

# Vedere Artificială (Computer Vision)

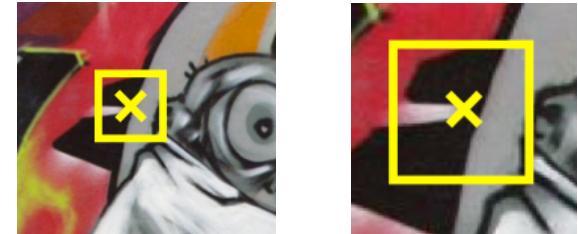
Bogdan Alexe

[bogdan.alexe@fmi.unibuc.ro](mailto:bogdan.alexe@fmi.unibuc.ro)

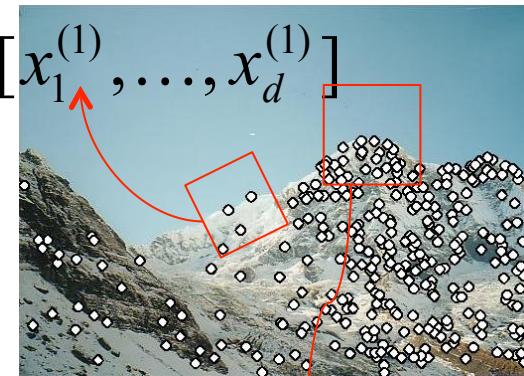
anul 2, master Informatică, semestrul I, 2019-2020

# Recapitulare – cursul trecut

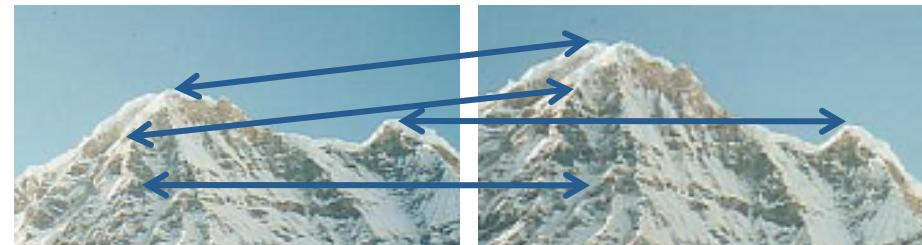
- 1) **Detectare:** detectori de puncte de interes invariante la scală: detectorul DoG, Harris-Laplace



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



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



# Comparație între detectoare de puncte de interes

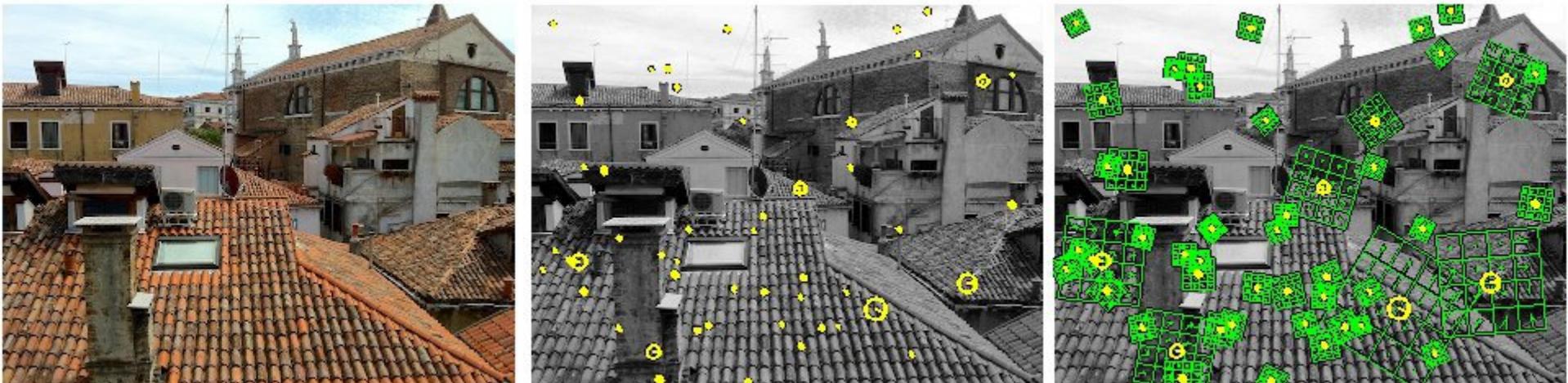
Table 7.1 Overview of feature detectors.

Feature Detector	Corner	Blob	Region	Rotation invariant	Scale invariant	Affine invariant	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		✓		✓	(✓)	(✓)	+	+	+	+

# 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  $16 \times 16$  pixeli în  $4 \times 4$  blocuri = 16 blocuri
  - calculează pentru fiecare bloc o histogramă de orientări ale pixelilor (intervalul  $0-360^0$  împărțit în 8 intervale)
  - valoare maximă a gradientilor 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  
(selectie aleatoare de 50  
de puncte)

Descriptori SIFT

# 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 2<sup>nd</sup> 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

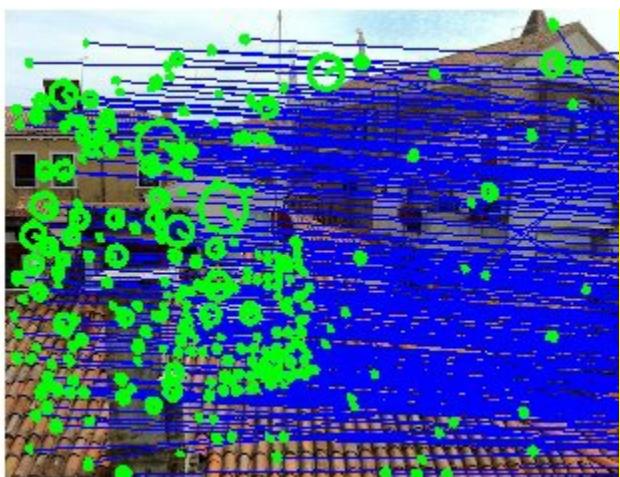
img1



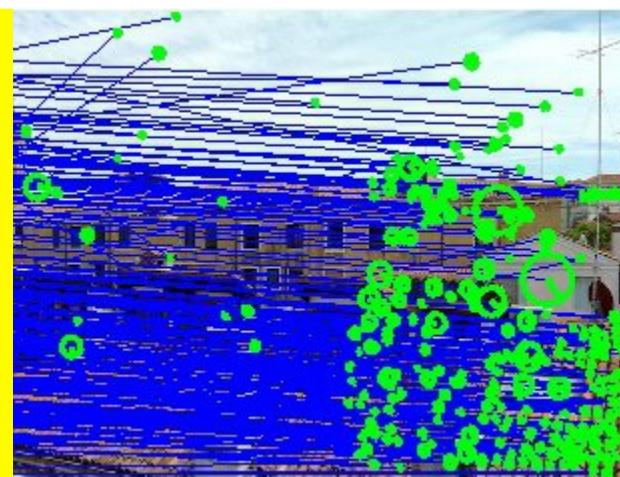
img2



img1

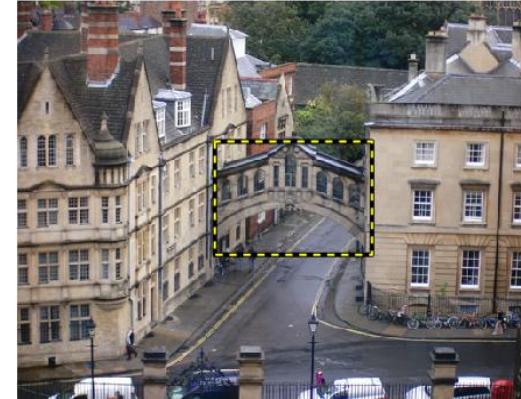
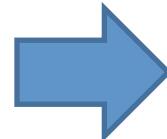
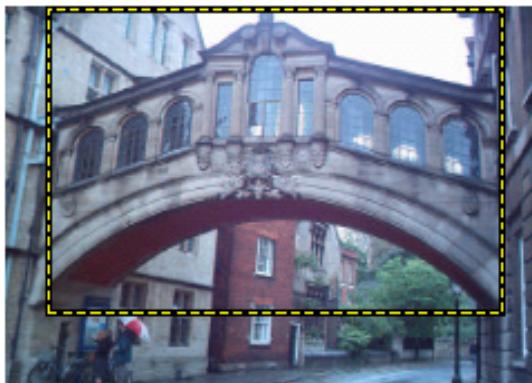


img2



# Planul cursului de azi

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



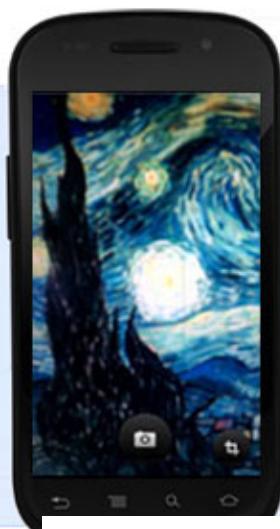


# Google Goggles

Use pictures to search the web.

[Watch a video](#)

<https://www.youtube.com/watch?v=Hhgfv0zPmH4>



## Get Google Goggles

Android (1.6+ required)

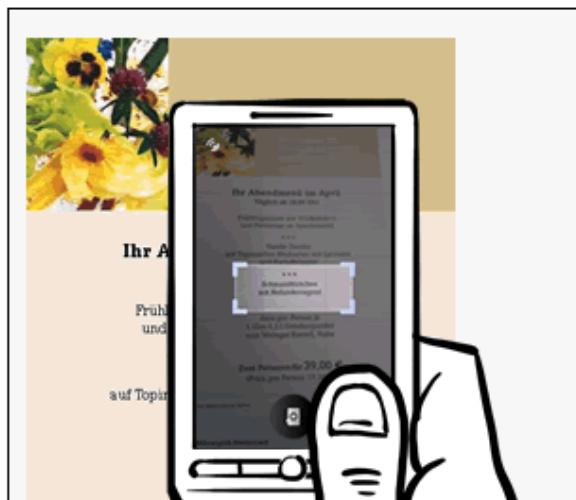
Download from [Android Market](#).

[Send Goggles to Android phone](#)

New! iPhone (iOS 4.0 required)

Download [from the App Store](#).

[Send Goggles to iPhone](#)



# Google Goggles (2009)



# Google Lens (2018)



# Recunoașterea sau regăsirea obiectelor specifice

Examplu I: căutare vizuală în video-uri

Interogare definită vizual

“Găsește acest ceas”



“Groundhog Day” [Rammis, 1993]



“Găsește acest loc”



# 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

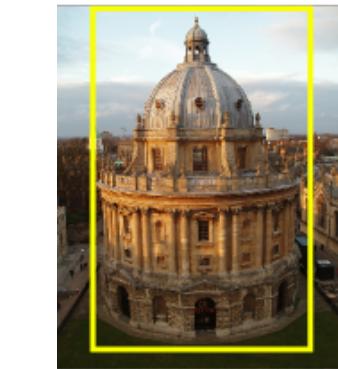
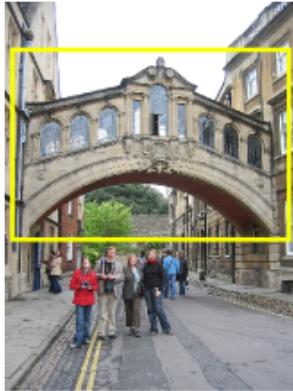
*We describe an approach to object and scene retrieval which searches for and localizes all the occurrences of a user outlined object in a video. The object is represented by a set of viewpoint invariant region descriptors so that recognition can proceed successfully despite changes in viewpoint, illumination 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.*

# Recunoașterea sau regăsirea obiectelor specifice

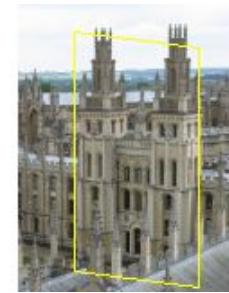
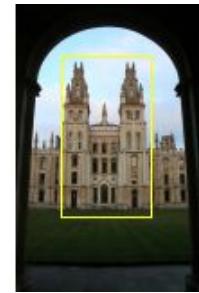
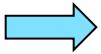
Examplu II: caută imagini pe web după locuri particulare



Găsește aceste locuri ...într-o bază de date cu imagini mare (1M) (landmarks)

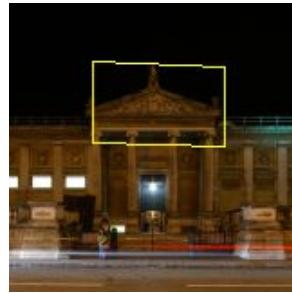
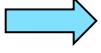
# Dificultate

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

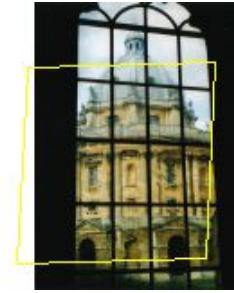


Scale

Viewpoint



Lighting



Occlusion

# Cursul trecut: Corespondențe pentru trăsături locale



Imagine 1

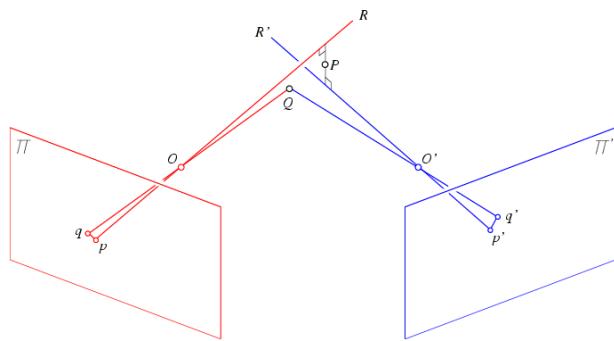


Imagine 2

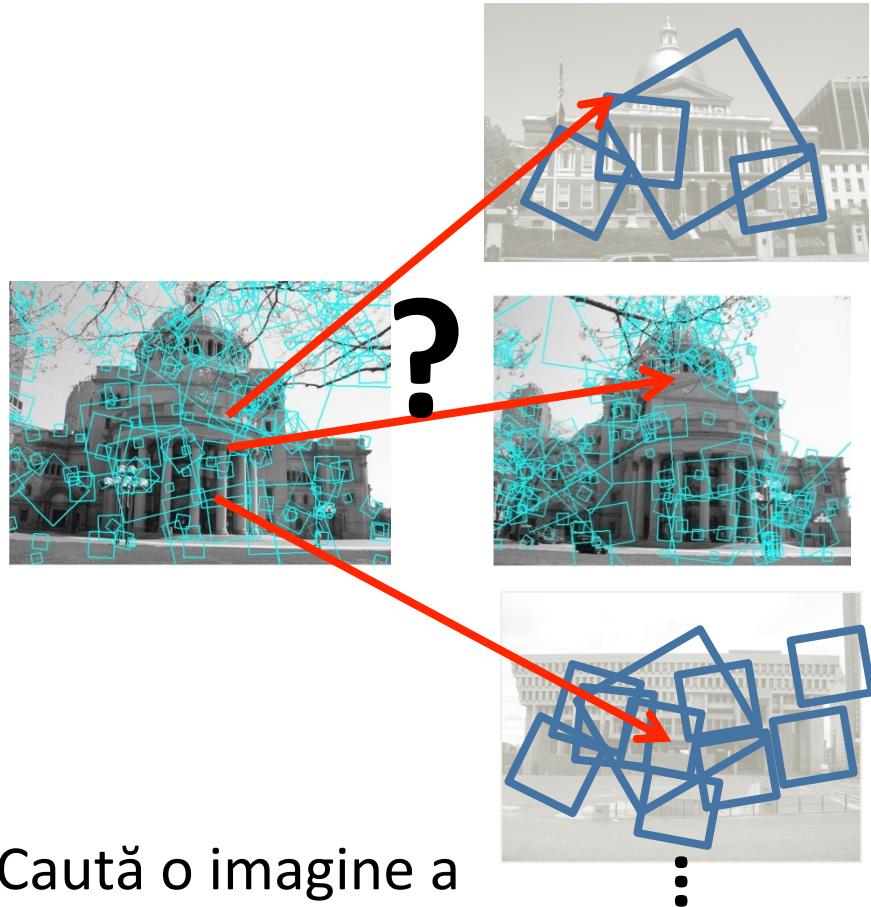
Pentru găsirea de trăsături locale corespondente, găsește patch-uri similare în înfățișare – calculează o distanță

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

# Corespondențe între imagini



vs

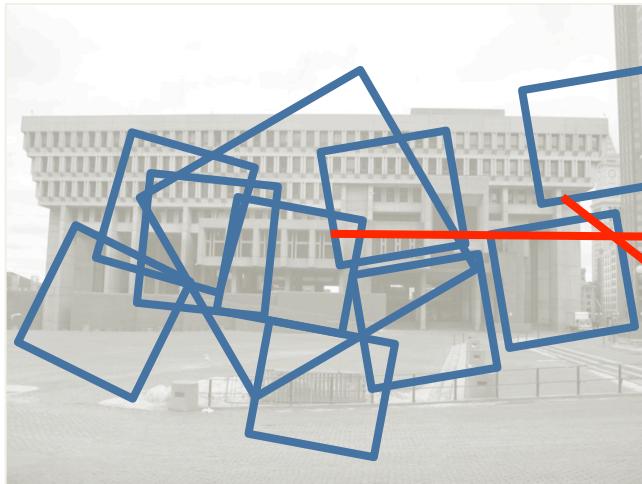


Corespondențe din  
mai multe imagini ale  
aceleiași scene

Caută o imagine a  
aceleiași scene

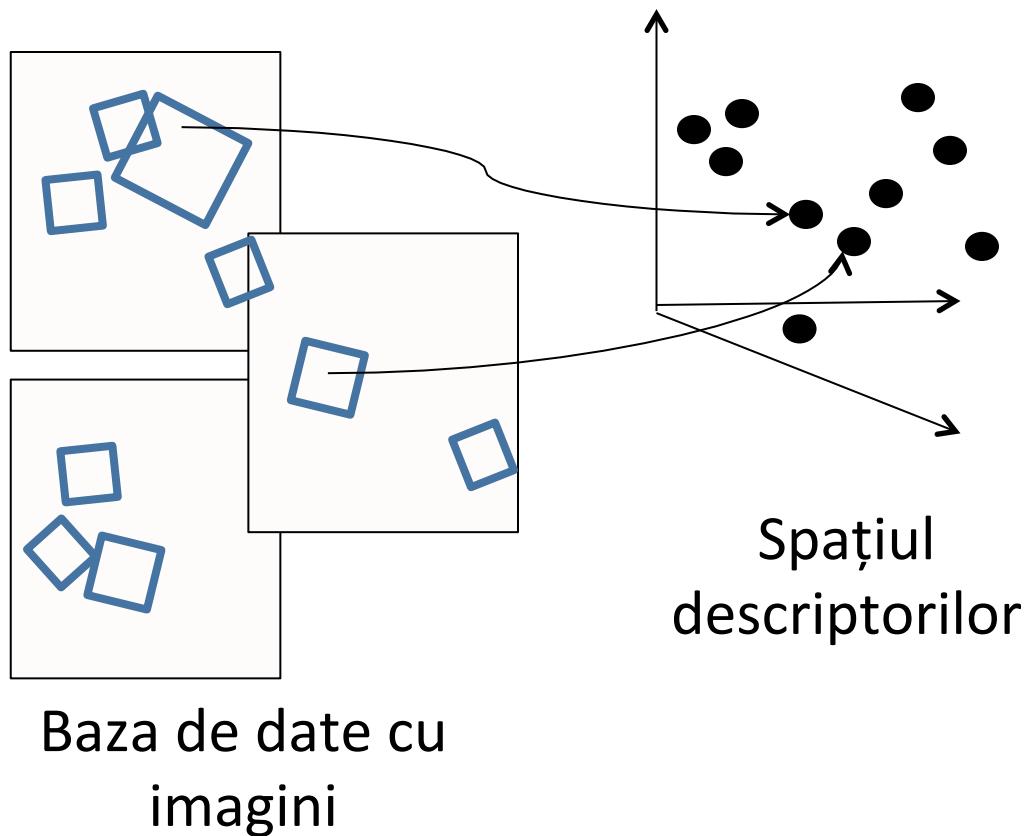
*Găsirea de corespondențe între trăsături locale invariante este folositoare și pentru recunoașterea de obiecte/scene.*

# Indexarea trăsăturilor locale



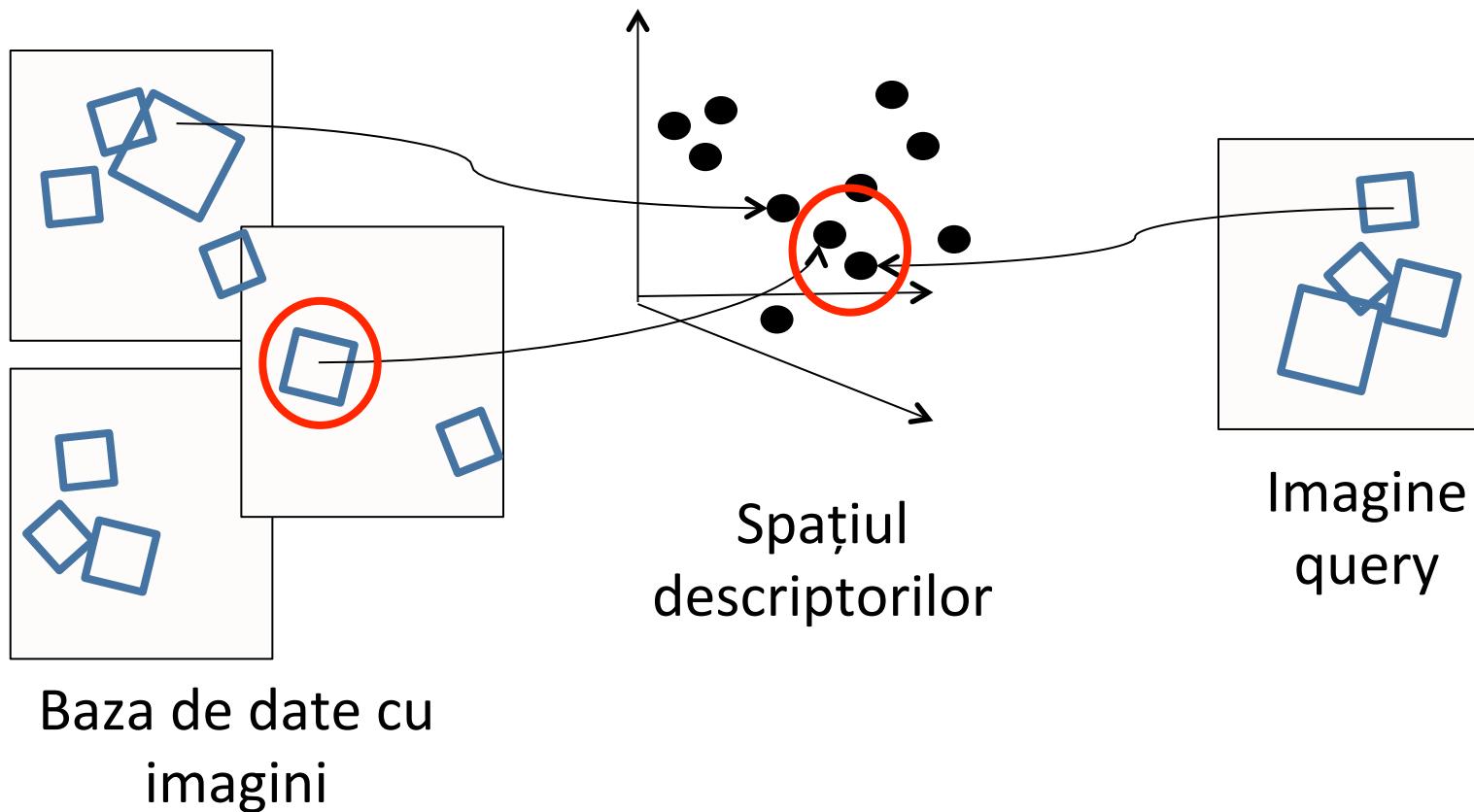
# Indexarea trăsăturilor locale

- 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

- Punțe 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 apropiati vecini
    - Kd-trees
    - Hashing

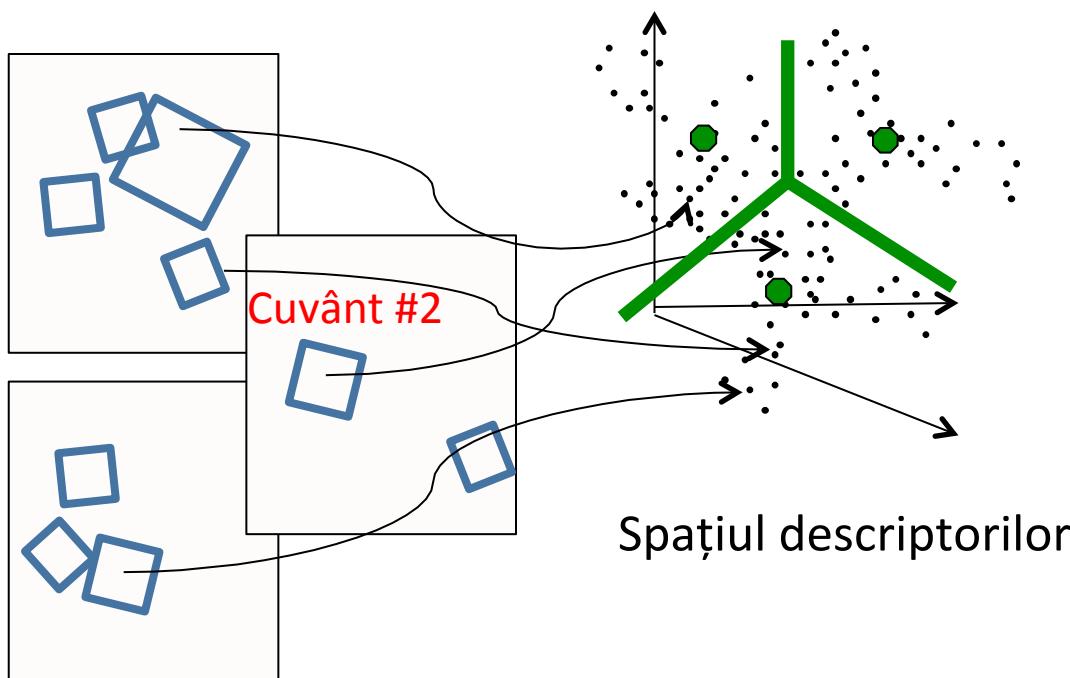
# Indexarea trăsăturilor locale: inverted file index

Index	
"Along I-75," From Detroit to Florida; <i>inside back cover</i>	
"Drive I-95," From Boston to Florida; <i>inside back cover</i>	
1929 Spanish Trail Roadway;	101-102,104
511 Traffic Information; 83	
A1A (Barrier Isl) - I-95 Access; 86	
AAA (and CAA); 83	
AAA National Office; 88	
Abbreviations,	
Colored 25 mile Maps; cover	
Exit Services; 196	
Travelogue; 85	
Africa; 177	
Agricultural Inspection Stns; 126	
Ah-Tah-Thi-Ki Museum; 180	
Air Conditioning, First; 112	
Alabama; 124	
Alachua; 132	
County; 131	
Alafia River; 143	
Alapaha, Name; 126	
Alfred B Macay Gardens; 106	
Alligator Alley; 154-155	
Alligator Farm, St Augustine; 169	
Alligator Hole (definition); 157	
Alligator, Buddy; 155	
Alligators; 100,135,138,147,156	
Anastasia Island; 170	
Anhica; 108-109,146	
Apalachicola River; 112	
Appleton Mus of Art; 136	
Aquifer; 102	
Arabian Nights; 94	
Art Museum, Ringling; 147	
Aruba Beach Cafe; 183	
Aucilla River Project; 106	
Babcock-Web WMA; 151	
Bahia Mar Marina; 184	
Baker County; 99	
Barefoot Mallmen; 182	
Barge Canal; 137	
Bee Line Expy; 80	
Belz Outlet Mall; 89	
Bernard Castro; 136	
Big "I"; 165	
Big Cypress; 155,158	
Big Foot Monster; 105	
Billie Swamp Safari; 160	
Blackwater River SP; 117	
Blue Angels	
Butterfly Center, McGuire; 134	
CAA (see AAA)	
CCC, The; 111,113,115,135,142	
Ca d'Zan; 147	
Caloosahatchee River; 152	
Name; 150	
Canaveral Natnl Seashore; 173	
Cannon Creek Airpark; 130	
Canopy Road; 106,160	
Cape Canaveral; 174	
Castillo San Marcos; 169	
Cave Diving; 131	
Cayo Costa, Name; 150	
Celebration; 93	
Charlotte County; 149	
Charlotte Harbor; 150	
Chautauqua; 116	
Chipley; 114	
Name; 115	
Choctawatchee, Name; 115	
Circus Museum, Ringling; 147	
Citrus; 88,97,130,136,140,180	
CityPlace, W Palm Beach; 180	
City Maps,	
Ft Lauderdale Expwys; 194-195	
Jacksonville; 163	
Kissimmee Expwys; 192-193	
Miami Expressways; 194-195	
Orlando Expressways; 192-193	
Pensacola; 26	
Tallahassee; 191	
Tampa-St. Petersburg; 63	
St. Augustine; 191	
Civil War; 100,108,127,138,141	
Clearwater Marine Aquarium; 187	
Collier County; 154	
Collier, Barron; 152	
Colonial Spanish Quarters; 168	
Columbia County; 101,128	
Coquina Building Material; 165	
Corkscrew Swamp, Name; 154	
Cowboys; 95	
Crab Trap II; 144	
Cracker, Florida; 88,95,132	
Crosstown Expy; 11,35,98,143	
Cuban Bread; 184	
Dade Battlefield; 140	
Dade, Maj. Francis; 139-140,161	
Daniel Beach Hurricane; 184	
Daniel Boone, Florida Walk; 117	
Daytona Beach; 172-173	
De Land; 87	
Driving Lanes; 85	
Duval County; 163	
Eau Gallie; 175	
Edison, Thomas; 152	
Eglin AFB; 116-118	
Eight Reale; 176	
Ellenton; 144-145	
Emanuel Point Wreck; 120	
Emergency Callboxes; 83	
Epiphytes; 142,148,157,159	
Escambia Bay; 119	
Bridge (I-10); 119	
County; 120	
Esterio; 153	
Everglade,90,95,139-140,154-160	
Draining of; 156,181	
Wildlife MA; 160	
Wonder Gardens; 154	
Falling Waters SP; 115	
Fantasy of Flight; 95	
Fayer Dykes SP; 171	
Fires, Forest; 166	
Fires, Prescribed; 148	
Fisherman's Village; 151	
Flagler County; 171	
Flagler, Henry; 97,165,167,171	
Florida Aquarium; 186	
Florida,	
12,000 years ago; 187	
Cavern SP; 114	
Map of all Expressways; 2-3	
Mus of Natural History; 134	
National Cemetery ; 141	
Part of Africa; 177	
Platform; 187	
Sheriff's Boys Camp; 126	
Sports Hall of Fame; 130	
Sun 'n Fun Museum; 97	
Supreme Court; 107	
Florida's Turnpike (FTP); 178,189	
25 mile Strip Maps; 66	
Administration; 189	
Coin System; 190	
Exit Services; 189	
HEFT; 76,161,190	
History; 189	
Names; 189	
Service Plazas; 190	
Spur SR91; 76	
Ticket System; 190	
Toll Plazas; 190	
Ford, Henry; 152	

- 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

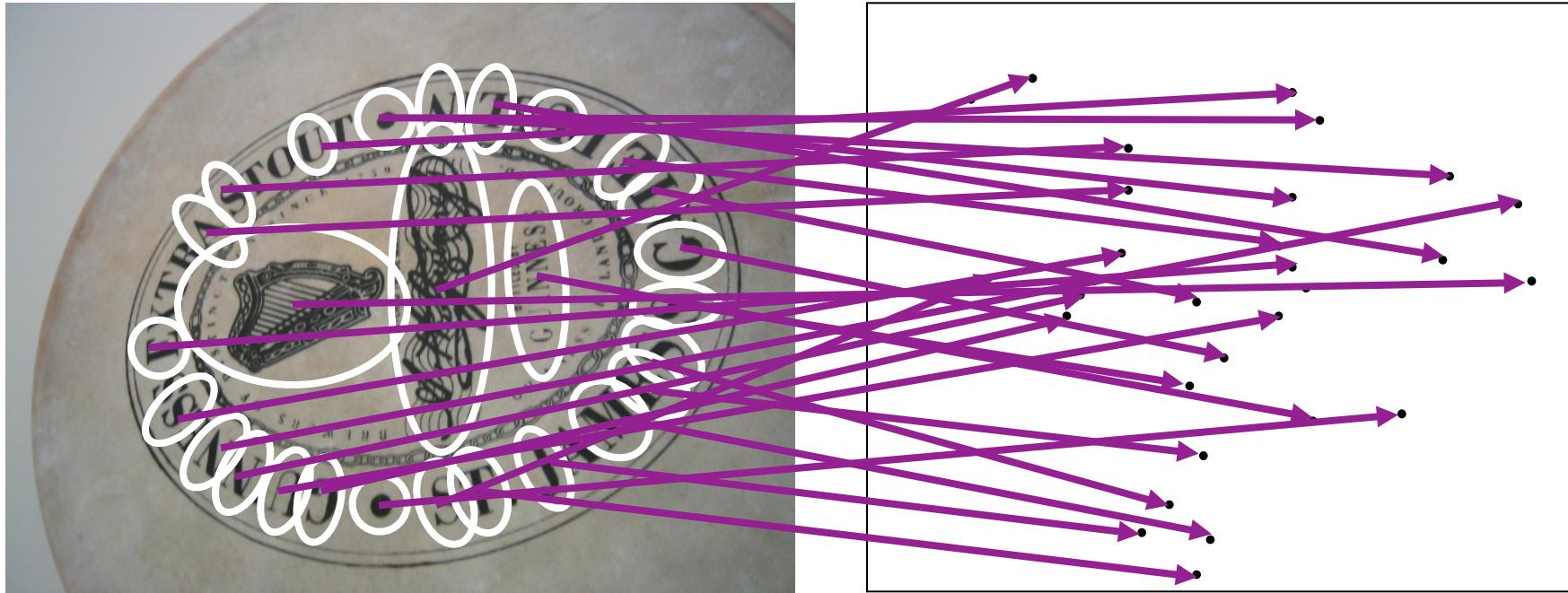
- Mapăm descriptorii în token-uri (cuvinte vizuale) partaționând spațiul descriptorilor



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

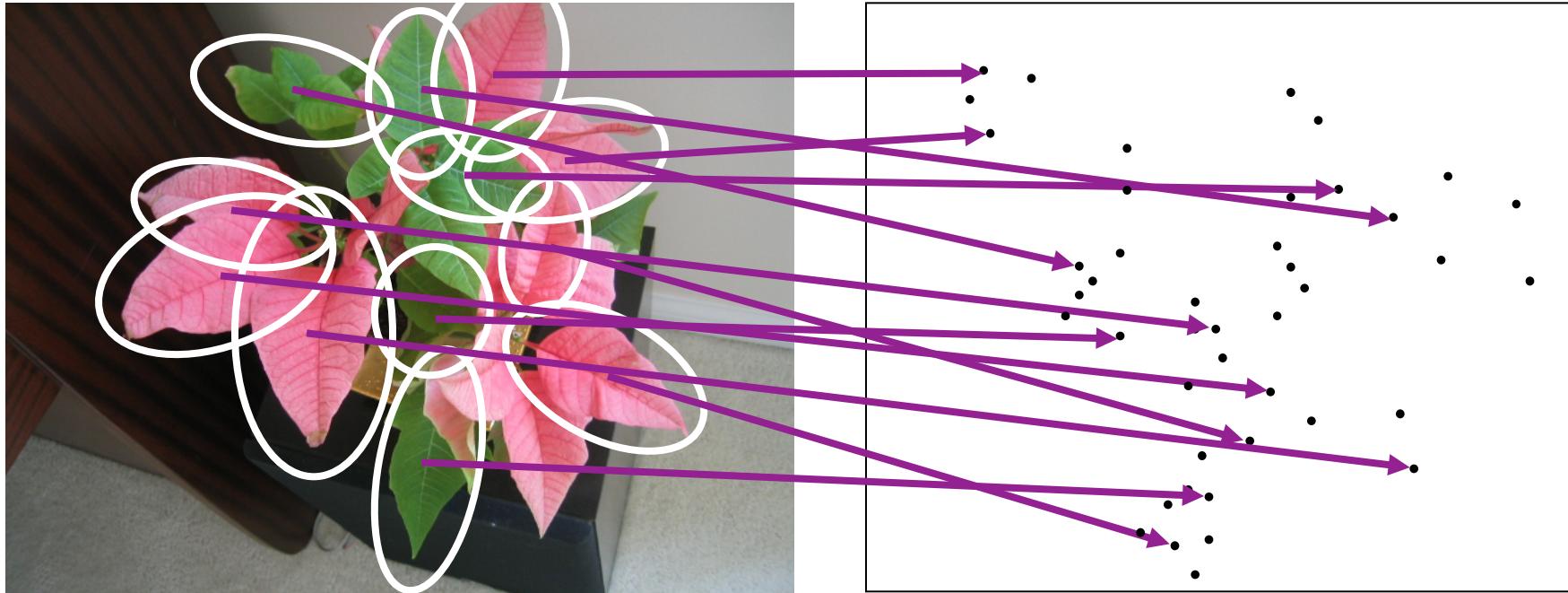
# Cuvinte vizuale: ideea de bază

- Extragem caracteristici locale dintr-un număr de imagini...

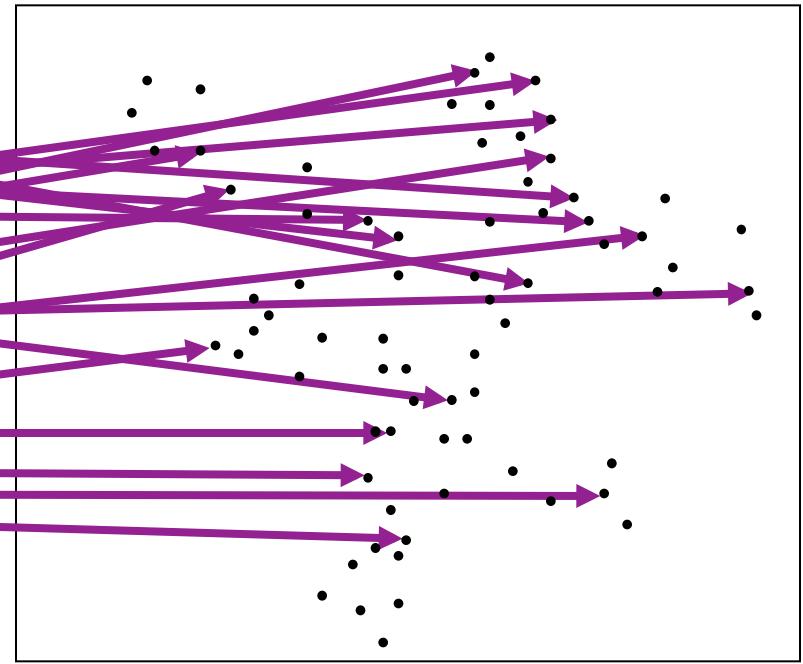
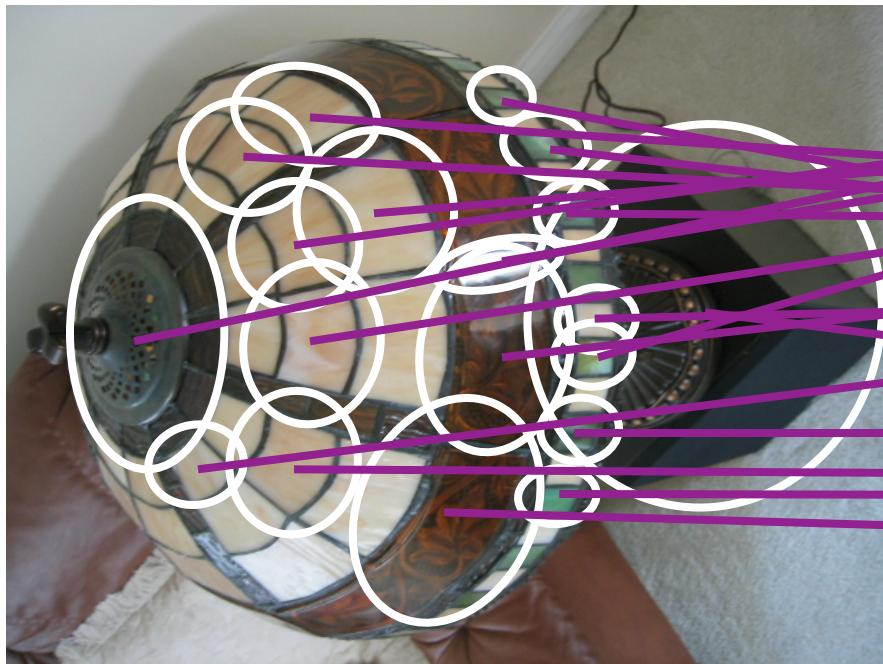


Spre exemplu, descriptori SIFT de dimensiune 128

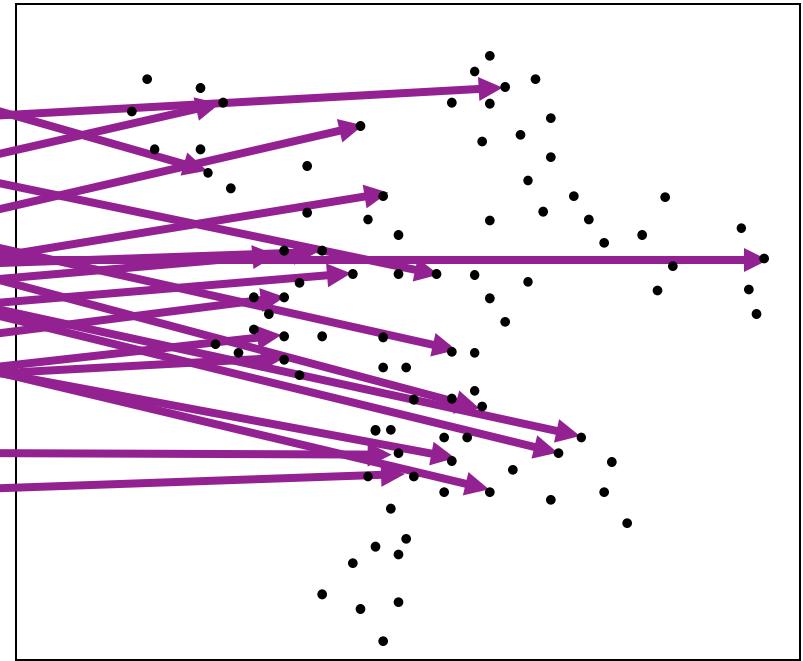
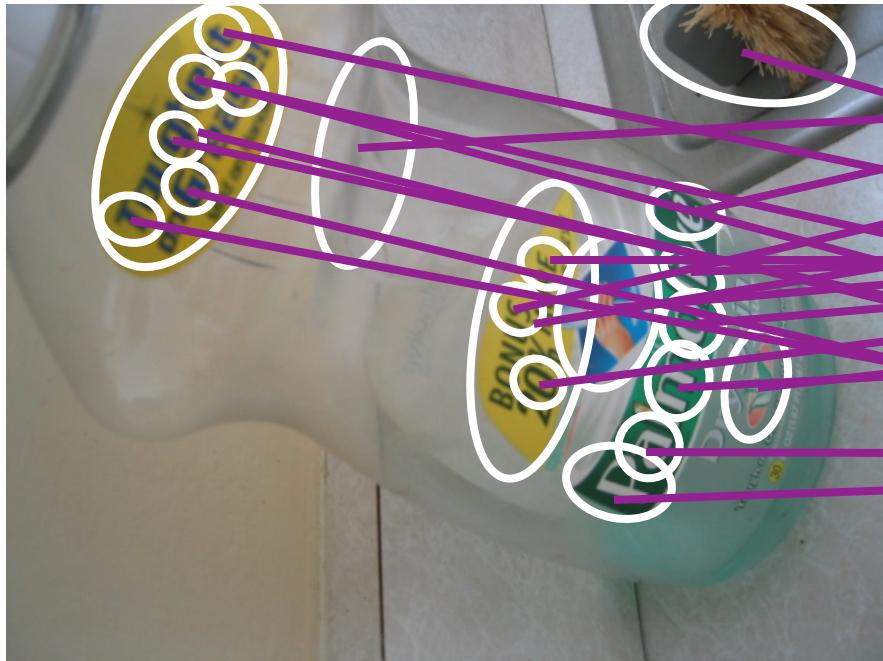
# Cuvinte vizuale: ideea de bază



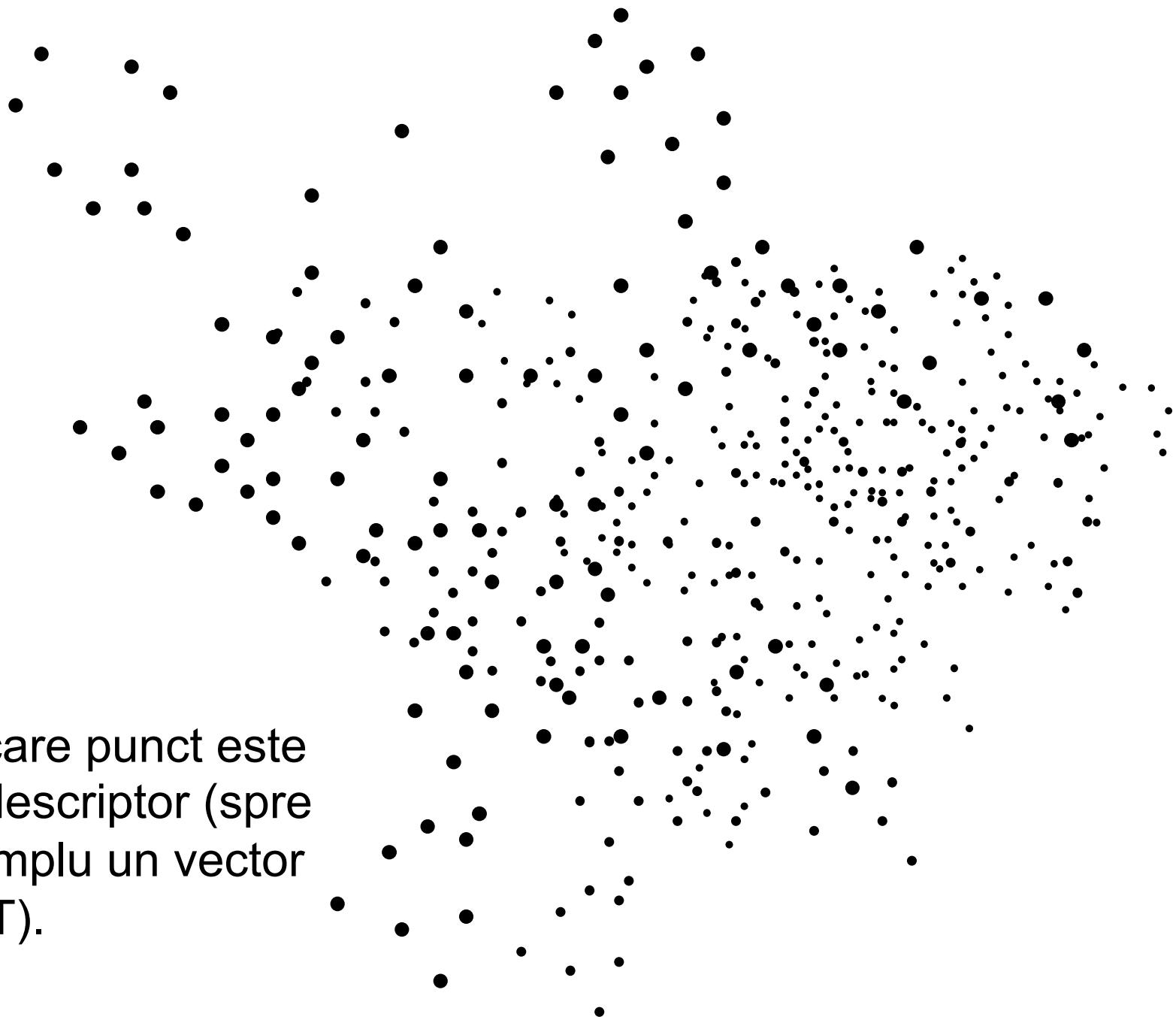
# Cuvinte vizuale: ideea de bază

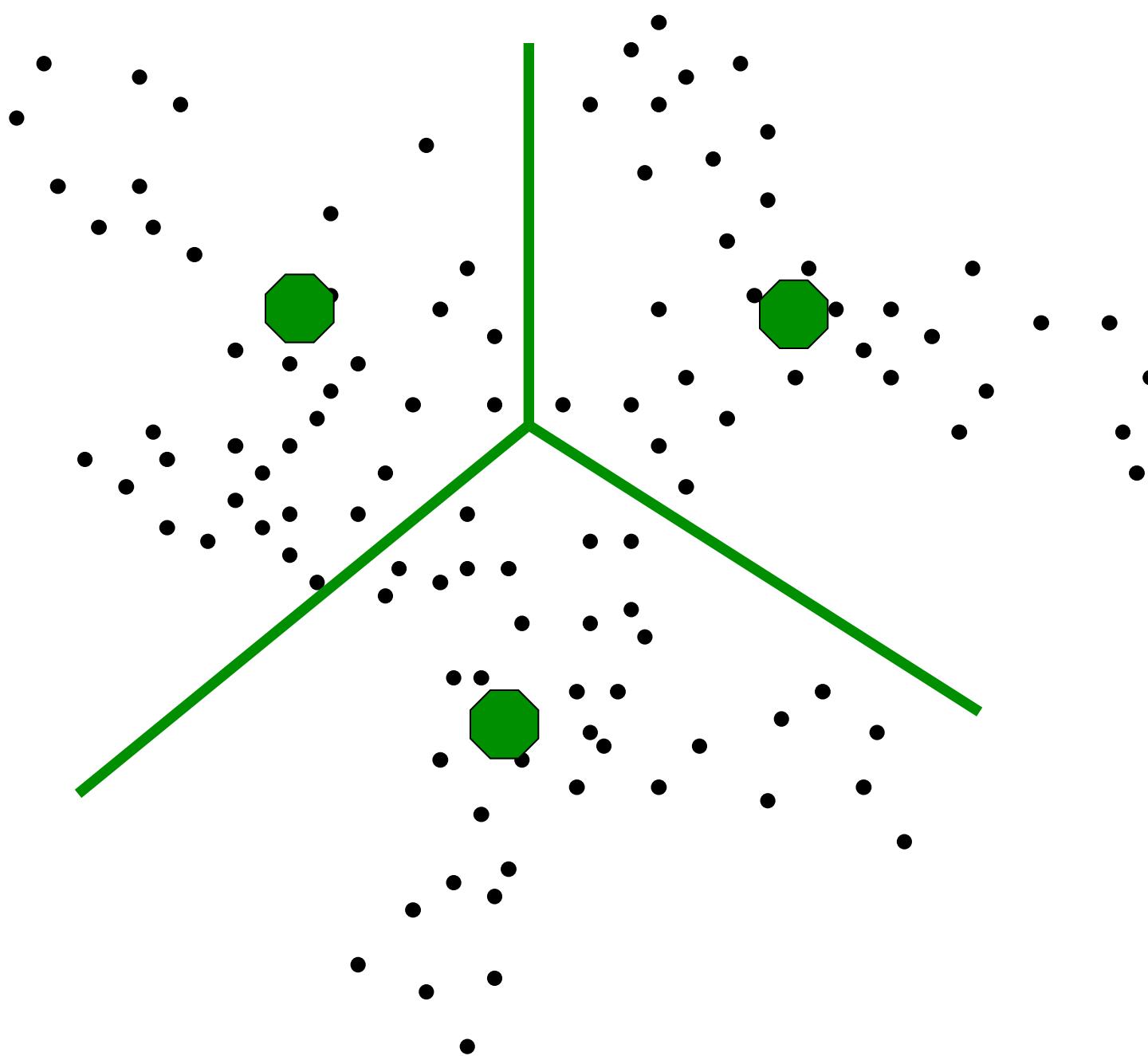


# Cuvinte vizuale: ideea de bază



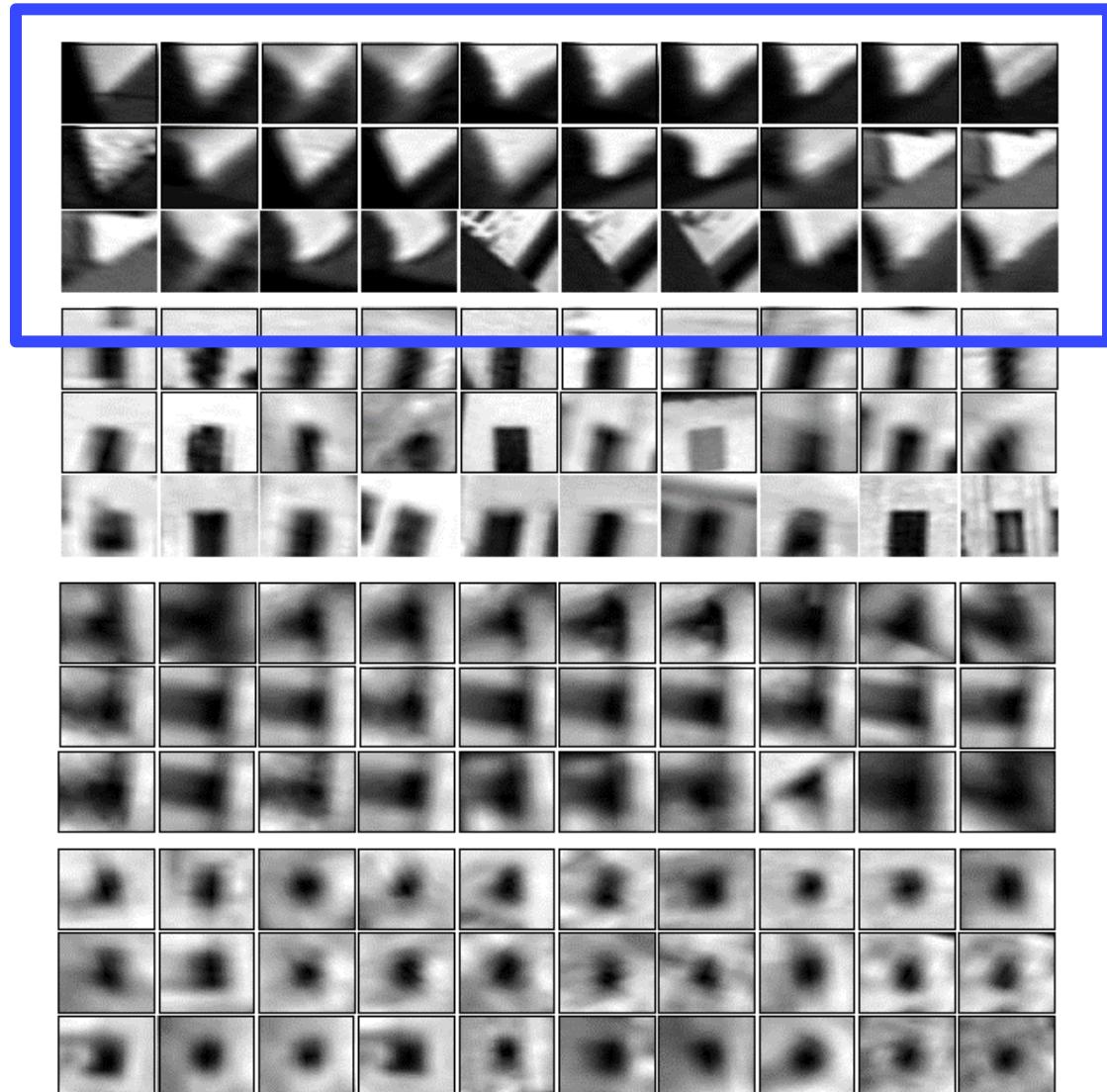
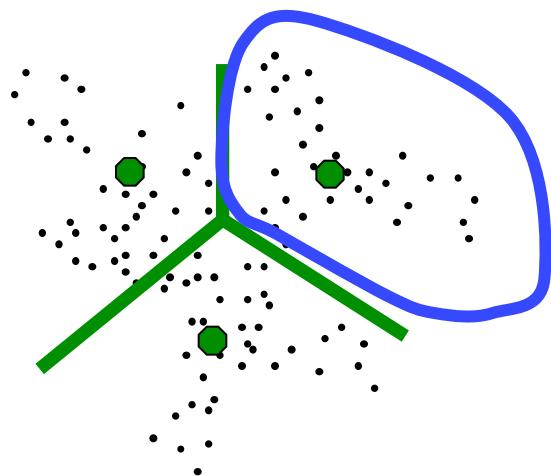
Fiecare punct este  
un descriptor (spre  
exemplu un vector  
SIFT).





# Cuvinte vizuale

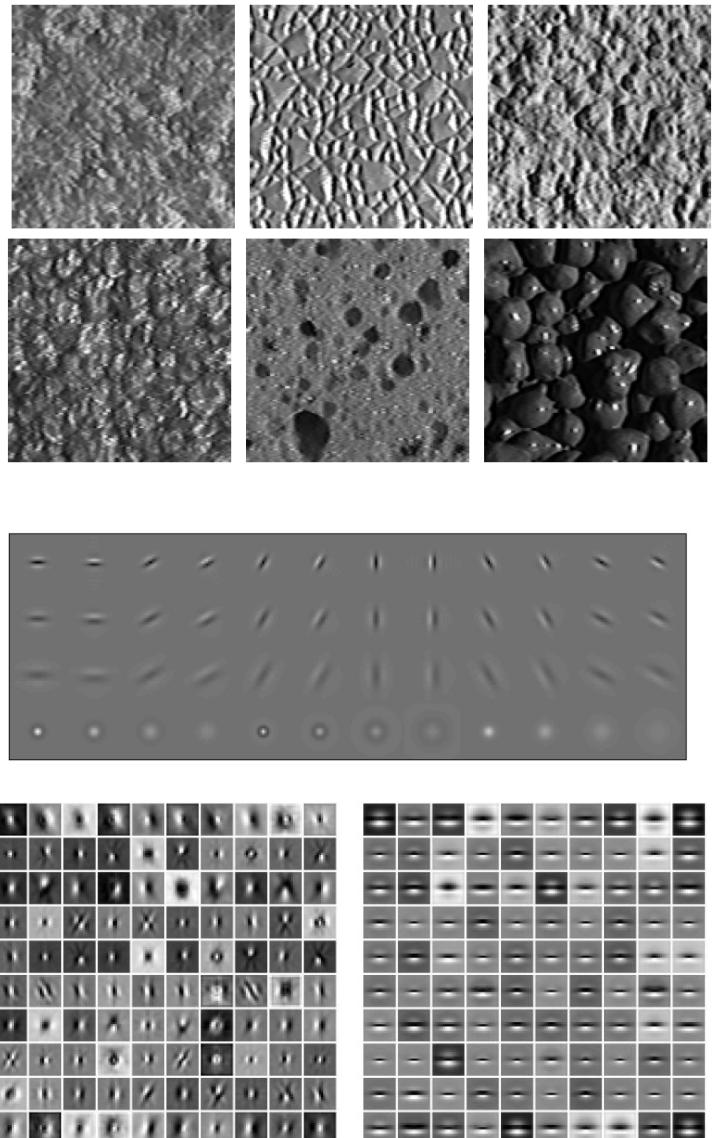
- Exemple: fiecare grup de patch-uri aparține aceluiași cuvânt vizual



Figură din articolul Sivic & Zisserman, ICCV 2003

# Cuvinte vizuale și textoni

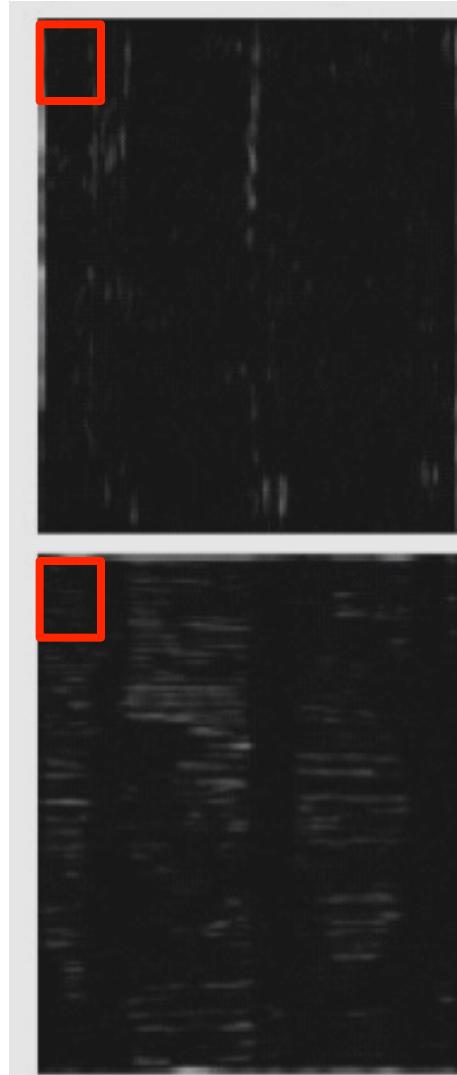
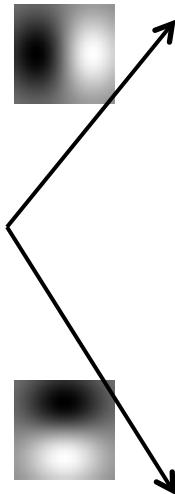
- Cuvintele vizuale au fost explorate întâi din perspectiva reprezentării texturii și a materialelor
- *Texton* = centrul unui cluster de răspunsuri obținute pe baza filtrării unor imagini
- Descrie textura și materialele pe baza distribuției textonilor



# Reprezentarea texturii: exemplu



imagine inițială



imagini filtrate  
(filtre pentru derive)

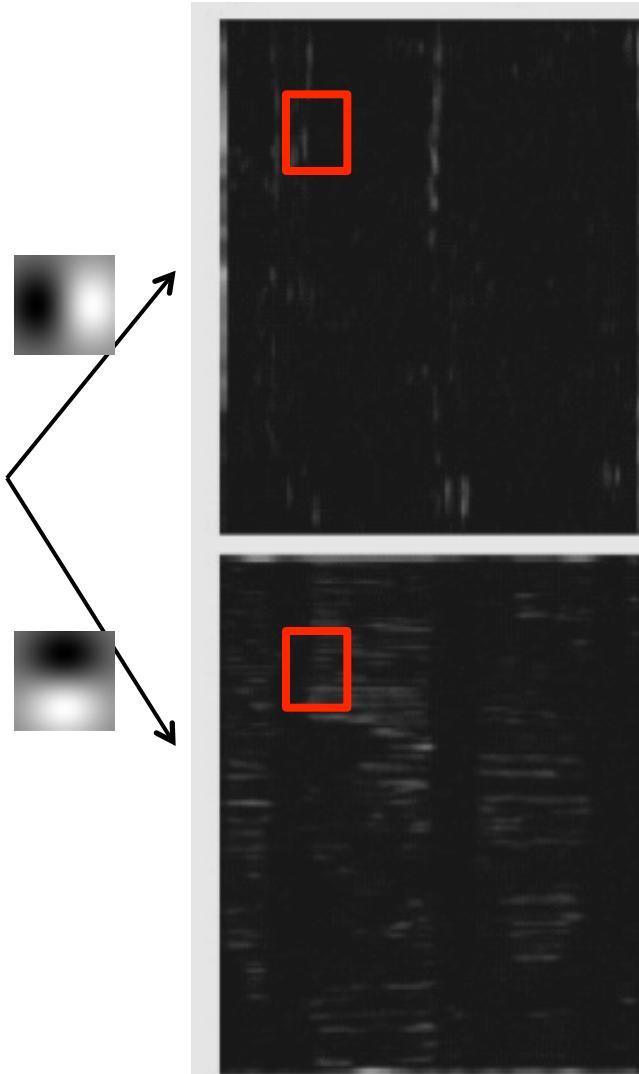
	<u>media</u> $d/dx$	<u>media</u> $d/dy$
fereastra #1	4	10
⋮	⋮	⋮

Statistică pentru a  
descrie pattern-urile în  
ferestre mici

# Reprezentarea texturii: exemplu



imagine inițială



imagini filtrate  
(filtre pentru deriveate)

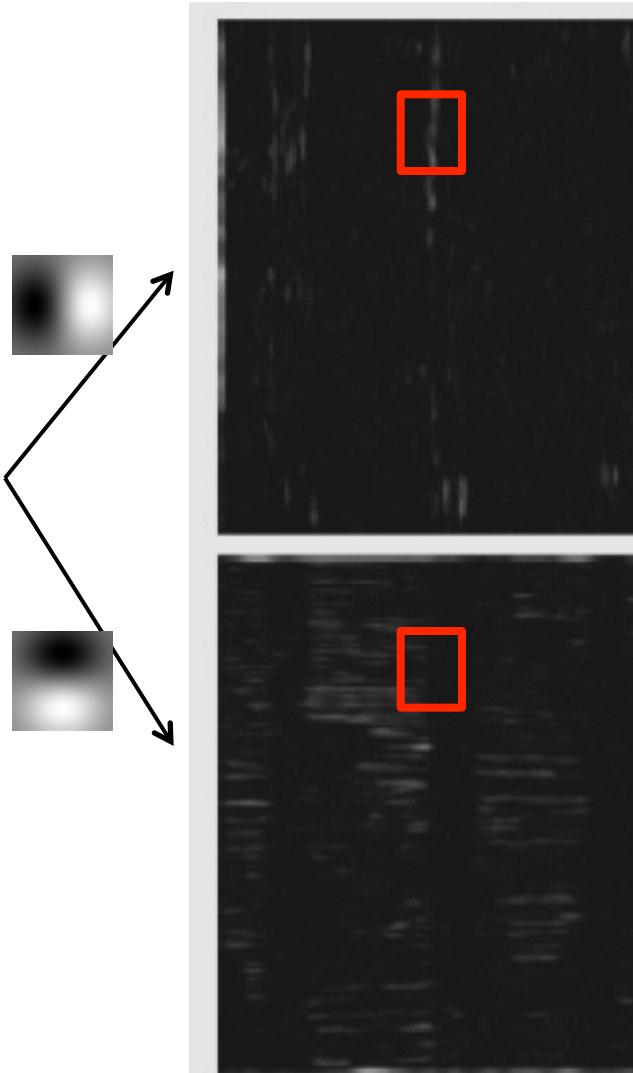
	<u>media</u> $d/dx$	<u>media</u> $d/dy$
fereastra #1	4	10
fereastra #2	25	7
⋮	⋮	⋮

Statistică pentru a  
descrie pattern-urile în  
ferestre mici

# Reprezentarea texturii: exemplu



Imagine inițială



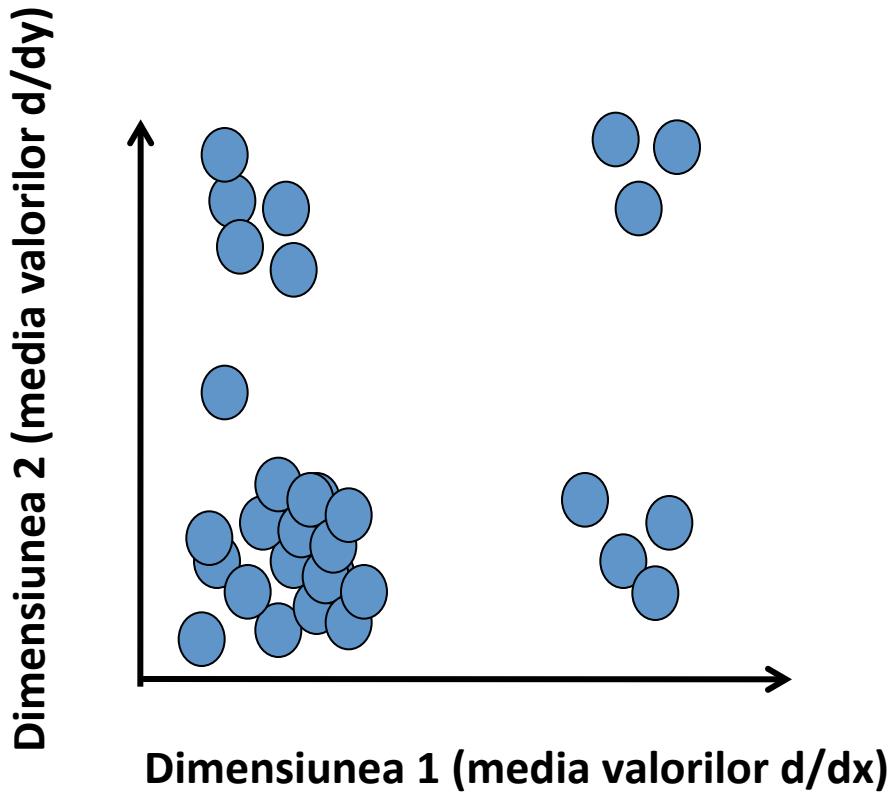
imagini filtrate  
(filtre pentru derive)

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

⋮

Statistici pentru a  
descrie pattern-urile în  
ferestre mici

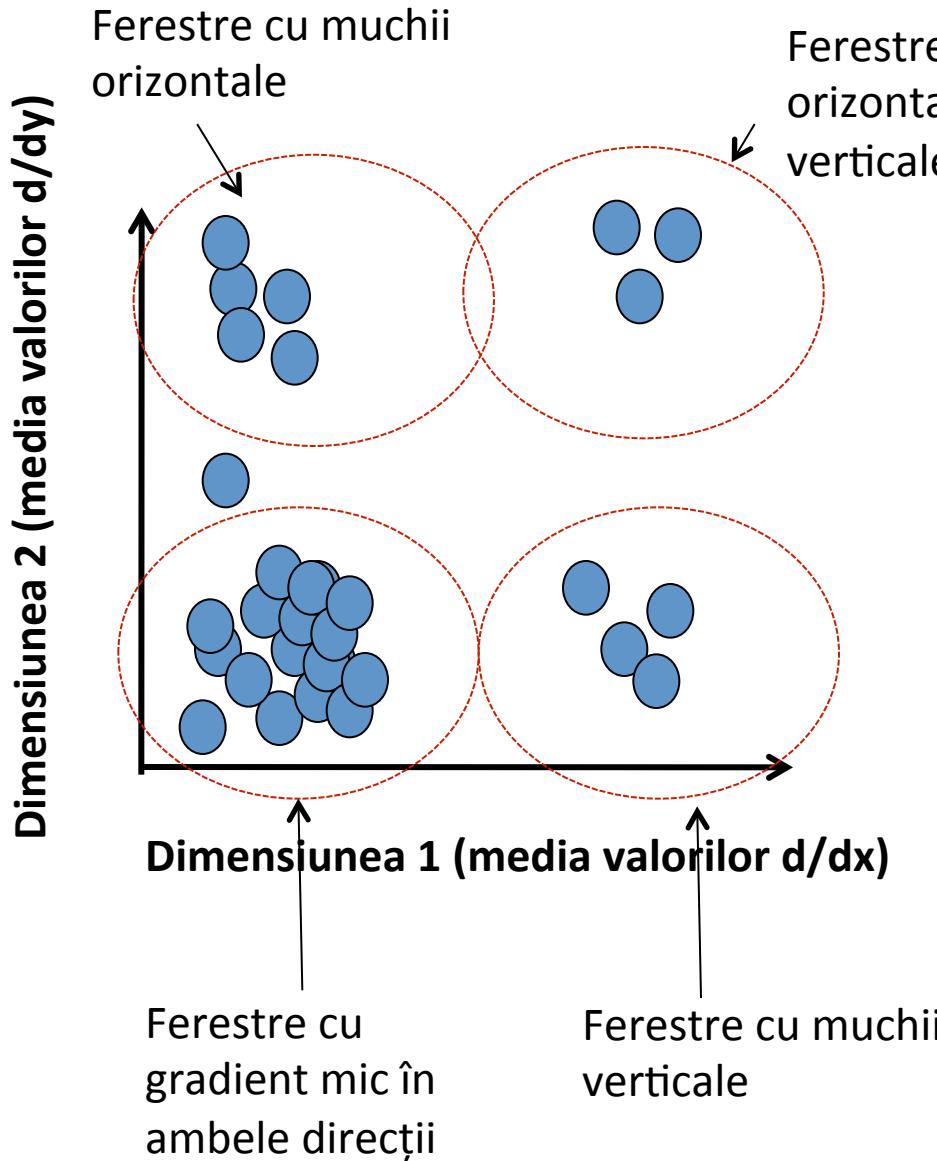
# Reprezentarea texturii: exemplu



	<u>media</u> $d/dx$	<u>media</u> $d/dy$
fereastra #1	4	10
fereastra #2	25	7
fereastra #3	18	20
⋮		

Statistici pentru a  
descrie pattern-urile în  
ferestre mici

# Reprezentarea texturii: exemplu



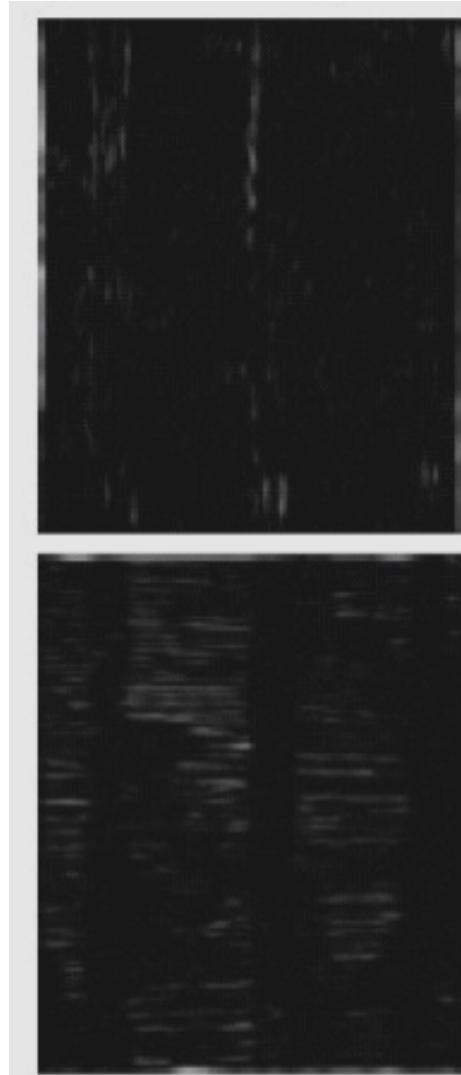
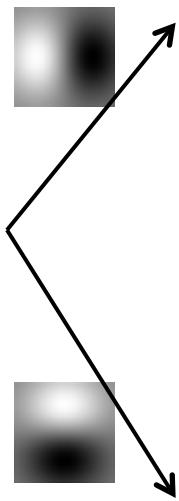
	<u>media</u> <u><math>d/dx</math></u>	<u>media</u> <u><math>d/dy</math></u>
fereastra #1	4	10
fereastra #2	25	7
fereastra #3	18	20
		⋮

Statistică pentru a descrie pattern-urile în ferestre mici

# Reprezentarea texturii: exemplu



Imagine inițială

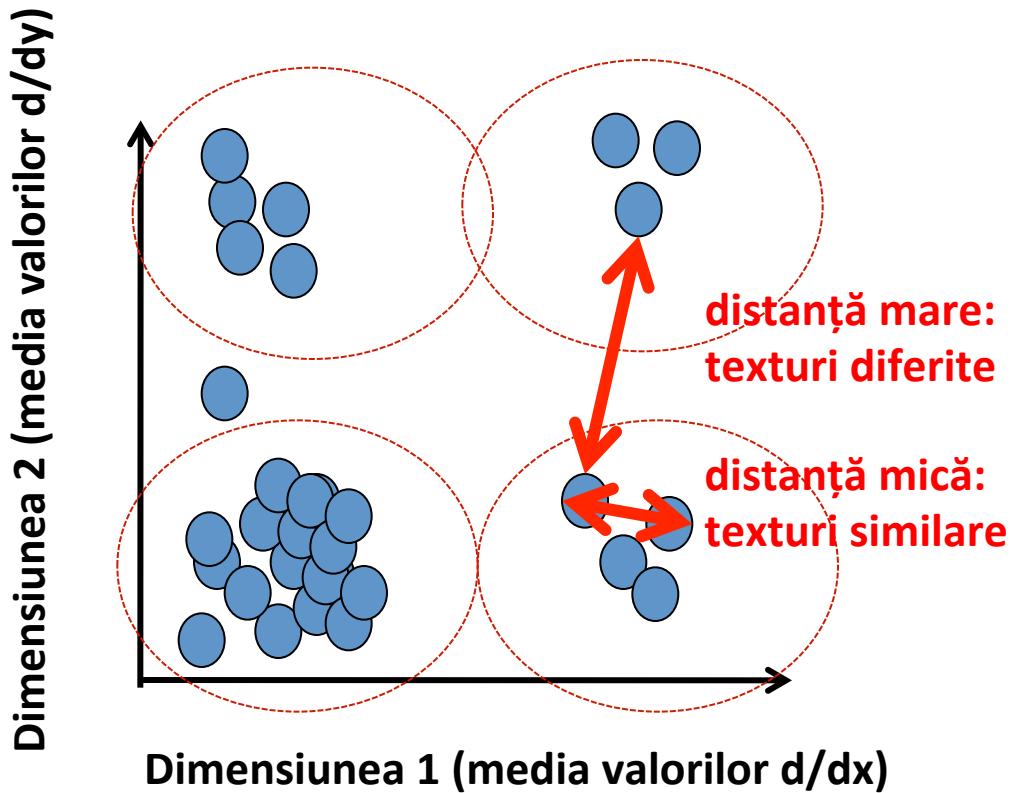


imagini filtrate  
(filtre pentru deriveate)



Vizualizare a tipurilor de textură.

# Reprezentarea texturii: exemplu



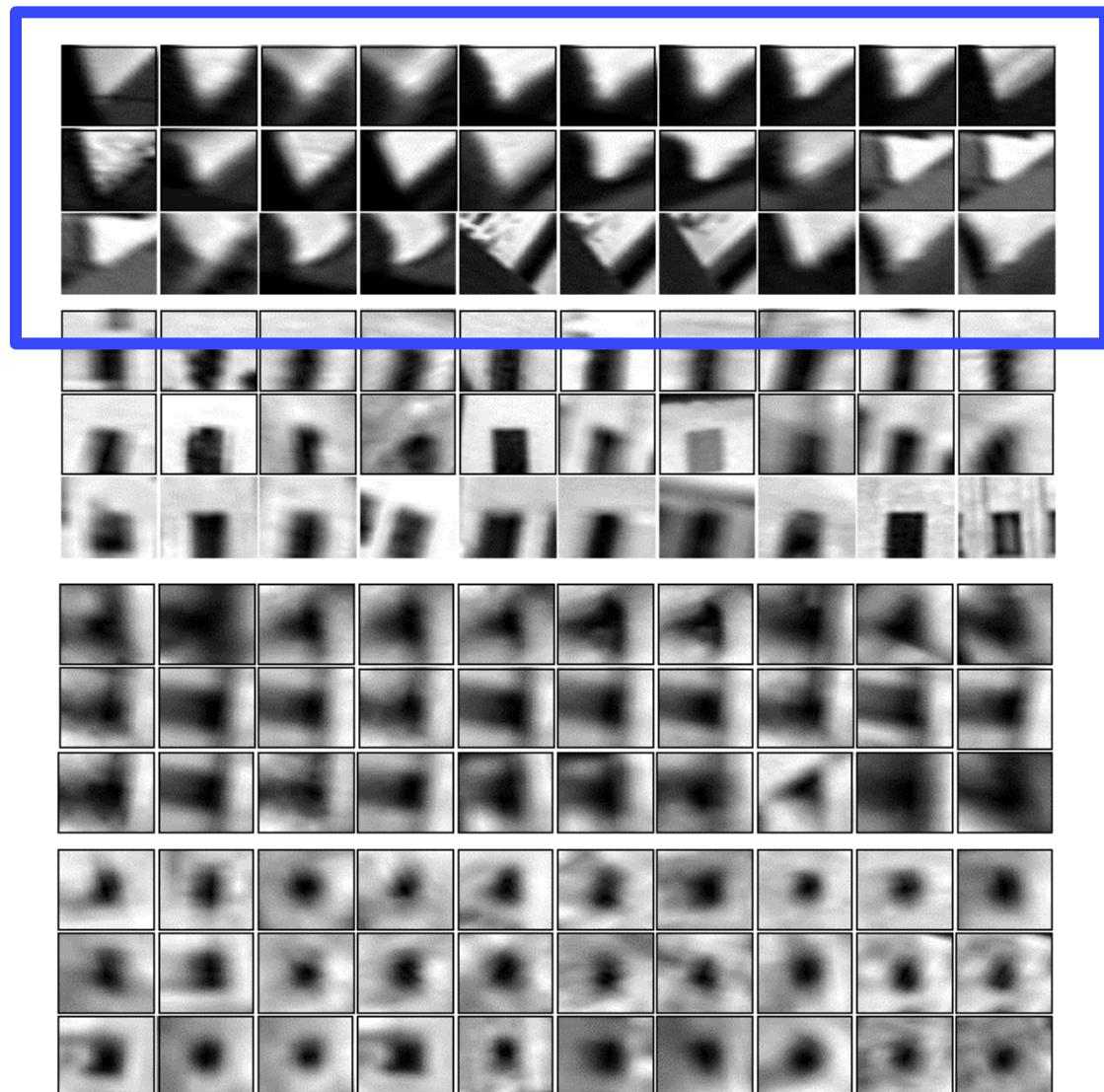
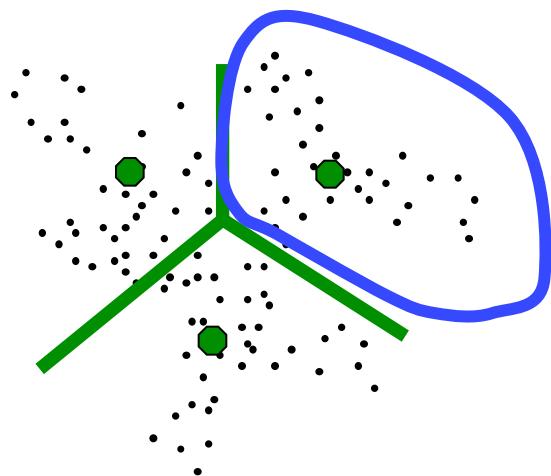
	media $d/dx$	media $d/dy$
fereastra #1	4	10
fereastra #2	25	7
fereastra #3	18	20

⋮

Statistici pentru a  
descrie pattern-urile în  
ferestre mici

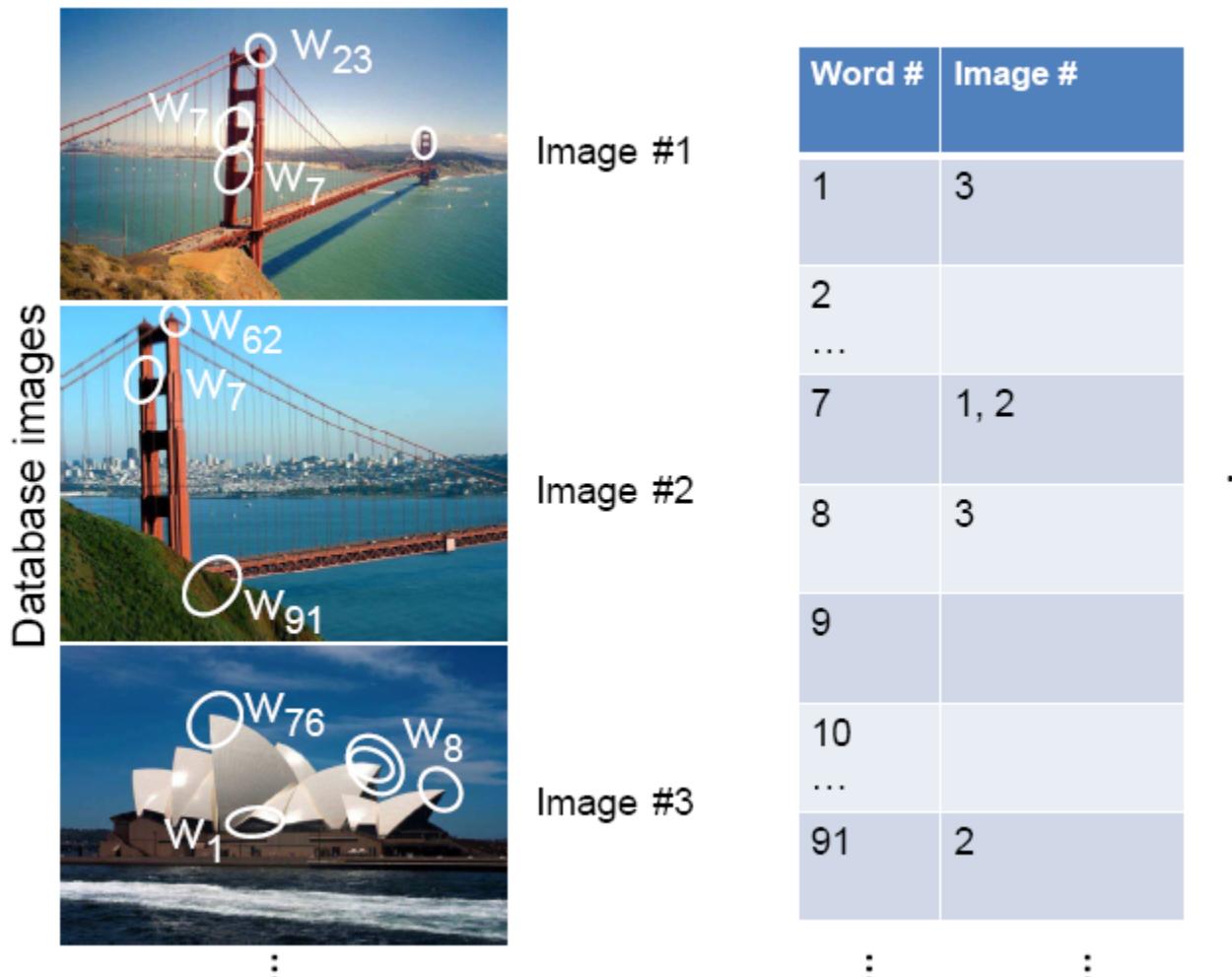
# Cuvinte vizuale

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



Figură din articolul Sivic & Zisserman, ICCV 2003

# Inverted file index



- 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 summarizăm 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 multime 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’**

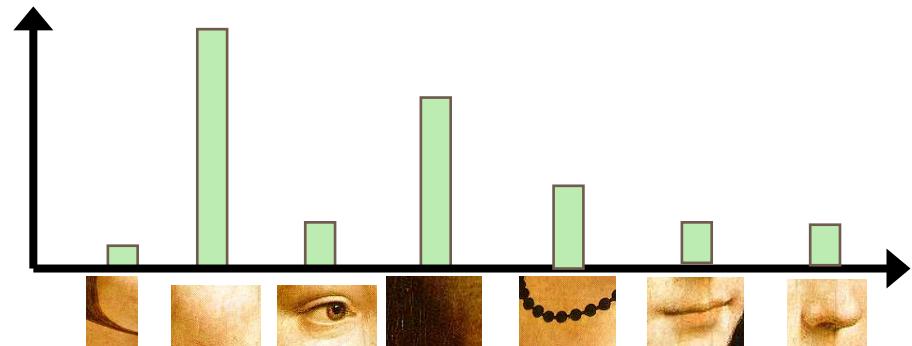
(sac de cuvinte vizuale)



**Imagine**

**Bag of ‘visual words’**

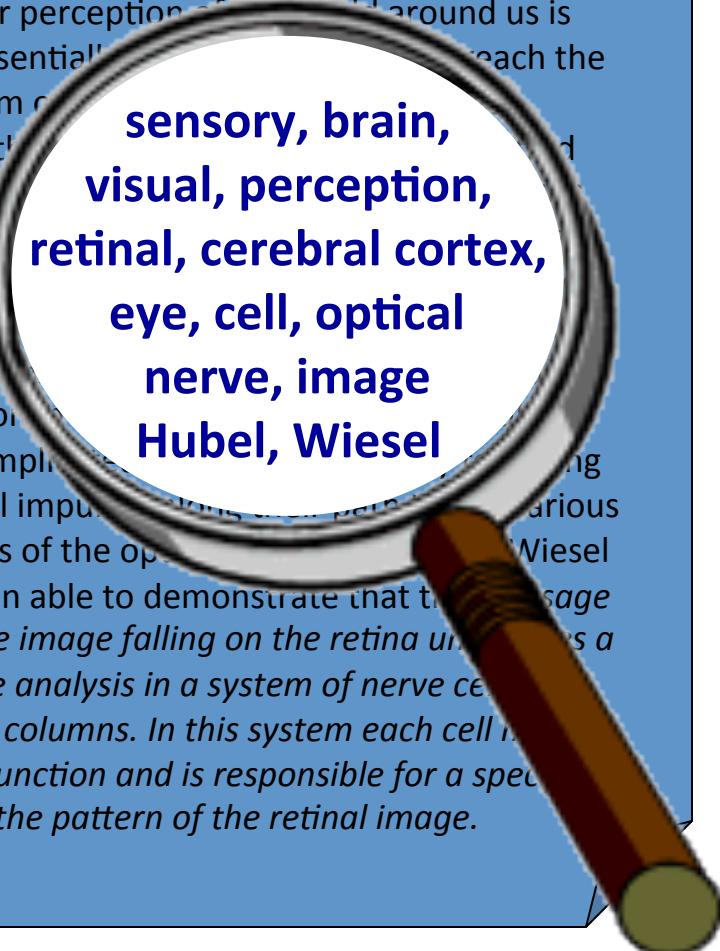
(sac de cuvinte vizuale)



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

# Analogia cu documentele

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 upon what we see. Light rays reach the brain from our eyes and are processed by the cerebral cortex. We have now learned that thought takes place in the brain. At one point by cerebra upon which the brain acts. Through now known that the process of perception is more complicated than was once believed. By analysing the visual impulses coming from the eye, various cell layers of the optic nerve have been able to demonstrate that the message about the image falling on the retina undergoes a step-wise analysis in a system of nerve cells stored in columns. In this system each cell has a specific function and is responsible for a specific detail in the pattern of the retinal image.



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

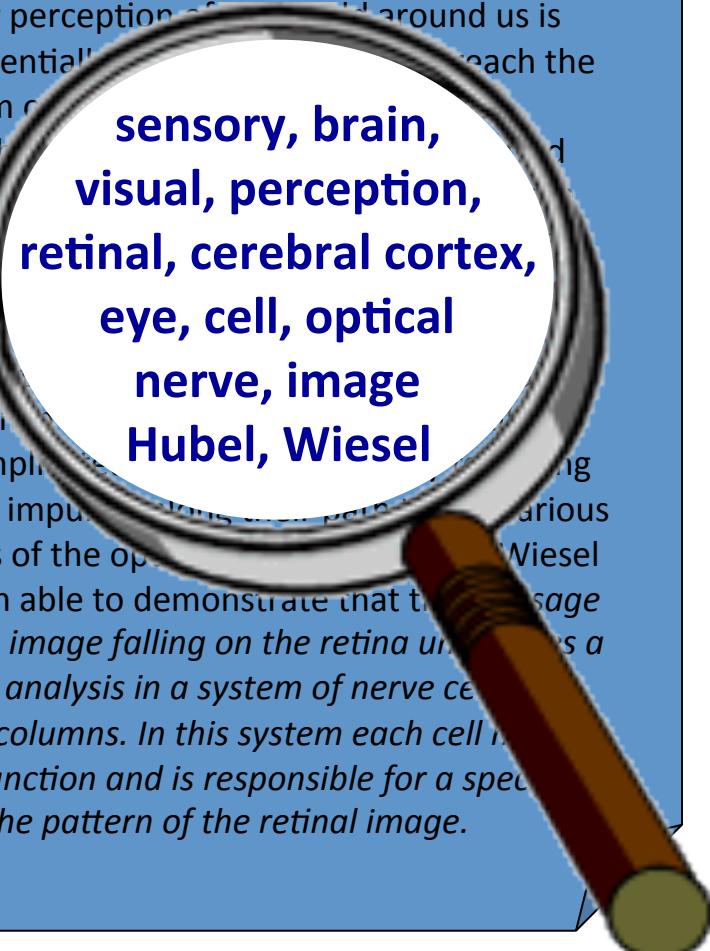
China is forecasting a trade surplus of \$90bn (£51bn) to \$100bn this year, a threefold increase on 2004's \$32bn. The Commerce Ministry said the surplus would jump to \$100bn by 2008. It predicted 30% growth in exports and a 18% rise in imports. The ministry further announced that China's central bank would deliberate the surpluses. One factor is the yuan's appreciation. Xiaochuan said the central bank would do more to boost the value of the yuan. The central bank stayed within a band of 2.2% of the value of the yuan against the US dollar in July and permitted it to move outside the band, but the US wants the yuan to be allowed to trade freely. However, Beijing has made clear that it will take its time and tread carefully, allowing the yuan to rise further in value.



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

# Modele de tip bag-of-words: reprezentarea documentelor

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 upon what reaches the brain from our eyes. We have thought that the point by which the cerebrum receives information upon what it sees is the optic nerve. Through now known that perception is a more complicated process than was at first imagined. The visual impulses pass through various cell layers of the optic nerve, and Wiesel and Hubel have been able to demonstrate that the message about the image falling on the retina undergoes a step-wise analysis in a system of nerve cells stored in columns. In this system each cell has a specific function and is responsible for a specific detail in the pattern of the retinal image.



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

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 visual system. Light reaches the brain from our eyes through the optic nerve and is processed by the cerebral cortex. We have thought that the visual system worked in this way since the point by which the optic nerve enters the brain was identified. However, it is now known that the visual system is much more complex than this. The visual system processes the visual input in parallel through various cell layers of the optic nerve. In 1961, Hubel and Wiesel have been able to demonstrate that the visual message about the image falling on the retina undergoes a step-wise analysis in a system of nerve cells that are stored in columns. In this system each cell has a specific function and is responsible for a specific detail in the pattern of the retinal 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 rise further next year, as predicted 30% jump in exports and a 18% rise in imports. The ministry also predicted a further a 18% rise in imports next year. China's deliberations over the surplus are likely to continue. One factor is that China's deliberate to allow the yuan to appreciate. XiaoChuan, the central bank's governor, said last week that more to be done to encourage the appreciation. He stayed within a range of 2% either side of the current value of the yuan. The Chinese government has allowed the yuan to appreciate by 2% in July and permitted it to do so again in August. The US wants the yuan to be allowed to trade freely. However, Beijing has made clear that it will take its time and tread carefully before allowing the yuan to rise further in value.

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

# 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

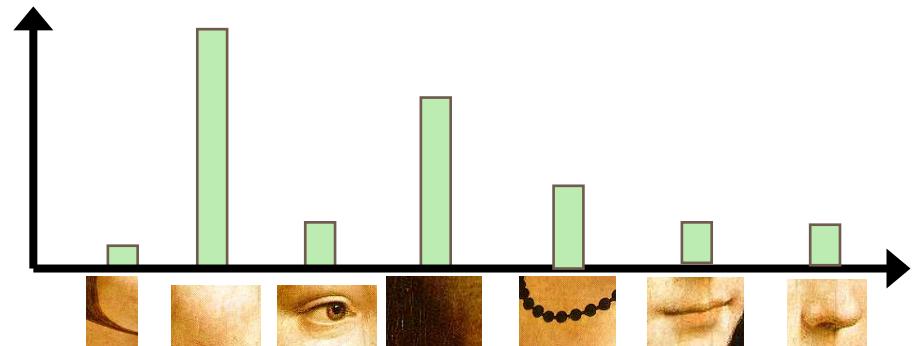
- Reprezentarea documentelor pe baza frecvențelor cuvintelor dintr-un dicționar



**Imagine**

**Bag of ‘visual words’**

(sac de cuvinte vizuale)



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

**Imagine**

**Bag of ‘visual words’**

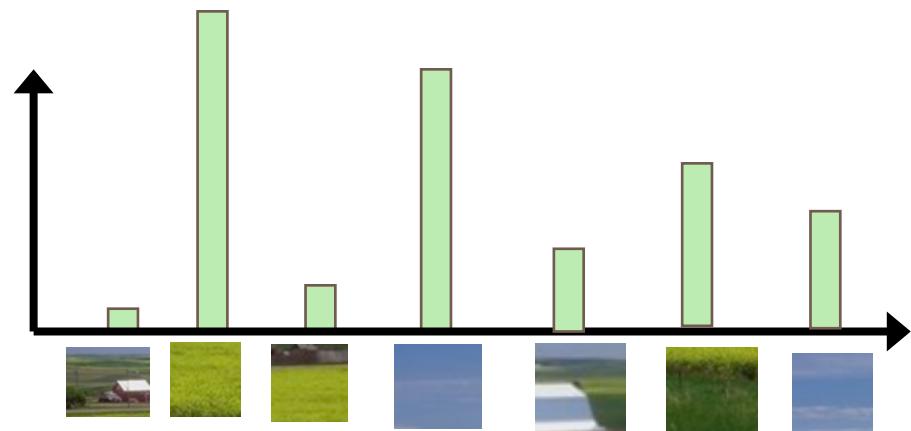
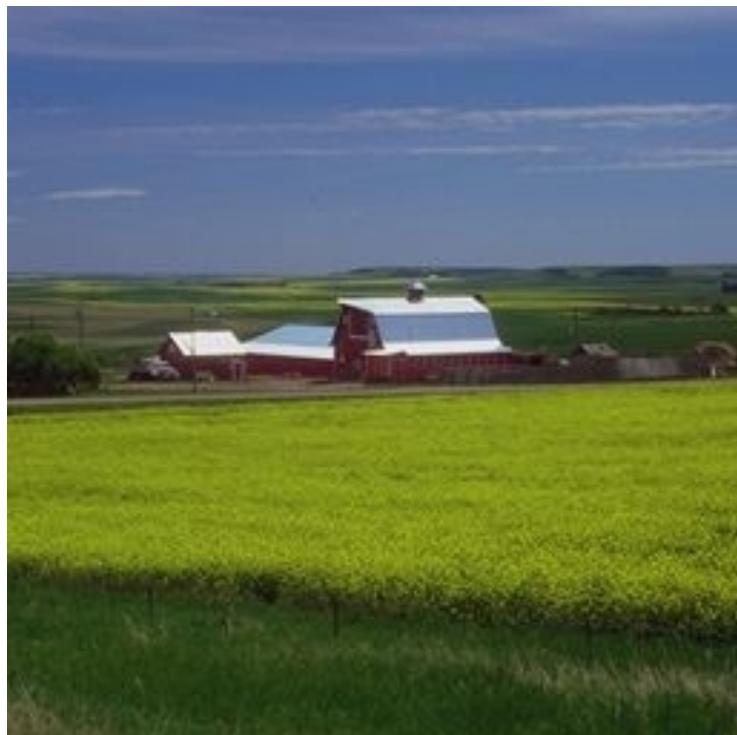
(sac de cuvinte vizuale)



**Imagine**

**Bag of ‘visual words’**

(sac de cuvinte vizuale)



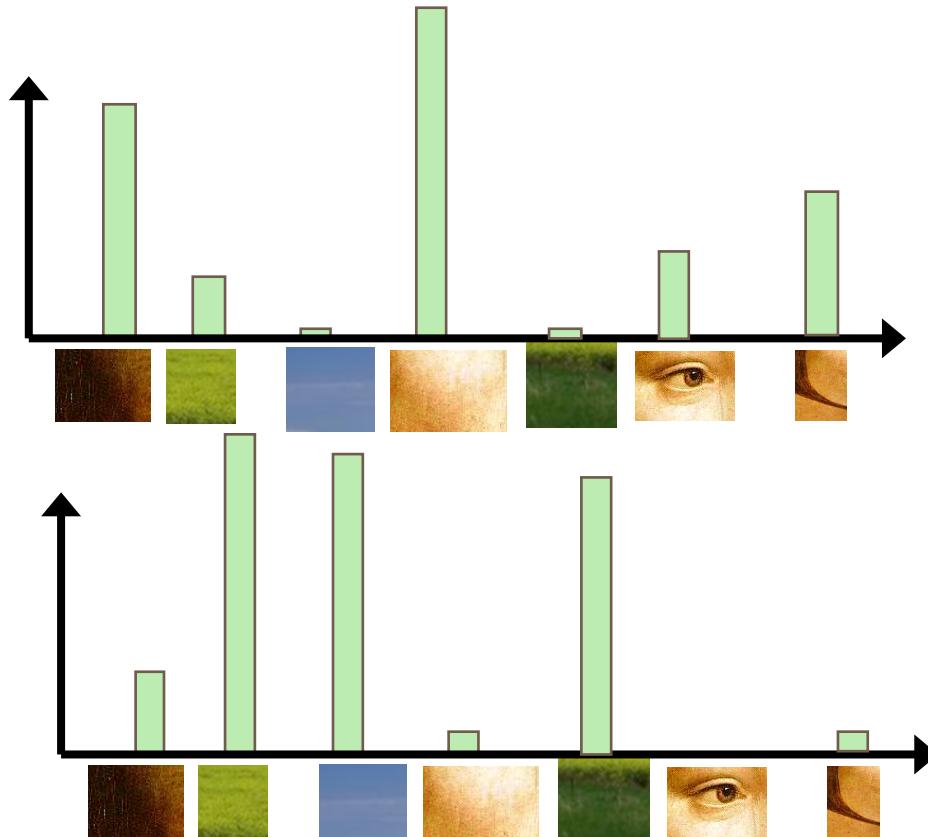
Idee de bază: reprezentăm o imagine ca o histogramă de pattern-uri prototip (cuvinte vizuale)

# Imagine

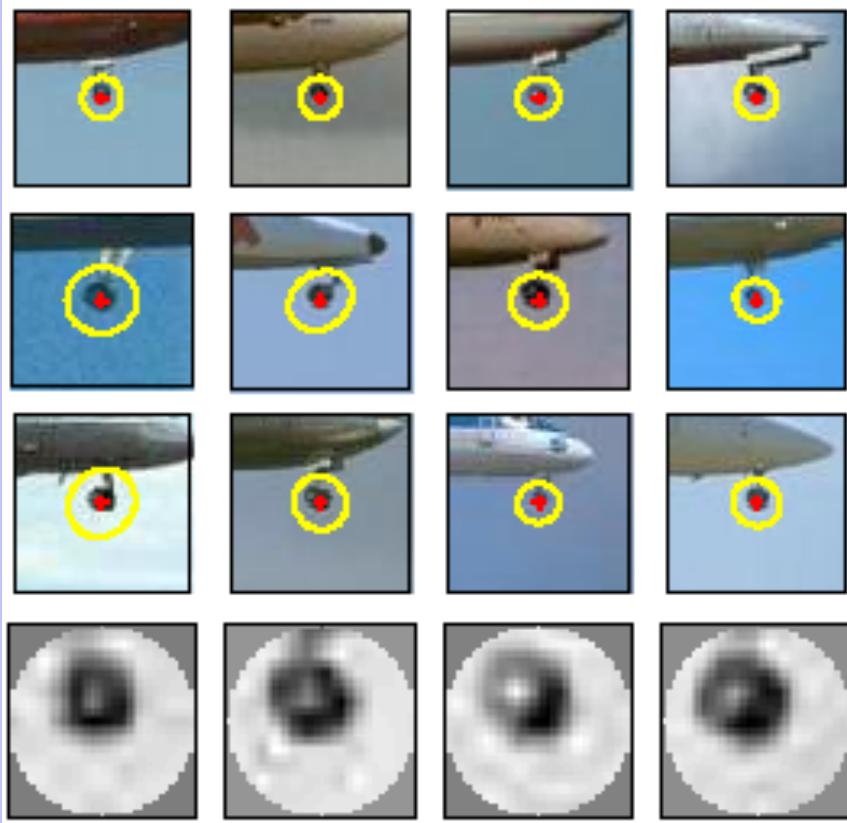
## Bag of ‘visual words’

(sac de cuvinte vizuale)

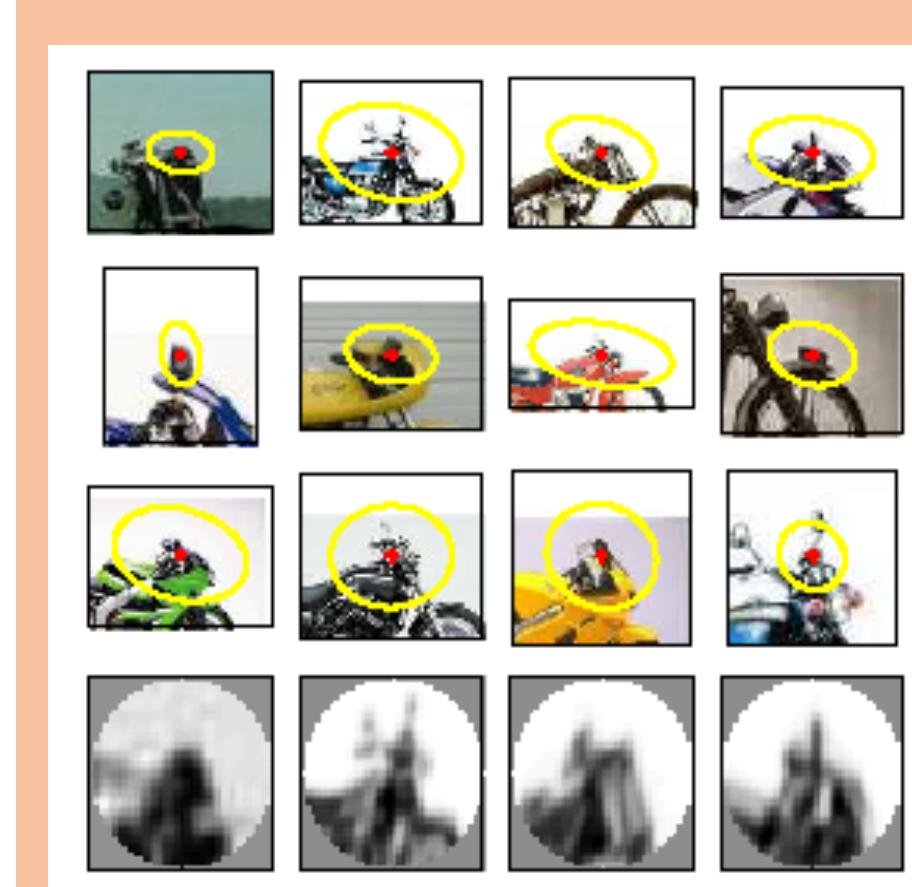
Dacă vreau să compar imagini între ele nu pot să am pattern-uri prototip specifice pentru fiecare imagine. Vreau să am aceleasi pattern-uri prototip pentru toate imaginile astfel încât comparația să fie posibilă.



# Visual words = cuvinte vizuale



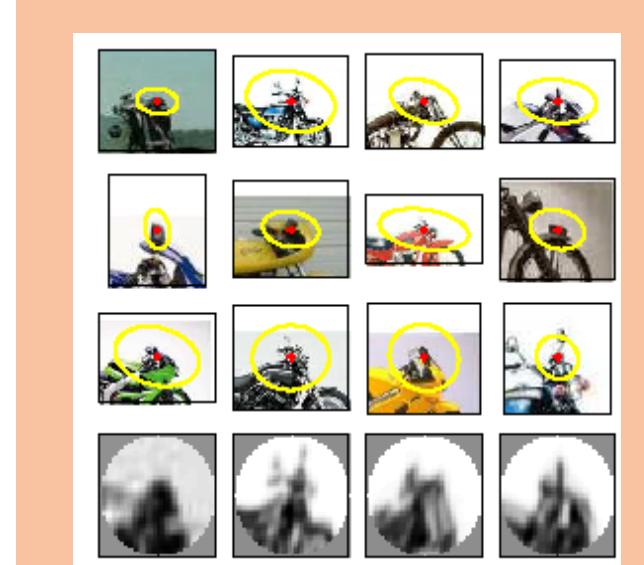
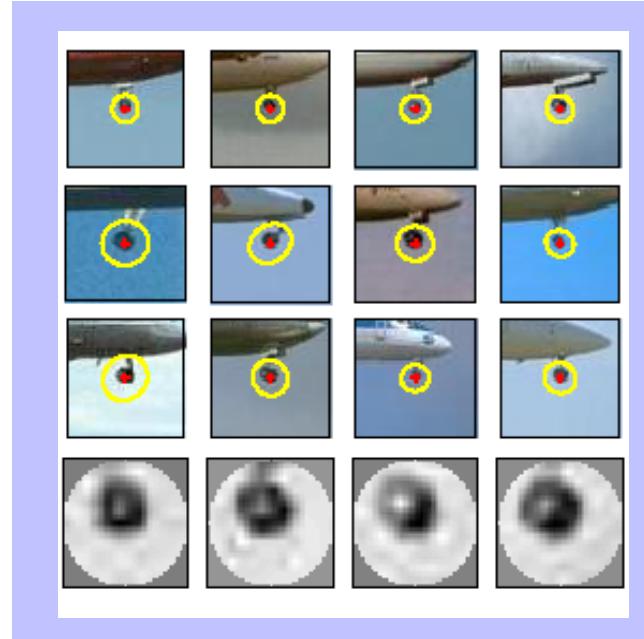
Subimaginea (patch-ul) care conține roata unui avion este diferită în fiecare imagine, dar foarte asemănătoare.



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
- Înfățișarea obiectelor variază foarte mult chiar și pentru aceeași clasă de obiecte
- Înfățișarea 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. Multime de imagini

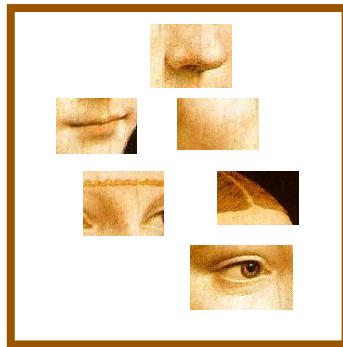


# Modelul ‘bag of visual words’

1. Multime de imagini



2. Extragem caracteristici din fiecare imagine



# Modelul ‘bag of visual words’

1. Multime de imagini



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



# Modelul ‘bag of visual words’

1. Multime de imagini



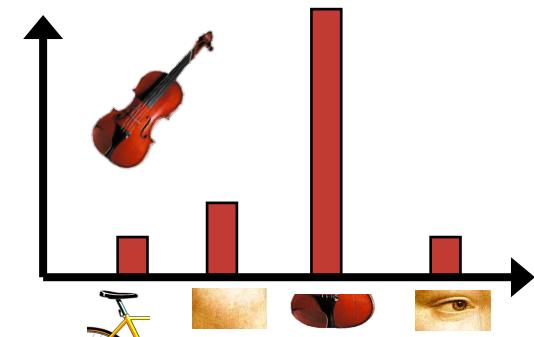
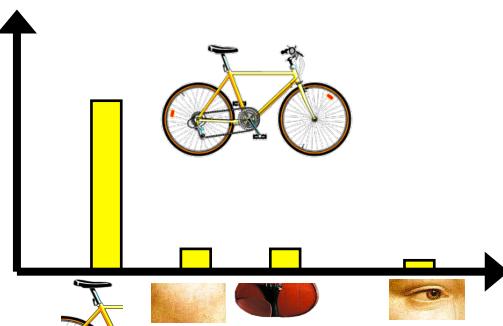
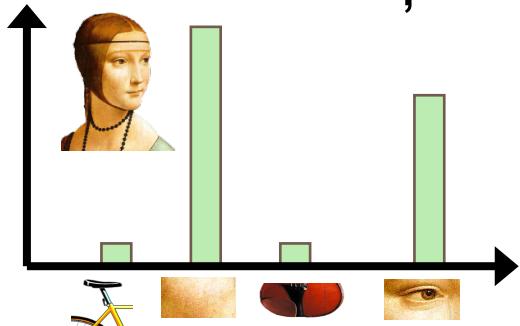
2. Extragem caracteristici din fiecare imagine
3. Învățăm un vocabular vizual = dicționar
4. Fiecare caracteristică este asignată ‘cuvântului vizual’ din dicționar cel mai apropiat

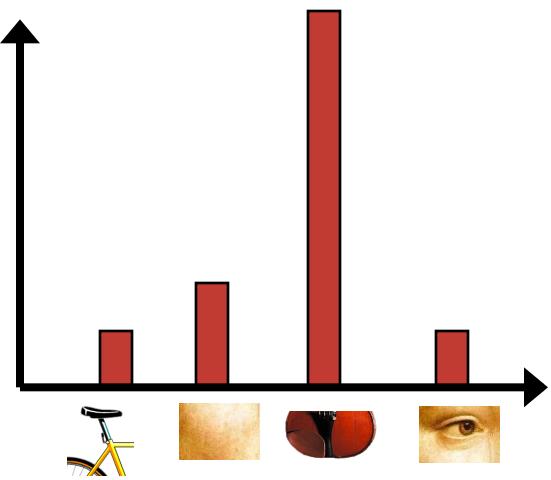
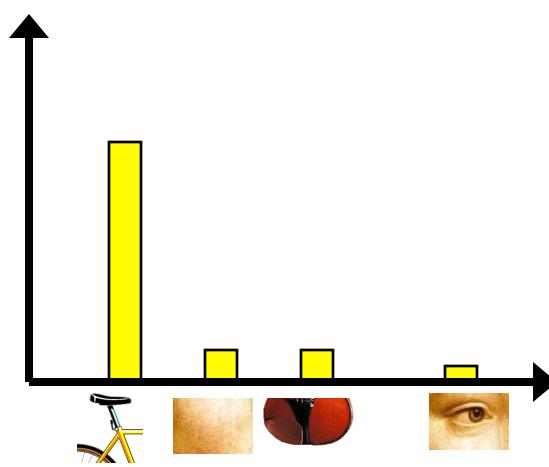
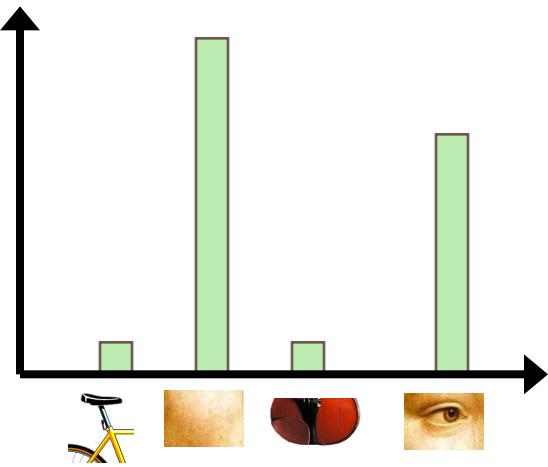
# Modelul ‘bag of visual words’

1. Multime de imagini



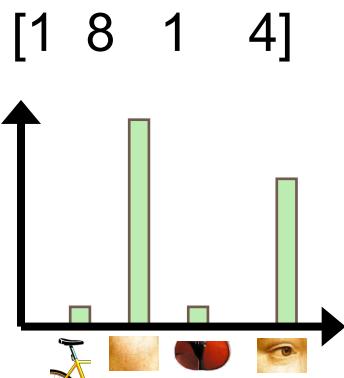
2. Extragem caracteristici din fiecare imagine
3. Învățăm un vocabular vizual = dicționar
4. Fiecare caracteristică este asignată ‘cuvântului vizual’ din dicționar cel mai apropiat
5. Reprezentăm imaginile pe baza unei histograme de frecvențe a ‘cuvintelor vizuale’



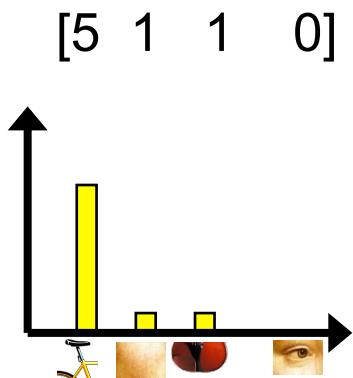


# Compararea reprezentărilor BOW

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



$\vec{d}_j$



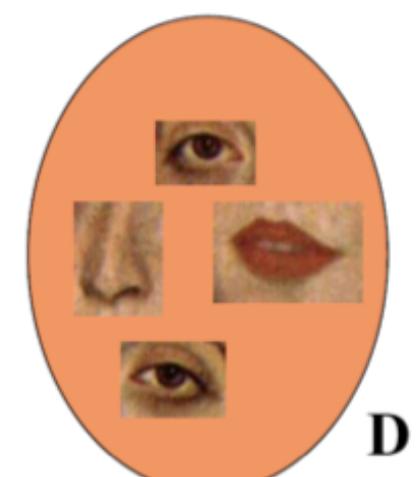
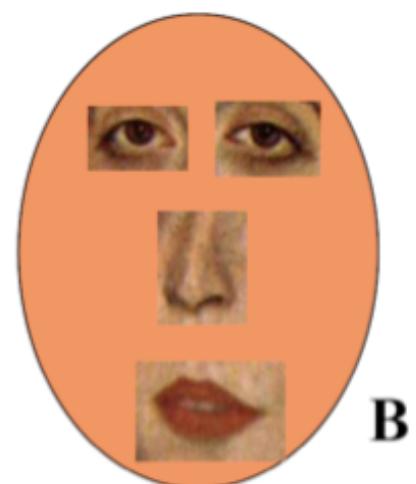
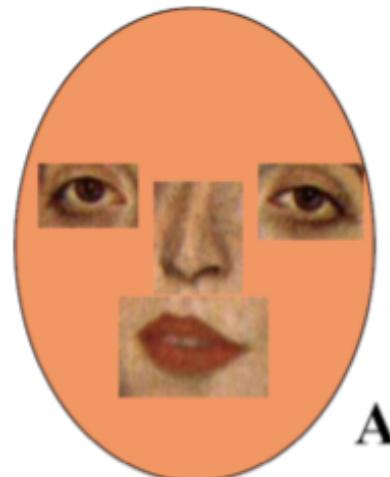
$\vec{q}$

$$\begin{aligned} sim(d_j, q) &= \frac{\langle d_j, q \rangle}{\|d_j\| \|q\|} \\ &= \frac{\sum_{i=1}^V d_j(i) * q(i)}{\sqrt{\sum_{i=1}^V d_j(i)^2} * \sqrt{\sum_{i=1}^V q(i)^2}} \end{aligned}$$

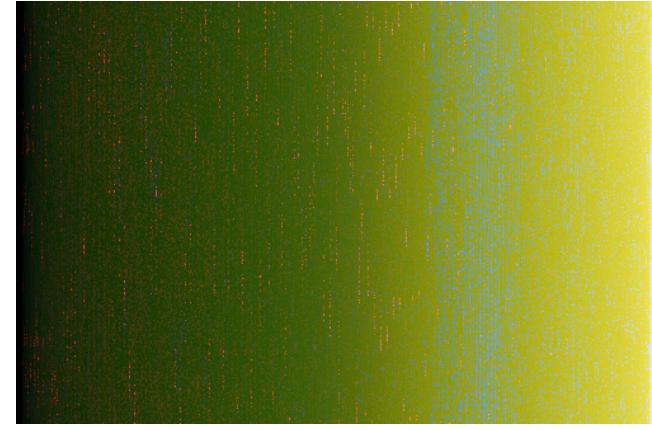
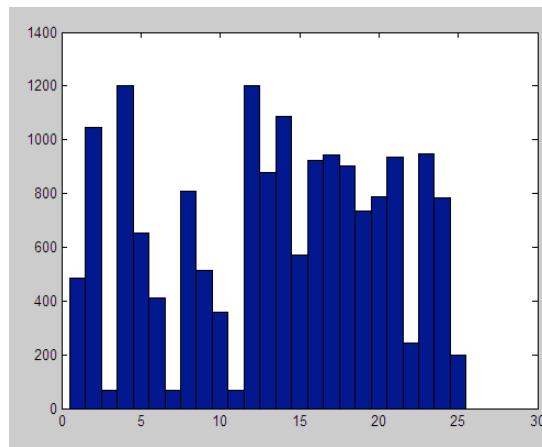
Pentru vocabulare cu V cuvinte

# 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?



# Poziția în imagine a caracteristicilor



Cele trei imagini au aceeași histogramă de culori.

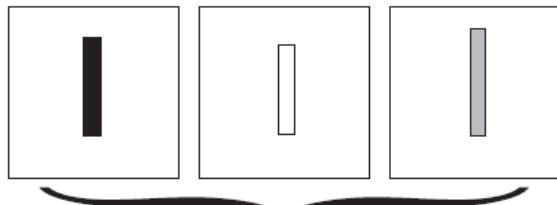
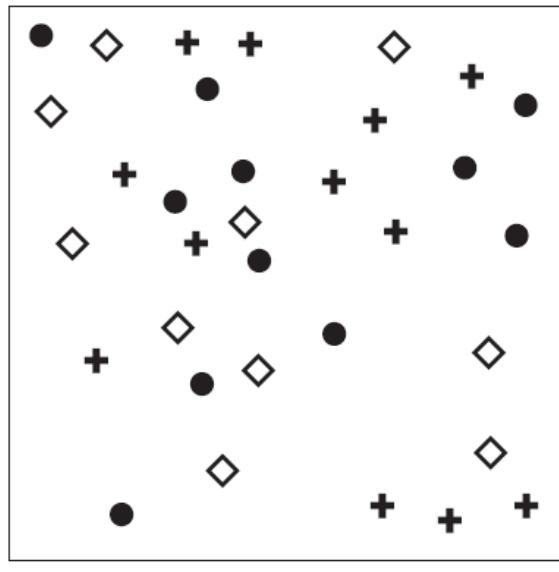
# Piramidă spațială



Calculează histograma pentru fiecare subimagine

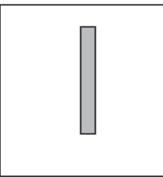
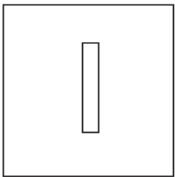
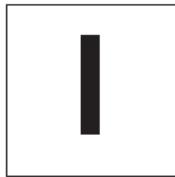
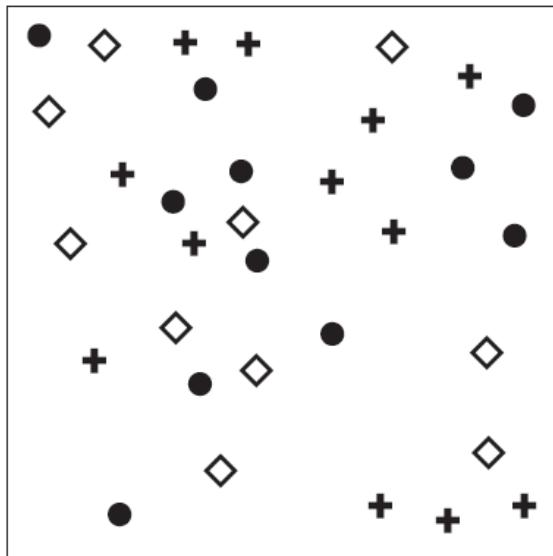
# Piramidă spațială

nivelul 0

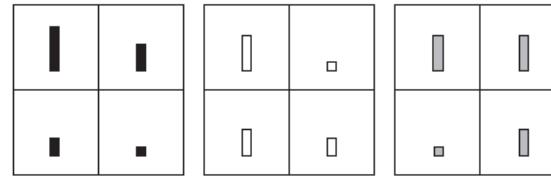
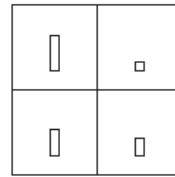
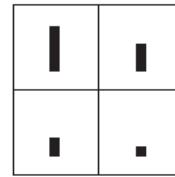
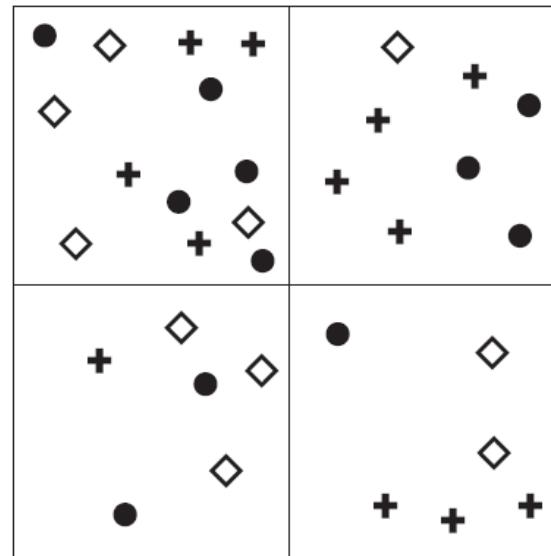


# Piramidă spațială

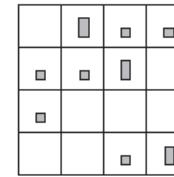
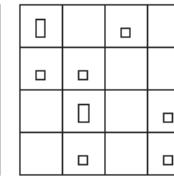
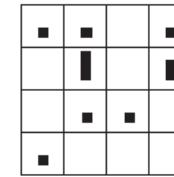
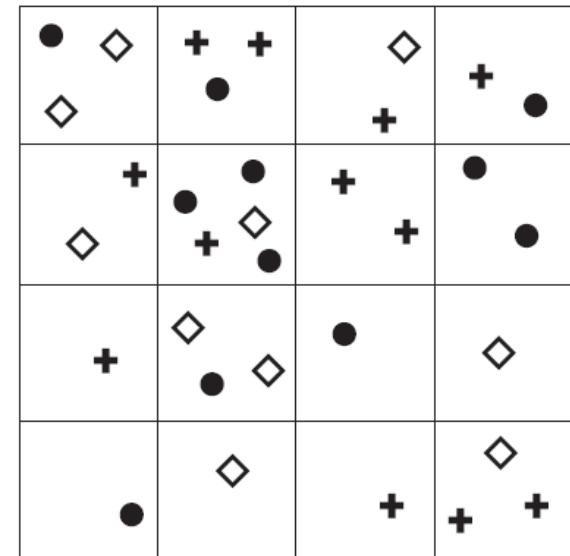
nivelul 0



nivelul 1



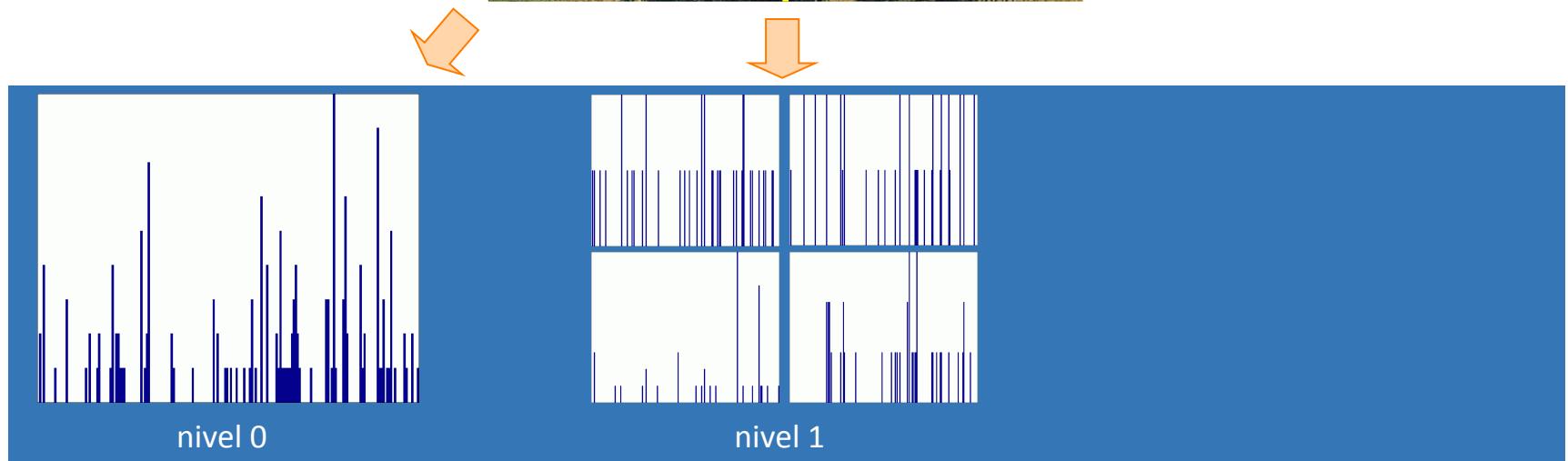
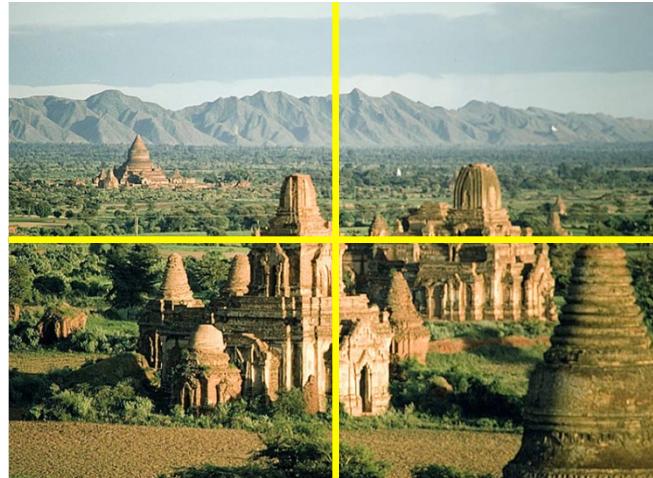
nivelul 2



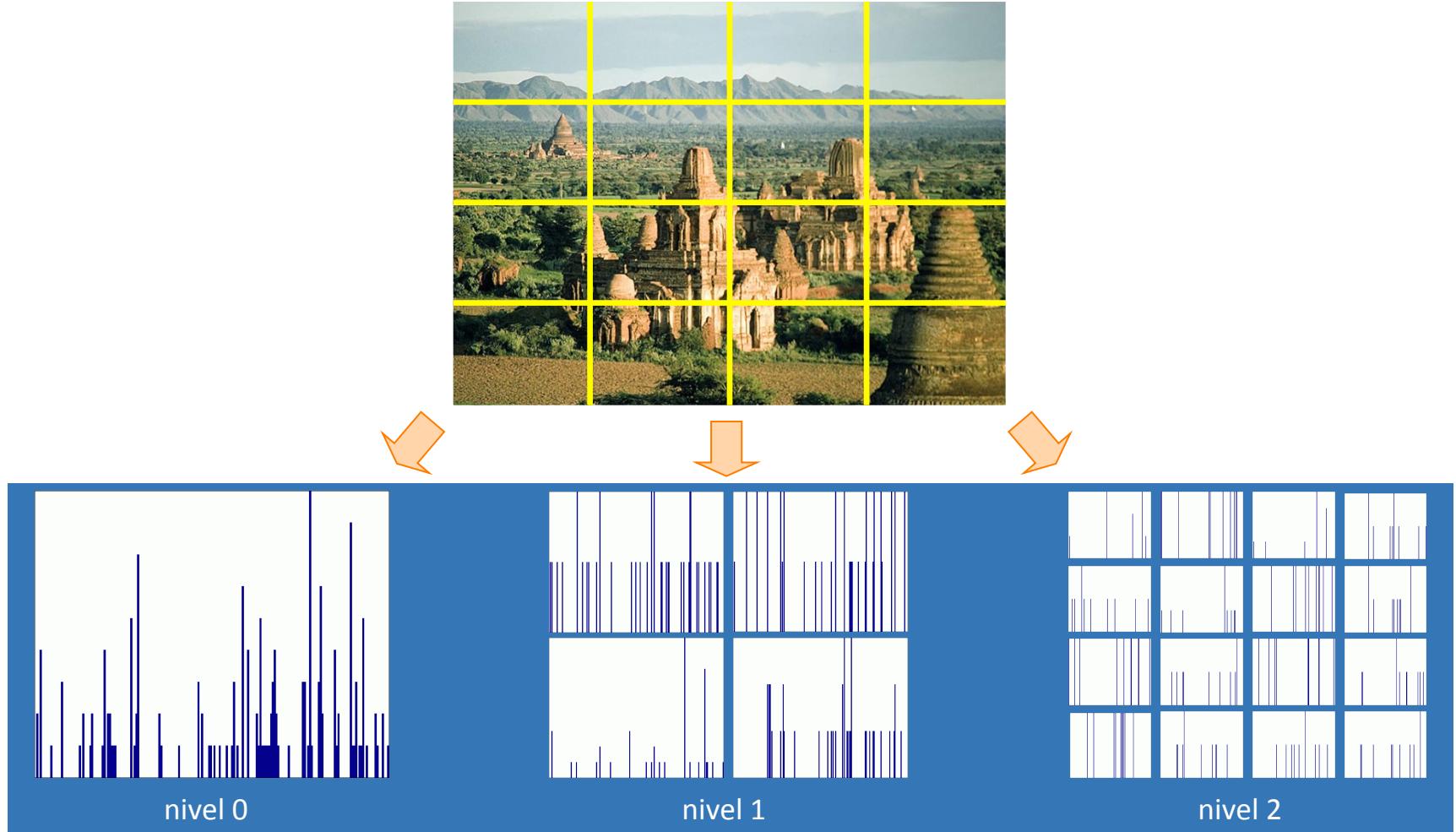
# Piramidă spațială



# Piramidă spațială



# Piramidă spațială



# Ponderarea tf-idf

- Term frequency – inverse document frequency
- 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)

Numărul de apariții ale cuvântului i în documentul d

numărul de cuvinte în documentul d

document – imagine  
cuvânt – cuvânt vizual

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i}$$

Numărul total de documente în baza de date

Numărul de documente din baza de date în care apare cuvântul i

# BOW pentru regăsire pe baza conținutului

Interogare definită vizual

“Găsește  
acest  
ceas”



“Groundhog Day” [Rammis, 1993]



“Găsește  
acest loc”



# retrieved shots

## Example



# Video Google System

1. Extragă 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>

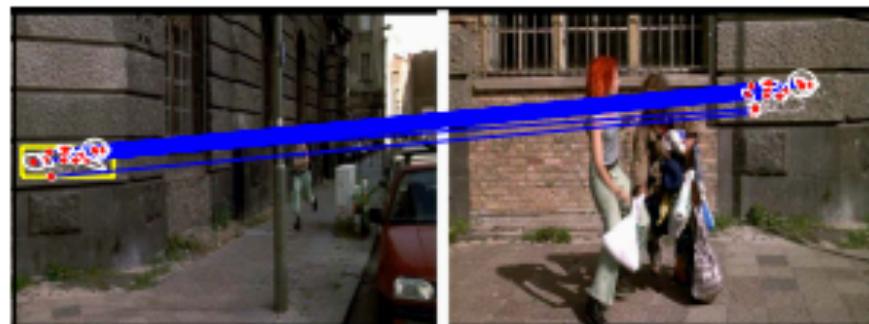
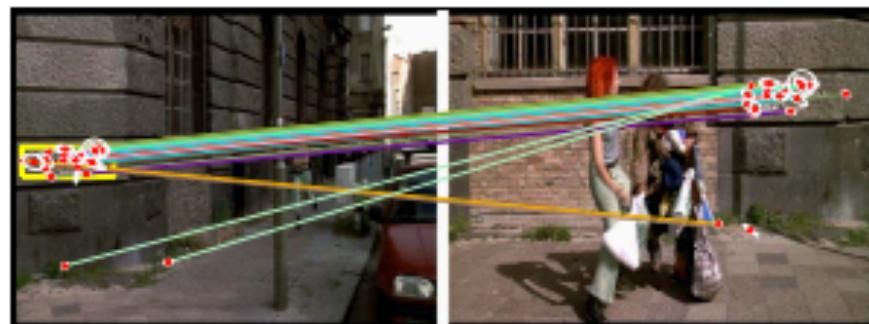
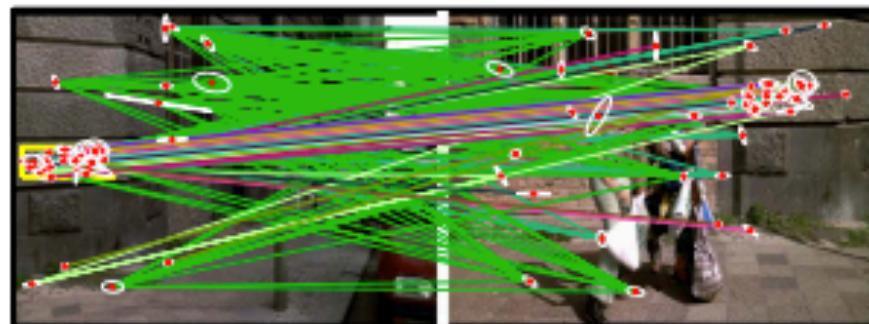


Query region



Retrieved frames

# Verificare spațială



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

-zonă de căutare de 15 vecini

-folosirea unui stop list: elimină cuvintele vizuale 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/>

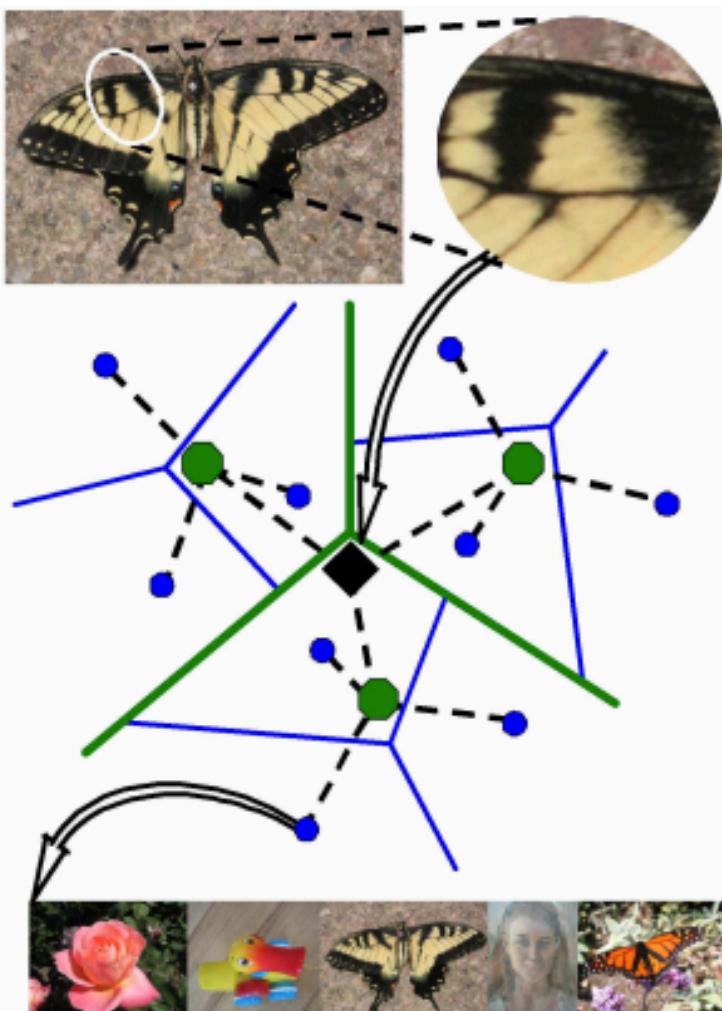
## Abstract

*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.*

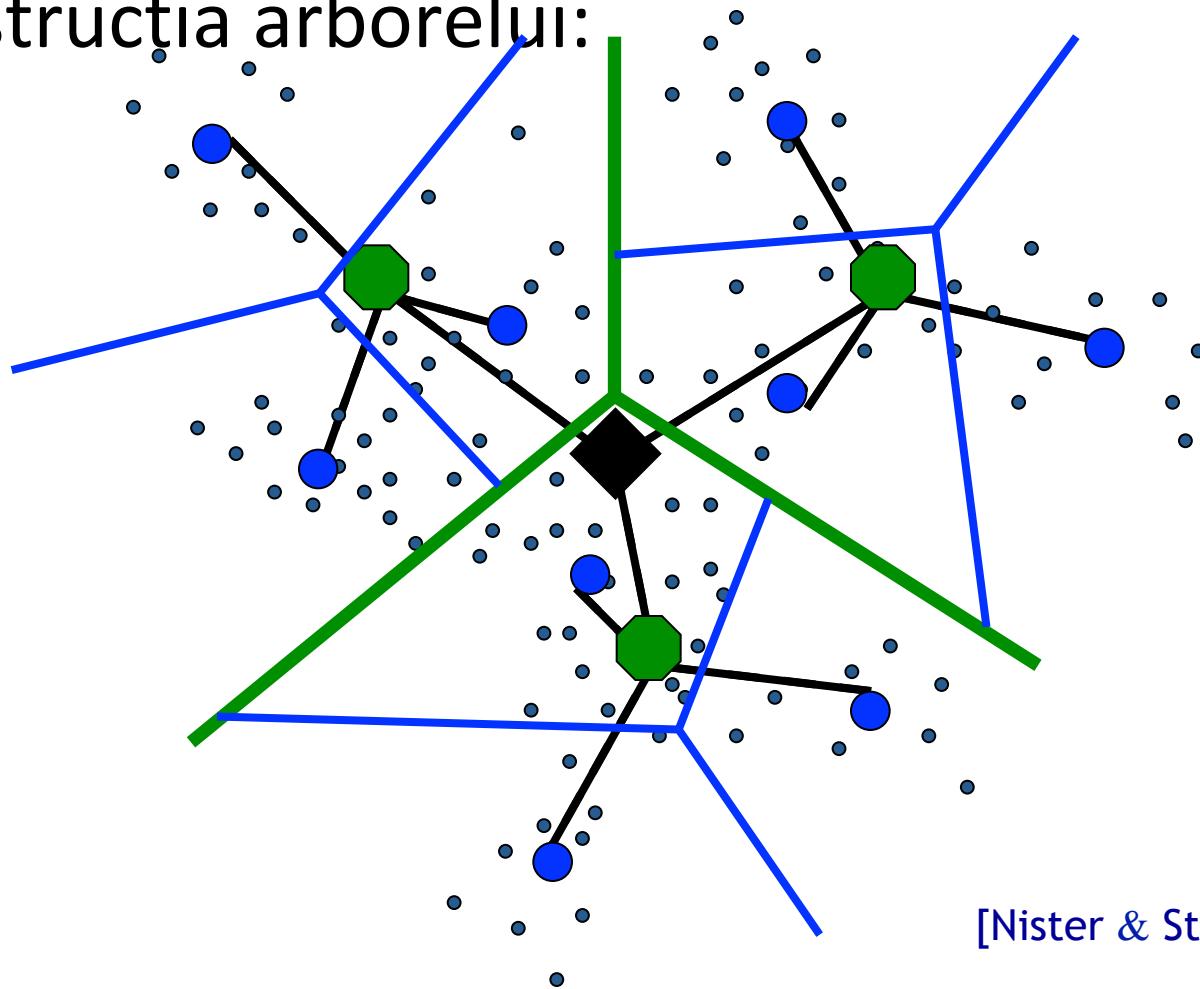
*The recognition quality is evaluated through retrieval on a database with ground truth, showing the power of the vocabulary tree approach, going as high as 1 million images.*

## 1. Introduction



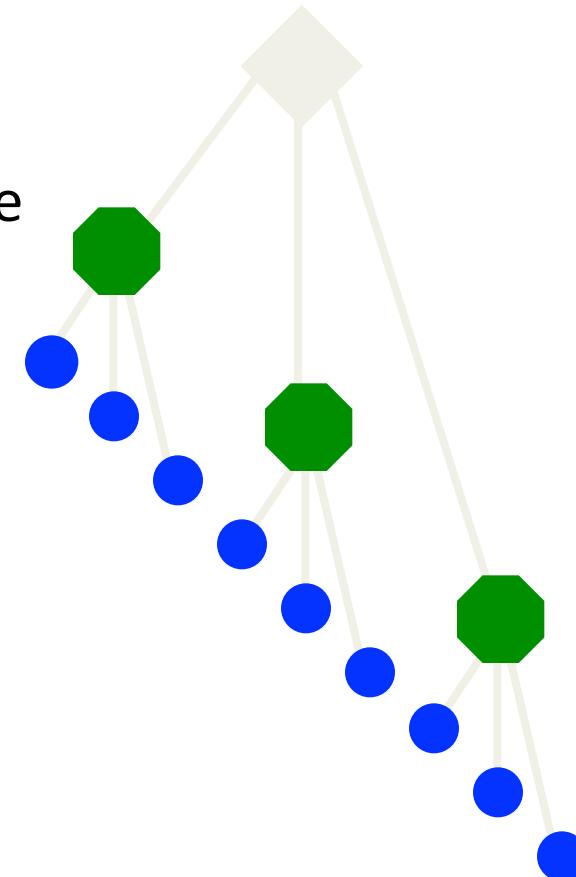
# Vocabular sub formă de arbore: clusterizare ierarhică pentru vocabulare mari

- Construcția arborelui:



# Vocabular sub formă de arbore

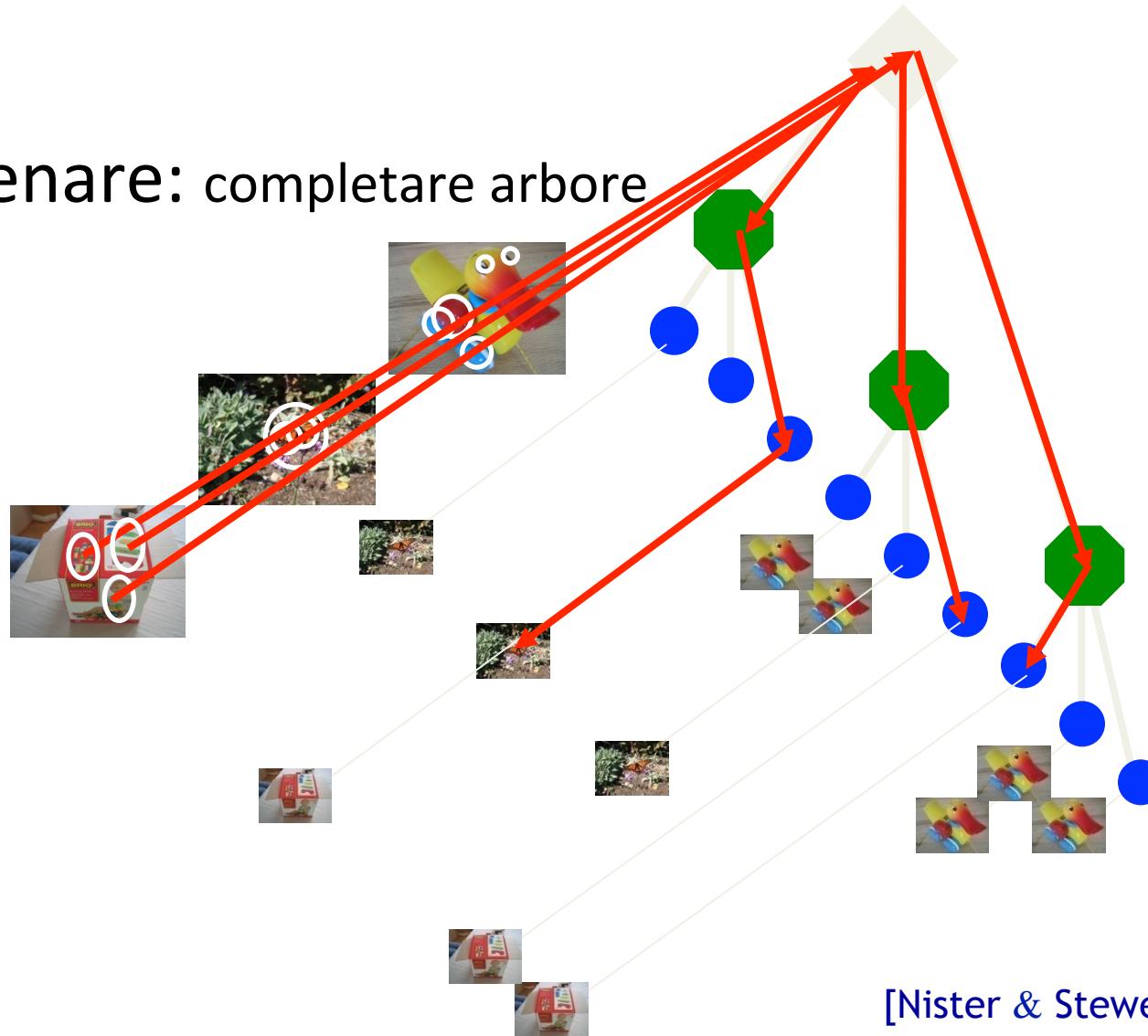
- Antrenare: completare arbore



[Nister & Stewenius, CVPR'06]

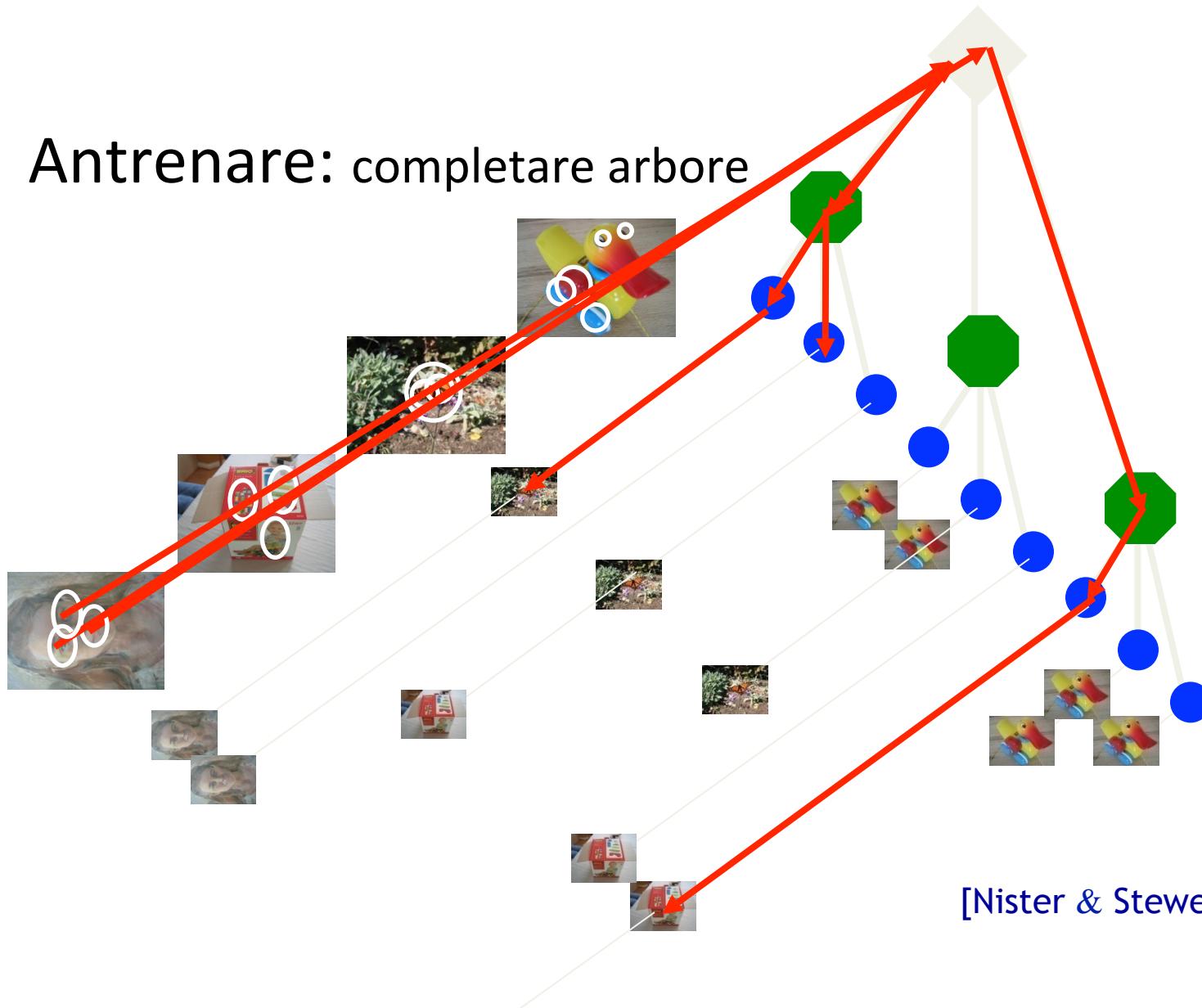
# Vocabular sub formă de arbore

- Antrenare: completare arbore

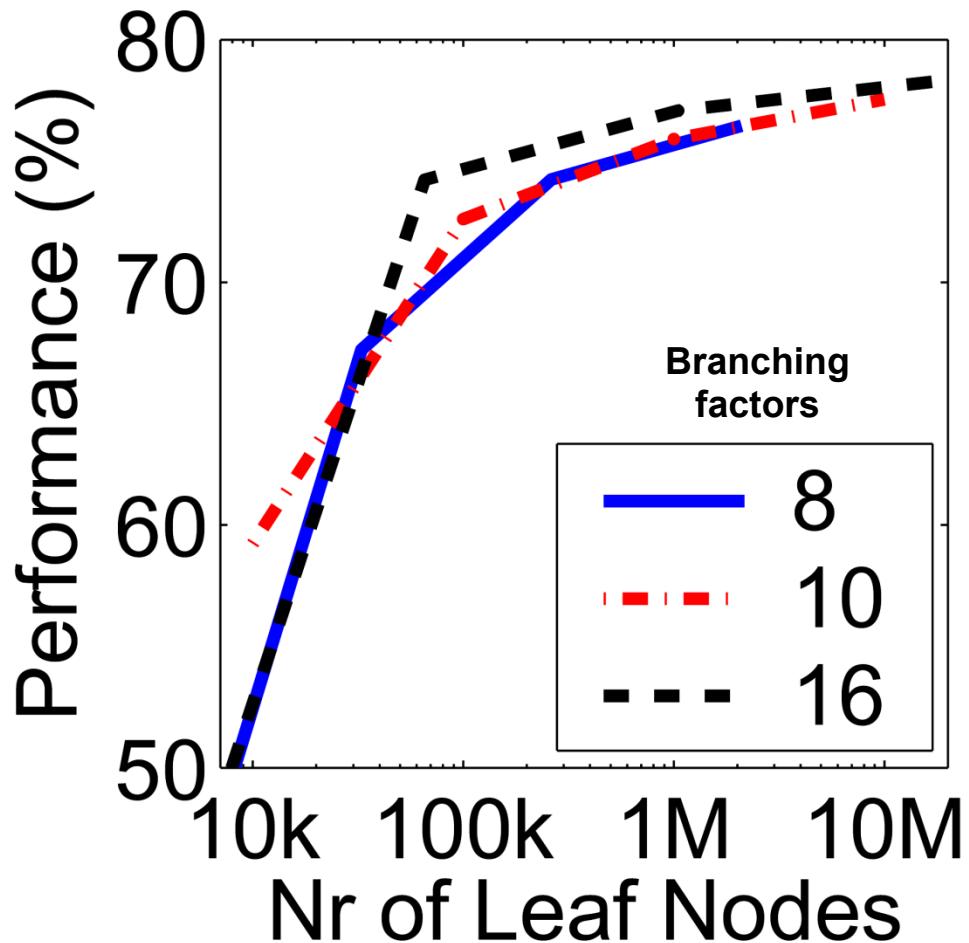


# Vocabular sub formă de arbore

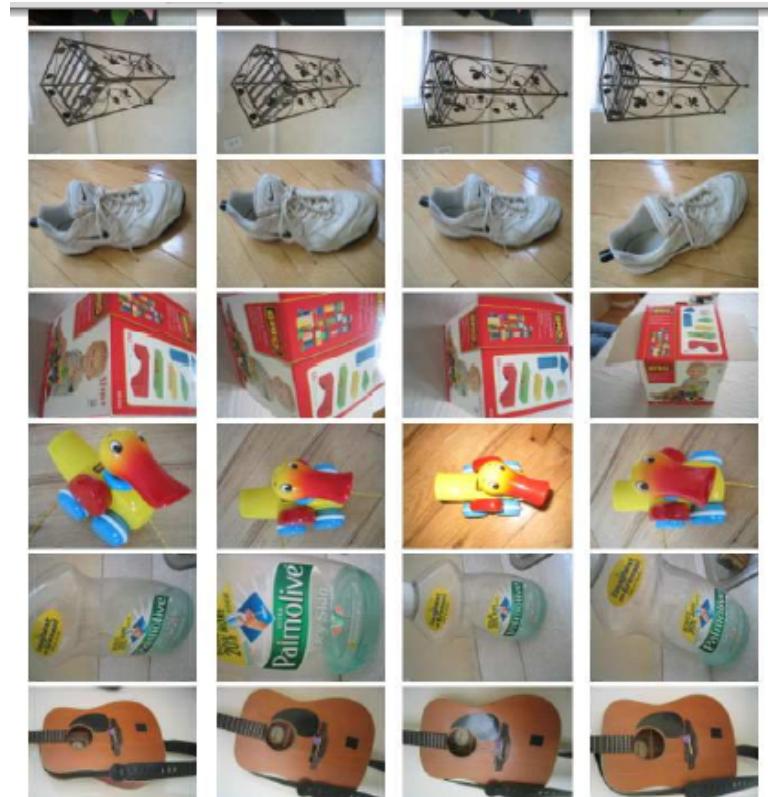
- Antrenare: completare arbore



# Dimensiunea vocabularului



Results for recognition task  
with 6347 images



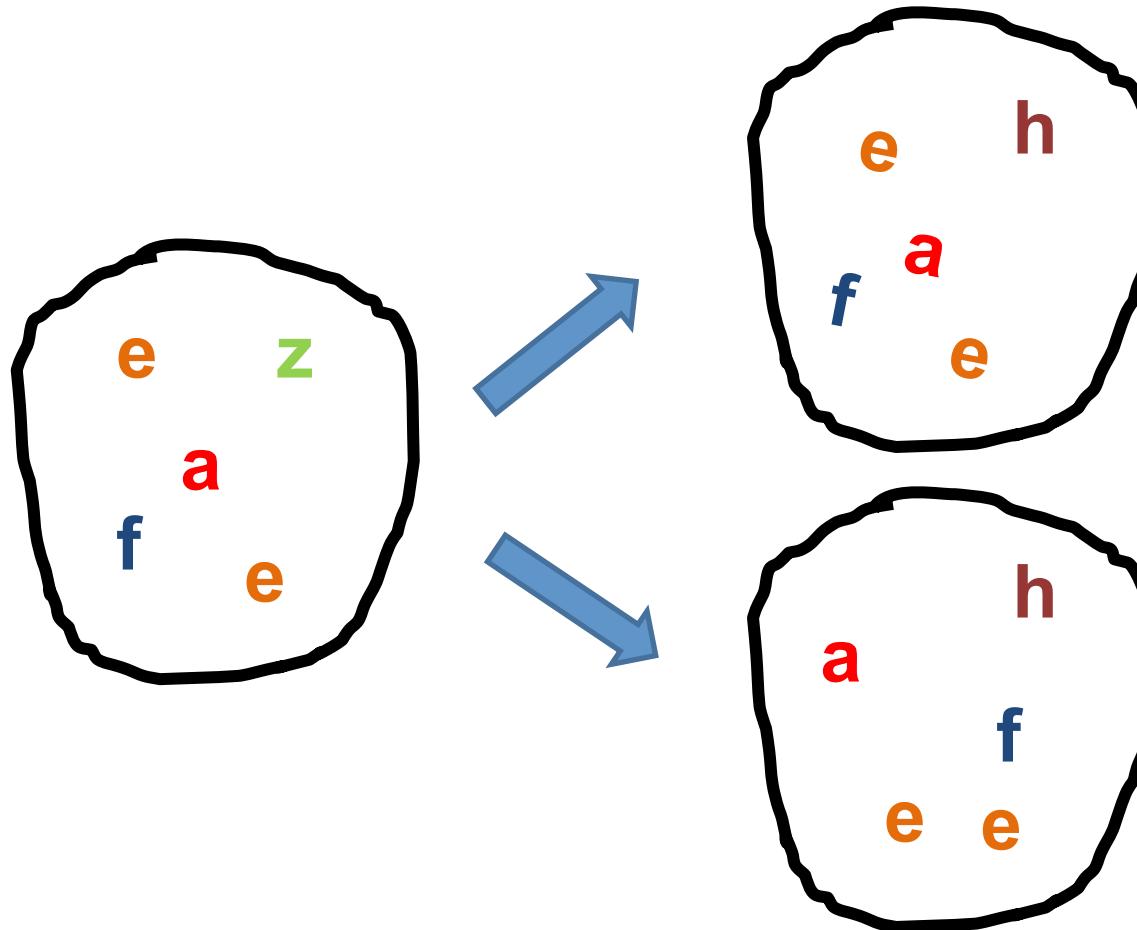
# Modelul BOW: pros / cons

- + 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 summarizăm 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 multime 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?*



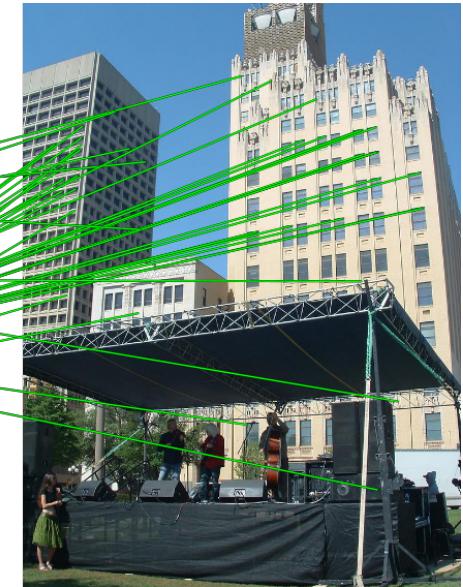
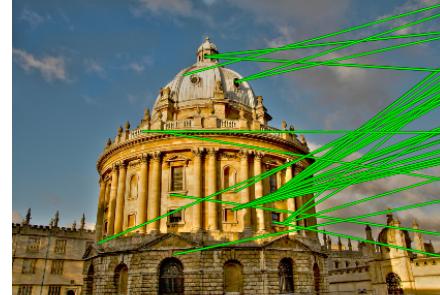
# Verificare spațială

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

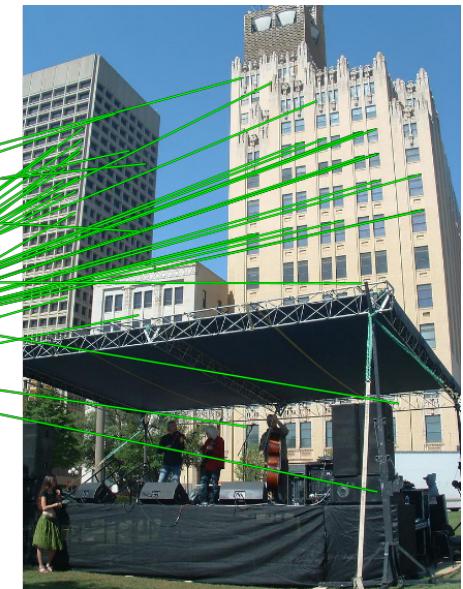
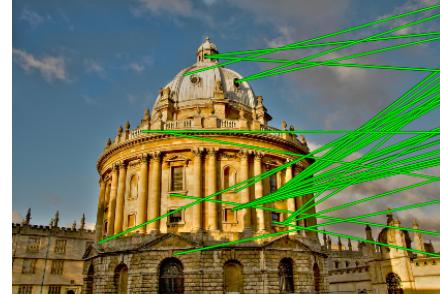
# Verificare spațială

Query



Imagine din baza de date  
cu similaritate mare BOW

Query



Imagine din baza de date  
cu similaritate mare BOW

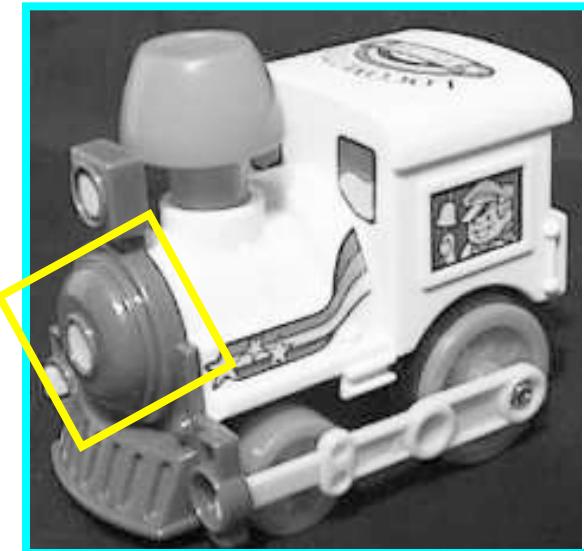
Numai o parte din corespondențe sunt mutual consistente.

# Verificare spațială: două strategii de bază

- Transformata Hough generalizată
  - fiecare corespondență 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.



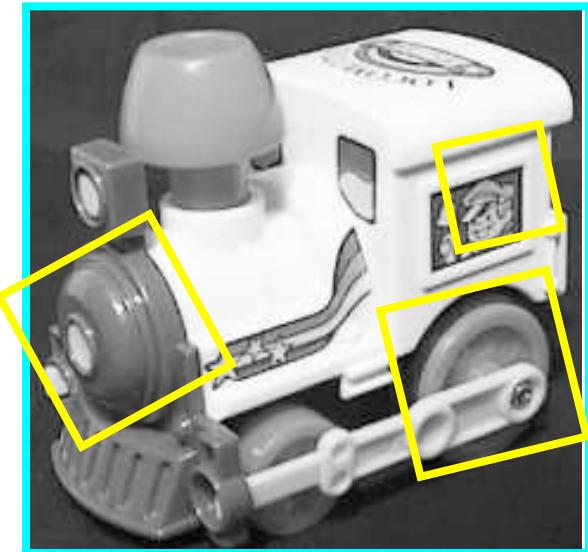
Model



Imagine test

# Transformata Hough generalizată

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



Model



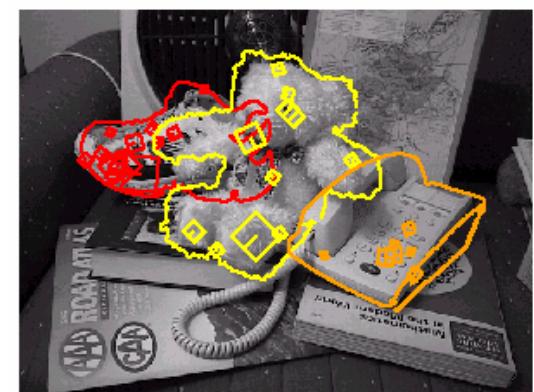
Imagine test

# 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 corespondență în imaginea test votează într-un spațiu Hough 4D asociat
  - Intervale de 30 de grade în orientare, factor 2 pentru mărime,  $0.25 * \text{diagonala imaginii}$  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 transformare afină

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

# Rezultate (Lowe'04)

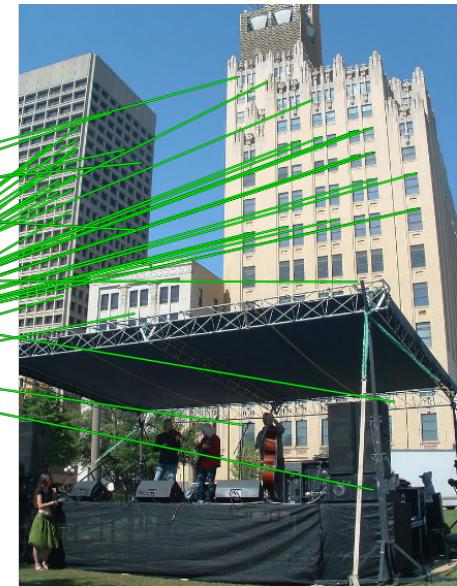
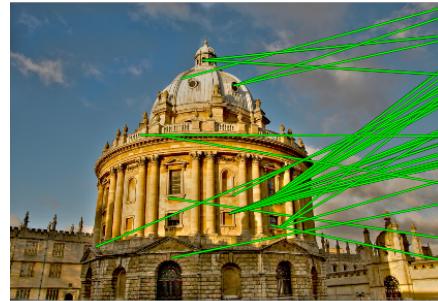


Background extras

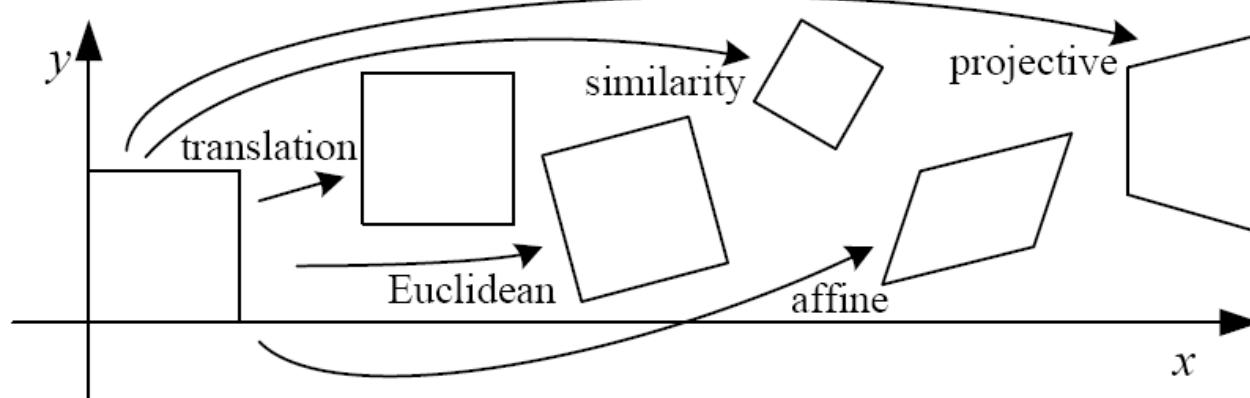
Obiecte găsite

Obiecte găsite  
chiar dacă sunt  
mascate

# Verificare cu RANSAC

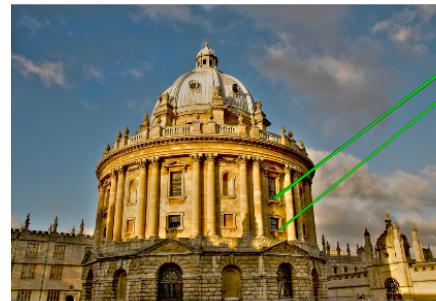
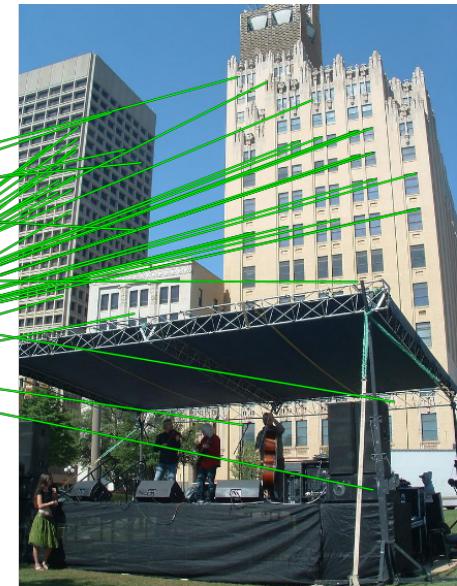
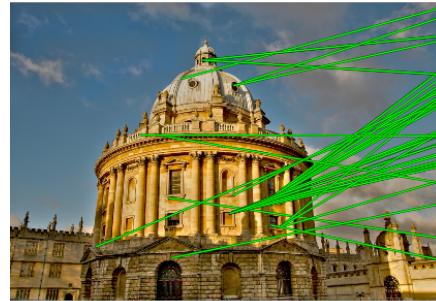
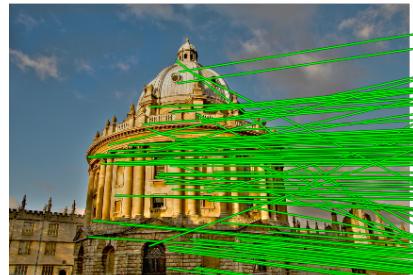


# Transformāri 2D



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

# Verificare cu RANSAC



# Probleme

- Cum summarizăm 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 multime de cuvinte vizuale pentru a putea recunoaște obiecte/scene specifice? Cum verificăm spațial acest lucru?
- Cum măsurăm rezultatele regăsirii pe baza unei imagini query?

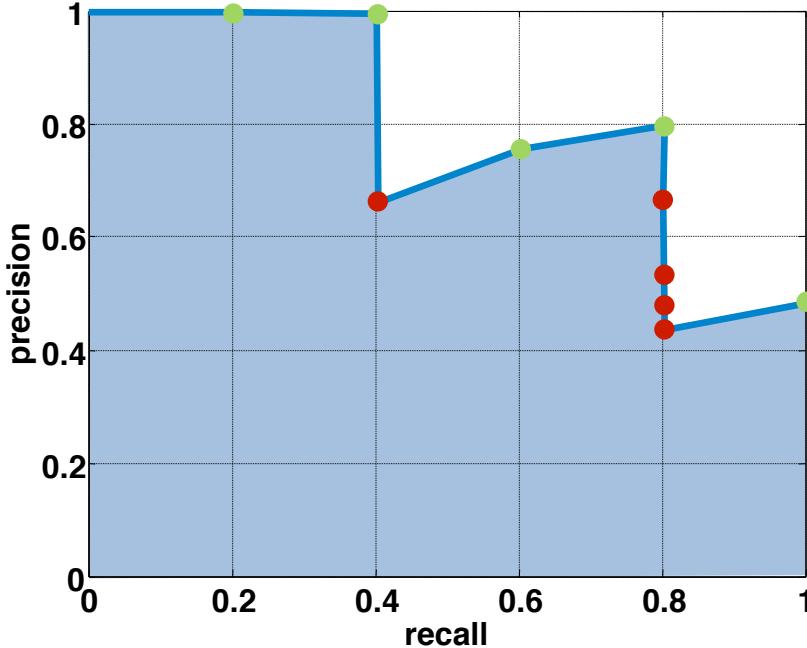
# Cum cuantificăm rezultatul?



Imagine query

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

$$\text{precizie} = \frac{\#\text{relevante}}{\#\text{returnate}}$$
$$\text{recall} = \frac{\#\text{relevante}}{\#\text{total relevante}}$$



Rezultate (ordonate):

