All You Should Know on Visual Recognition Pipelines: from theory to iCub

Sean Ryan Fanello
iCub Facility – Istituto Italiano di Tecnologia
Motivations

Human Robot Interaction is a **new** and **natural** application for object recognition in robotics settings, where strong cues are often available, allowing object detectors to be avoided.

Recognition as a tool for complex tasks: grasp, manipulation, affordances, pose


**Introduction**

**Single Instance Recognition**

1. Image ➔ Recognition System ➔ Pattern Recognition & Machine Learning

**Object Categorization**

1. Image ➔ Recognition System ➔ A book

**Image Retrieval**

1. Image ➔ Recognition System ➔ "Same" Image Representations

*All You Should Know on Visual Recognition Pipelines: from theory to iCub - S.R. Fanello*
Introduction

- Geometric Information
- Invariance to Image Transformation
- Discriminative Power
- Real-Time

Still an open problem…
We want to convert an image into a compact feature vector \( \mathbf{v} \in \mathbb{R}^n \)

It must be robust to viewpoint, illumination, occlusion etc. Ideally an **Invariant Representation**

Invariant to: scale, occlusions, lighting, view-point

Example: **Color Histogram**

Nice invariance properties, but not *informative*

Trade-off between invariance and information

More Data $\iff$ Invariance is not needed
Global Descriptors

Color Histogram
Swain, Ballard, “Color indexing”, IJCV’91.

GIST of a Scene
Oliva, Torralba, “Modeling the shape of the scene: a holistic representation of the spatial envelope”, IJCV’01.

CENTRIST: Census Transform hIStogram

Highly efficient to compute and to match, but they lack of description power. Not suitable for discriminative tasks.
A set of local descriptors is extracted. These features are invariant to geometric transformations. Most widely used: SIFTs

8 orientations of the gradient in 4x4 spatial grid. 128 dimensions

Notice: Local features are invariant when both keypoint detector and descriptor are used. In object categorization detectors do not perform well due to intra-class variability
Current state-of-the-art recognition pipelines use hierarchical image representations.

Main Idea:
Starting from low-level descriptors (e.g., SIFT) we compute higher order statistics with a sequence of coding-pooling operators.

BOW-like systems have the first layer hand crafted (e.g., SIFT)
Pro: Efficient, easy to implement
Cons: No invariance

Deep Architectures (e.g., HMAX) try to learn also the first layer
Pro: Invariant to small transformation, Higher (?) accuracy
Cons: TOO slow
Local Descriptors: $x_1, \ldots, x_M$ are extracted from the image. Keypoint detectors are not suitable, a dense grid better catches image statistics.

These descriptors are not invariant to scale, rotation and translation etc.

Main Assumption: Enough Data is available
**Dictionary Learning:**

We want to relate local descriptors to a common basis, called dictionary. Intuitively, a dictionary $D = [\mu_1, \ldots, \mu_K]$ represents all the relevant descriptors and it is able to describe the content of any image. The dictionary is learned using K-Means algorithm with $K=$dictionary size on a large set of SIFTs (~1M)

$$\min_{D,U} \|X - DU\|_F^2$$

s.t. $\text{Card}(u_i) = 1, |u_i| = 1, u_i \geq 0, \forall i = 1, \ldots, T$
Visual Recognition Pipelines

**Coding Operators**
The coding stage maps the input features $x_1, \ldots, x_M$ into an overcomplete space.

**Based on the Reconstruction Error**

$$u_i = \arg \min_u \|x - Du\|^2 + \lambda R(u)$$

s.t. $C(u)$

**Higher Order Statistics**

$$u_k = \frac{1}{M \sqrt{\pi k}} \sum_{i=1}^{M} q_{ik} \Sigma_k^{-\frac{1}{2}}(x_i - \mu_k)$$

$$v_k = \frac{1}{M \sqrt{2\pi k}} \sum_{i=1}^{M} q_{ik} [(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k) - 1]$$
Coding Operators

Vector Quantization (VQ)

\[
\min_{u_i} \| x_i - Du_i \|^2 \\
\text{s.t. Card}(u_i) = 1, |u_i| = 1, u_i \geq 0
\]

Sparse Coding (SC)

\[
\min_{u_i} \| x_i - Du_i \|^2 + \lambda \| u_i \|_1
\]

Locality-constrained Linear Coding (LLC)

\[
\min_{\bar{u}_i} \| x_i - \bar{D}\bar{u}_i \|^2 \\
\text{s.t. } 1^T\bar{u}_i = 1
\]

They all minimize the reconstruction error:

\[
u_i = \arg \min_u \| x - Du \|^2 + \lambda R(u) \\
\text{s.t. } C(u)
\]
Visual Recognition Pipelines

Image → Local Coding-Pooling → Coding Stage → Pooling Stage → Classifier

Spatial Pyramid Representation
Image partitioned in $2^l \times 2^l$ segments

Example: Max Pooling
$h_{s,j} = \max_{i \in Y_s} u_{i,j} \quad \forall j = 1, \ldots, K$

Pooling Operators
They aggregate local descriptors into a single one.

21 Spatial Bins
Visual Recognition Pipelines

**Learning Stage**
After the pooling stage we get a descriptor \( z \in \mathbb{R}^{KS} \) where \( K \) is the dictionary size and \( S \) the number of cells of the spatial pyramid.

Multi-class classification problem that can be solved with a standard One-vs-All paradigm. Linear Classifiers are more suitable for Real-Time tasks.

\[
\text{Minimize} \quad \lambda \Omega(W) + \frac{1}{n} \sum_{i=1}^{n} L(y_i, W^T x_i)
\]

Non-linear feature mapping can be used to approximate non-linear Kernels
A. Vedaldi, A. Zisserman – Efficient additive kernels via explicit feature maps. CVPR 2010
Recap

**Local Descriptors**
Dense Grid of SIFT descriptors with fixed scale and grid spacing

**Dictionary Learning**
A dictionary is learned with K-Means (K=dictionary size) on a large set of SIFTs

**Coding-Pooling**
Features are mapped into an overcomplete space and then aggregated together. The image is transformed into a single vector \( \mathbf{v} \in \mathbb{R}^n \)

**Linear Classification**
Use your favorite classification method (GURLS!) with One-vs-All paradigm
Generalizing the Pipeline

Image

i-th Object Class

e.g. SIFT, HOG, Patch

Coding Stage

e.g. VQ, SC, LLC, VLAD, SV, FK

Pooling Stage

e.g. SPR

Classifier

e.g. SVM, RLS, MRF, BN

Main Modifications

• Image Representation from stereo depth
• Motion and Appearance
• Temporal Pooling
• Video Segmentation

Action/Gesture Recognition

RGB and DEPTH STREAMS

RGB

Depth via Stereo Vision

Frame Representation

Motion

Appearance

Coding/Pooling Stage

Action Learning & Recognition

Classifier

Video Segmentation

All You Should Know on Visual Recognition Pipelines: from theory to iCub - S.R. Fanello
iCubWorld Single Instance Recognition

Object Categorization between Computer Vision & Robotics

Main Difficulty: **structured clutter**, therefore the context cannot be exploited.

**iCubWorld available at:** [http://www.iit.it/it/projects/data-sets.html](http://www.iit.it/it/projects/data-sets.html)
Object Categorization with iCub

Sean Ryan Fanello, Ugo Pattacini, Nicoletta Noceti, Giorgio Metta, Francesca Odone

iCub Facility - Istituto Italiano di Tecnologia
DIBRIS - Università degli Studi di Genova

Example of application with visual recognition capabilities

Classifier (GURLS!)

Class \( v \in \mathbb{R}^n \)

Your Yarp Module

Image \( v \in \mathbb{R}^n \)

SparseCoder

Build your own app
SparseCoder Module

Where?
$ICUB_ROOT/contrib/src/sparseCoder$

Dependencies
SIFT GPU, OpenCV 2.X, Yarp

Docs:

Coding Methods:
Best Code Entries, Sparse Coding, Bag of Words

Spatial Pyramid Matching:
Yes, customizable pyramid levels

Dictionary Learning:
With K-Means

More Functionalities:
PCA on SIFTs, Power Normalization, Sparse & Dense SIFTs
Thank You!
Any Question?