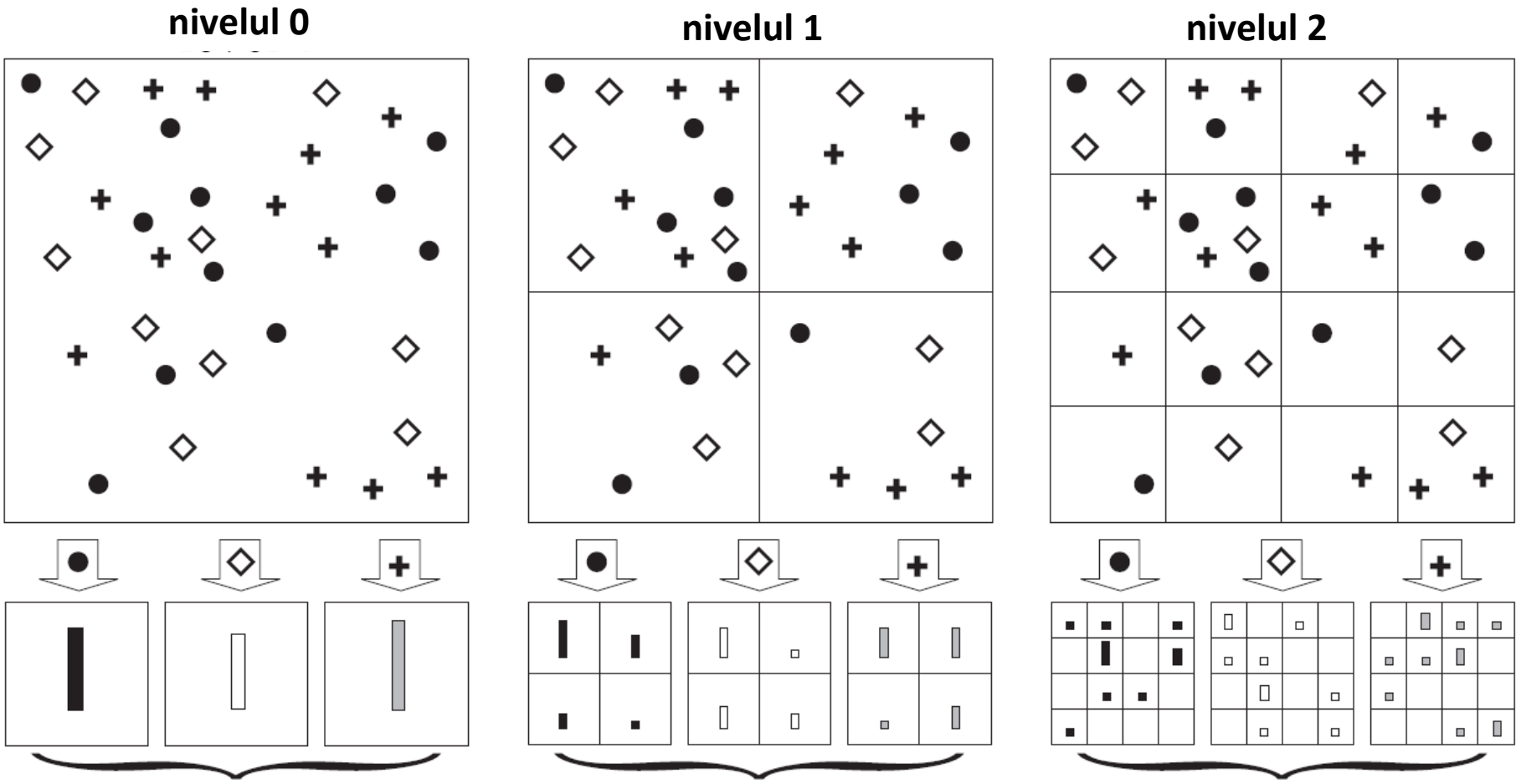
Piramidă spațială

Calculează histograma pentru fiecare subimagine

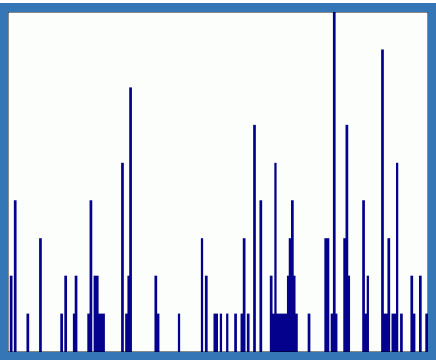
Piramidă spațială



Piramidă spațială



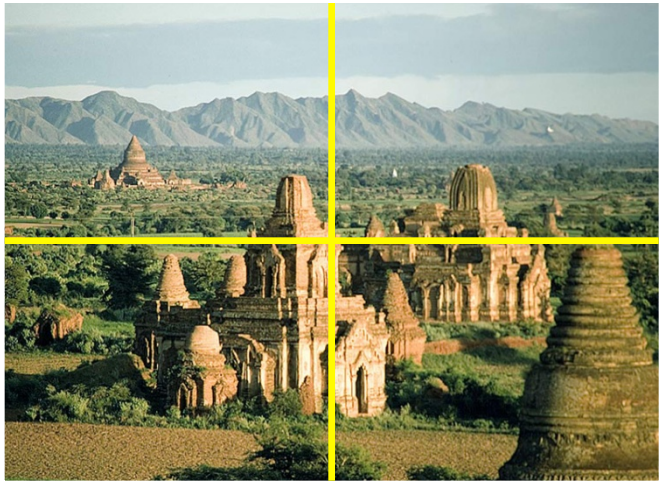
Piramidă spațială

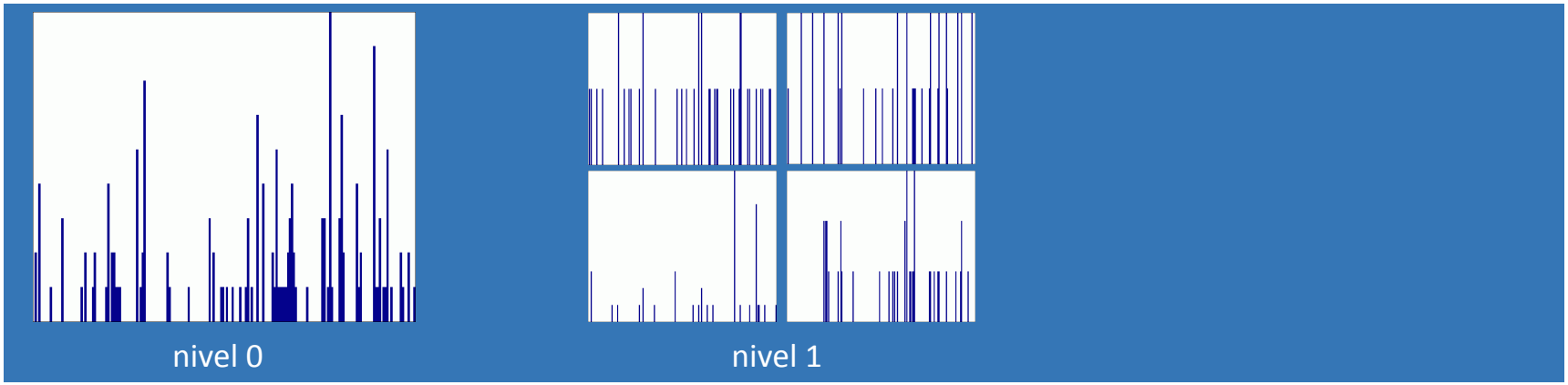


nivel 0

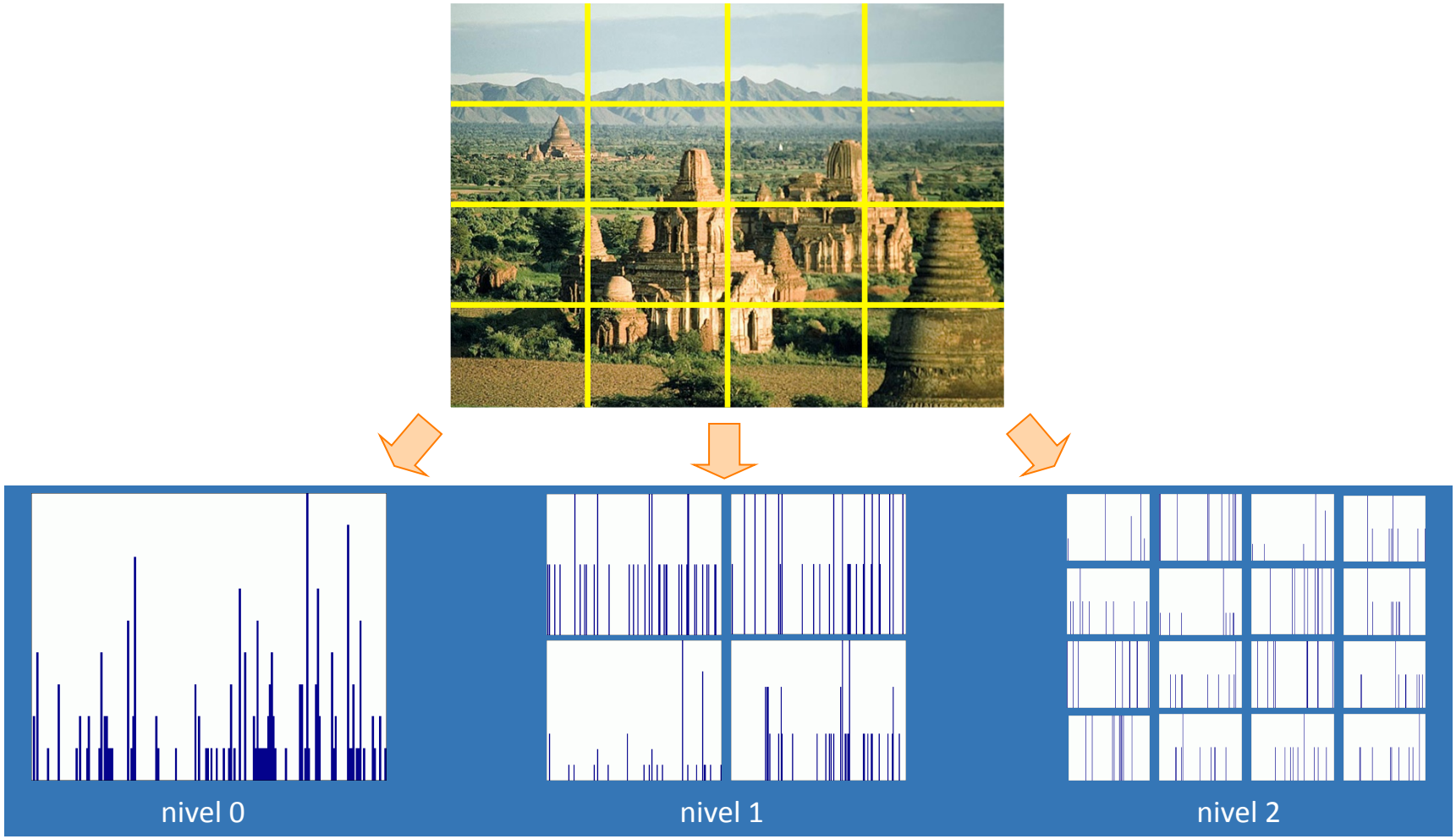
Slide adaptat după S. Lazebnik

Piramidă spațială





Piramidă spațială

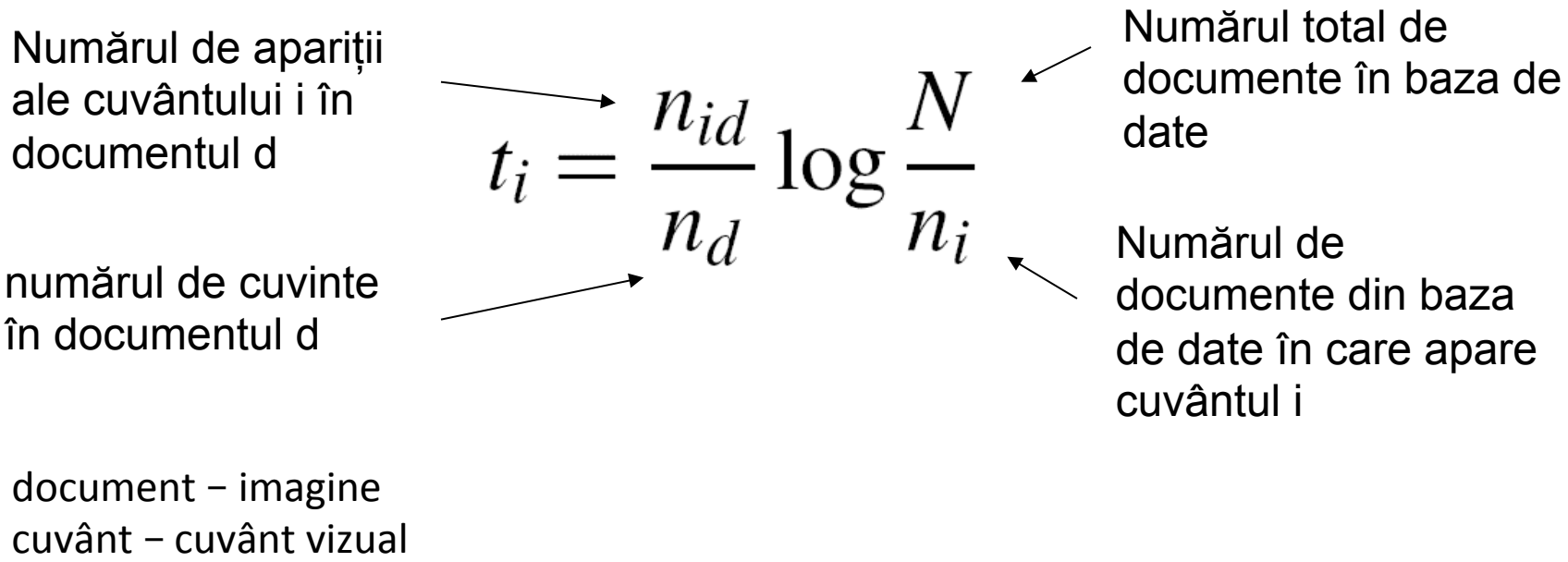


Slide adaptat după S. Lazebnik

Ponderarea tf-idf

- **T**erm **f**requency – **i**nverse **d**ocument **f**requency
- Descrie o imagine pe baza frecvențelor fiecărui cuvânt, micșorând ponderea celor care apar prea des în baza de date

(ponderarea standard în regăsirea de text)



BOW pentru regăsire pe baza conținutului

Interogare definită vizual

"Găsește acest ceas"

"Găsește acest loc"

"Groundhog Day" [Rammis, 1993]

Sivic & Zisserman, 2003

Example

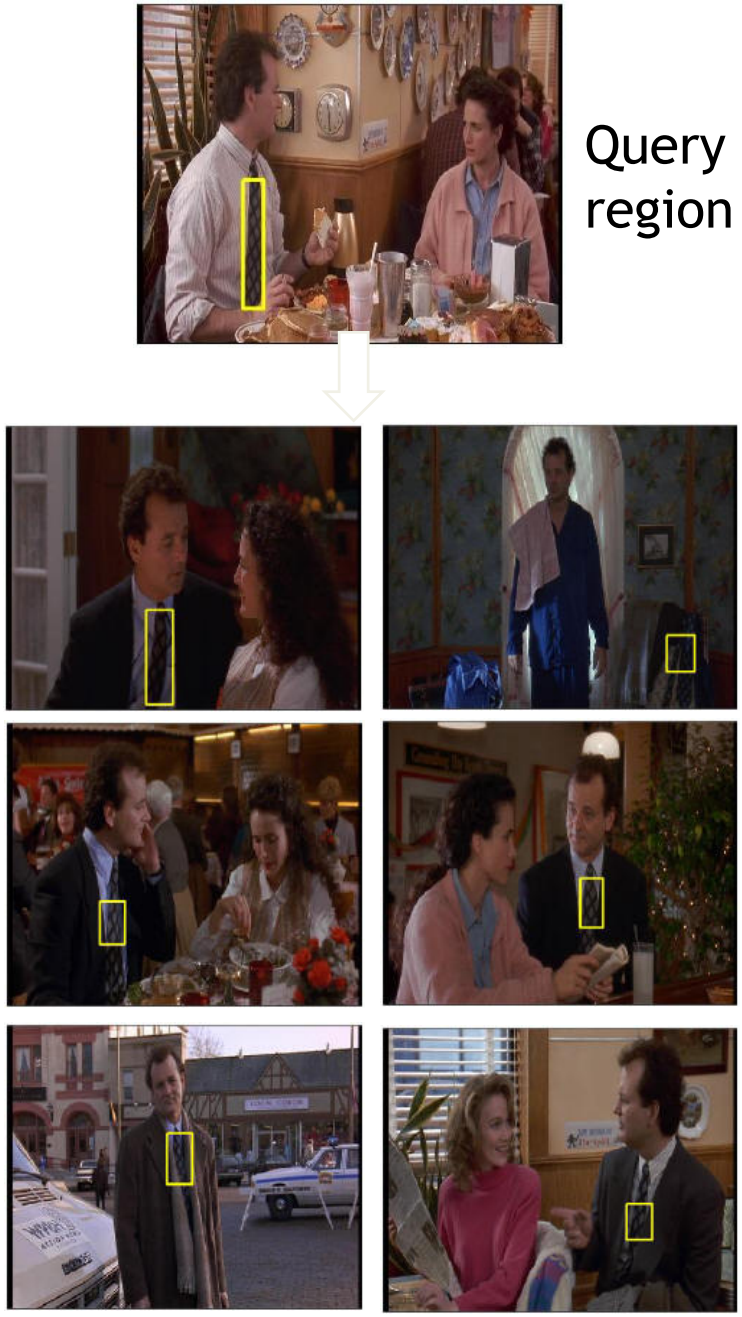
retrieved shots

Sivic & Zisserman, 2003

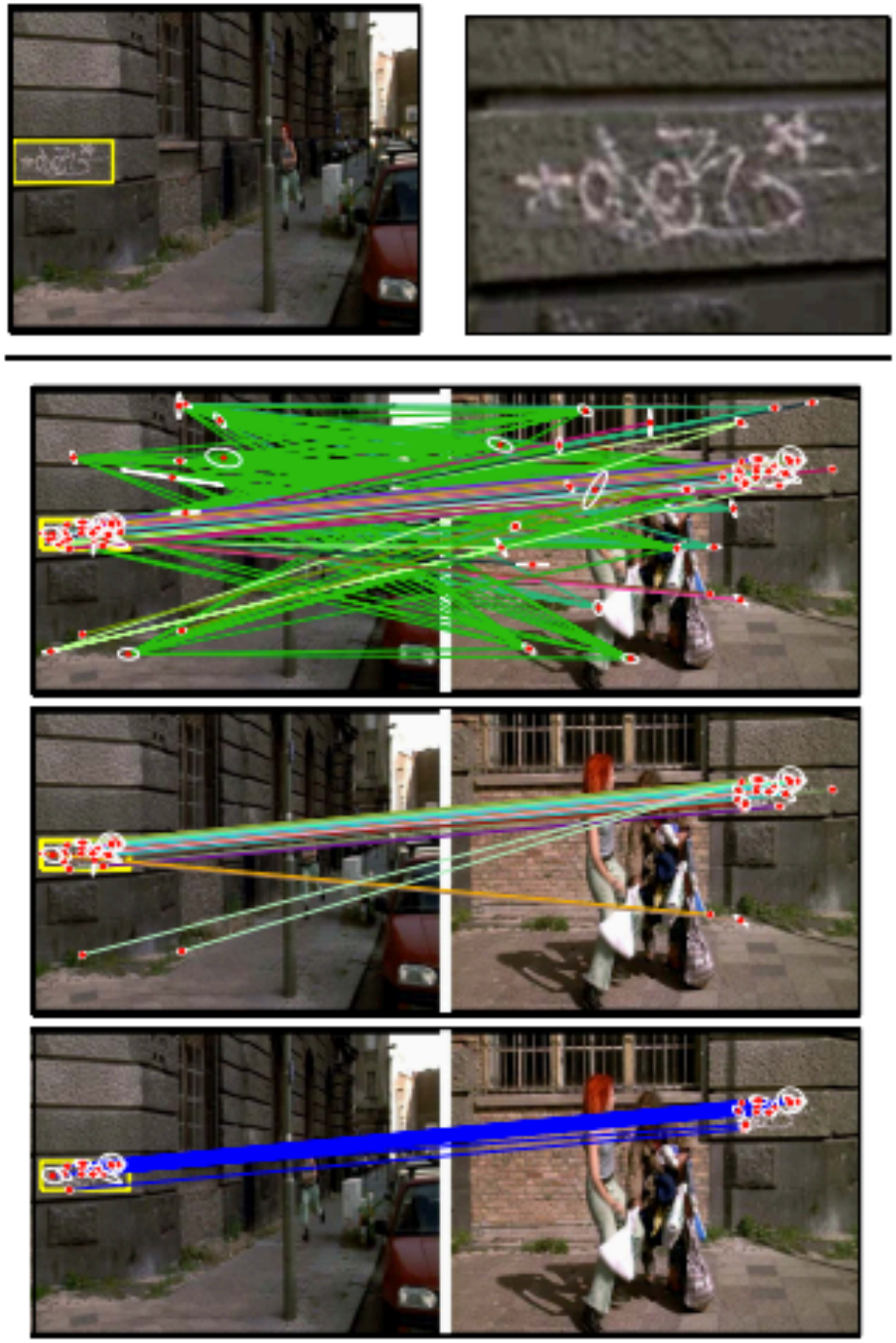
Video Google System

1. Extrage toate cuvintele vizuale din regiunea query
2. Aplică Inverted file index pentru găsirea de imagini relevante candidat
3. Compară frecvențele cuvintelor (tf-idf)
4. Verificare spațială

- Demo online at : http://www.robots.ox.ac.uk/~vgg/research/vgoogle/index.html



Sivic & Zisserman, 2003



Verificare spațială

-verifică că vecinii au corespondențe în același frame și eventual aceeași aranjare spațială prin estimarea unei transformări afine

-zonă de căutare de 15 vecini

-folosirea unui stop list: elimină cuvintele vzuale cele mai frecvente, care apar într-un număr mare de imagini − pot conduce la corespondențe false

Figure 6: Matching stages. Top row: (left) Query region and (right) its close-up. Second row: Original word matches. Third row: matches after using stop-list, Last row: Final set of matches after filtering on spatial consistency.

Scalable Recognition with a Vocabulary Tree

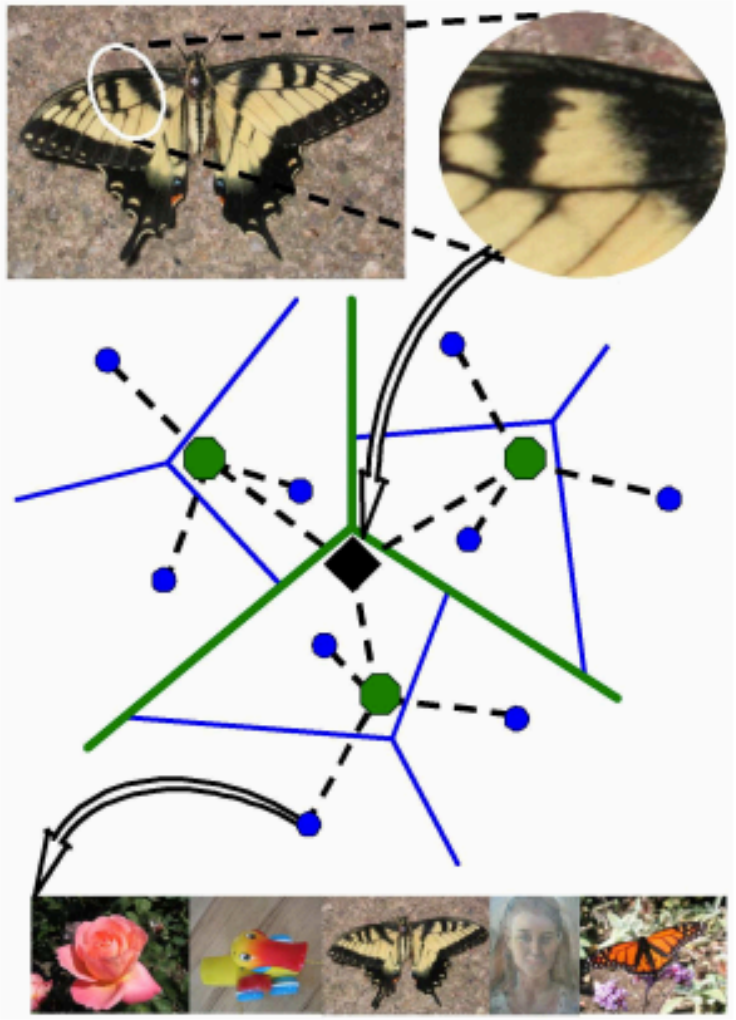
David Nistér and Henrik Stewénius
Center for Visualization and Virtual Environments
Department of Computer Science, University of Kentucky
http://www.vis.uky.edu/~dnister/ http://www.vis.uky.edu/~stewe/

Abstract

A recognition scheme that scales efficiently to a large number of objects is presented. The efficiency and quality is exhibited in a live demonstration that recognizes CD-covers from a database of 40000 images of popular music CD's.

The scheme builds upon popular techniques of indexing descriptors extracted from local regions, and is robust to background clutter and occlusion. The local region descriptors are hierarchically quantized in a vocabulary tree. The vocabulary tree allows a larger and more discriminatory vocabulary to be used efficiently, which we show experimentally leads to a dramatic improvement in retrieval quality. The most significant property of the scheme is that the tree directly defines the quantization. The quantization and the indexing are therefore fully integrated, essentially being one and the same.

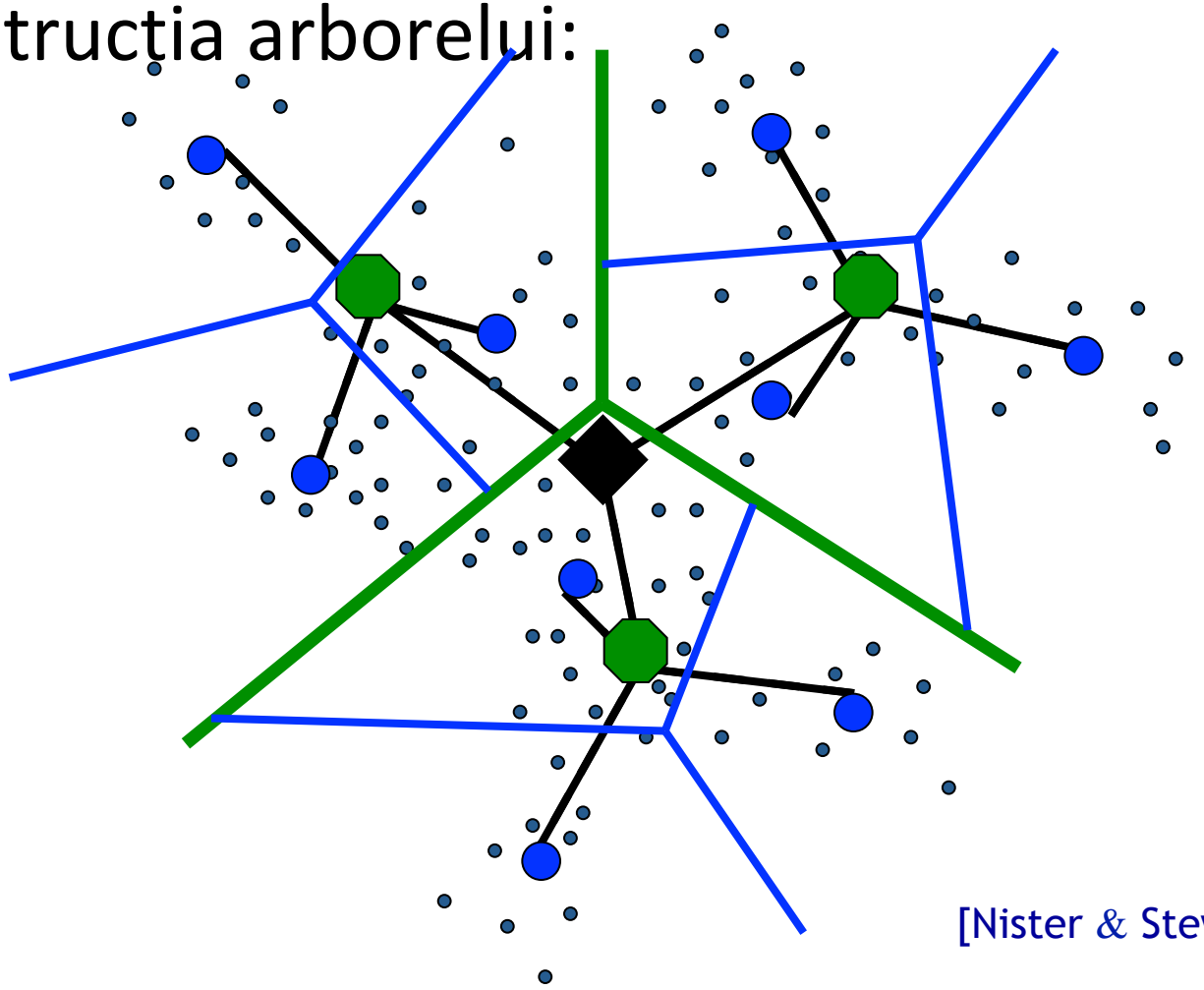
The recognition quality is evaluated through retrieval on a database with ground truth, showing the power of the vocabulary tree approach, going as high as 1 million images.



1. Introduction

Vocabular sub formă de arbore: clusterizare ierarhică pentru vocabulare mari

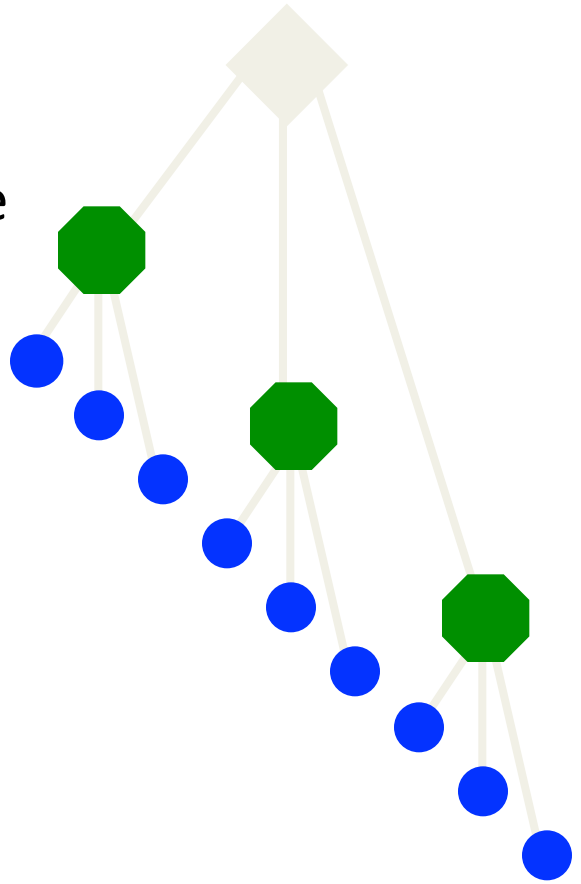
- Construcția arborelui:



[Nister & Stewenius, CVPR'06]

Vocabular sub formă de arbore

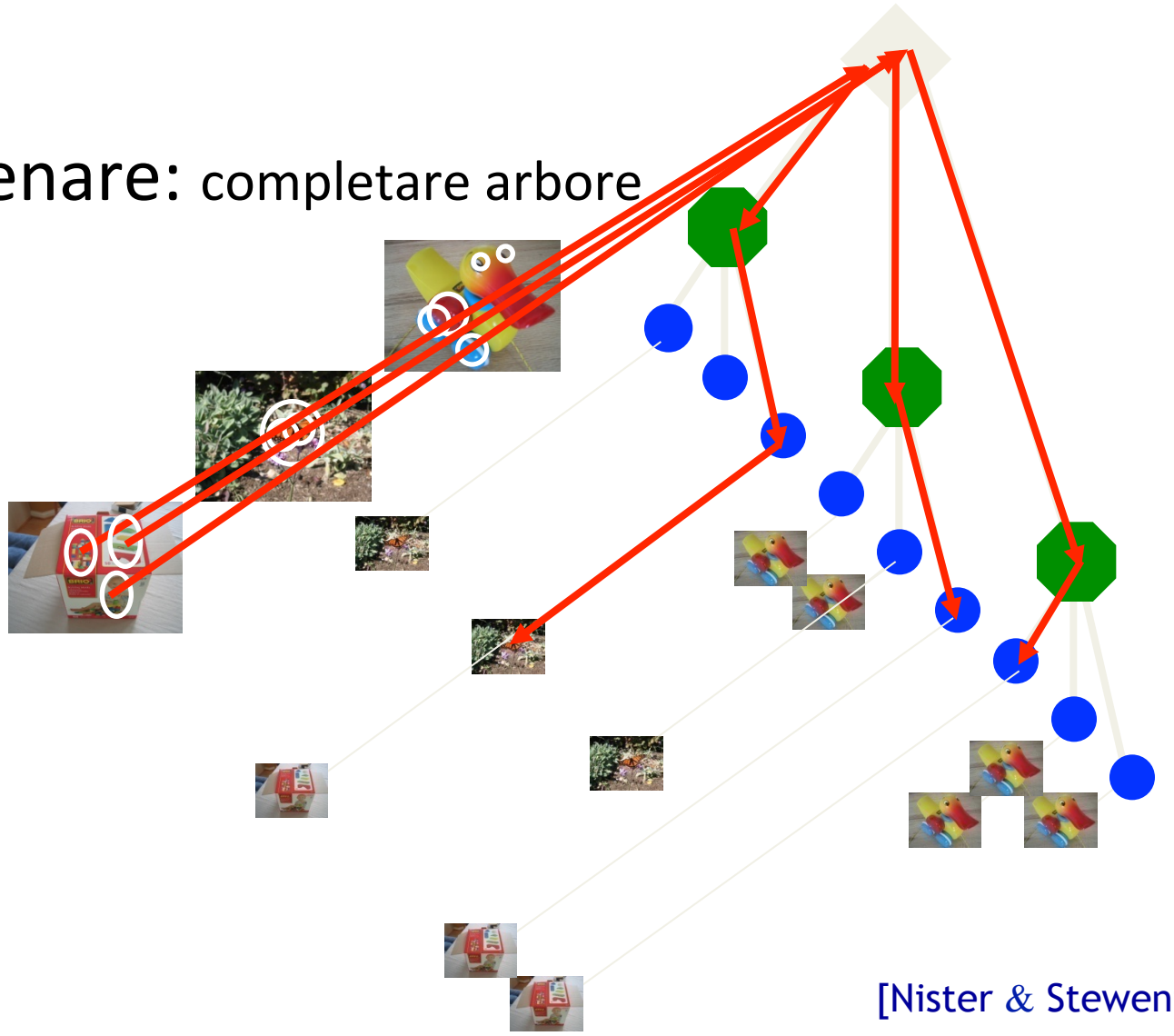
- Antrenare: completare arbore



[Nister & Stewenius, CVPR'06]

Vocabular sub formă de arbore

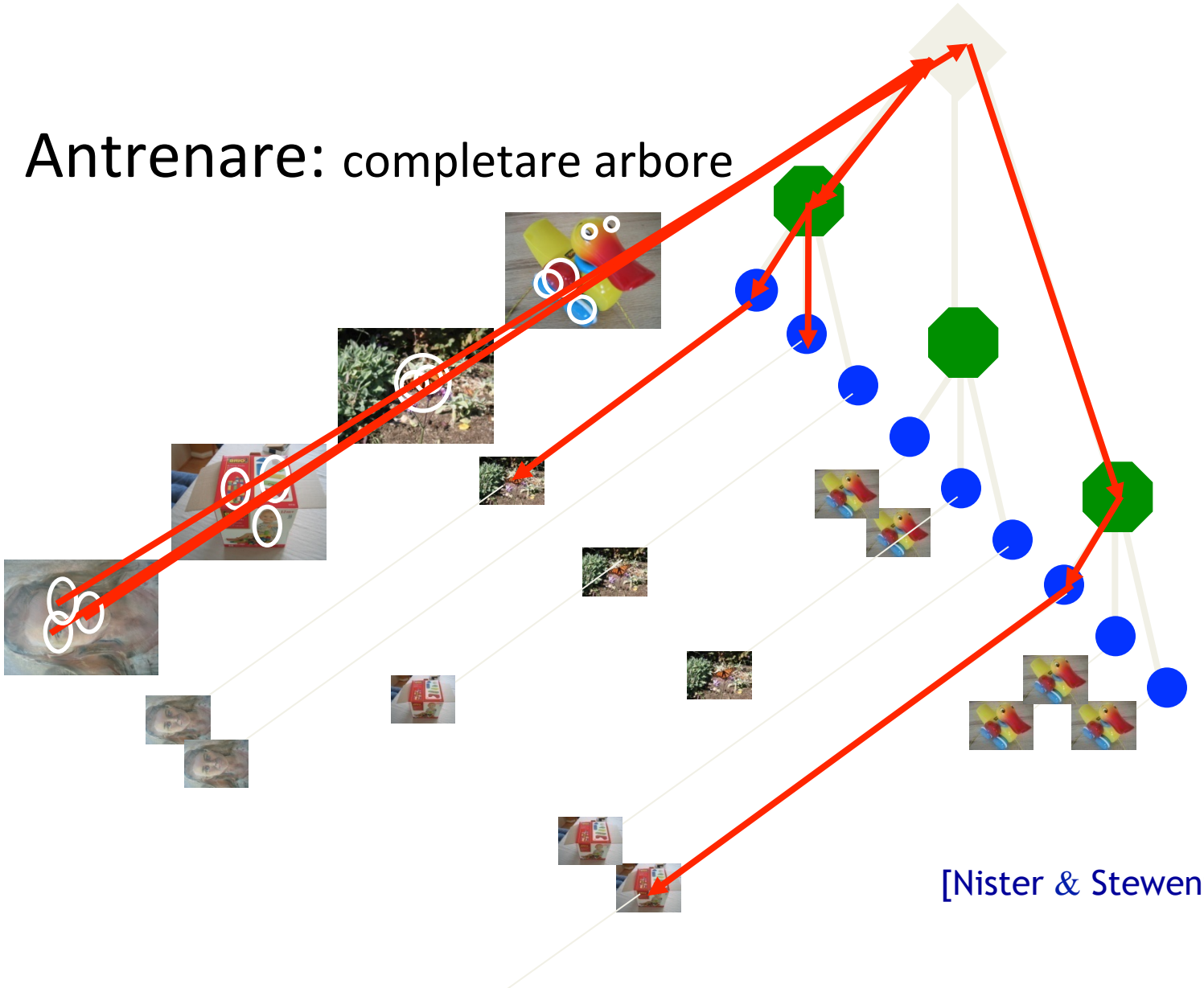
- Antrenare: completare arbore



[Nister & Stewenius, CVPR'06]

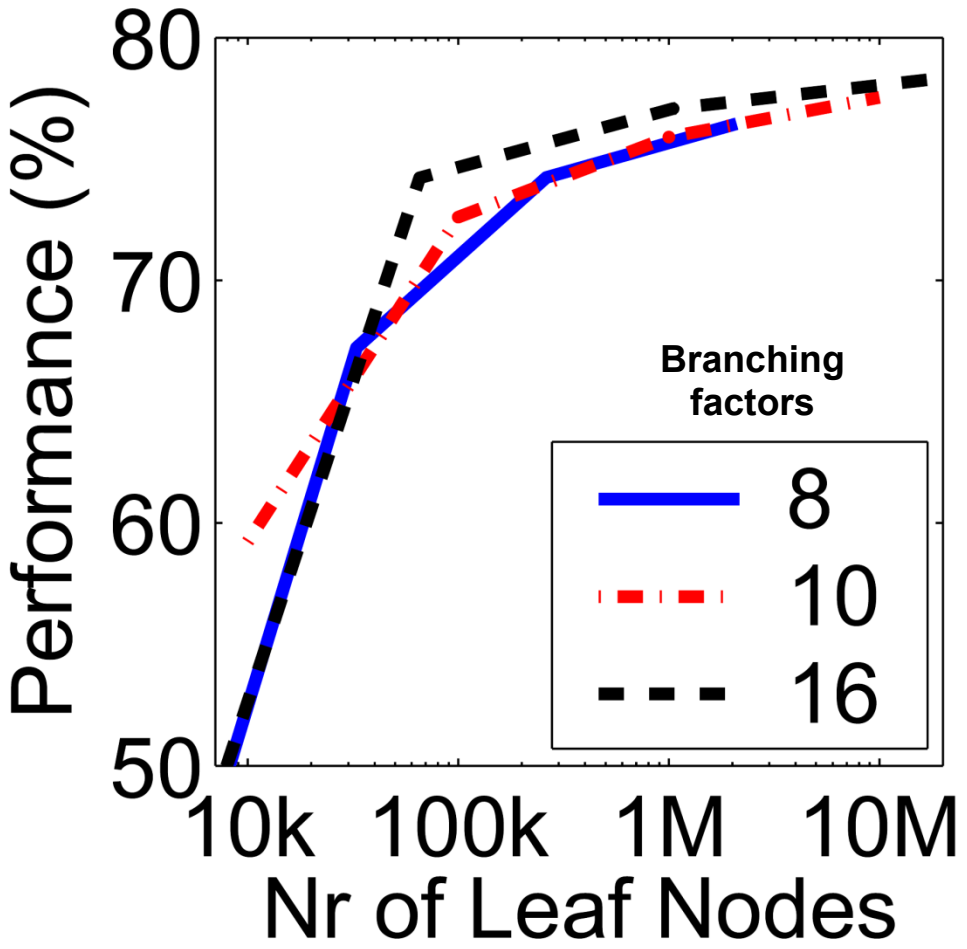
Vocabular sub formă de arbore

- Antrenare: completare arbore

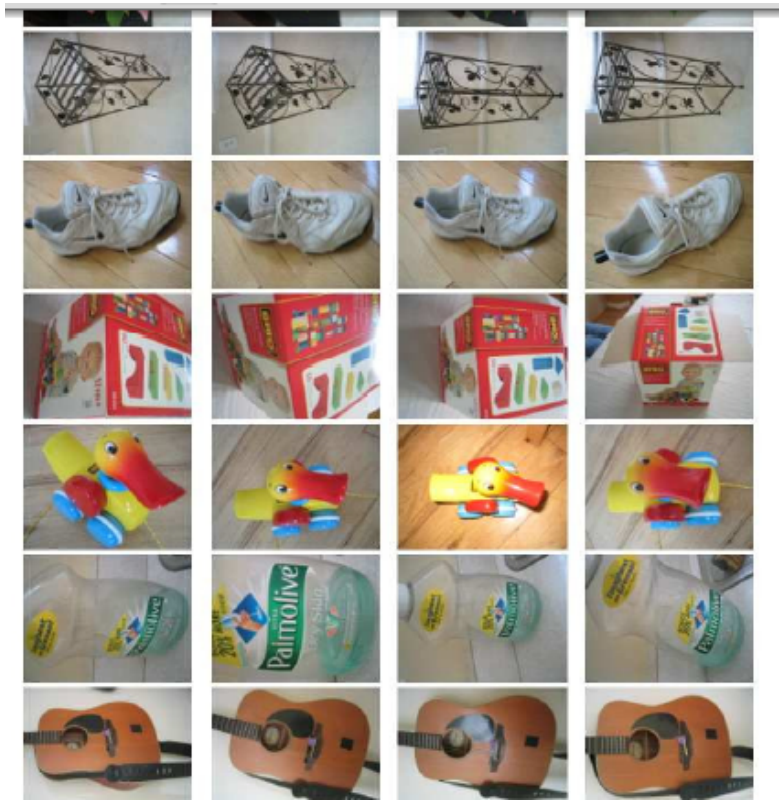


[Nister & Stewenius, CVPR'06]

Dimensiunea vocabularului



Results for recognition task with 6347 images



Nister & Stewenius, CVPR 2006

Modelul BOW: pros / cons

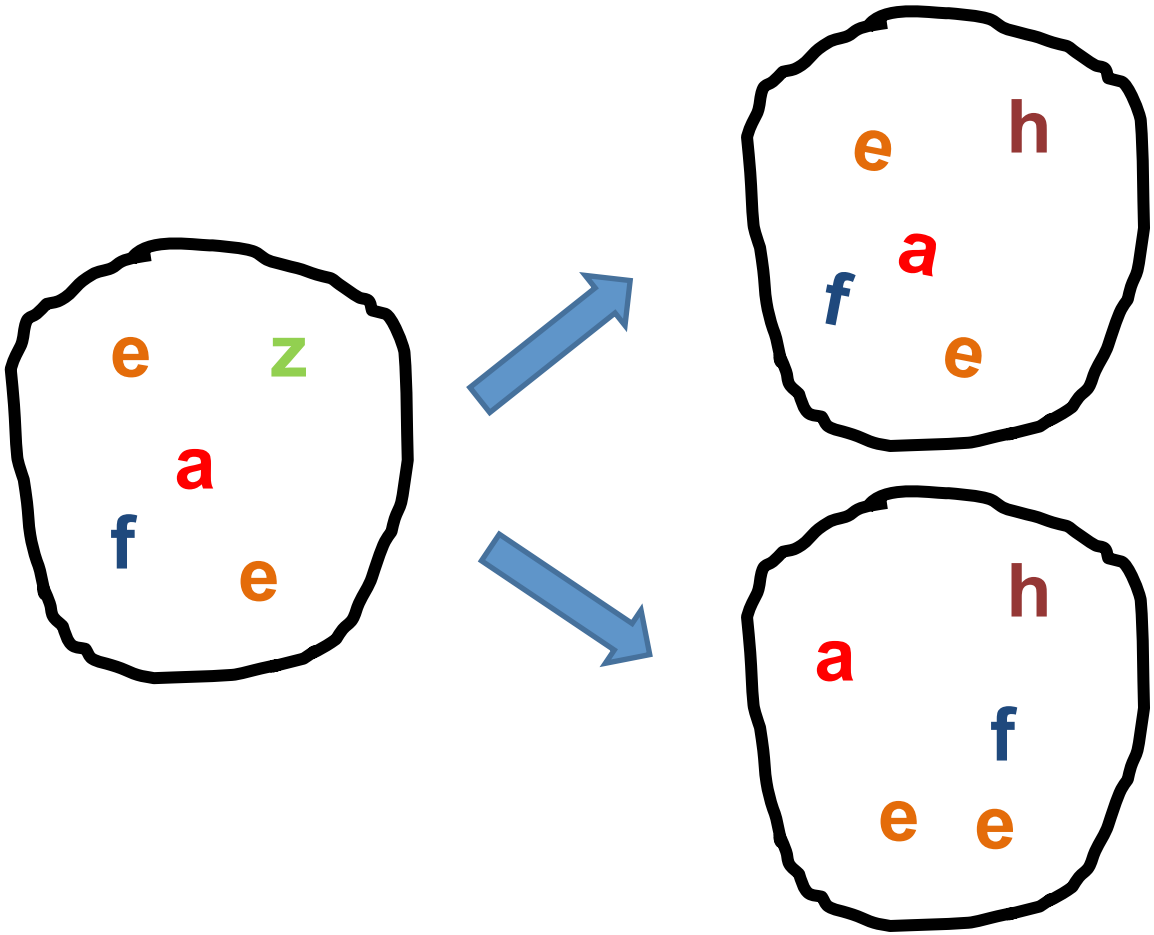
+ flexibil la geometrie / deformări / unghiul camerei
+ reprezentarea compactă a conținutului vizual al unei imagini
+ în practică furnizează rezultate foarte bune

- modelul de bază ignoră geometria – trebuie realizată o verificare spațială/encodată de descriptori
- reprezentarea globală a unei întregi imagini amestecă background-ul cu foreground-ul
- construcția vocabularului optim nu e clară

Probleme

- Cum sumarizăm conținutul vizual al unei imagini? Putem compara conținutul vizual a două imagini?
- Cât de mare ar trebui să fie vocabularul vizual? Cum ar trebui realizată eficient cuantizarea spațiului de trăsături?
- Este suficient ca două regiuni să aibă aceeași mulțime de cuvinte vizuale pentru a putea recunoaște obiecte/scene specifice? Cum verificăm spațial acest lucru?
- Cum măsurăm rezultatele regăsirii pe baza unei imagini query?

Care imagine se potrivește mai bine?

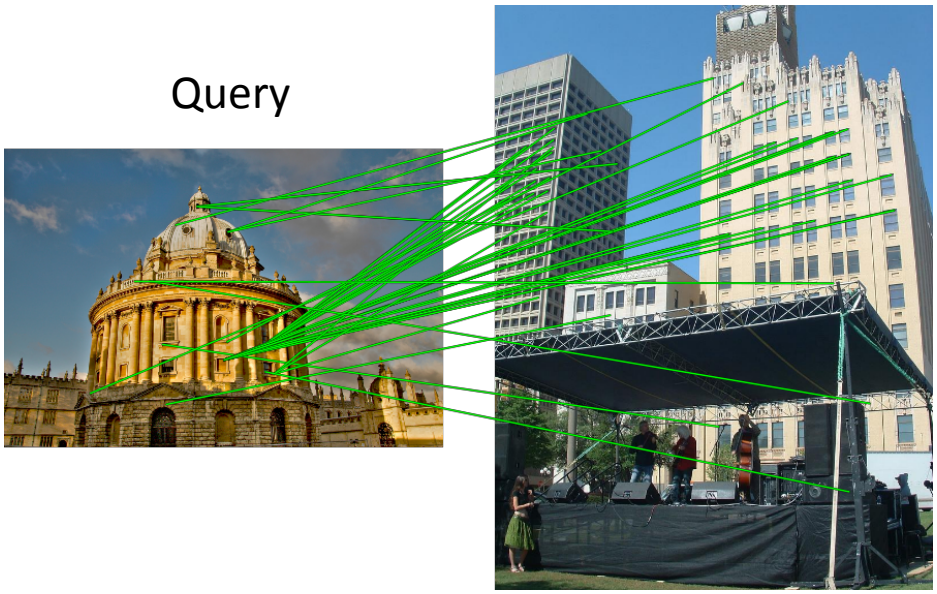


Verificare spațială

Query

Imagine din baza de date
cu similaritate mare BOW

Query



Imagine din baza de date
cu similaritate mare BOW

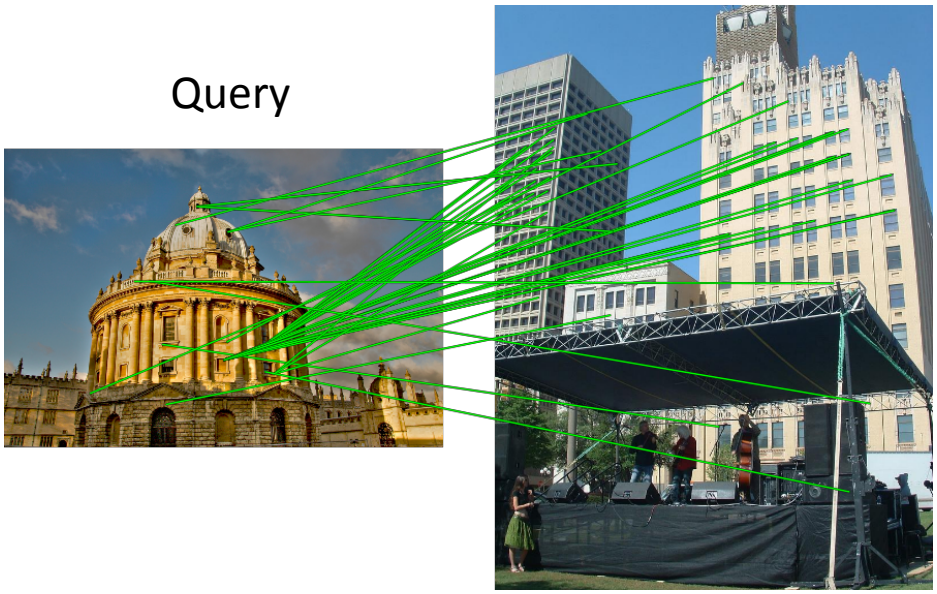
Ambele perechi de imagini au multe cuvinte vizuale în comun

Verificare spațială

Query

Imagine din baza de date
cu similaritate mare BOW

Query



Imagine din baza de date
cu similaritate mare BOW

Numai o parte din corespondențe sunt mutual consistente.

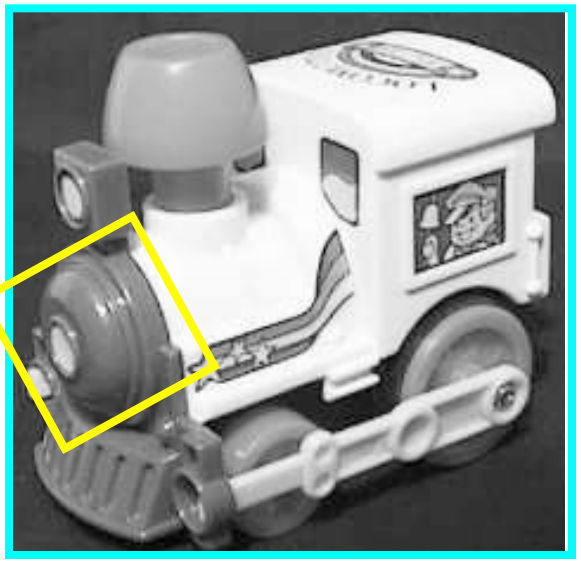
Verificare spațială: două strategii de bază

- Transformata Hough generalizată
  - fiecare corespondeță votează pentru poziție, mărime, orientare în spațiul Hough parametrizat al modelului obiectului
  - verificarea parametrilor cu număr de voturi mare

- RANSAC
  - estimează o transformare geometrică și verifică numărul de puncte inlier (numărul de corespondențe care verifică transformarea geometrică)

Transformata Hough generalizată

- Dacă folosim trăsături locale (ca SIFT) care sunt invariante la mărime, rotație și translație, atunci fiecare trăsătură găsită drept corespondență votează pentru o anumită mărime, translație și orientare a modelului în imagine.

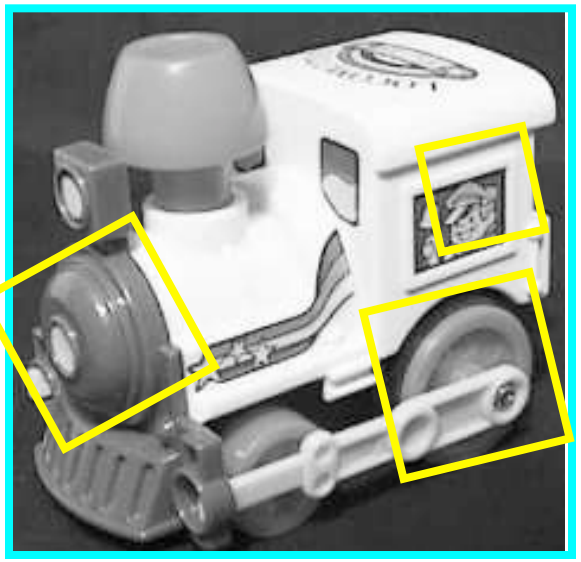


Model

Imagine test

Transformata Hough generalizată

- O ipoteză generată de o singură corespondență poate fi insuficientă;
- Fiecare corespondeță **votează** pentru a ipoteză în spațiul Hough asociat



Model

Imagine test

Detalii de implementare (Lowe'04)

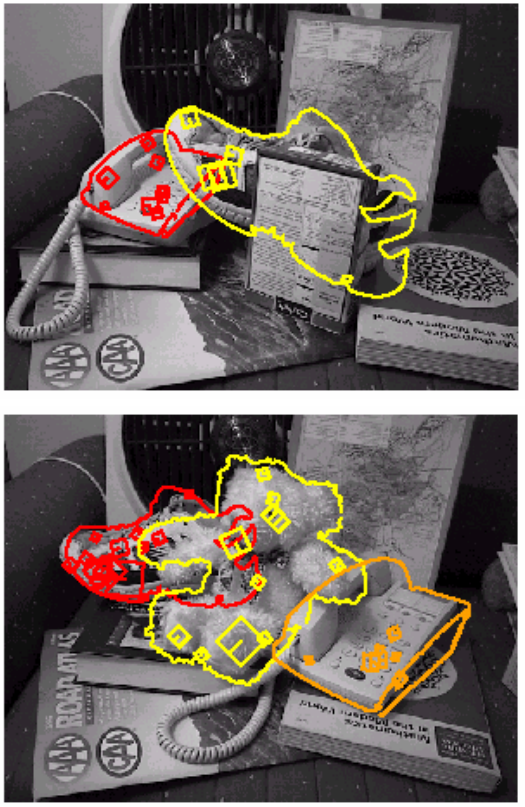
- **Antrenare:** pentru fiecare trăsătură a modelului, înregistrează poziția în 2D, mărimea și orientarea modelului (relativ la un cadru normalizat − bounding-box)
- **Testare:** fiecare trăsătură SIFT găsită drept corespondeță în imaginea test votează într-un spațiu Hough 4D asociat
  - Intervale de 30 de grade în orientare, factor 2 pentru mărime, 0.25 *diagonala imaginii pentru poziție
  - O trăsătură votează pentru cele mai apropiate 2 intervale în fiecare dimensiune (16 voturi)
- Găsește toate punctele din matricea de acumulare Hough care au cel puțin 3 puncte și realizează o verificare geometrică
  - Estimează o trasformare afină

David G. Lowe. **"Distinctive image features from scale-invariant keypoints."** *IJCV* 60 (2), pp. 91-110, 2004.

Rezultate (Lowe'04)

Background extras

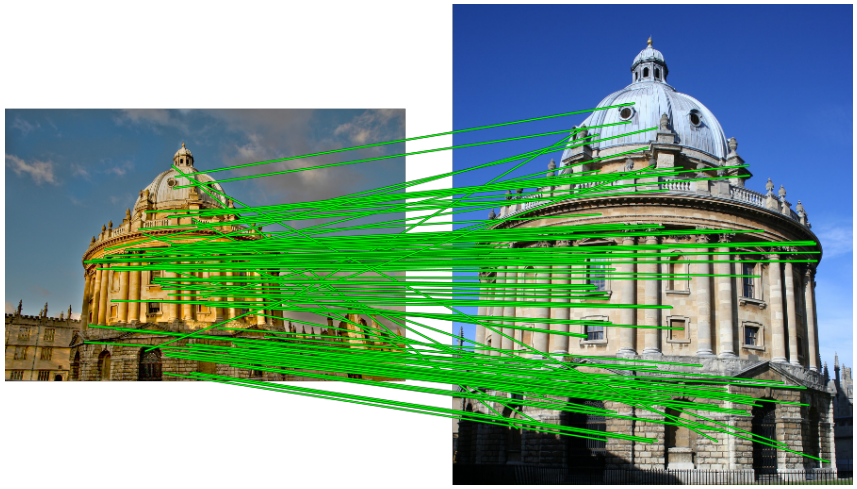
Obiecte găsite

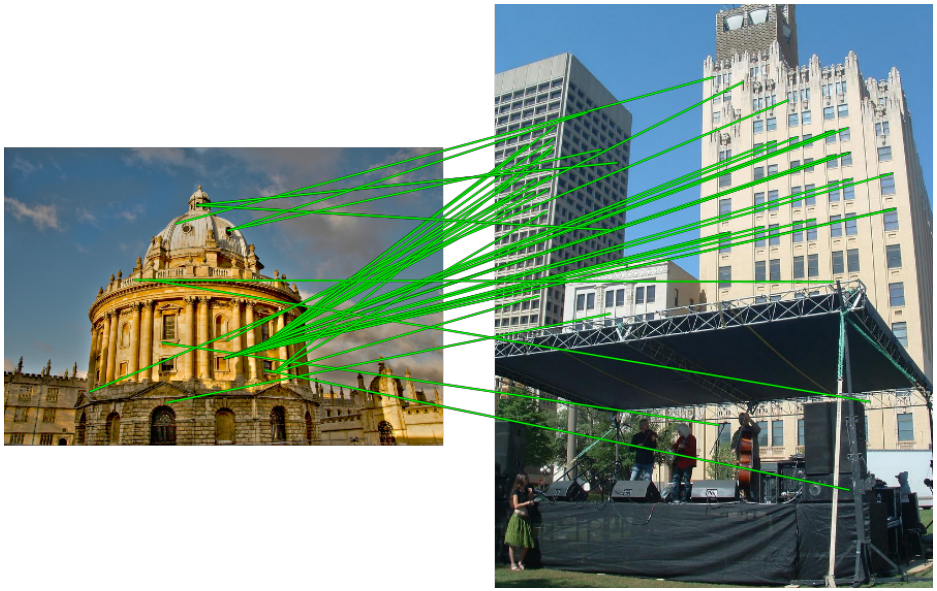


Obiecte găsite chiar dacă sunt mascate

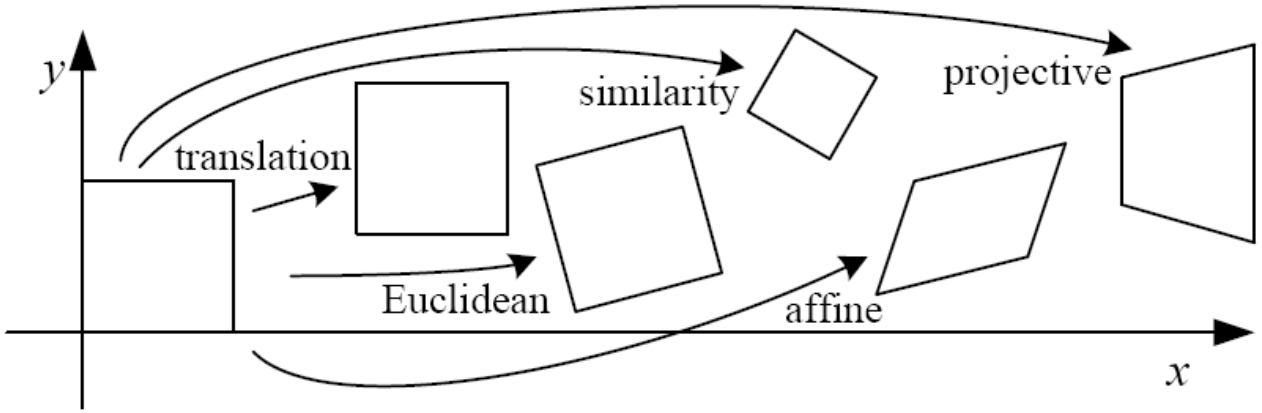
[Lowe]

Verificare cu RANSAC



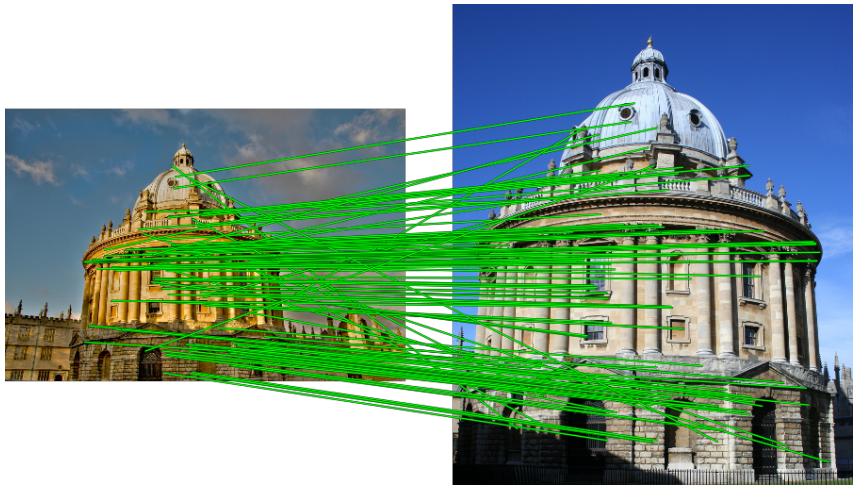


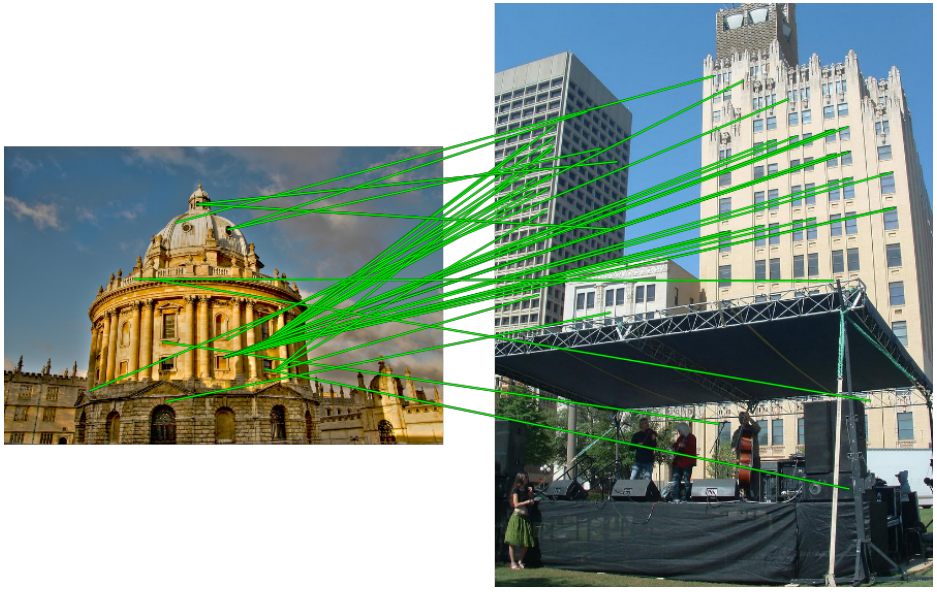
Transformări 2D

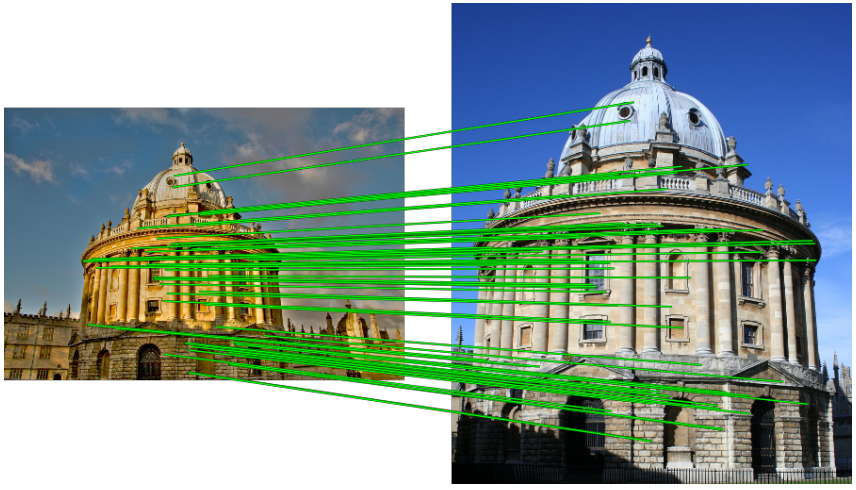


Name	Matrix	# D.O.F.	Preserves:	Icon
translation	$\left[\begin{array}{c\|c} \boldsymbol{I} & \boldsymbol{t} \end{array}\right]_{2\times 3}$	2	orientation $+\cdots$	
rigid (Euclidean)	$\left[\begin{array}{c\|c} \boldsymbol{R} & \boldsymbol{t} \end{array}\right]_{2\times 3}$	3	lengths $+\cdots$	
similarity	$\left[\begin{array}{c\|c} s\boldsymbol{R} & \boldsymbol{t} \end{array}\right]_{2\times 3}$	4	angles $+\cdots$	
affine	$\left[\begin{array}{c} \boldsymbol{A} \end{array}\right]_{2\times 3}$	6	parallelism $+\cdots$	
projective	$\left[\begin{array}{c} \tilde{\boldsymbol{H}} \end{array}\right]_{3\times 3}$	8	straight lines	

Verificare cu RANSAC







Probleme

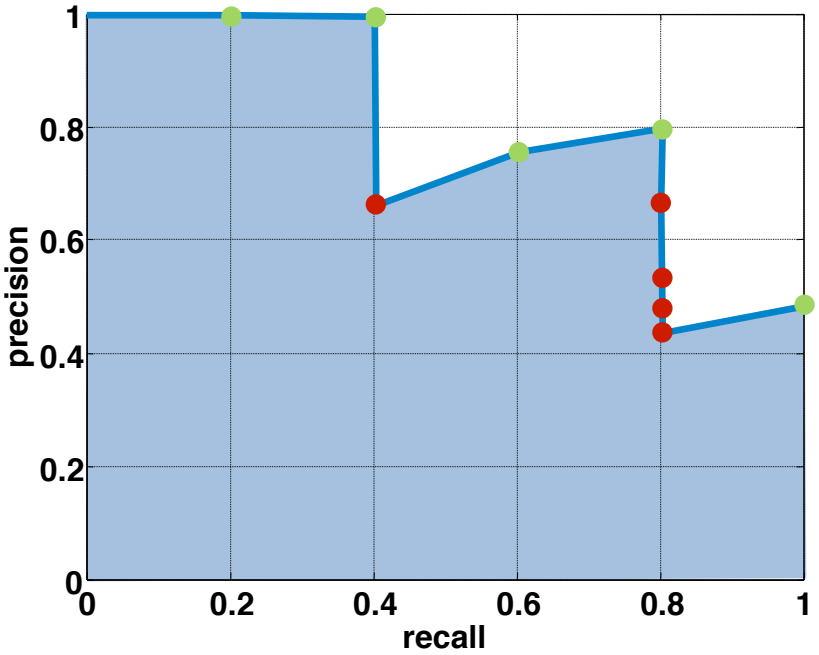
- Cum sumarizăm conținutul vizual al unei imagini? Putem compara conținutul vizual a două imagini?
- Cât de mare ar trebui să fie vocabularul vizual? Cum ar trebui realizată eficient cuantizarea spațiului de trăsături?
- Este suficient ca două regiuni să aibă aceeași mulțime de cuvinte vizuale pentru a putea recunoaște obiecte/scene specifice? Cum verificăm spațial acest lucru?
- Cum măsurăm rezultatele regăsirii pe baza unei imagini query?

Cum cuantificăm rezultatul?

Imagine query

Baza de date: 10 imagini
Relevante (total): 5 imagini

precizie = #relevante / #returnate
recall = #relevante / #total relevante



Rezultate (ordonate):