Deep Representation for Recognition

ALI - HOSSEIN

Baselines

- What systems do you use as baseline for a new task?

- Some Representations
  - SIFT
  - HOG
  - GIST
  - LBP
  - BOW
  - Shape context
  - Contours

Baselines

- A Classifier
  - SVM
  - Random Forest
  - Boosting
  - Logistic Regression
  - MKL

Baselines

- Some Encoding of Geometry
  - Spatial Pyramid Matching (SPM)
  - Image Gridding
  - DPM
  - Joint features
  - Geometry encoded features (Pose estimation random forests)

Baselines

- And some are more sophisticated
  - Feature combination
  - Fusion of different classifiers
  - Segmentation
  - Hiearchical Representation
  - Mixture Modelling
  - Latent Structures
  - Strong Supervision

What if…

- There is one representation which used in a simple learning machinery (e.g. linear SVM)
  - beats all simpler baselines
  - better or on a par with more sophisticated baselines
  - for many (if not all) recognition tasks
  - No magic…

Deep Representation (OverFeat)

- Apparently there is one around!

- OverFeat: A CNN trained on ILSVRC 2013 for the task of Object Image Classification

- third to the last layer (4096 dimensional)

- very fast to compute on GPU in O(mili-sec)

Layer	1	2	3	4	5	6	7	Output 8
Stage	conv + max	conv + max	conv	conv	conv + max	full	full	full
# channels	96	256	512	1024	1024	3072	4096	1000
Filter size	11x11	5x5	3x3	3x3	3x3	-	-	-
Conv. stride	4x4	1x1	1x1	1x1	1x1	-	-	-
Pooling size	2x2	2x2	-	-	2x2	-	-	-
Pooling stride	2x2	2x2	-	-	2x2	-	-	-
Zero-Padding size	-	-	1x1x1x1	1x1x1x1	1x1x1x1	-	-	-
Spatial input size	231x231	24x24	12x12	12x12	12x12	6x6	1x1	1x1

The time of low scoring baselines is gone!

Image Classification (Pascal VOC Object)

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog
GHM[11]	76.7	74.7	53.8	72.1	40.4	71.7	83.6	66.5	52.5	57.5	62.8	51.1
AGS[13]	82.2	**83.0**	58.4	76.1	**56.4**	**77.5**	**88.8**	69.1	**62.2**	61.8	64.2	51.3
NUS[33]	82.5	79.6	64.8	73.4	54.2	75.0	77.5	79.2	46.2	**62.7**	41.4	74.6
CNN-SVM	**88.5**	81.0	**83.5**	**82.0**	42.0	72.5	85.3	**81.6**	59.9	58.5	**66.5**	**77.8**

Table 1: Pascal VOC 2007 Image Class

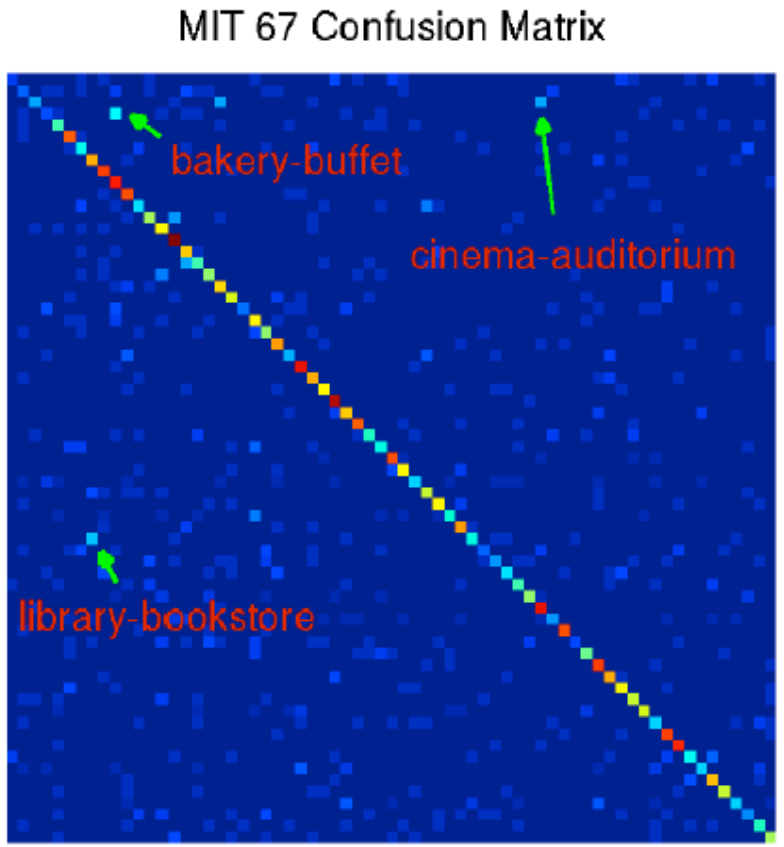
r	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
5	57.5	62.8	51.1	81.4	71.5	86.5	36.4	55.3	60.6	80.6	57.8	64.7
2	61.8	64.2	51.3	**85.4**	**80.2**	**91.1**	48.1	61.7	**67.7**	86.3	70.9	71.1
2	**62.7**	41.4	74.6	85.0	76.8	**91.1**	53.9	61.0	67.5	83.6	70.6	70.5
9	58.5	**66.5**	**77.8**	81.8	78.8	90.2	**54.8**	**71.1**	62.6	**87.2**	**71.8**	**73.9**

2007 Image Classification Results

Image Classification (MIT 67 Indoor Scenes)

method	mean Accuracy
ROI + Gist[31]	26.05
DPM[25]	30.40
Object Bank[20]	37.60
RBow[26]	37.93
BoP[18]	46.10
miSVM[21]	46.40
D-Parts[34]	51.40
IFV[18]	60.77
MLrep[12]	**64.03**
CNN-SVM	58.44

Table 2: Results on MIT 67 indoor scenes dataset



OK, maybe obvious…

Fine-grained Recognition (CUB Bird 200)

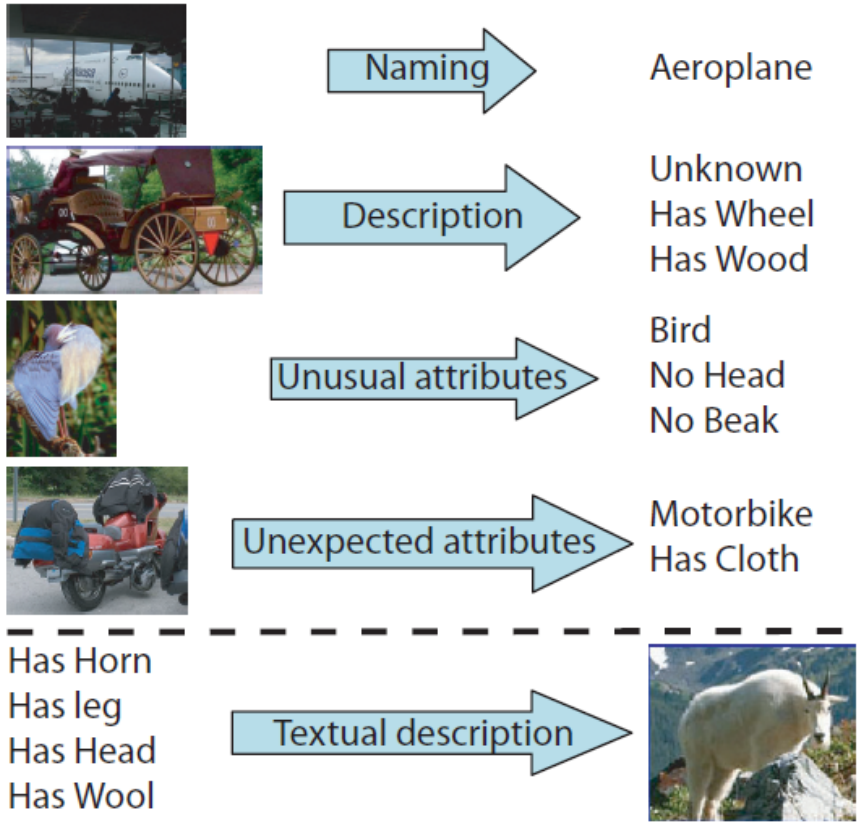
method	mean Accuracy
Sift+Color+SVM[35]	17.31
CNN-SVM	**53.29**
Pose pooling kernel[37]	28.2
RF[36]	19.2
DPD[38]	50.98
Poof[7]	**56.78**

Fine-grained Recognition (Oxford 102 flowers)

Can we push it further?!

method	mean Accuracy
HSV [22]	43.0
SIFT internal [22]	55.1
SIFT boundary [22]	32.0
HOG [22]	49.6
HSV+SIFTi+SIFTb+HOG(MKL) [22]	72.8
BOW(4000) [6]	65.5
SPM(4000) [6]	67.4
FLH(100) [6]	72.7
BiCos seg [9]	79.4
[3] w/o seg	76.7
[3]	**80.66**
CNN w/o seg	74.7

Attribute Detection (Pascal Object Attributes)



method	within category	across category	mAUC (mAP)
Farhadi et. al	83.4	-	73.0
Latent Model	62.16	79.88	-
Attribute Feedback			76.0
CNN-SVM	91.67	82.23	89.04 (53.59)

Attribute Detection (H3D human attributes)

method	male	long hair	glasses	hat	tshirt	long slvs	shorts	jeans	long pnts	mAP
SPM[8]	68.1	40.0	25.9	35.3	30.6	58.0	31.4	39.5	84.3	45.91
Poselets[8]	82.4	**72.5**	**55.6**	60.1	51.2	74.2	45.5	54.7	90.3	65.18
DPD[38]	**83.7**	70.0	38.1	**73.4**	49.8	78.1	64.1	**78.1**	93.5	69.88
CNN-SVM	83.0	67.6	39.7	66.8	**52.6**	**82.2**	**78.2**	71.7	**95.2**	**70.78**

Wow! what else?!

Visual Similarity (Face Verification)

Let’s see if we can break it down

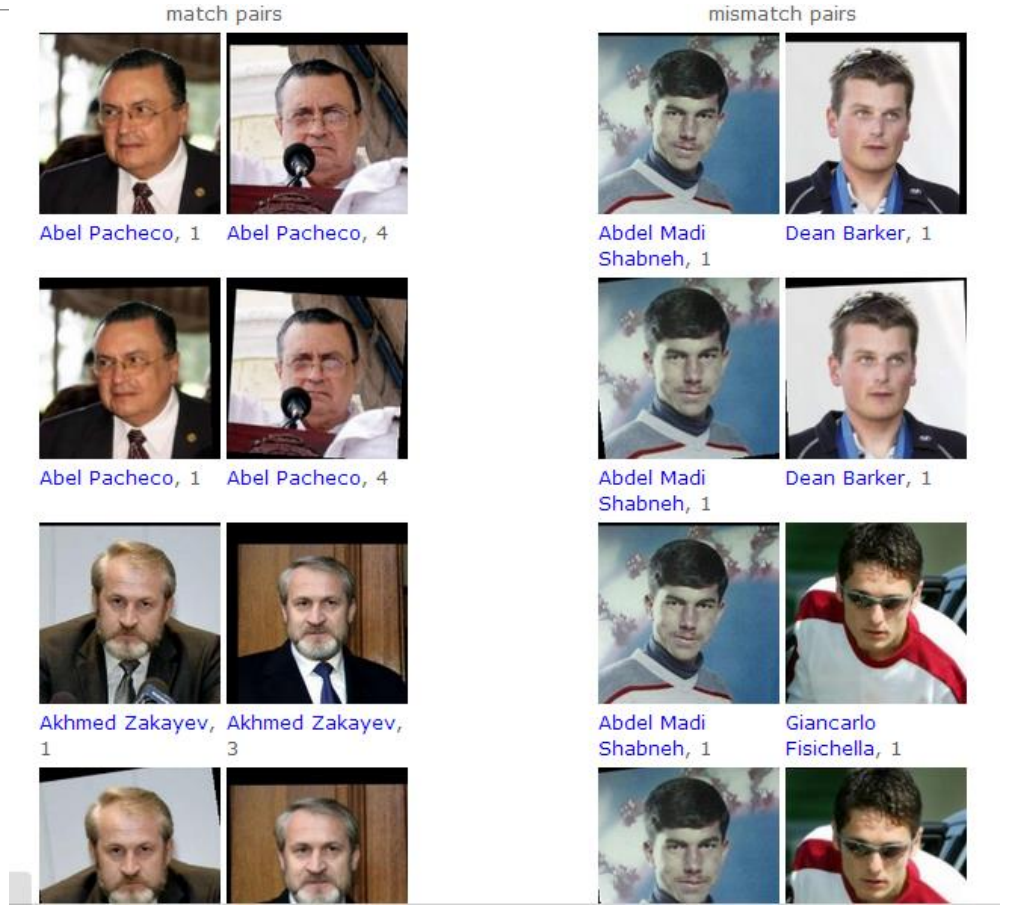


Image Instance Retrieval (Buildings)

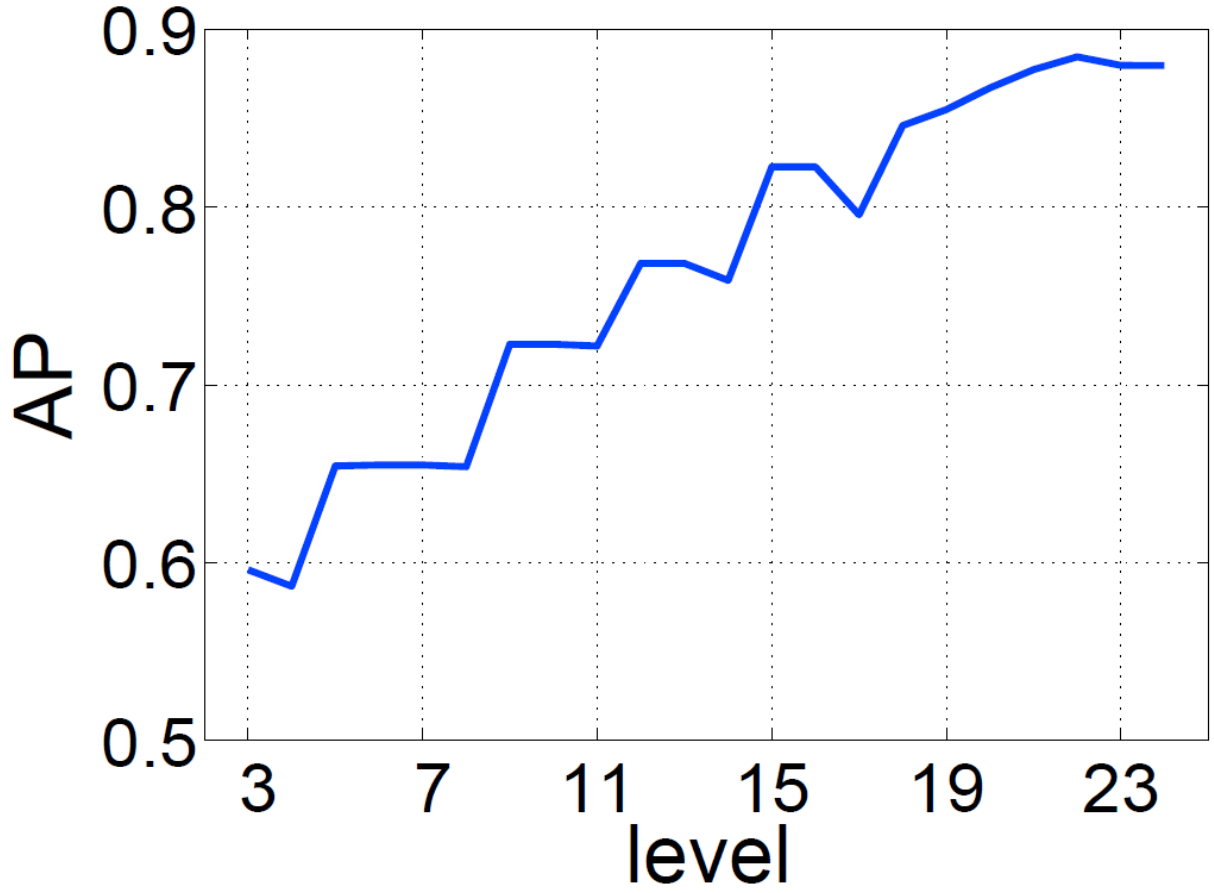
Nope!

Image Instance Retrieval (Sculptures)

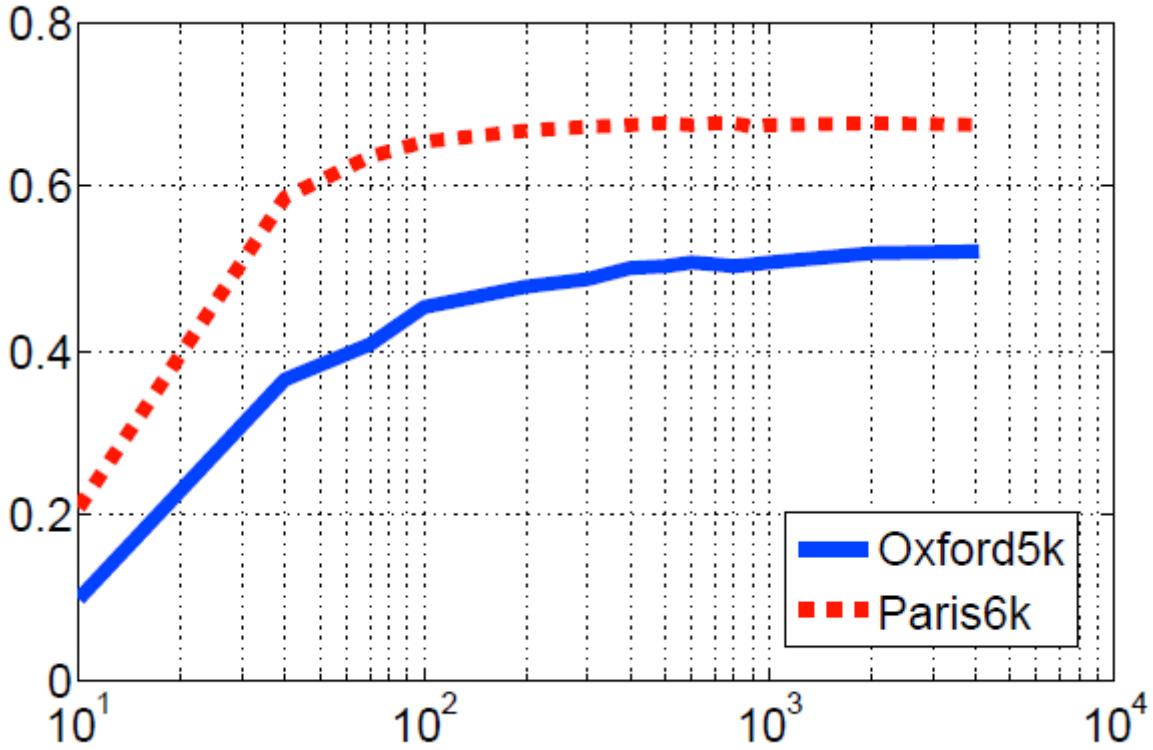
Image Instance Retrieval (General Objects)

	Oxford5k [29]	Paris6k[30]	sculpture6k[4]	Holidays [15]	UKBench[23]
VLAD 64D + SSR [16]	0.304 [16]	0.521	-	0.556[16]	3.28 [16]
VLAD 64D + innorm [5]	**0.555**[5]	0.642	-	**0.646**	**3.38**
BoW 200kD	0.364[16]	0.460[30]	0.086[4]	0.540	2.81 [16]
IFV 64D [28]	0.418 [5]	-	-	0.626	3.35 [16]
BoB	N/A	N/A	0.253[4]	N/A	N/A
Deep Rep	0.520	**0.676**	**0.269**	**0.646**	3.05

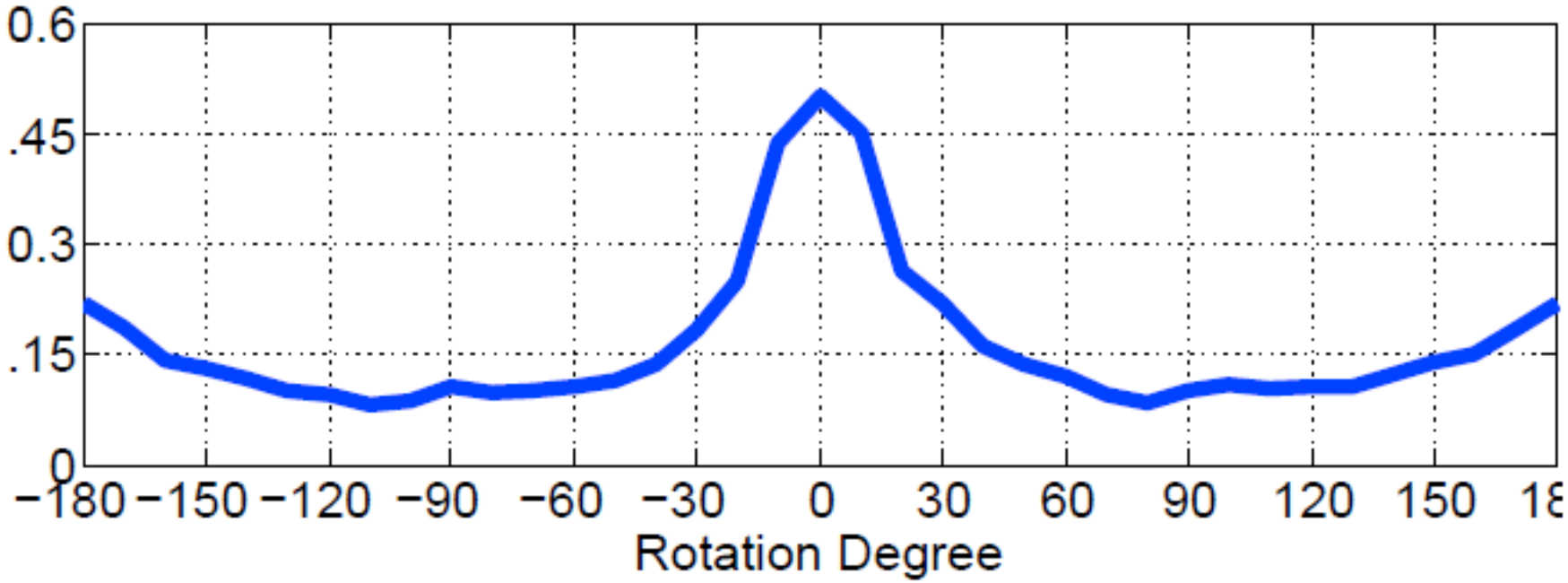
Different Levels



Even Lower Dimensions…



Rotation invariant?



Well…

Let’s give some credit to computer vision….