Adapted Vocabularies for Generic Visual Categorization

Florent Perronnin, Christopher Dance, Gabriela Csurka, and Marco Bressan

ECCV 2006

Presented by: Sainath Shenoy

Overview

- Generic Visual categorization
- Related Work
- Applications
- Technical approach
- Examples, results, performance
- Conclusion

Generic Visual Categorization

A **Pattern classification** problem which consists of assigning one or multiple **labels** to an image, based on its **semantic content**

Generic: cope with various object and scene types

plane coast interior

Challenge: handle variations in view, lighting, occlusion, and typical object and scene variations

Related Work

Geometric object models

Fergus, Perona, Zisserman (2003)

Object models = geometric constellations of parts,

Parts characterized by location, scale, and appearance,

But difficult to handle variable appearance and view points.

Identify common low level features to categorize efficiently

Torralba, Murphy, Freeman (2004)

Joint boosting for common low level features shared across different classes

Elegant approach to transferable learning (similar appearance)

But boosting is very costly for numerous classes

Related Work (2)

Use a universal vocabulary for categorization

Csurka, Dance, Fan, Willamowski and Bray (2004), Bishop and Ulusoy (2005), FeiFei and Perona (2005)

Based on analogy to text categorization

Defines visual vocabulary

Computes bags of key patches / visual words

Categorizes these bags

But adapted vocabularies allow for better performance

Applications

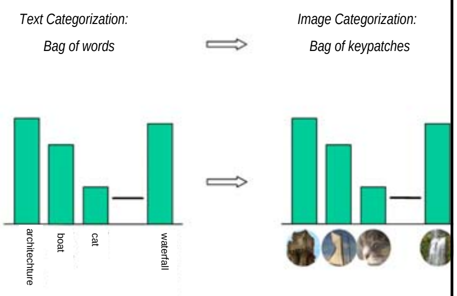
Tagging images with content:

- Web image retrieval
- Images in documents
- Photographic archives
- Consumer photo albums

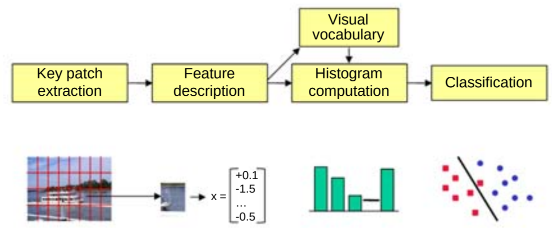
Assisting other processing:

- Image enhancement
- Image selection (illustration)

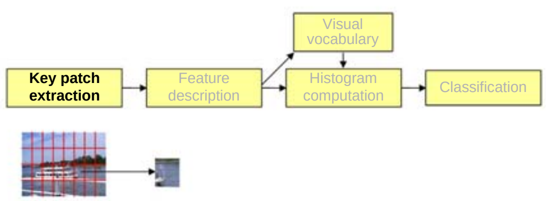
Technical approach:
Analogy to text categorization



Technical approach: system design



Technical approach: key patch extraction



Regular grid to **extract patches** at different scales capturing characteristic & distinctive information about the represented scene or objects.

Per image: ~ 500 patches

Technical approach:
key patch extraction

Key patch extraction:

Previously through key point detectors locating particular points in an image (e.g. important gray scale variations)

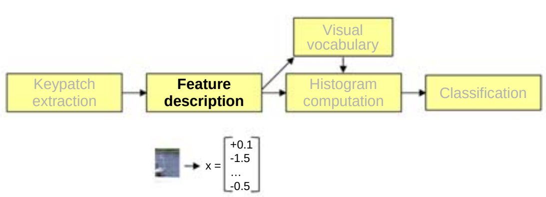
Worked well for object recognition
(handles occlusion, scale and view point variations)

Now applying a regular grid at different scales to the image

Fixed number of patches for each image

Better performance with respect to accuracy
Better performance with respect to speed

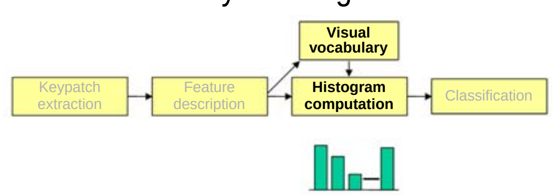
Technical approach: feature description



Feature descriptors are high-dimensional vectors capturing relevant appearance information, but ignoring irrelevant view, noise, lighting variations.

Feature descriptors describe orientation histograms within the patch region.
Per patch: one 128 dimensional descriptor, reduced to 50 using PCA

Technical approach:
visual vocabulary & histograms

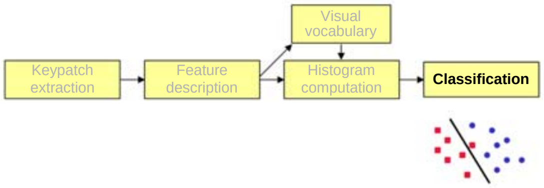


Feature descriptors are mapped to a **visual vocabulary**.

A **Histograms** counts the number of occurrences of the different *visual words* in each image.

Per image: one 1000 - 2000 dimensional histograms

Technical approach: system design



Classification is based on conventional machine learning approaches, particularly SVM, Boosting, Bayesian models.

Per category: one classifier

Technical approach:visual vocabulary

Purpose:

- Provide "mid-level" image representation
- Bridge semantic gap between low-level features and high-level concepts

Vocabulary Construction: *learned* from the training set

- Previous approaches:
  - Range from ***universal*** vocabularies (less accurate) to ***class specific*** vocabularies (costly)
  - Use ***hard*** or ***soft*** assignment of patch descriptors to visual words
- ***Their approach***:
  - First learn a universal vocabulary (MLE)
  - Then adapt it efficiently to obtain class specific vocabulary (MAP)
  - Use soft assignment of patch descriptors to visual words (GMMs)

Universal Vocabulary Training: MLE

E Step

$$\gamma_t(i) = p(i|x_t, \lambda^u) = \frac{w_i^u p_i^u(x_t)}{\sum_{j=1}^{N} w_j^u p_j^u(x_t)}.$$

M Step

$$\hat{w}_i^u = \frac{1}{T}\sum_{t=1}^{T}\gamma_t(i)$$

$$\hat{\mu}_i^u = \frac{\sum_{t=1}^{T}\gamma_t(i)x_t}{\sum_{t=1}^{T}\gamma_t(i)}$$

$$(\hat{\sigma}_i^u)^2 = \frac{\sum_{t=1}^{T}\gamma_t(i)x_t^2}{\sum_{t=1}^{T}\gamma_t(i)} - (\hat{\mu}_i^u)^2$$

Class Vocabulary Adaptation: MAP

E Step

$$\gamma_t(i) = p(i|x_t, \lambda^a).$$

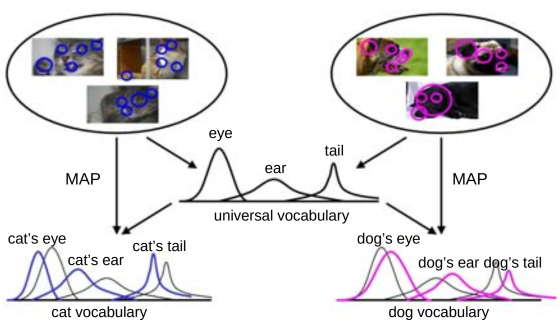
M Step

$$\hat{w}_i^a = \frac{\sum_{t=1}^{T}\gamma_t(i) + \tau_i^w}{T + \sum_{i=1}^{N}\tau_i^w},$$

$$\hat{\mu}_i^a = \frac{\sum_{t=1}^{T}\gamma_t(i)x_t + \tau_i^m \mu_i^u}{\sum_{t=1}^{T}\gamma_t(i) + \tau_i^m},$$

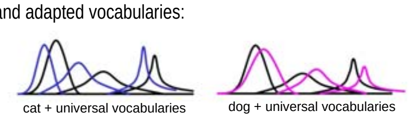
$$(\hat{\sigma}_i^a)^2 = \frac{\sum_{t=1}^{T}\gamma_t(i)x_t^2 + \tau_i^s\left((\sigma_i^u)^2 + (\mu_i^u)^2\right)}{\sum_{t=1}^{T}\gamma_t(i) + \tau_i^s} - (\hat{\mu}_i^a)^2 .$$

Universal and adapted vocabularies (1/2)



Universal and adapted vocabularies (2/2)

For each class, form a new vocabulary by merging the universal and adapted vocabularies:



Assumption:

- If an image belongs to a class, it is best described by the adapted vocabulary of this class
- If not, it is best described by the universal vocabulary

Technical approach: histograms

Why histograms?

- Bag of words histogram has worked well for text categorization.
- Easy to discard clutter (irrelevant words) which constitute the majority in most text (and image) categorization applications
- Simple and removes dependencies on word order

Bi-partite histograms:

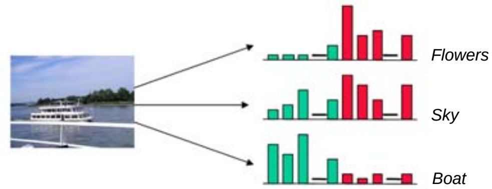
- Characterization of an image by comparing the relevance of the class-specific vocabularies and the universal vocabulary

Benefits:

- Significant improvement of the categorization accuracy

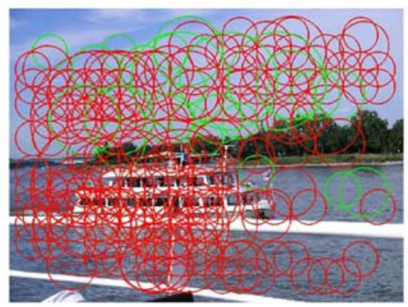
Technical approach: categorization

For each image and for each class, compute **bi-partite** histograms:

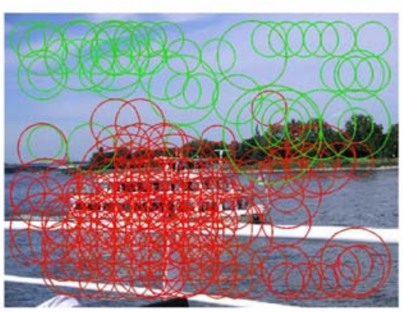


Key feature: **separate class-relevant information from irrelevant one**

Example: boat image



Patch relevance on *Flowers* category



Patch relevance on *Sky* category

Patch relevance on *Boat* category

Eight category results

Experiment:

Train XRCE categorizer on XRCE images.

Test on ***independent*** user images.

	% correct
Amusement Park	92.5 %
Boat	88.8 %
New York City	75.5 %
Sunrise&Sunset	90.0 %
Surfing	69.3 %
Tennis	93.6 %
Underwater	88.2 %
Waterfalls	90.3 %
average	86.0%

Eight category results

Problems:

Images often rather multi- than mono-label,
e.g. ***Surfing*** images often contain ***boat***s

Category concepts not always concordant

difference between training and test set

27 categories results (RevealThis)

Various categories relevant for *Travel***:**

AmusementPark, Animal, Archaeology, ArtsObjects, Beach, Boat, Buildings, Coast, Countryside, Desert, Face, Flowers, Interior, Map, Mountain, Painting, Persons, Plane, SkyActivity, TerrainSports, Train, Trees, Underwater, Vehicle, WaterActivity, Waterscape, WinterActivity

Results obtained with 5-fold Cross-Validation

on *homogeneous* **set:**

Correct rate between 50% (***Animal, ArtsObjects***)

and > 80% (***Interior, Map***)

Overall 65% correct rate

Typical (obvious) confusions between classes:

Desert ↔ ***Beach, Persons*** ↔ ***Face***

Results on 4 (very different) categories

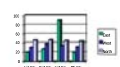
Categories relevant

e.g. within scientific books or articles

Maps (97%), ***Tables*** (96 %),

Graphics (94 %), ***Charts*** (92 %).

	G1	G2	G3
Y1	12	19	55
Y2	55	44	11



3-fold Cross-Validation Results

Overall 95% correct rate.

Comparison

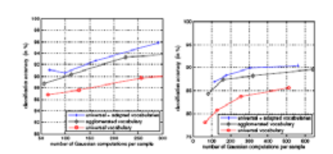


Fig. 4. Comparison of the proposed approach (universal + adapted vocabularies) with the two baseline systems (universal vocabulary and agglomerated vocabulary) on (a) the LAVA7 database and (b) the Wang database

Conclusion

A generic visual categorizer that:

Scales well with the number of categories added

Performs well (low error rate and run-time) on diverse generic categories without task dependent “tweaking” or manual operations with training data

Is extensibly engineered around a simple text-categorization analog

Work in Progress

Consider multi-label images

Integrate color information

References

Florent Perronnin, Christopher Dance, Gabriela Csurka, and Marco Bressan (ECCV 2006)

Fergus, Perona, Zisserman (2003)

Torralba, Murphy, Freeman (2004)

Csurka, Dance, Fan, Willamowski and Bray (2004),
Bishop and Ulusoy (2005), FeiFei and Perona (2005)

Slides and material taken from

http://www.igd.fraunhofer.de/igd-a7/mir2005/Day1/04-Session2/02-Jutta-Willamowski/Jutta-Willamowski.pdf