iV&L Net Training School 2015

Tutorial on Datasets, Annotations and Metrics in Computer Vision

iV&LSS 2015

integrating Vision & Language

Dr. Sergio Escalera Guerrero
ChaLearn, Computer Vision Center (UAB) and University of Barcelona
Head of Human Pose Recovery and Behavior Analysis Group
Vice-chair of International Association on Pattern Recognition IAPR-TC12: Visual and multimedia systems
sergio@maia.ub.es

Leuven, Juve 1-4, 2015

INDEX

How many CV DBs are available?

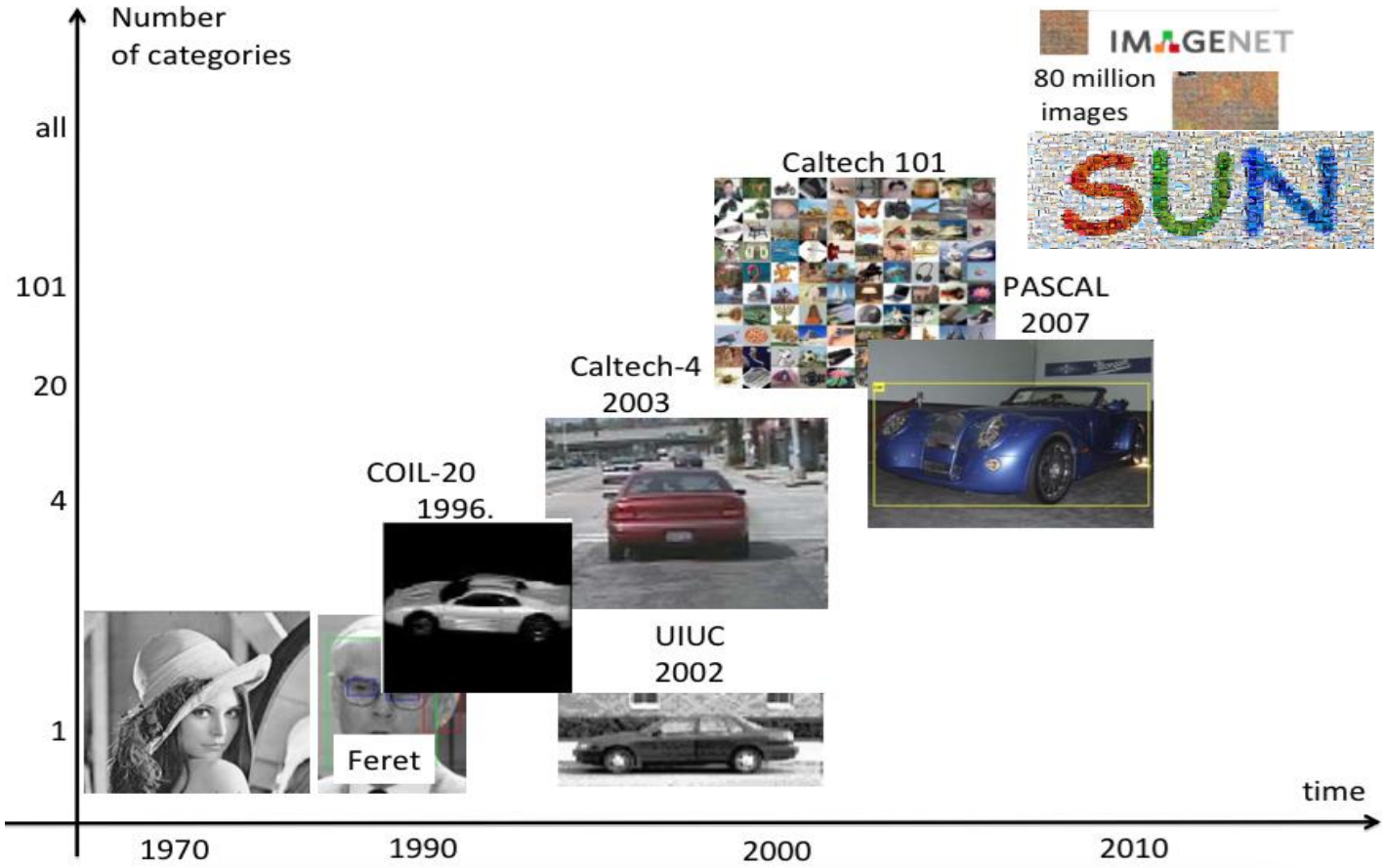
- Countless
  - Several research papers propose new datasets
  - Most of them are used for comparative among works in the field
  - Some of them are then selected for future challenges:
    - E.g. OpenCV CVPR 2015 Vision challenge:
      - **Vision Challenge**
      - OpenCV is launching a community-wide challenge to update and extend the OpenCV library. An award pool of $50,000 will be provided to the best performing algorithms in the following 11 CV application areas:

•image segmentation
•image registration
•human pose estimation
•SLAM (Simultaneous localization and mapping)
•multi-view stereo matching
•object recognition

•face recognition
•gesture recognition
•action recognition
•text recognition
•tracking

Computer vision datasets

- Image/video datasets have increased in size in a dramatic way recently.



From: The Promise and Perils of Benchmark Datasets and Challenges. David Forsyth, Alyosha Efros, Fei-Fei Li, Antonio Torralba and Andrew Zisserman. Frontiers in Computer Vision Workshop, CVPR 2011.

Which is the recognition tasks in CV databases?

OpenCV challenge is very representative of the different topics:

- image segmentation
- image registration
- human pose estimation
- SLAM
- multi-view stereo matching
- object recognition
- face recognition
- gesture recognition
- action recognition
- text recognition
- tracking

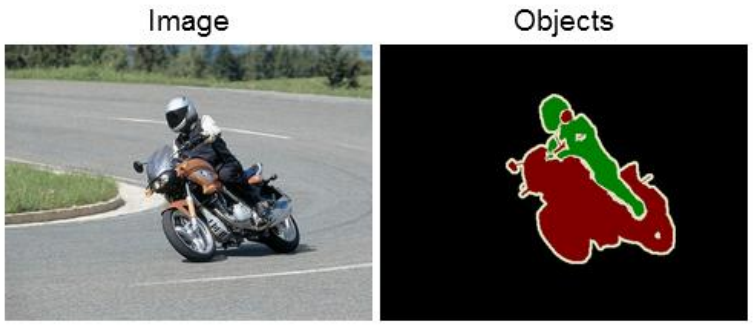
- Which is the input data:
  - RGB
  - Other visual representations: Depth, Thermal
  - Multi-modal multi-disciplinary in several cases: in combination with inertial sensors or audio, involving signal processing, pattern recognition, machine learning, computer vision, NLP, etc.

Examples of data for different computer vision topics

» image segmentation

Segmentation Competition

- **Segmentation**: Generating pixel-wise segmentations giving the class of the object visible at each pixel, or "background" otherwise.

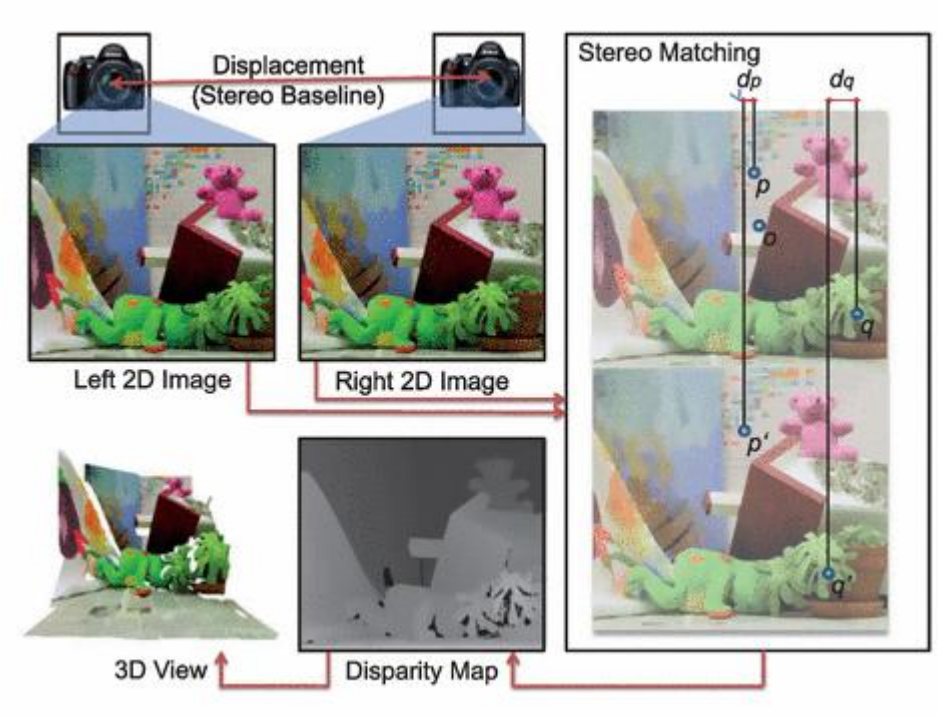


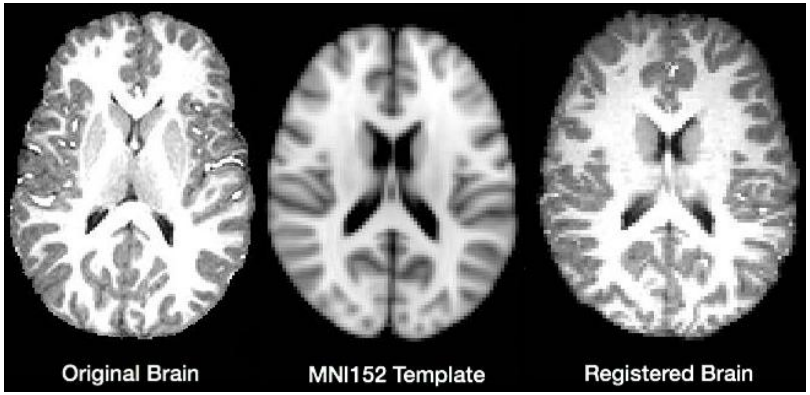
http://pascallin.ecs.soton.ac.uk/challenges/VOC/

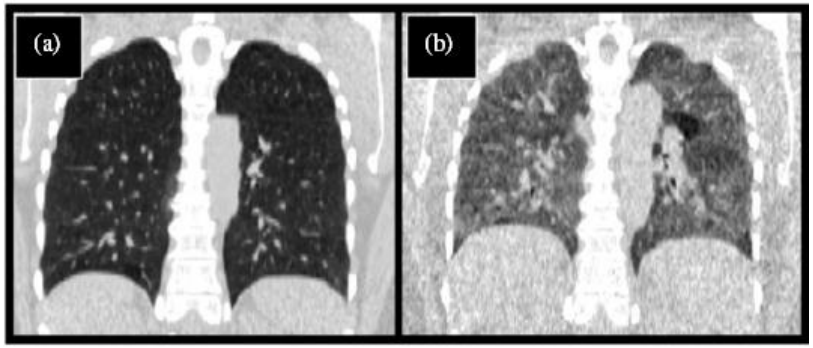
http://grand-challenge.org/All_Challenges/

Examples of data for different computer vision topics

» image registration and multi-view stereo matching







http://www.dir-lab.com/ReferenceData.html

Examples of data for different computer vision topics

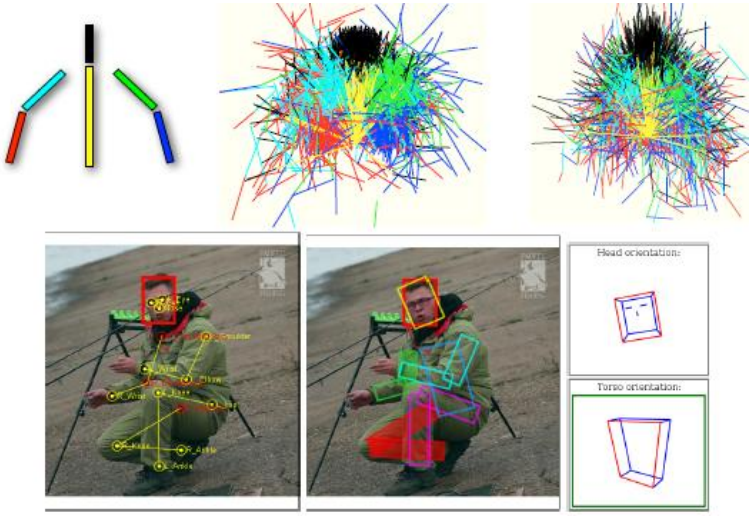
Human Layout Pascal 2007-2010 competitions

Person Layout: Predicting the bounding box and label of each part of a person (head, hands, feet).

Image | Person Layout

http://pascallin.ecs.soton.ac.uk/challenges/VOC/

Human Pose Estimation: New Benchmark and State of the Art Analysis CVPR 2014



Dataset	#training	#test	img. type
Full body pose datasets			
Parse [16]	100	205	diverse
LSP [12]	1,000	1,000	sports (8 types)
PASCAL Person Layout [6]	850	849	everyday
Sport [21]	649	650	sports
UIUC people [21]	346	247	sports (2 types)
LSP extended [13]	10,000	-	sports (3 types)
FashionPose [2]	6,530	775	fashion blogs
J-HMDB [11]	31,838	-	diverse (21 act.)
Upper body pose datasets			
Buffy Stickmen [8]	472	276	TV show (Buffy)
ETHZ PASCAL Stickmen [3]	-	549	PASCAL VOC
Human Obj. Int. (HOI) [23]	180	120	sports (6 types)
We Are Family [5]	350 imgs.	175 imgs.	group photos
Video Pose 2 [18]	766	519	TV show (Friends)
FLIC [17]	6,543	1,016	feature movies
Sync. Activities [4]	-	357 imgs.	dance / aerobics
Armlets [9]	9,593	2,996	PASCAL VOC/Flickr
MPII Human Pose (this paper)	**28,821**	**11,701**	diverse (491 act.)

Examples of data for different computer vision topics

» object recognition

Visual Object Classes Challenge 2012 (VOC2012)

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

Classification/Detection Competitions

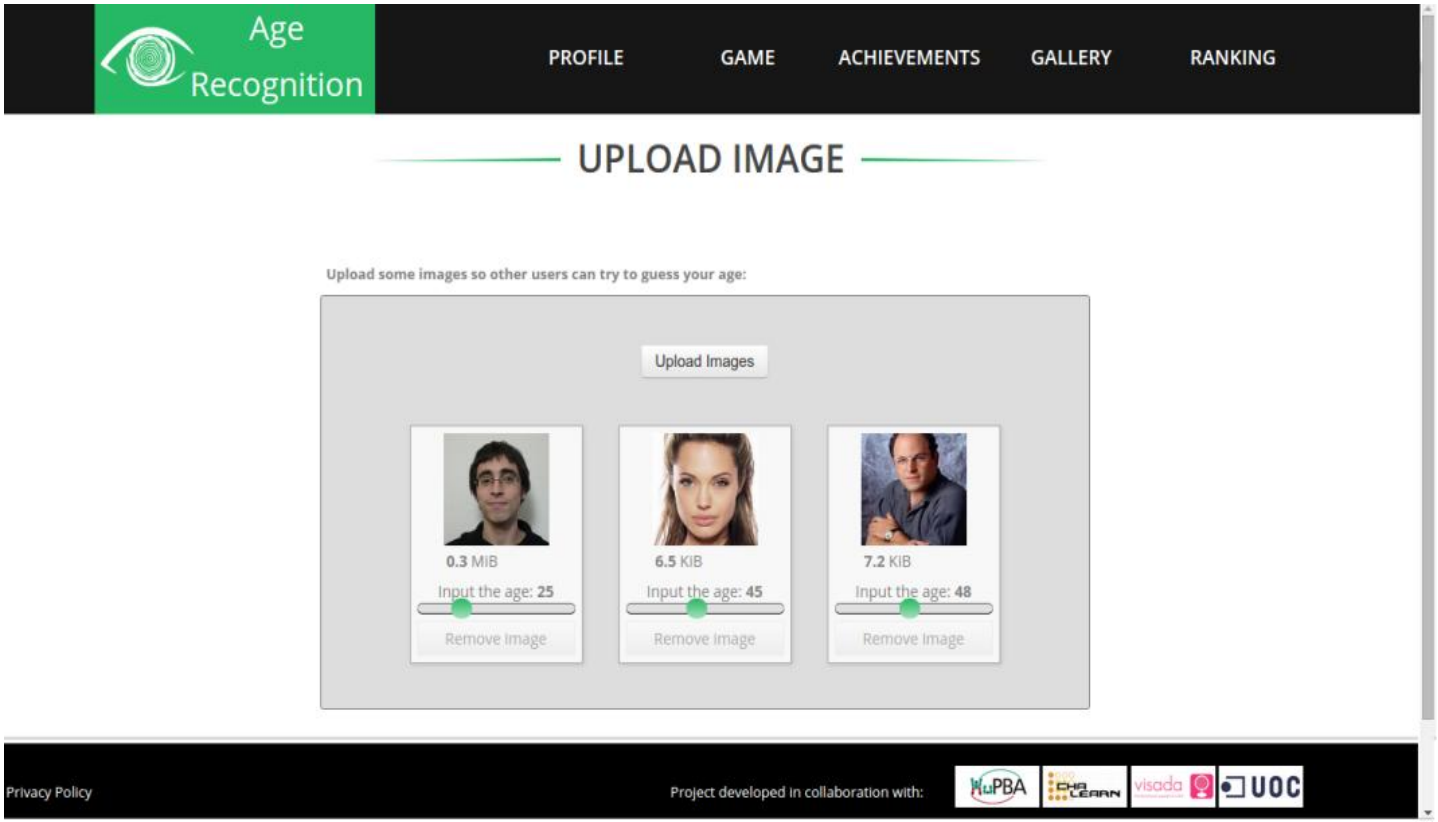
1. **Classification**: For each of the twenty classes, predicting presence/absence of an example of that class in the test image.
2. **Detection**: Predicting the bounding box and label of each object from the twenty target classes in the test image.

20 classes

Examples of data for different computer vision topics

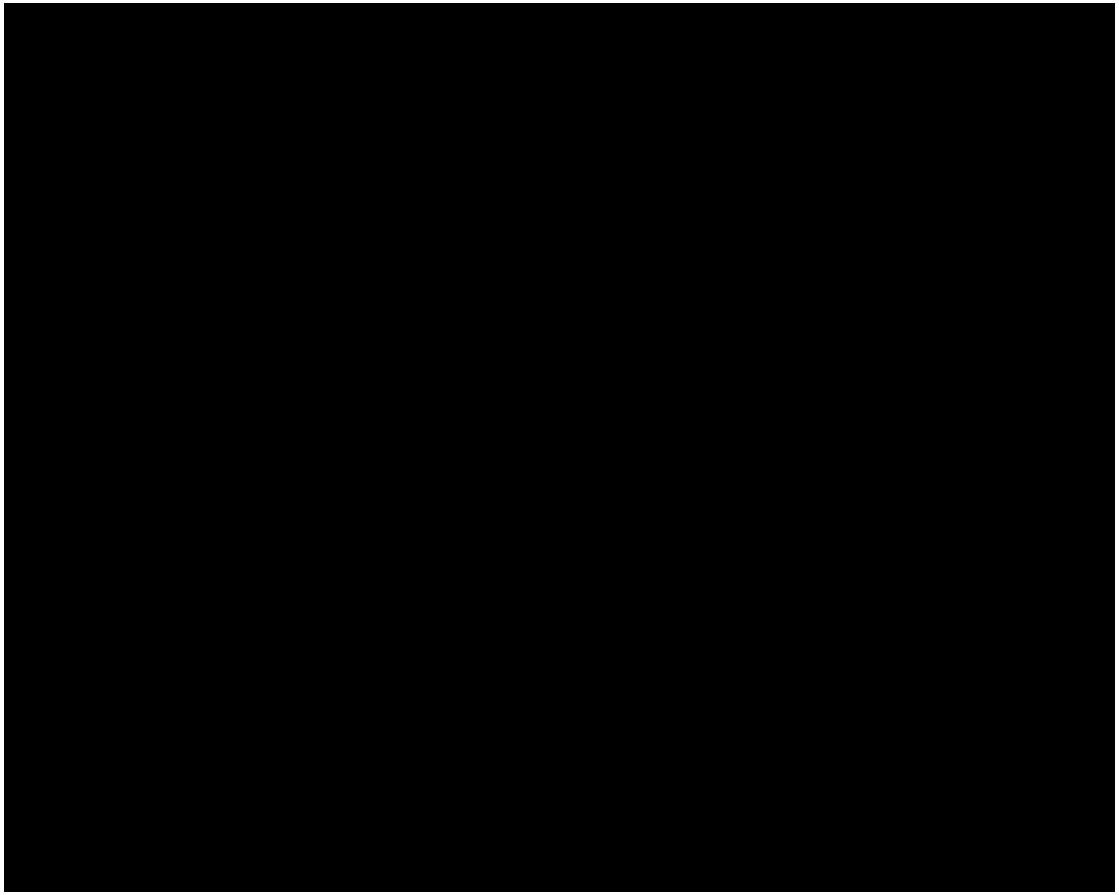
» face recognition (identification, age estimation, gender recognition, facial expression analysis)

ChaLearn Looking at People
http://gesture.chalearn.org/



Examples of data for different computer vision topics

» gesture recognition



http://gesture.chalearn.org/

Examples of data for different computer vision topics

» text recognition


TECHNOLOGIST
Andreas Tzortzis

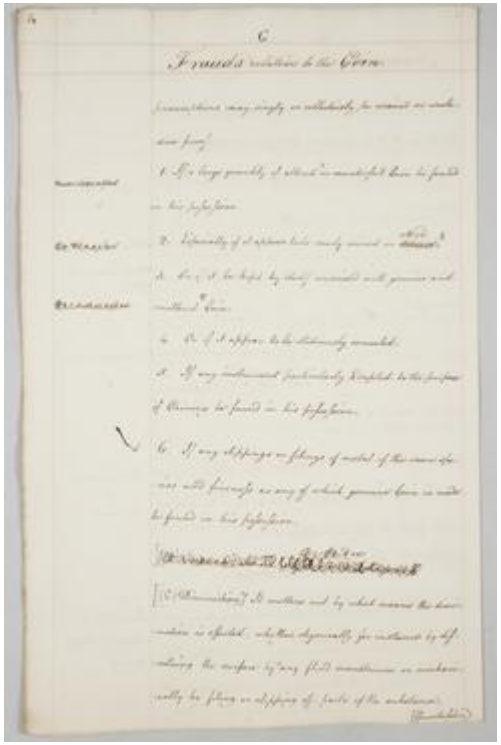
Steal These Books Online

Pictures That Move In 3-D

Paulo Coelho is championing a book-promotion model that is gaining momentum among fellow, if lesser known, authors.

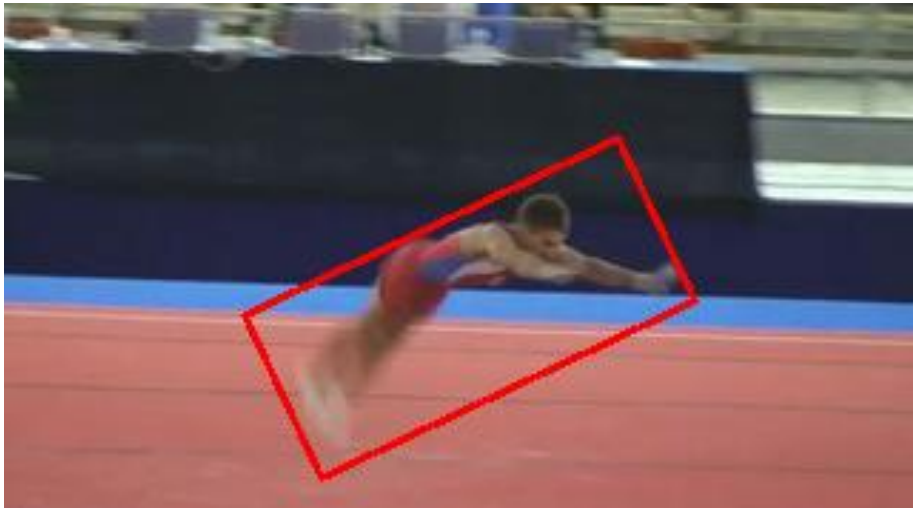
To Our Readers

Our Eyes and Ears. One veteran correspondent returns to Baghdad to evaluate the war, and we honor another as she prepares to take her final bow



Examples of data for different computer vision topics

» Tracking

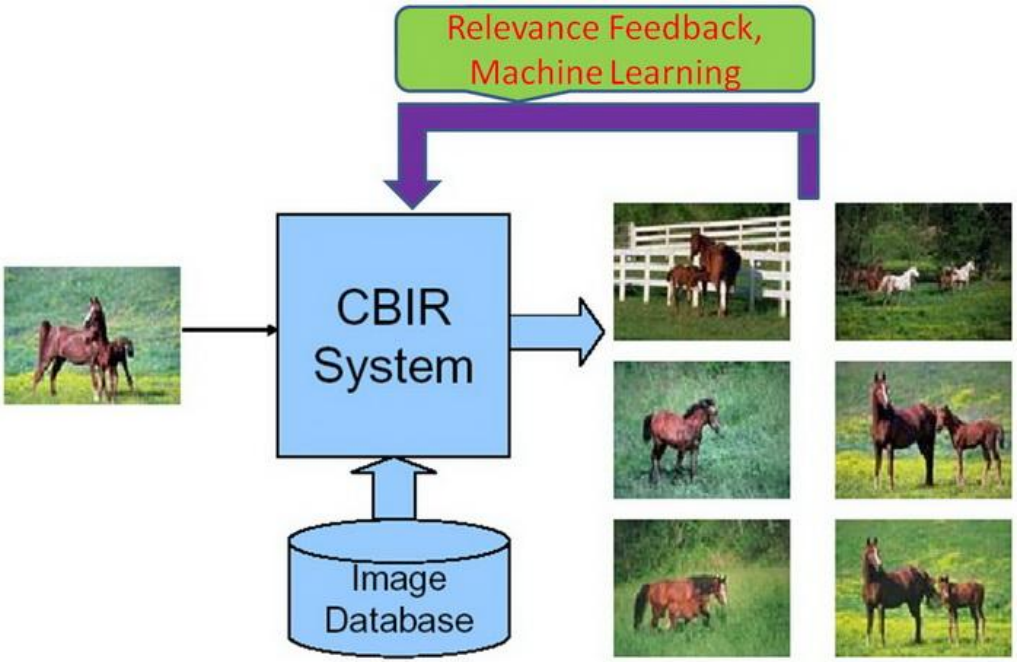


frameN: X1, Y1, X2, Y2, X3, Y3, X4, Y4

http://www.votchallenge.net/vot2014/

Examples of data for different computer vision topics

» Image retrieval



Microsoft Research

MSR-Bing Image Retrieval Challenge (IRC)

http://research.microsoft.com/en-us/projects/irc/

INDEX

image segmentation

Segmentation Competition

- **Segmentation**: Generating pixel-wise segmentations giving the class of the object visible at each pixel, or "background" otherwise.

Image

Objects

Annotation: label at pixel level
Don't care regions: boundaries

- Per-pixel classification

$$performance = \frac{number\_of\_corrected\_classified\_pixels}{total\_number\_of\_pixels}$$

- Per-class classification

$$performance = \frac{1}{C}\sum_{i=1}^{C}\frac{number\_of\_corrected\_classified\_pixels\_class\_i}{total\_number\_of\_pixels\_class\_i}$$

- Overlapping (jaccard index)

$$performance = \frac{1}{C}\sum_{i=1}^{C}\frac{TP}{TP+FP+FN}$$

(Intersection over the union)

image registration and multi-view stereo matching

Annotation: label at pixel level
Don't care regions: boundaries

There is no clear standard

- *Gary E. Christensen, Xiujuan Geng, Jon G. Kuhl, Joel Bruss, Thomas J. Grabowski, Imran A. Pirwani, Michael W. Vannier, John S. Allen, Hanna Damasio, Introduction to the Non-rigid Image Registration Evaluation Project (NIREP), Biomedical Image Registration, Lecture Notes in Computer Science Volume 4057, 2006, pp 128-135*

- Some used measurements:

  - Overlapping (jaccard index)

$$performance = \frac{1}{C}\sum_{i=1}^{C}\frac{TP\_i}{TP\_i + FP\_i + FN\_i}$$

(Intersection over the union)

  - Correlation based on intensity values

human pose estimation

Annotation: bounding boxes for limbs

- **"PCP"** metric: considers a body part to be localized correctly if the estimated body segment endpoints are within 50% of the ground-truth segment length from their true locations.
- **"PCPm"** metric: uses 50% of the mean ground-truth segment length over the entire test set as a matching threshold for applying "PCP".

Annotation: joint coordinates

- **"PCK"** metric: measures accuracy of the localization of the body joints. The threshold for matching of the joint position to the ground-truth is defined as a fraction of the person bounding box size.
- **"PCKh"** metric: as "PCK" but define the matching threshold as 50% of the head segment length.

http://www.robots.ox.ac.uk/~vgg/publications/papers/ferrari08.pdf

http://ps.is.tuebingen.mpg.de/publications/168/get_file

Annotations and metrics

object recognition

Annotation: bounding box

- **Detection vs classification**

- Detection

$$performance = \frac{1}{C}\sum_{i=1}^{C}\frac{TP\_i}{TP\_i + FP\_i + FN\_i}$$

Where a hit for TP_i should satisfy $\frac{\cap}{\cup} > \Theta$

- Multi-class classification

$$performance = \frac{number\_of\_corrected\_classified\_samples}{total\_number\_of\_samples}$$

~ sum(Diag (confusion matrix)) / (sum(sum(confusion matrix))

Face Analysis

- Face detection: detection
  Annotation: bounding box
- Identification: multi-class classification
  Annotation: label per face/image
- Gender Recognition: binary classification
  Annotation: label per face/image
- Verification: binary classification
  Annotation: label per face/image
- Age estimation:
  - Regression deviation

    Annotation: label/labels/levels per face/image
- Facial expression analysis:
  - multi-class classification
  - Regression in some cases
  - Multi-label

Multi-label Evaluation Metrics: Label-Based

Basic Strategy:

Calculate classic single-label metric on each label *independently*, and then combine metric values over all labels.

Label-based multi-label metrics are easy to compute, but ignore the relationships between different labels!

Exact Match: is the most strict metric, indicating the percentage of samples that have all their labels classified correctly.

Multi-label Evaluation Metrics: Instance-Based

Basic Strategy:

Calculate metric value for each instance by addressing relationships among different class labels (especially the *ranking quality*), and then return the mean value over all instances.

Popular instance-based multi-label metrics [Schapire & Singer, MLJ00]:

Given the learned predictor *h*(·) or *f*(·,·), and a test set $T = \{(\boldsymbol{x}_i, Y_i | 1 \leq i \leq t)\}$

(1) *Hamming loss*

$$\text{hloss}_T(h) = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{q}|h(\boldsymbol{x}_i \Delta Y_i)|$$

Evaluates how many times an instance-label pair is misclassified (is the percentage of the wrong labels to the total number of labels. As a loss metric, 0 is better), being *q* a normalization value (e.g. the number of labels, cardinality of *Y*).

(3) *Coverage*

$$\text{coverage}_T(f) = \frac{1}{t}\sum_{i=1}^{t}\max_{y \in Y_i} rank_f(\boldsymbol{x}_i, y) - 1$$

Evaluates how many steps are needed, on average, to go down the label list to cover all proper labels of the instance.

[$rank_f(\boldsymbol{x}_i, y)$ returns the rank of *y* derived from $f(\boldsymbol{x}_i, y)$]

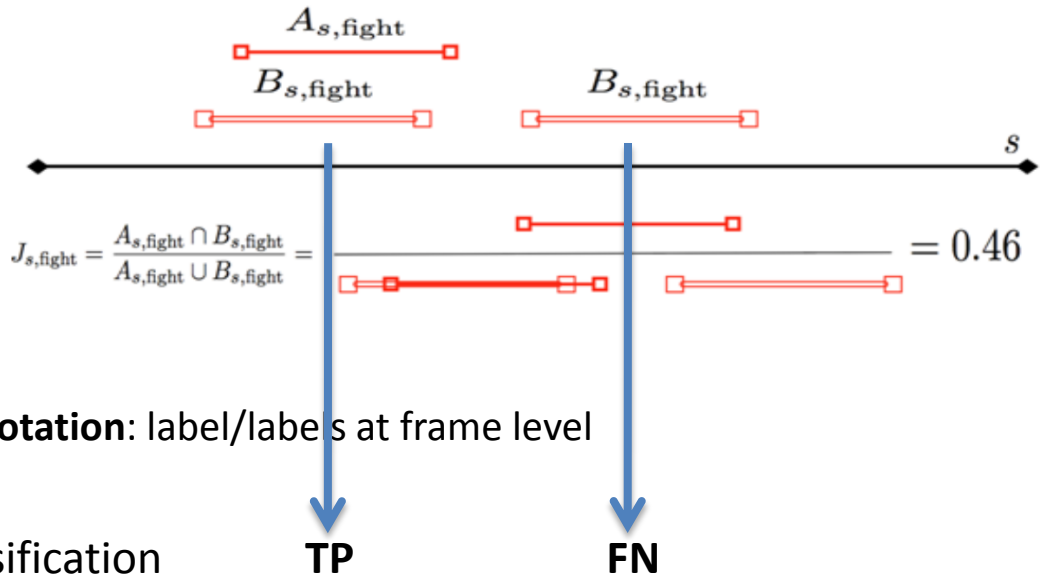
Useful metric for image retrieval (related also to taxonomy analysis)

gesture recognition (classification vs spotting)

- Multi-class classification **Annotation**: label per segmented sequence

- Spotting (can be multi-label depending on behavior taxonomy)

$$J_{s,n} = \frac{A_{s,n} \bigcap B_{s,n}}{A_{s,n} \bigcup B_{s,n}},$$

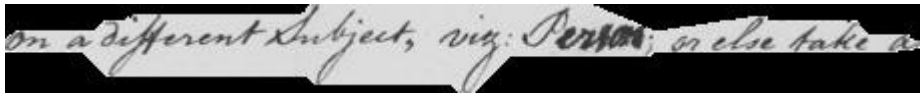


Annotation: label/labels at frame level

Or even estandar classification

text recognition

Annotation: words locations and ground truth words



- WER – Word Error Rate (based on Levensthein distance)

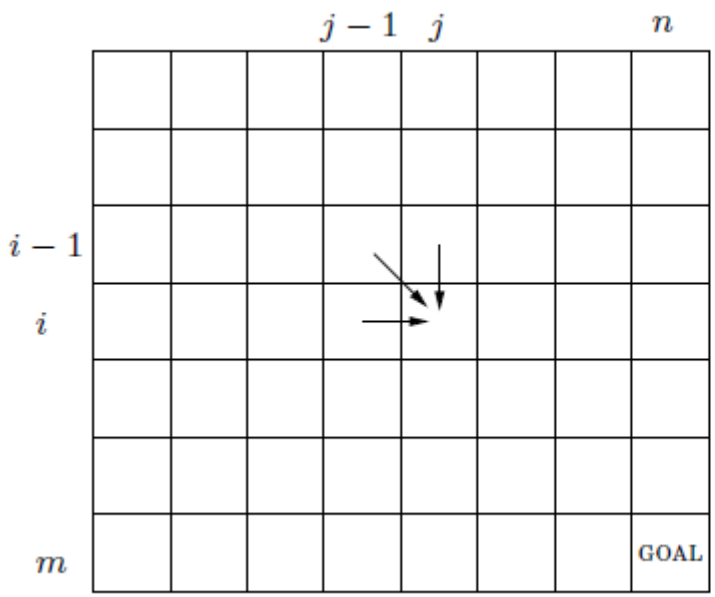
$$WER = \frac{S + B + I}{N}$$

- *S* number of substitutions
- *B* number of deletions
- *I* number of insertions
- *N* number of sentence words

$$WER(i,j) = \min \begin{cases} WER(i-1,j) + 1 \\ WER(i,j-1) + 1 \\ WER(i-1,j-1) + \Delta(i,j) \end{cases}$$

text recognition

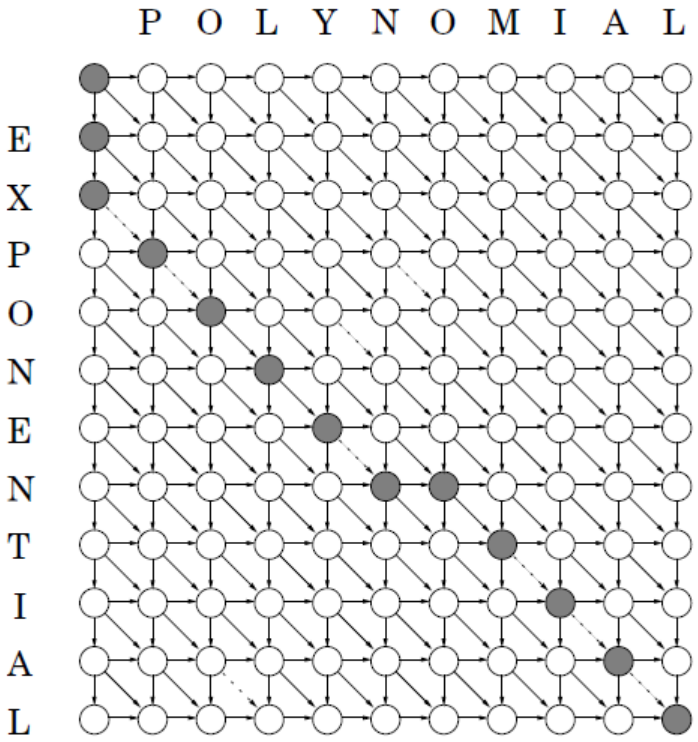
$$E(i,j) = \min\{E(i-1,j)+1, E(i,j-1)+1, E(i-1,j-1)+\texttt{diff}(i,j)\}$$



		P	O	L	Y	N	O	M	I	A	L
	0	1	2	3	4	5	6	7	8	9	10
E	1	1	2	3	4	5	6	7	8	9	10
X	2	2	2	3	4	5	6	7	8	9	10
P	3	2	3	3	4	5	6	7	8	9	10
O	4	3	2	3	4	5	5	6	7	8	9
N	5	4	3	3	4	4	5	6	7	8	9
E	6	5	4	4	4	5	5	6	7	8	9
N	7	6	5	5	5	4	5	6	7	8	9
T	8	7	6	6	6	5	5	6	7	8	9
I	9	8	7	7	7	6	6	6	6	7	8
A	10	9	8	8	8	7	7	7	7	6	7
L	11	10	9	8	9	8	8	8	8	7	6

text recognition

$$E(i,j) = \min\{E(i-1,j)+1, E(i,j-1)+1, E(i-1,j-1)+\texttt{diff}(i,j)\}$$




```
E X P O N E N - T I A L
- - P O L Y N O M I A L
```

Tracking

Annotation: trajectory coordinates

- **Trajectories**: accumulated distance for detected localitions

- **Bounding boxes**:
  - Overlap
  - Hit per tracked bounding box based on overlapping threshold

Usage of training – validation – test sets

- Partitions:
  - Training used for learning methods
  - Validation useful for tuning parameters, support generalization, and avoid or delay the appearance of overfitting
  - Test only used for final generalization performance
  - Use of many splits of the data:
    - N-fold cross-validation
    - Random vs stratified
  - Confidence interval is useful to analyze the stability of the results
  - Statistical significance analysis

Janez Demsar, Statistical Comparisons of Classifiers over Multiple Data Sets, Journal of Machine Learning Research, Volume 7, 12/1/2006, Pages 1-30

Which aspects of the data set are interesting

Looking for generalization:

-Ground truth quality

- don't care regions
- Inter-labeller varability
- etc.

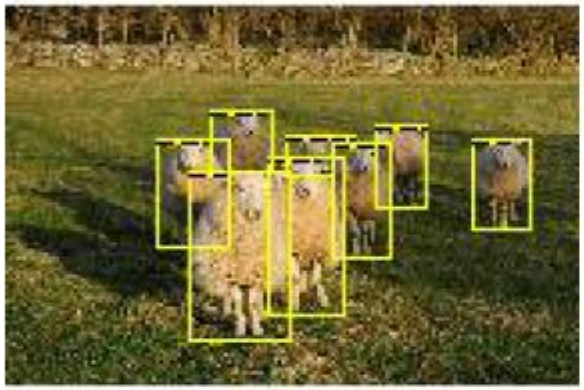
-Training set variability

-Amount of data

INDEX

The 2005-2012 Visual Object Challenges

A. Zisserman, C. Williams, M. Everingham, L. v.d. Gool



Classification: is there an X in this image?,

Detection: where are the X's ?

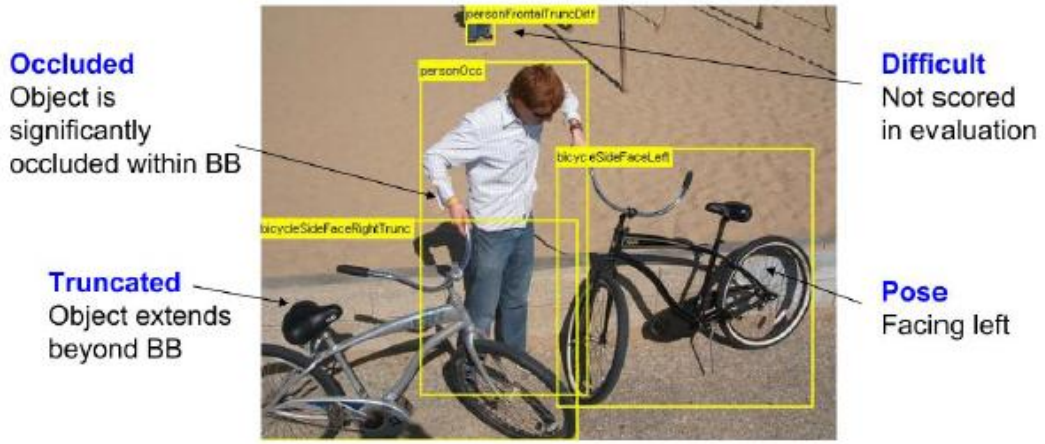
Segmentation: which pixels belong to X ?

The challenge organizer perspective

Selecting the data

real images from flickr (no selection / cleaning)

Annotation, how



Annotation, who

- Annotation parties
- The Amazon turk

Experimental setting

- At least 500 images per object
- Equally divided among training/validation and test
- Increased along years (enables to measure progress)

Software supplied

- Includes baseline classifier/detector/segmenter
- Generates precision-recall curve and computes accuracy scores
- On train/validation/test and other datasets

Means that results on VOC can be consistently compared in publications

Evaluation

Option 1

Release test data and annotation (most liberal) and participants can assess performance

Cons: open to abuse

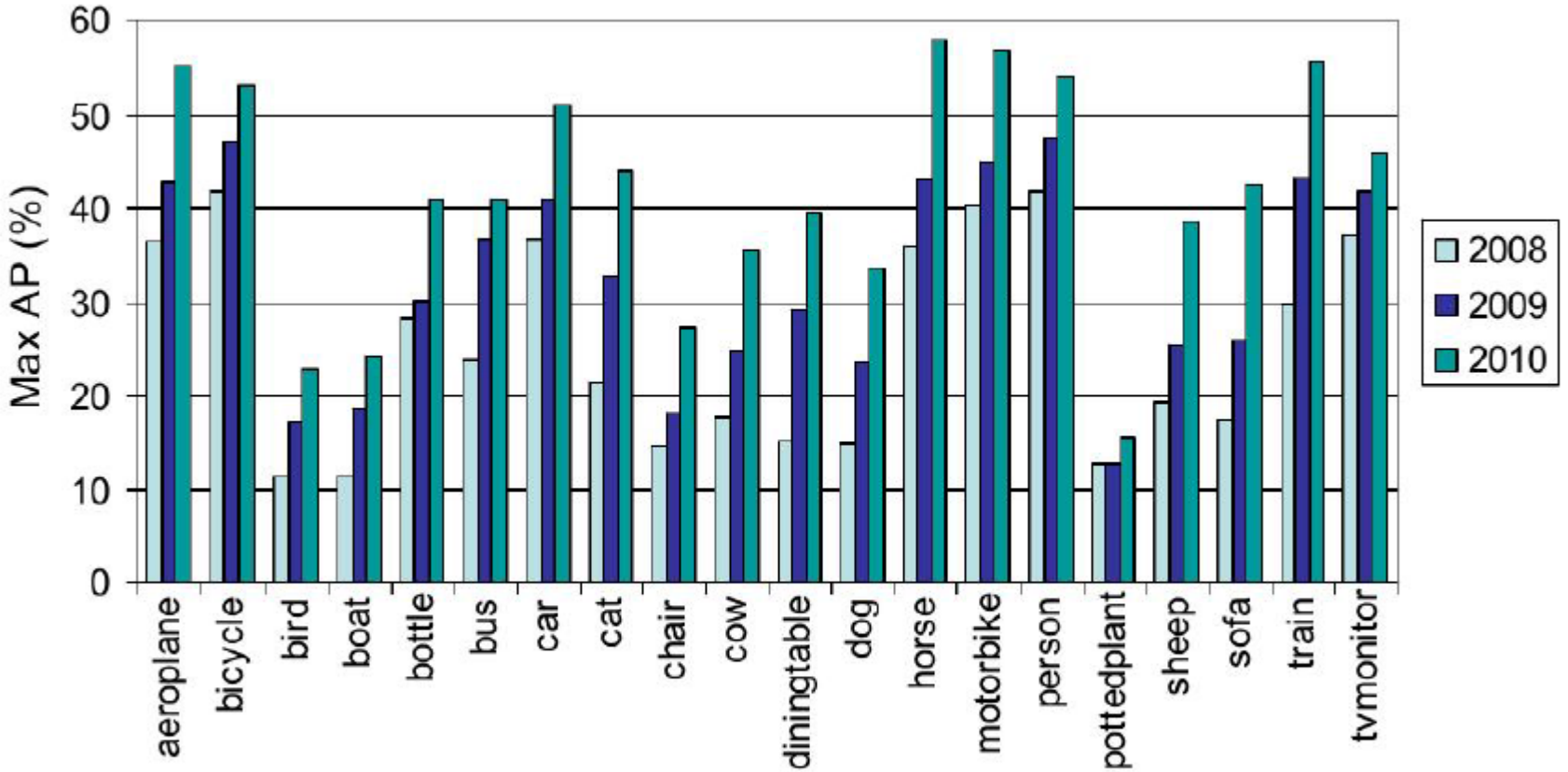
Option 2

Release test data, but test annotation withheld - participants submit results and organizers assess performance (use an evaluation server)

Option 3

No release of test data - participants have to submit software and organizers run this and assess performance

Detection Challenge: Progress 2008-2010



- Results on 2008 data improve for best 2009 and 2010 methods for all categories, by over 100% for some categories
  - Caveat: Better methods or more training data?

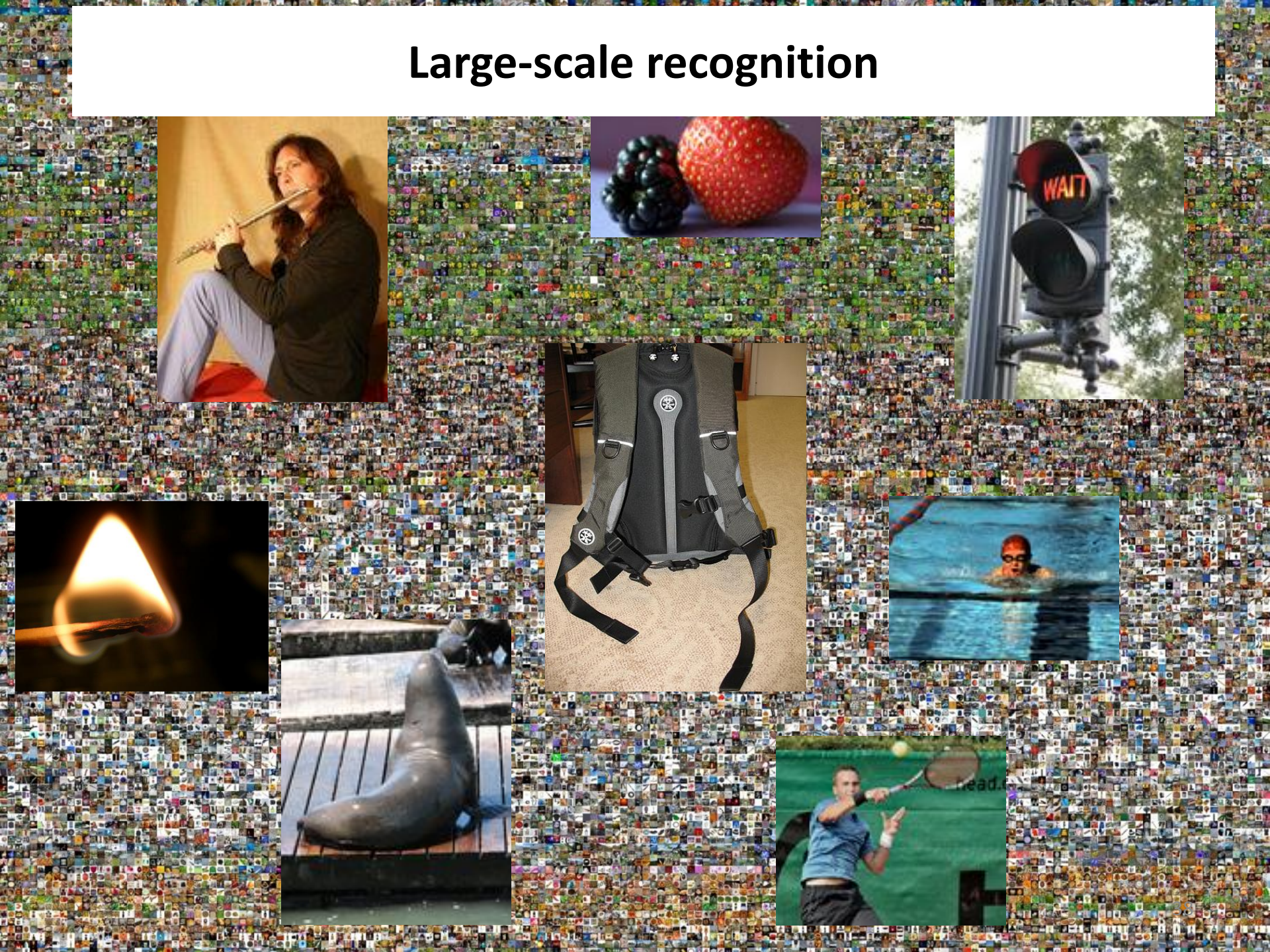
INDEX

Large Scale Visual Recognition Challenge (ILSVRC)

http://image-net.org/challenges/LSVRC/

Thx to: Olga Russakovsky, Stanford University

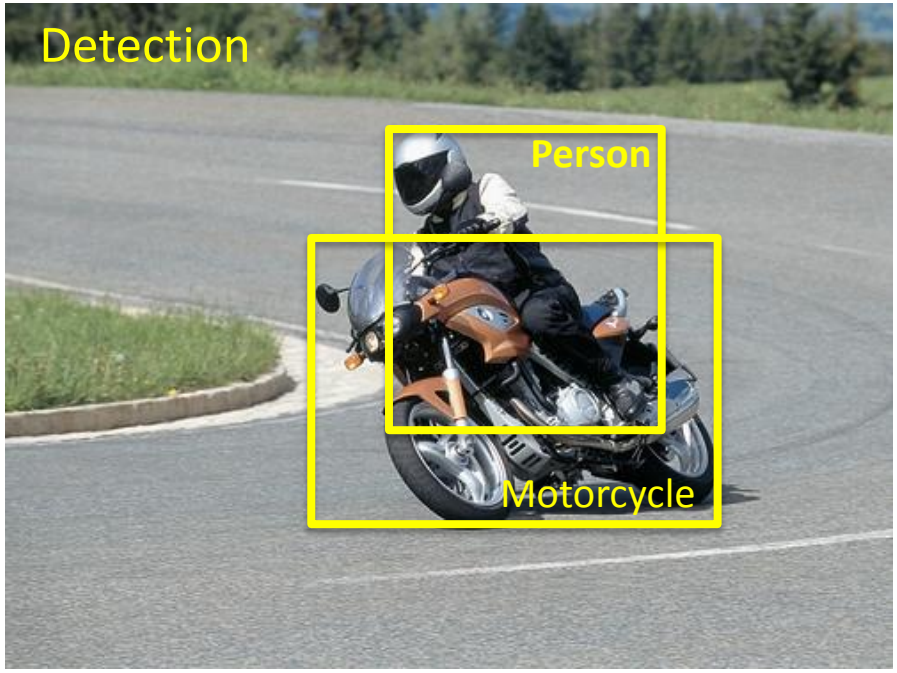
Large-scale recognition

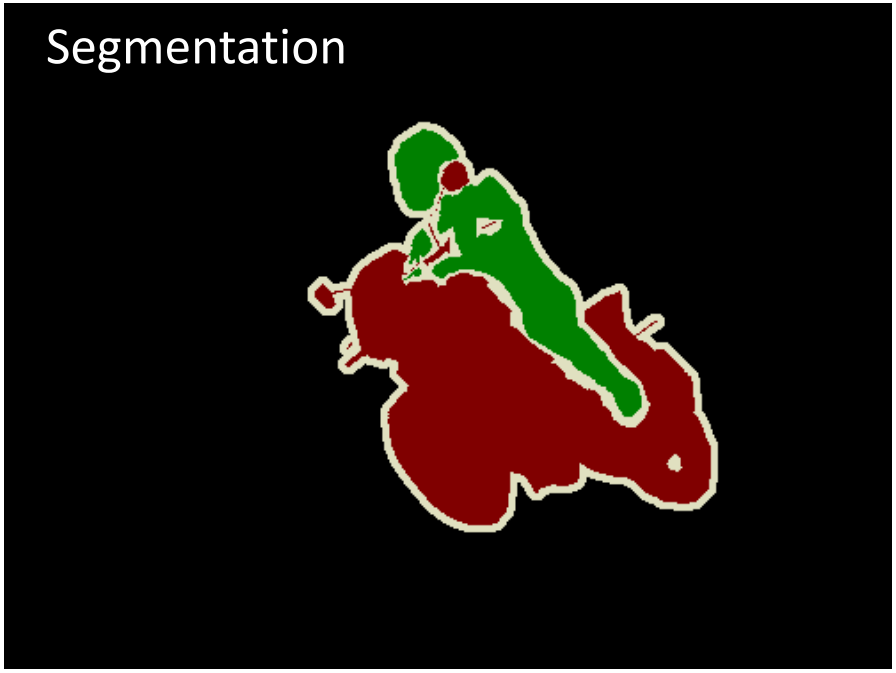


PASCAL VOC 2005-2012

20 object classes **22,591 images**

Classification: person, motorcycle





Action: riding bicycle

Everingham, Van Gool, Williams, Winn and Zisserman.
The PASCAL Visual Object Classes (VOC) Challenge. IJCV 2010.

IMAGENET ILSVRC 2010-2014

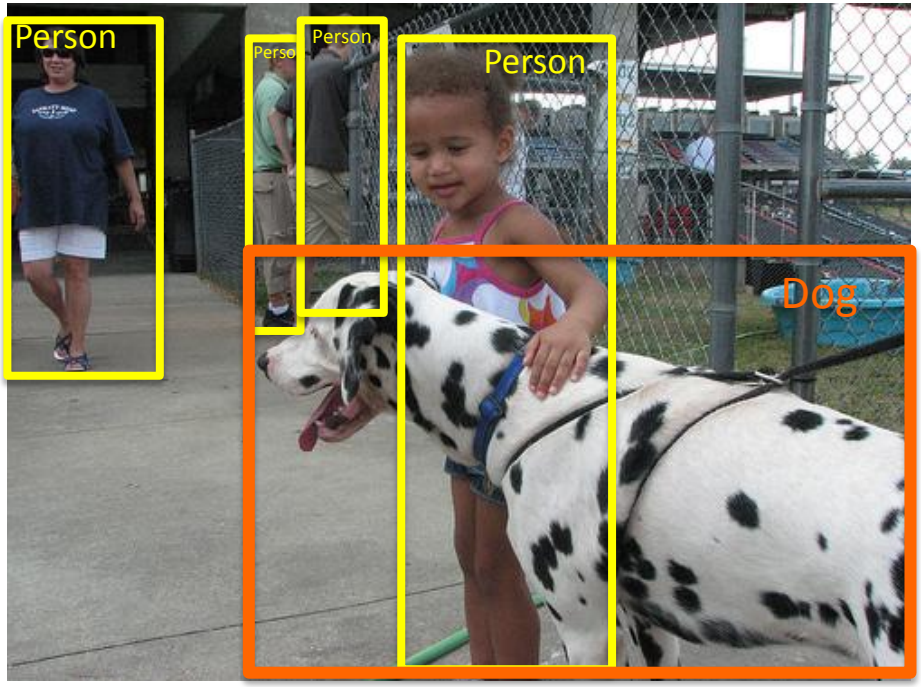
~~20 object classes~~ ~~22,591 images~~

200 object classes **517,840 images**

DETECTION

1000 object classes **1,431,167 images**

CLASSIF.



http://image-net.org/challenges/LSVRC/

ILSVRC types of image annotations

Image classification

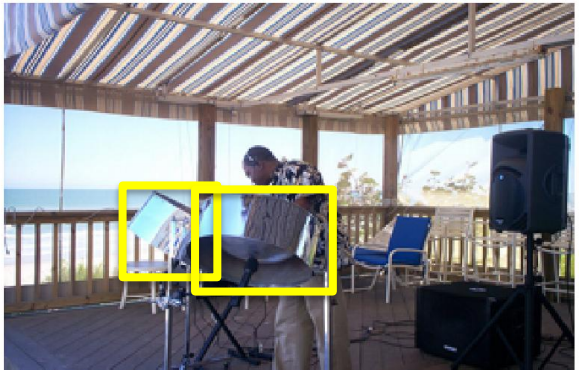
- one object class per image
- no bounding boxes

Steel drum

Single-object localization

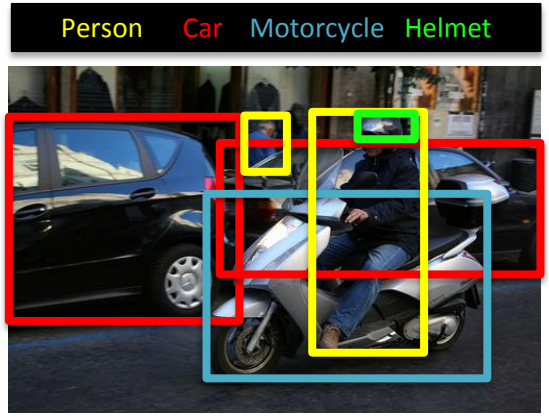
- one object class per image
- bounding boxes around all instances of this class

Steel drum



Object detection

- all target object classes
- bounding boxes around all instances



Statistics of ILSVRC2014 released annotated images:

1000 object classes	1000 object classes	200 object classes
1,331,167 images	573,966 images	81,799 images
	657,231 bounding boxes	228,981 bounding boxes

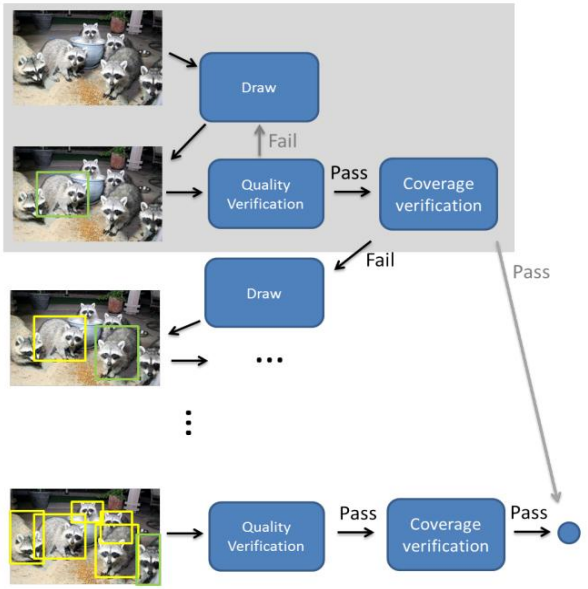
ILSVRC large-scale annotation

amazon mechanical turk™
Artificial Artificial Intelligence

IMAGENET

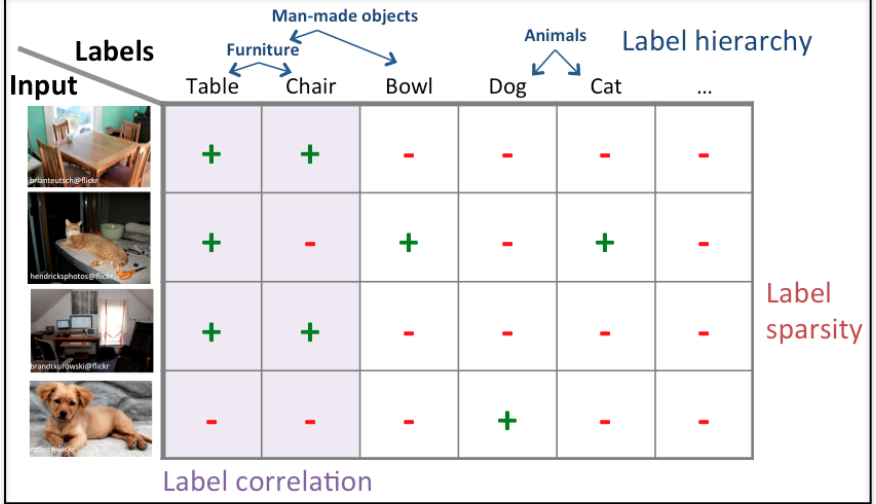
J. Deng, W. Dong, R. Socher, L.-J. Li, L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. CVPR, 2009

Multi-step pipeline for bounding box annotation



H. Su, J. Deng, L. Fei-Fei. Crowdsourcing Annotation for Visual Object detection. AAAI Hcomp workshop 2012.

Hierarchical multi-label annotation



J. Deng, O. Russakovsky, J. Krause, M. Bernstein, A. Berg, L. Fei-Fei. Scalable multi-label annotation. CHI, 2014

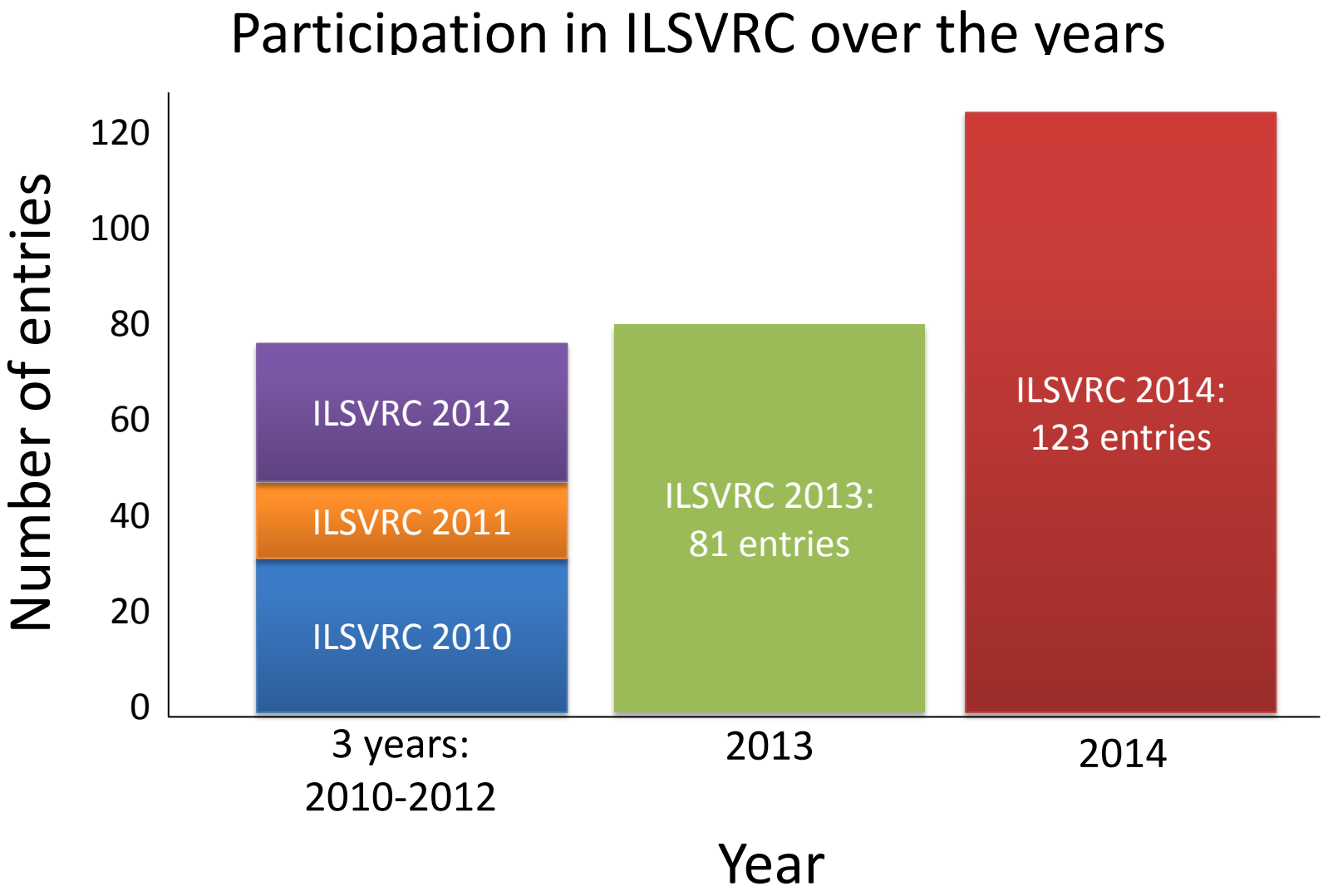
O. Russakovsky*, J. Deng*, et al. ImageNet Large Scale Visual Recognition Challenge. http://arxiv.org/abs/1409.0575

Challenge procedure every year

1. Training data released: images and annotations
2. Test data released: images only (annotations hidden)
3. Participants train their models on train data
4. Submit text file with predictions on test images
5. Evaluate and release results, and run a workshop

http://image-net.org/challenges/LSVRC/2014/eccv2014

IMAGENET


Participation in ILSVRC over the years
Number of entries
120
100
80
60
40
20
0
ILSVRC 2012
ILSVRC 2011
ILSVRC 2010
ILSVRC 2013:
81 entries
ILSVRC 2014:
123 entries
3 years:
2010-2012
2013
2014
Year

IMAGENET

Fast computational resource: GPUs

Powerful algorithm: Convolutional neural network

Diverse large-scale annotated data: ILSVRC

A. Krizhevsky et al. ImageNet Classification with Deep Convolutional Neural Networks. NIPS 2012

ILSVRC image classification task

Steel drum

Output:
Scale
T-shirt
Steel drum
Drumstick
Mud turtle

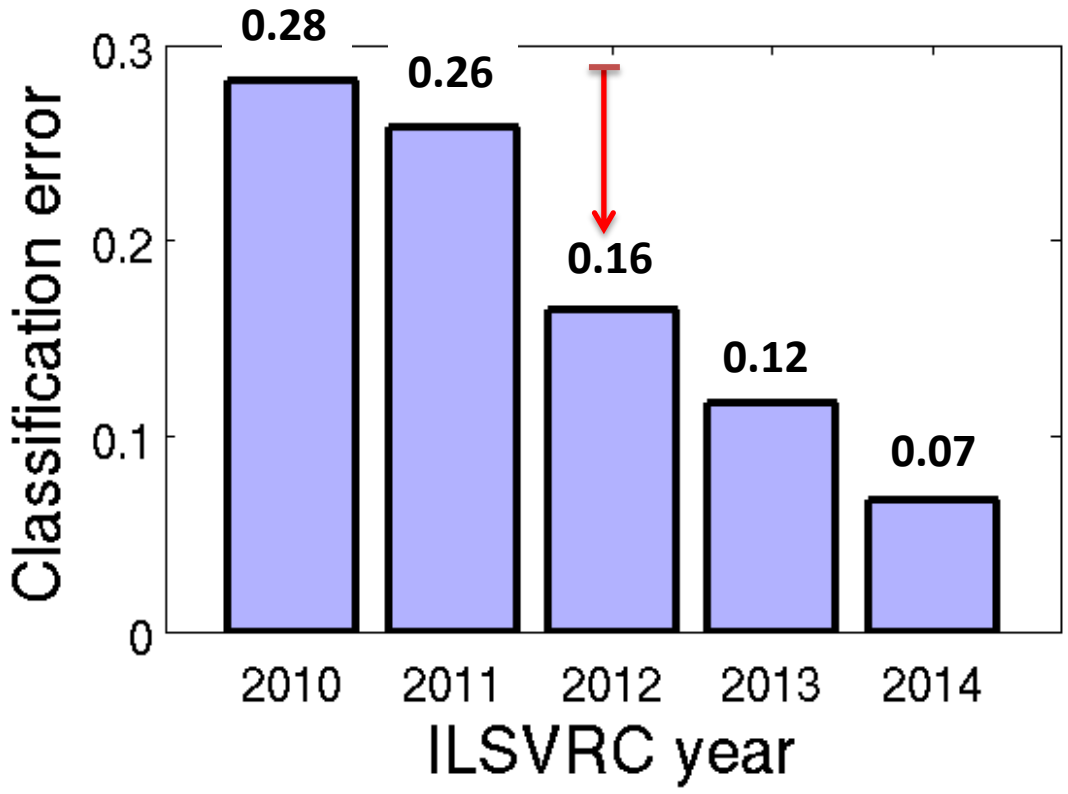
✔

Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

✗

$$\text{Error} = \frac{1}{100{,}000} \sum_{\substack{\text{100,000} \\ \text{images}}} 1[\text{incorrect on image i}]$$

Deep learning impact on ILSVRC classification accuracy



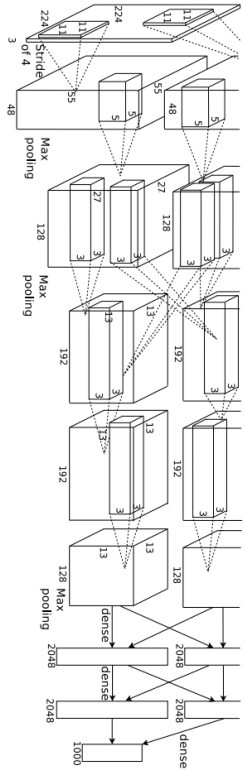
Massive drop in error with a deep learning method

Russakovsky* and Deng* et al., ImageNet Large Scale Visual Recognition Challenge, http://arxiv.org/abs/1409.0575

Deep learning is here to stay

Year 2012

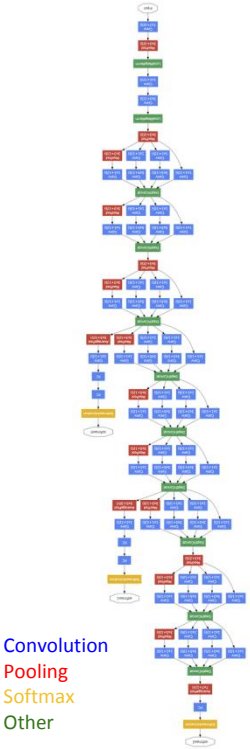
SuperVision



[Krizhevsky NIPS 2012]

Year 2014

GoogLeNet

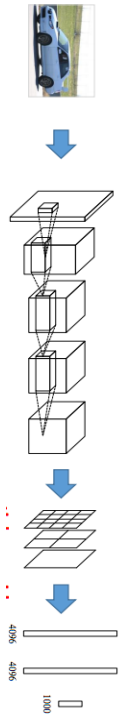


[Szegedy arxiv 2014]

VGG

[Simonyan arxiv 2014]

MSRA



[He arxiv 2014]

35/36 teams used deep learning

20/36 teams used open-source Caffe implementation

(Highest *accuracy* in percent of any method in ILSVRC 2012-2014)

Easiest and hardest categories for image classification

Easiest

red fox (100) hen-of-the-woods (100) ibex (100) goldfinch (100) flat-coated retriever (100)





tiger (100) hamster (100) porcupine (100) stingray (100) Blenheim spaniel (100)





… and 111 more categories with 100% accuracy!

Hardest

muzzle (71) hatchet (68) water bottle (68) velvet (68) loupe (66)





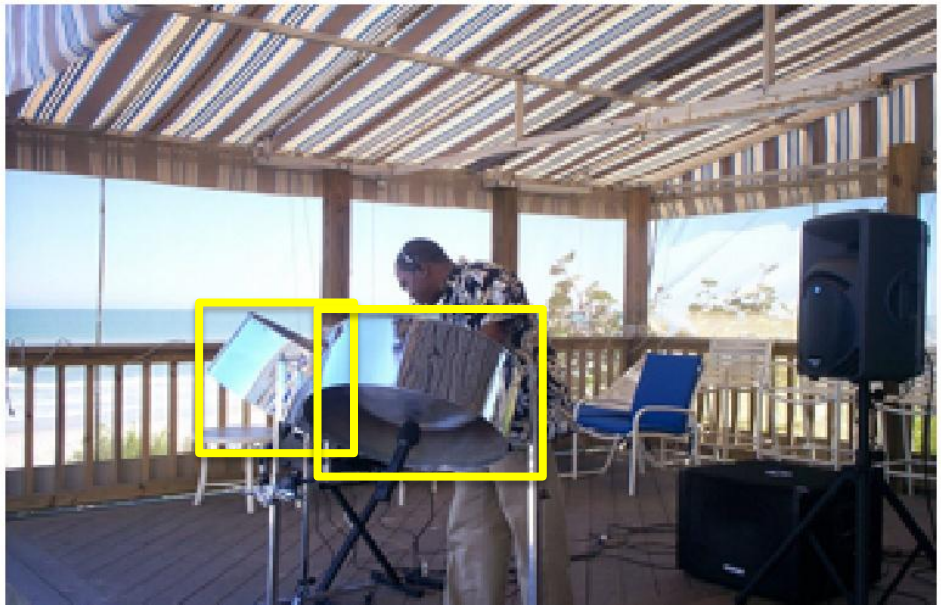
hook (66) spotlight (66) ladle (65) restaurant (64) letter opener (59)





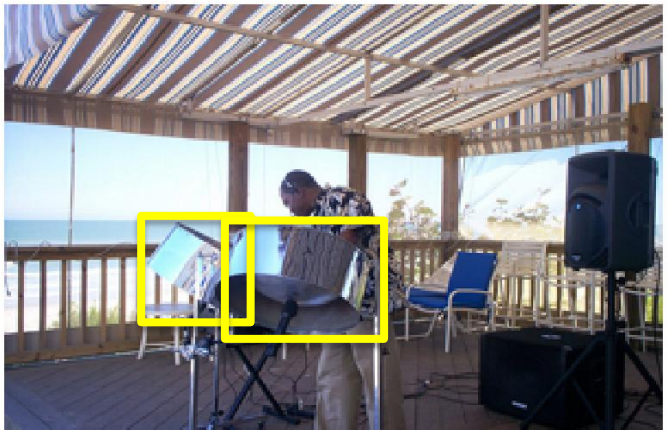
ILSVRC single-object localization task

Steel drum

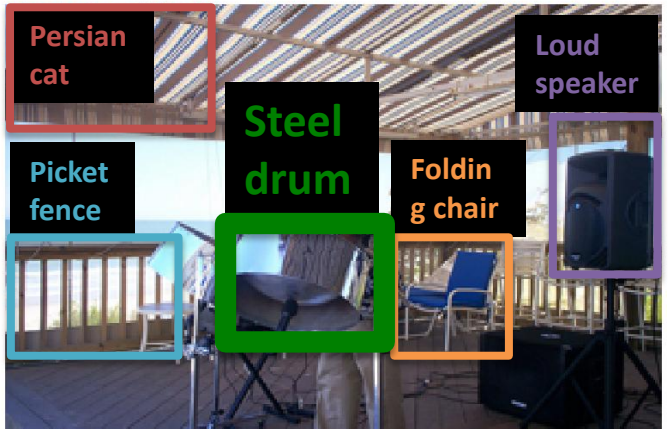


ILSVRC single-object localization task

Steel drum

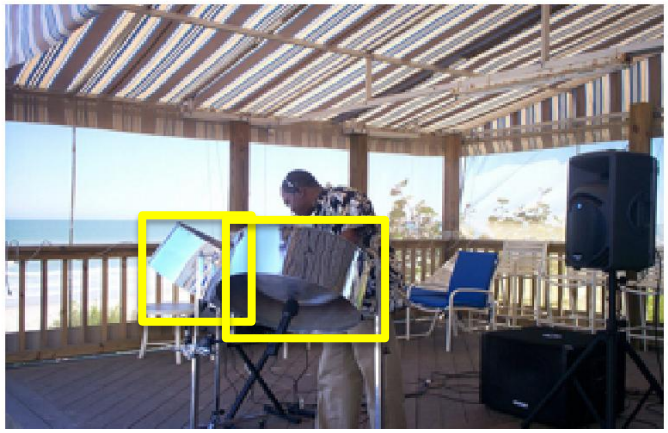


Output

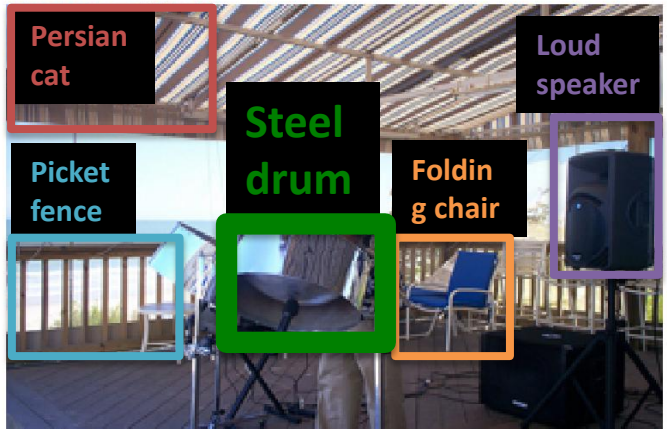


ILSVRC single-object localization task

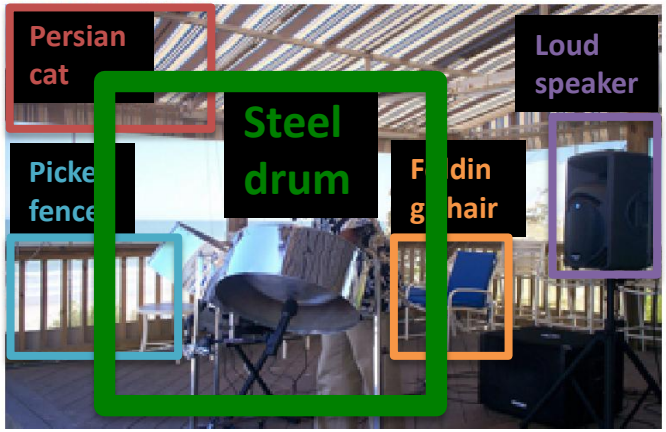
Steel drum



Output



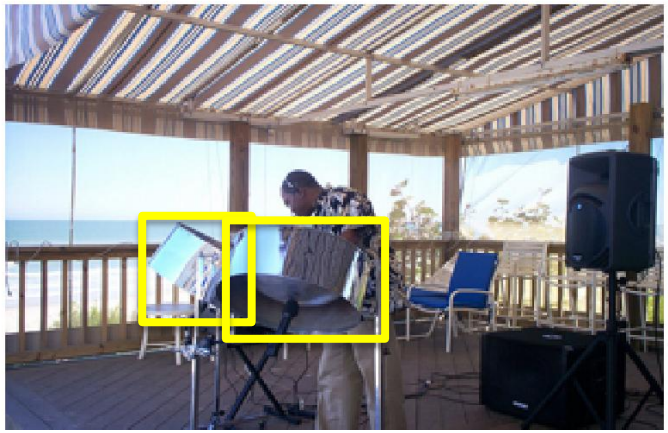
Output (bad localization)



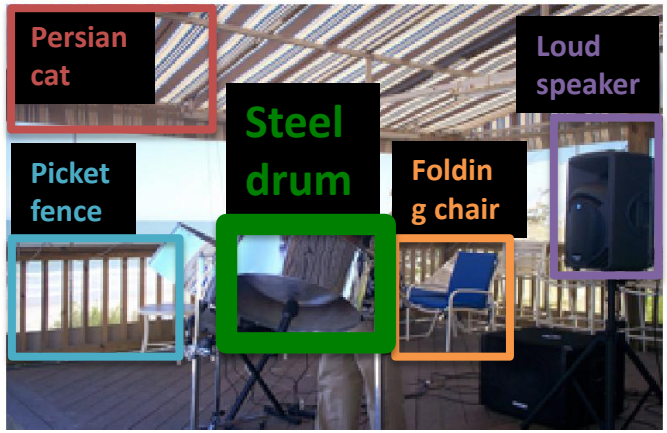
Output (bad classification)

ILSVRC single-object localization task

Steel drum

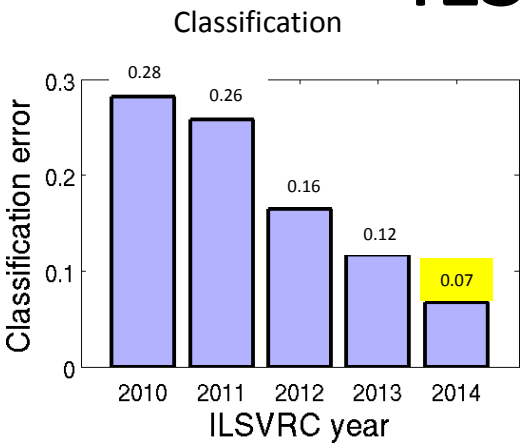


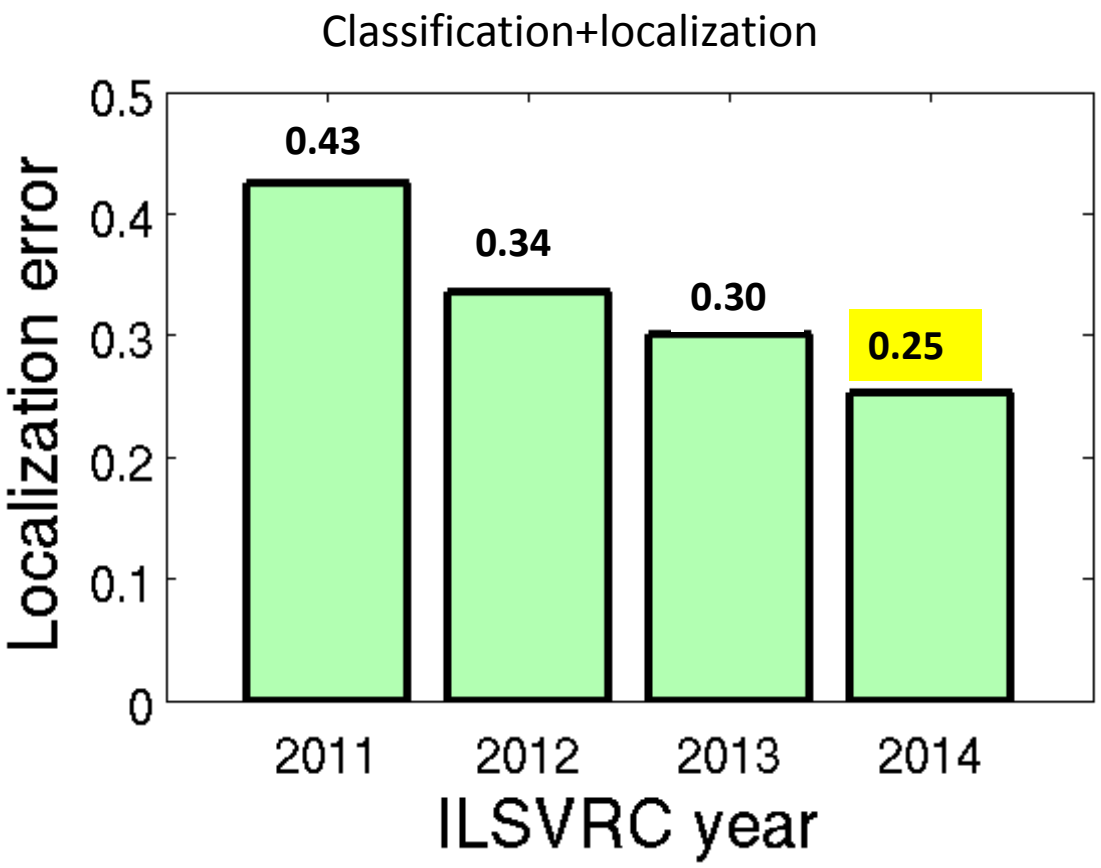
Output



$$\text{Error} = \frac{1}{100{,}000} \sum_{100{,}000 \text{ images}} 1[\text{incorrect on image i}]$$

ILSVRC over the years





Russakovsky* and Deng* et al., ImageNet Large Scale Visual Recognition Challenge, http://arxiv.org/abs/1409.0575

(Highest *accuracy* in percent of any method in ILSVRC 2012-2014)

Easiest and hardest categories for single-object localization

Easiest

Leonberg (100) ruffed grouse (100) ruddy turnstone (100) giant schnauzer (99) tiger (99)





Maltese dog (99) Japanese spaniel (99) Tibetan mastiff (99) hare (99) African hunting dog (99)





Hardest

horizontal bar (41) flagpole (38) hook (37) lakeside (36) letter opener (36)





spotlight (35) wing (35) ladle (28) pole (27) space bar (23)

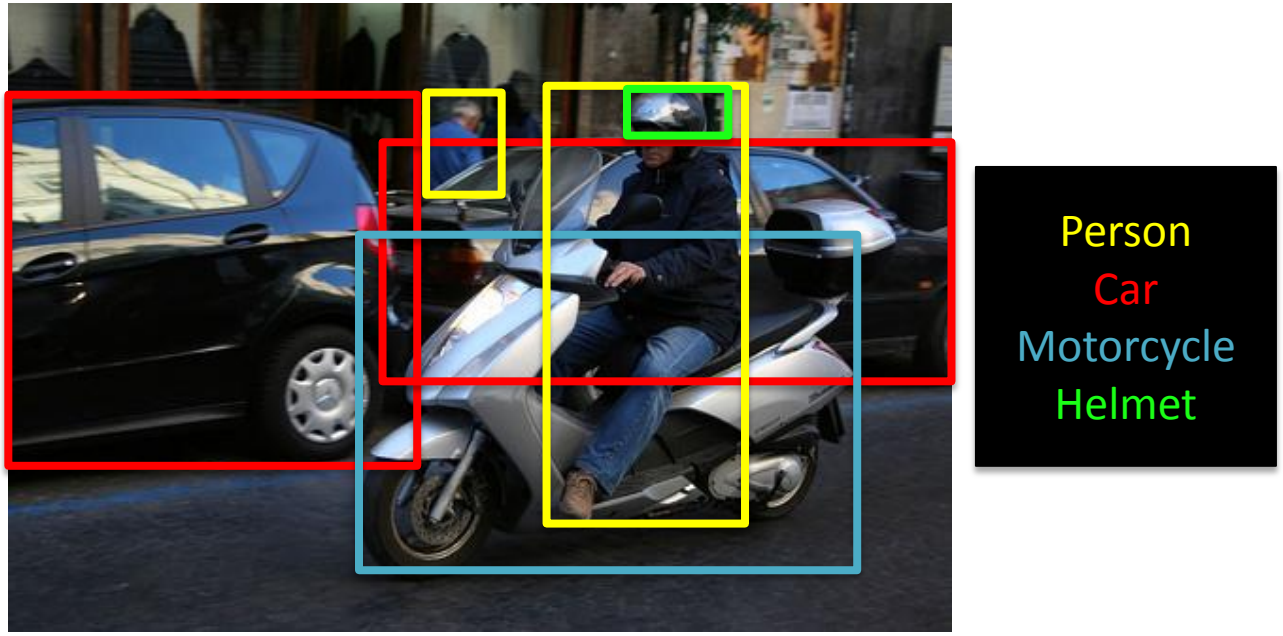




Russakovsky* and Deng* et al., ImageNet Large Scale Visual Recognition Challenge, http://arxiv.org/abs/1409.0575

ILSVRC object detection task

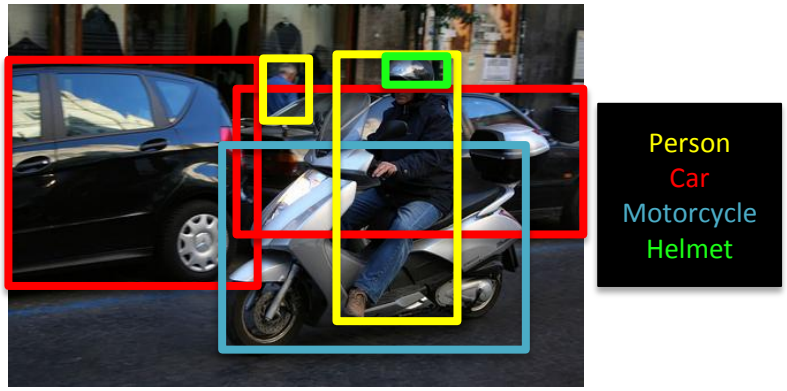
Fully annotated 200 object classes across 120,000 images



Allows evaluation of generic object detection in cluttered scenes at scale

ILSVRC object detection task

All instances of all target object classes expected to be localized on all test images

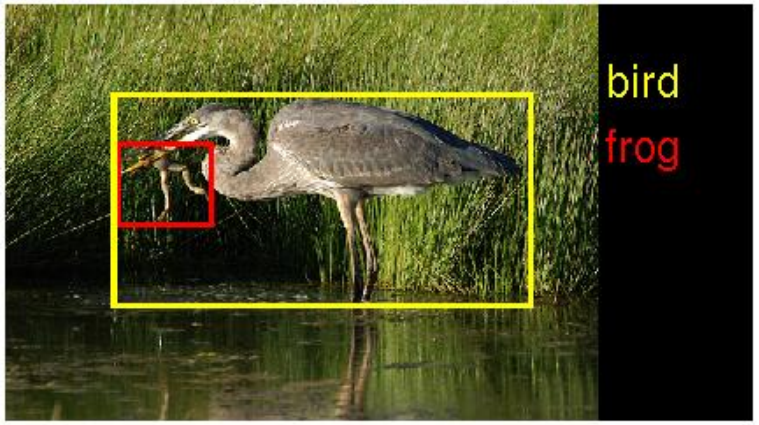


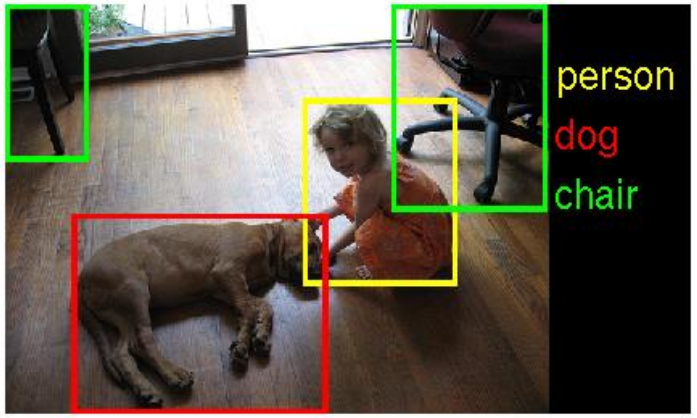
Evaluation modeled after PASCAL VOC:

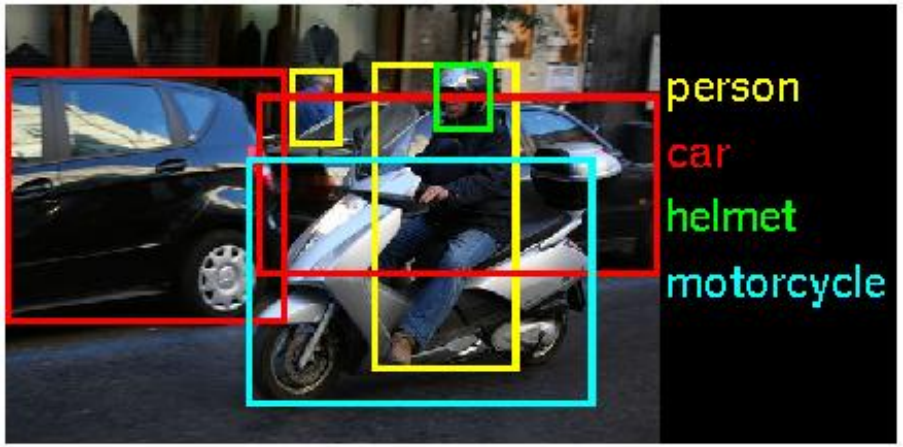
- Algorithm outputs a list of bounding box detections with confidences
- A detection is considered correct if overlap with ground truth is big enough
- Evaluated by average precision per object class
- Winners of challenge is the team that wins the most object categories

Everingham, Van Gool, Williams, Winn and Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. IJCV 2010.

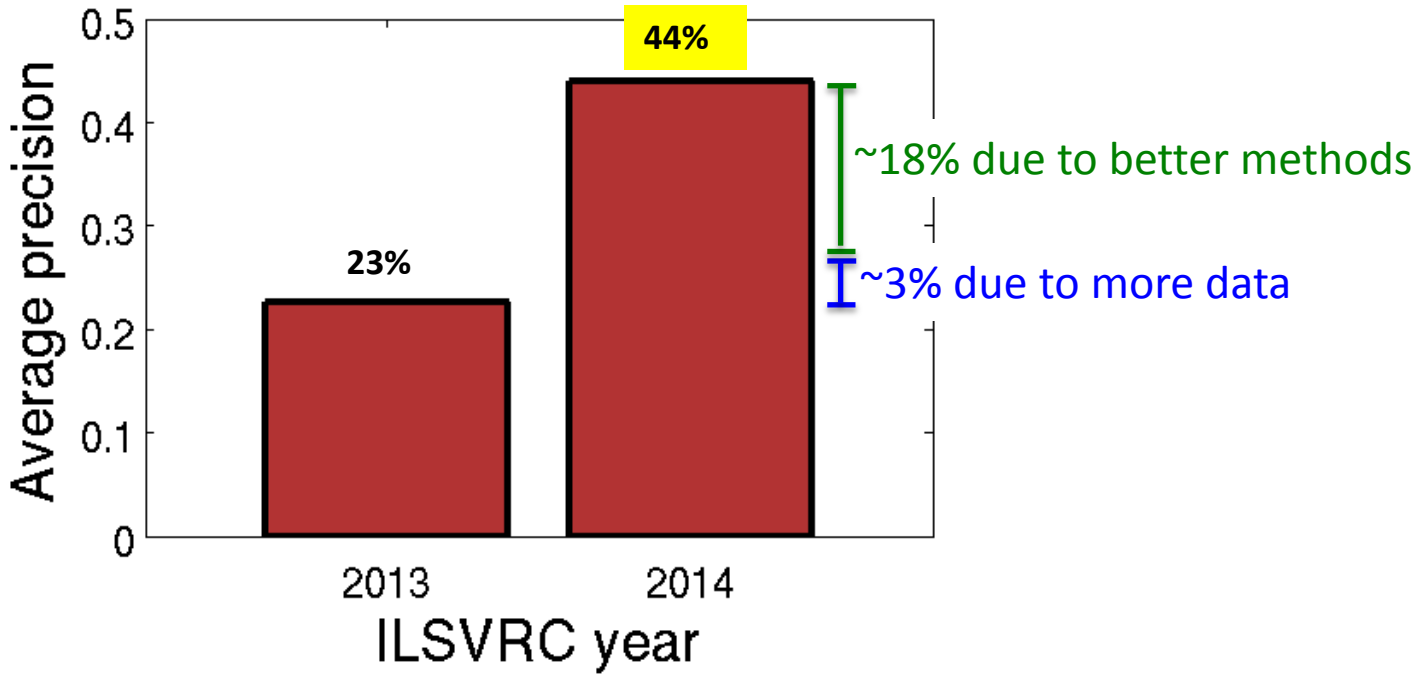
ILSVRC object detection data







ILSVRC detection since 2013



1.9x increase in object detection average precision in one year

Russakovsky* and Deng* et al., ImageNet Large Scale Visual Recognition Challenge, http://arxiv.org/abs/1409.0575

IMAGENET

(Highest average precision in percent of any method in ILSVRC 2012-2014)

Easiest and hardest categories for object detection

Easiest

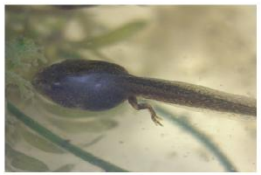
butterfly (93)

dog (84)

volleyball (83)

rabbit (83)

frog (82)



basketball (80)

snowplow (80)

bird (78)

tiger (77)

zebra (77)

Hardest

lamp (15)

flute (15)

horizontal bar (14)

spatula (13)

nail (13)

ski (12)

microphone (11)

rubber eraser (10)

ladle (9)

backpack (8)

Russakovsky* and Deng* et al., ImageNet Large Scale Visual Recognition Challenge, http://arxiv.org/abs/1409.0575

Q: So how well do current methods work on large-scale object recognition?

- Well, they work much better than last year!
- Work very well on
  - Classifying and detecting animals
  - Recognizing objects with distinctive patterns
- Don’t work as well on
  - Thin objects
  - Untextured objects

Russakovsky* and Deng* et al., ImageNet Large Scale Visual Recognition Challenge, http://arxiv.org/abs/1409.0575

What is human accuracy on ILSVRC2014 classification?

Human vs computer accuracy on ILSVRC2014 classification

- Compared expert human annotators with winning GoogLeNet entry

	Annotator 1
Total number of images	1500
GoogLeNet classification error	6.8%
Human classification error	?

Russakovsky* and Deng* et al., ImageNet Large Scale Visual Recognition Challenge, http://arxiv.org/abs/1409.0575

Human vs computer accuracy on ILSVRC2014 classification

- Compared expert human annotators with winning GoogLeNet entry

	Annotator 1
Total number of images	1500
GoogLeNet classification error	6.8%
Human classification error	5.1%

Russakovsky* and Deng* et al., ImageNet Large Scale Visual Recognition Challenge, http://arxiv.org/abs/1409.0575

Human vs computer accuracy on ILSVRC2014 classification

- Compared expert human annotators with winning GoogLeNet entry

	Annotator 1
Total number of images	1500
GoogLeNet classification error	6.8%
Human classification error	5.1%

- Annotator 1 achieved better accuracy than GoogLeNet by 1.7%
- Task required *significant* amount of training for humans

Russakovsky* and Deng* et al., ImageNet Large Scale Visual Recognition Challenge, http://arxiv.org/abs/1409.0575

Human vs computer accuracy on ILSVRC2014 classification

- Compared expert human annotators with winning GoogLeNet entry

	Annotator 1	Annotator 2
Total number of images	1500	258
GoogLeNet classification error	6.8%	5.8%
Human classification error	5.1%	12.0%

- Annotator 1 achieved better accuracy than GoogLeNet by 1.7%
- Task required *significant* amount of training for humans

Russakovsky* and Deng* et al., ImageNet Large Scale Visual Recognition Challenge, http://arxiv.org/abs/1409.0575

Q: Are current methods close to human-level classification accuracy?

Current methods are not as good as humans yet,

but

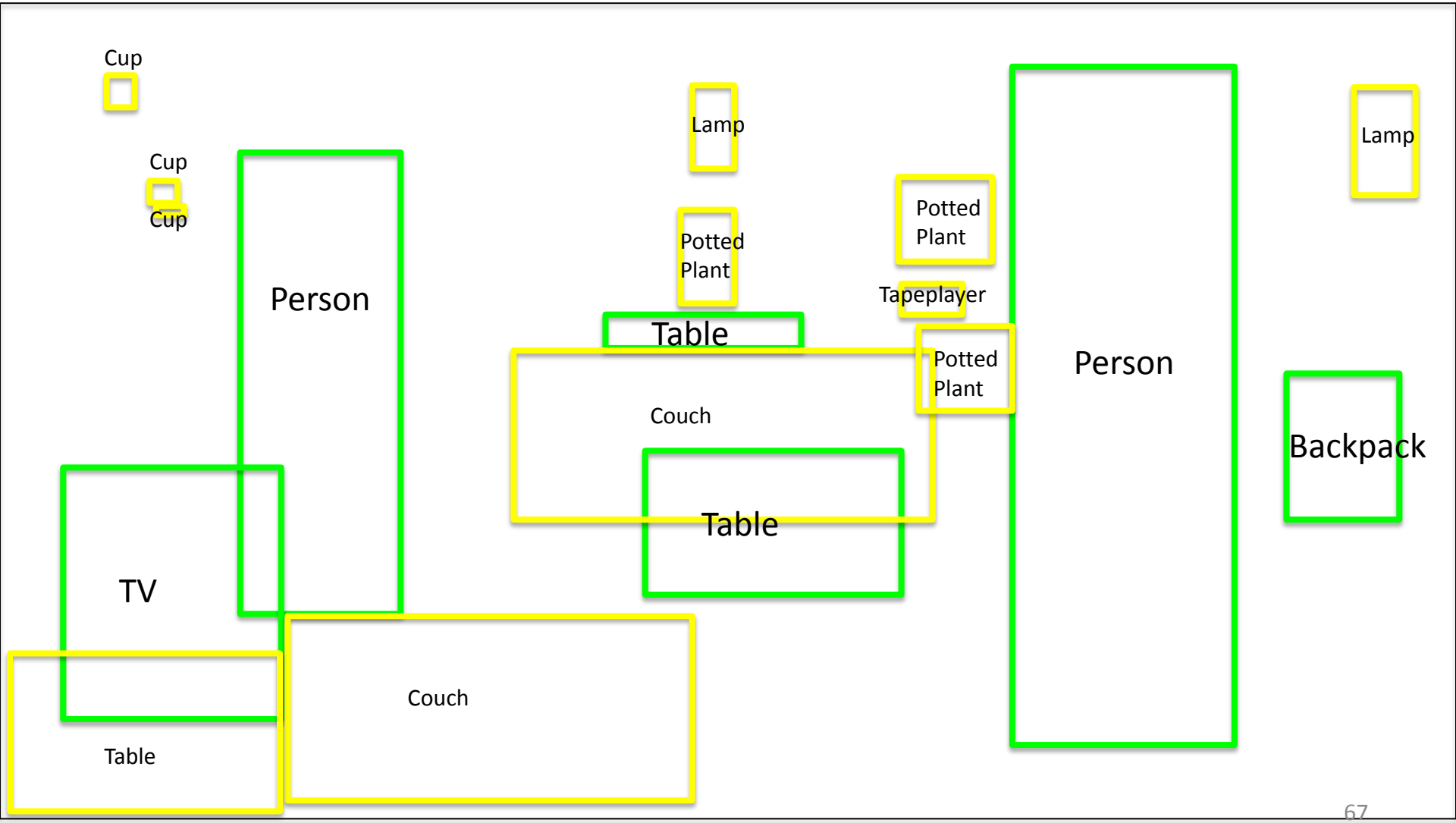
current methods are better than non-domain-expert humans on fine-grained classification!

Russakovsky* and Deng* et al., ImageNet Large Scale Visual Recognition Challenge, http://arxiv.org/abs/1409.0575

IMAGENET

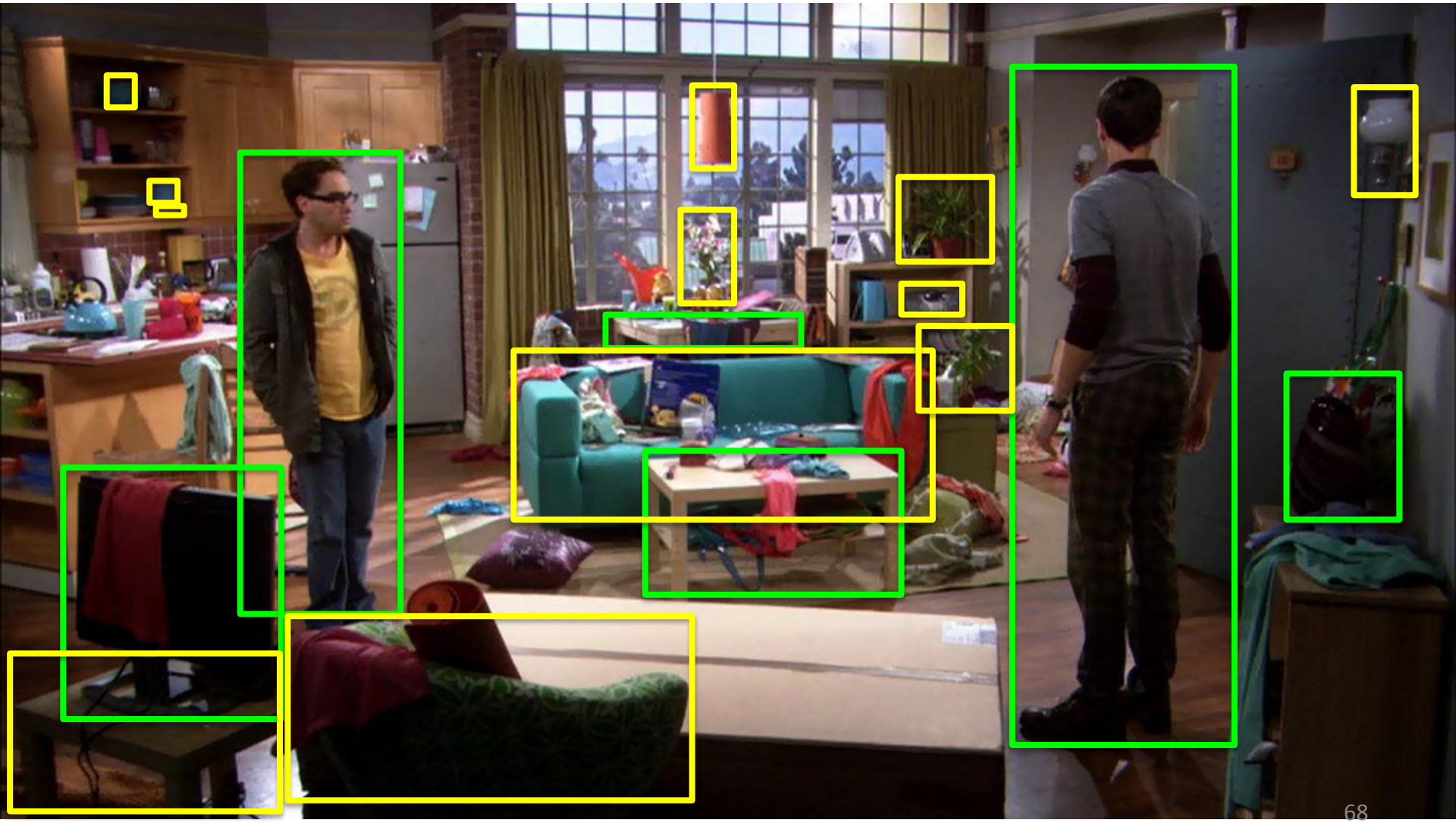
Future: MORE OBJECTS
MORE CONTEXT INFORMATION

ILSVRC object detection: *all instances of the 200 target objects*



IMAGENET

Future: MORE OBJECTS MORE CONTEXT INFORMATION



INDEX

1. Computer Vision Datasets
2. Annotations and metrics
3. PASCAL Challenges
4. IMAGENET Challenges
5. **ChaLearn Challenges**
6. Other large scale CV DBs

ChaLearn Looking at People

ChaLearn http://www.chalearn.org/

Mission:

Machine Learning is the science of building hardware or software that can achieve tasks by learning from examples. The examples often come as {input, output} pairs. Given new inputs a trained machine can make predictions of the unknown output.

Examples of machine learning tasks include:

- automatic reading of handwriting
- assisted medical diagnosis
- automatic text classification (classification of web pages; spam filtering)
- financial predictions

We organize challenges to stimulate research in this field. **The web sites of past challenges remain open** for post-challenge submission as ever-going benchmarks.

ChaLearn is a tax-exempt organization under section 501(c)(3) of the US IRS code. DLN: 17053090370022.

ChaLearn Looking at people (multimedia datasets, http://gesture.chalearn.org/)

ChaLearn Looking at People Challenges and Workshops

CVPR 2011	Workshop and Challenge on Multi-modal Sign Language Recognition
CVPR 2012	Workshop and Challenge on Multi-modal Sign Language Recognition
ICPR 2012	Workshop and Challenge on Multi-modal Sign Language Recognition
ICMI 2013	Workshop and Challenge on Multi-modal gesture recognition
ECCV 2014	Workshop and Challenge on multi-modal gesture spotting, human pose recovery, action/interaction spotting
CVPR 2015	Workshop and Challenge on human pose recovery, action/interaction spotting, cultural event recognition
ICCV 2015	Workshop and Challenge on age estimation, action spotting and cultural event recognition
~ 2016	Workshop and Challenge on Multi-modal speed interviews analysis

And so on! Let us know about your opinion! sergio@maia.ub.es

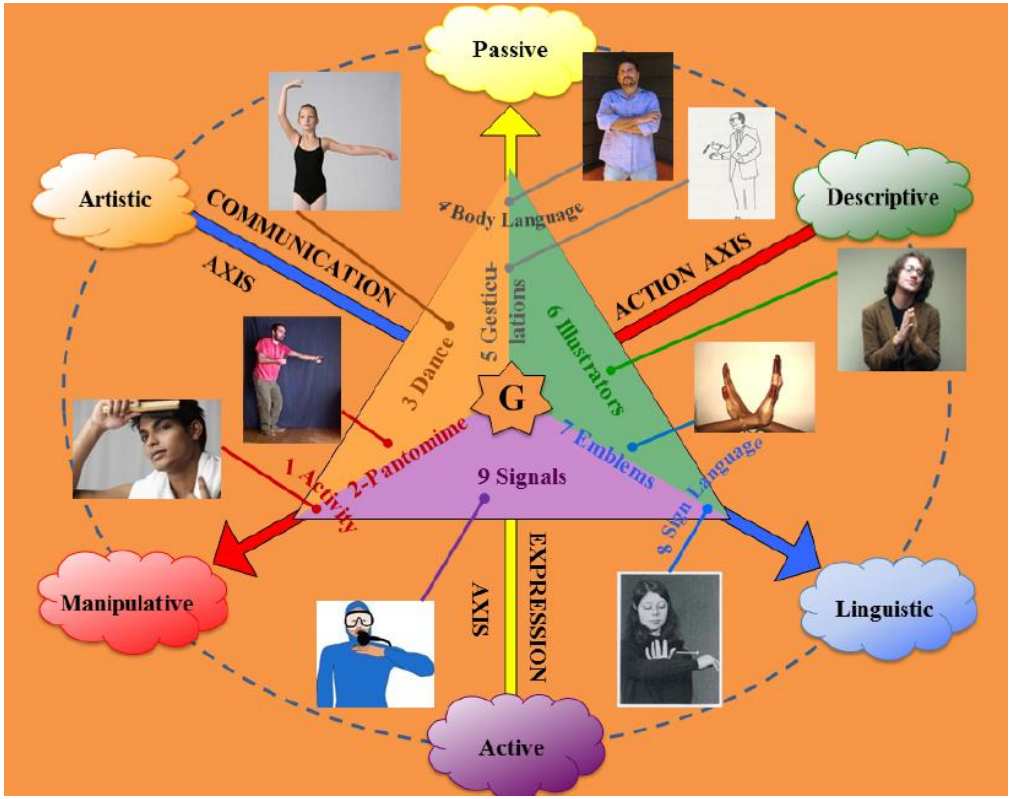
ChaLearn Looking at People

ChaLearn Looking at People Challenges and Workshops

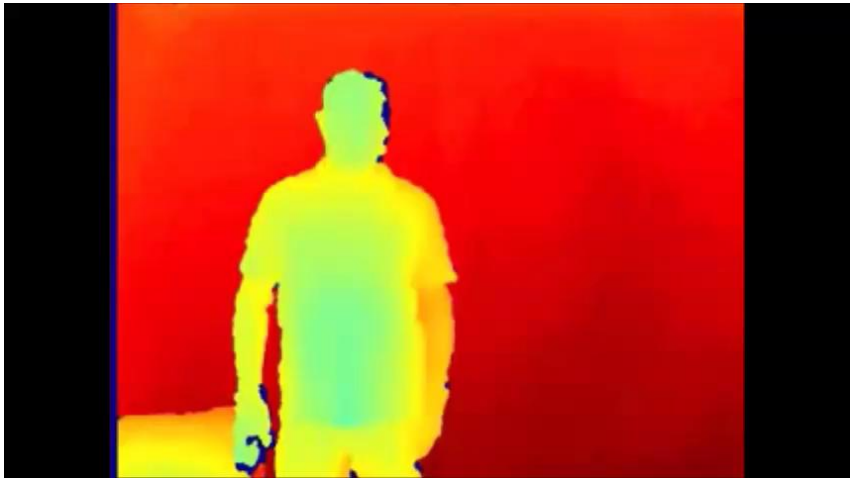
CVPR 2011 Workshop and Challenge on Multi-modal Sign Language Recognition

CVPR 2012 Workshop and Challenge on Multi-modal Sign Language Recognition

ICPR 2012 Workshop and Challenge on Multi-modal Sign Language Recognition



Microsoft Research



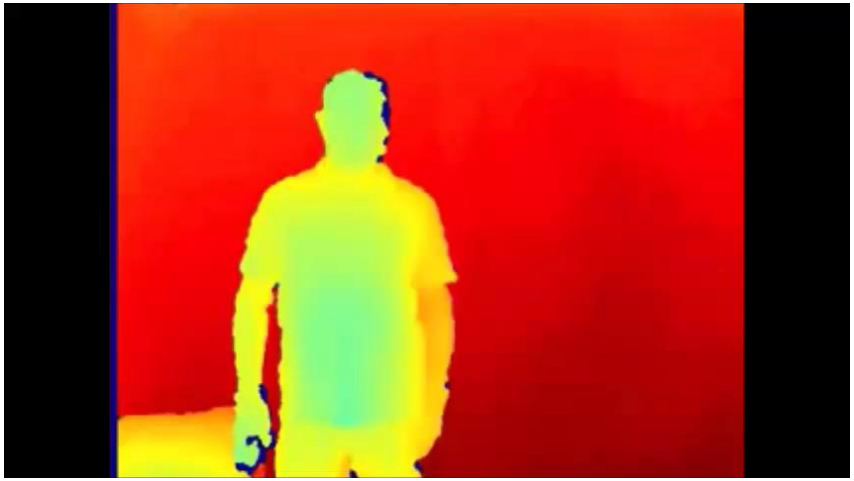
ChaLearn Looking at People

ChaLearn Looking at People Challenges and Workshops

CVPR 2011	Workshop and Challenge on Multi-modal Sign Language Recognition
CVPR 2012	Workshop and Challenge on Multi-modal Sign Language Recognition
ICPR 2012	Workshop and Challenge on Multi-modal Sign Language Recognition

Microsoft Research

1. Body language gestures (like scratching your head, crossing your arms).
2. Gesticulations performed to accompany speech.
3. Illustrators (like Italian gestures).
4. Emblems (like Indian Mudras).
5. Signs (from sign languages for the deaf).
6. Signals (like referee signals, diving signals, or Marshalling signals to guide machinery or vehicle).
7. Actions (like drinking or writing).
8. Pantomimes (gestures made to mimic actions).
9. Dance postures.



Evaluation metric: levenstein edition distance

ChaLearn Looking at People Challenges and Workshops

ICMI 2013 Workshop and Challenge on Multi-modal gesture recognition

Multi-modal ChaLearn Gesture Recognition Challenge and Workshop

http://gesture.chalearn.org/
sunai.uoc.edu/chalearn

Web of the competition
Data

The challenge features a **quantitative evaluation** of automatic gesture recognition from a multi-modal dataset recorded with **Kinect** (providing RGB images of face and body, depth images of face and body, skeleton information, joint orientation and audio sources), **including 13,858 Italian gestures from near 30 users.**

The emphasis of this edition of the competition will be on multi-modal automatic learning of a **vocabulary of 20 types of Italian anthropological/cultural gestures performed by different users**, with the aim of **performing user independent continuous gesture recognition combined with audio information**.

ChaLearn Looking at People Challenges and Workshops

ICMI 2013 Workshop and Challenge on Multi-modal gesture recognition

Gesture categories (1/2)

(1) *Vattene*

(2) *Viene qui*

(3) *Perfetto*

(4) *E un furbo*

(5) *Che due palle*

(6) *Che vuoi*

(7) *Vanno d'accordo*

(8) *Sei pazzo*

(9) *Cos hai combinato*

(10) *Nonme me friega niente*

ChaLearn Looking at People Challenges and Workshops

ICMI 2013 Workshop and Challenge on Multi-modal gesture recognition

Gesture categories (2/2)

(11) *Ok*

(12) *Cosa ti farei*

(13) *Basta*

(14) *Le vuoi prendere*

(15) *Non ce ne piu*

(16) *Ho fame*

(17) *Tanto tempo fa*

(18) *Buonissimo*

(19) *Si sono messi d'accordo*

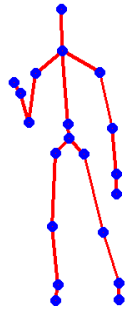
(20) *Sono stufo*

ChaLearn Looking at People Challenges and Workshops

ICMI 2013 Workshop and Challenge on Multi-modal gesture recognition

Data and modalities



- Framerate 20FPS
- RGB: 640x480
- Depth: 640x480
- Audio: Kinect 20 michropone array
- Users: 27
- Italians: 81%

- Total number of sequences: 956 ϵ [1,2] min.
- Total number of gestures: 13,858
- Total number of frames: 1.720.800
- Noisy gestures

Data structure information: *S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athistos, H.J. Escalante, "Multi-modal Gesture Recognition Challenge 2013: Dataset and Results", ICMI 2013*.

ChaLearn Looking at People Challenges and Workshops

ICMI 2013 Workshop and Challenge on Multi-modal gesture recognition

Easy and challenging aspects of the data.

Easy
Fixed camera
Near frontal view acquisition
Within a sequence the same user
Gestures performed mostly by arms and hands
Camera framing upper body
Several available modalities: audio, skeletal model, user mask, depth, and RGB
Several instances of each gesture for training
Single person present in the visual field

Challenging
Within each sequence:
Continuous gestures without a resting pose
Many gesture instances are present
Distracter gestures out of the vocabulary may be present in terms of both gesture and audio
Between sequences:
High inter and intra-class variabilities of gestures in terms of both gesture and audio
Variations in background, clothing, skin color, lighting, temperature, resolution
Some parts of the body may be occluded
Different Italian dialects

Evaluation metric: levenstein edition distance

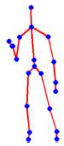
ChaLearn Looking at People Challenges and Workshops

ICMI 2013 Workshop and Challenge on Multi-modal gesture recognition

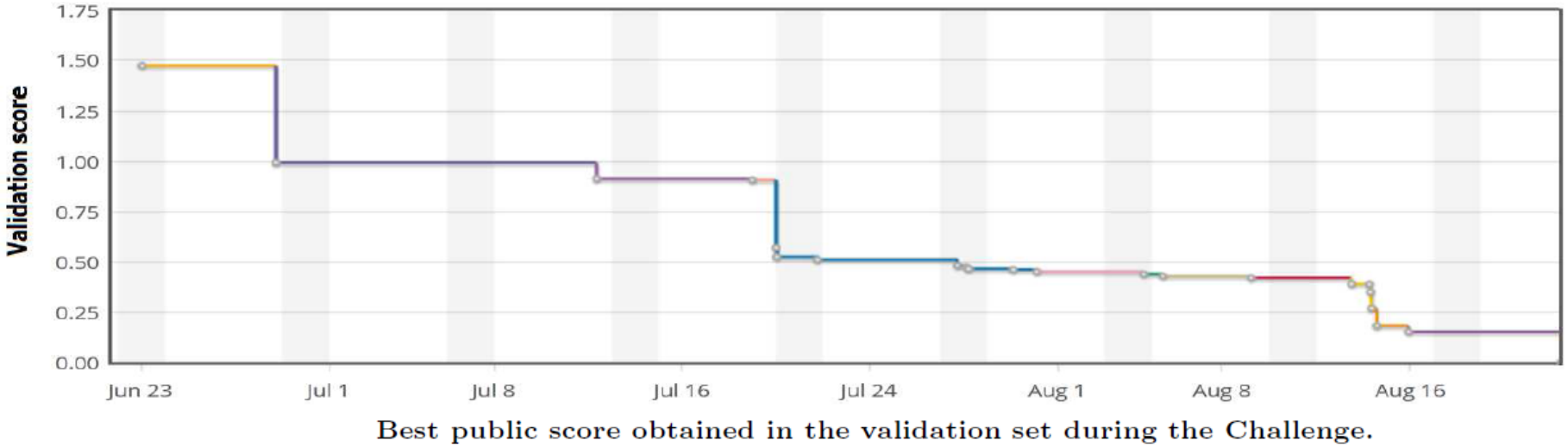

RGB


Depth


User mask


Skeletal model

+Audio **Levenstein score**



Best public score obtained in the validation set during the Challenge.

kaggle

ChaLearn Looking at People Challenges and Workshops
ICMI 2013 Workshop and Challenge on Multi-modal gesture recognition

- Participation

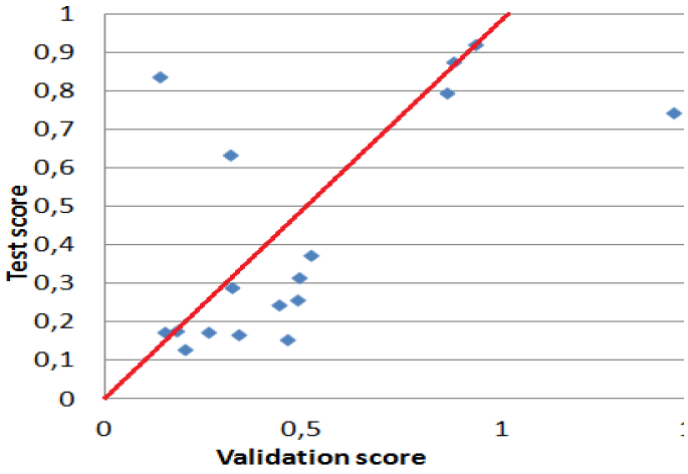
• The challenge attracted high level of participation, with a total of **54 teams and near 300 total number of entries**.
• Finally, **17 teams successfully submitted their prediction in final test set, while providing also their code for verification and summarizing their method by means of a fact sheet questionnaire**.
• After verifying the codes and results of the participants, the final scores of the top rank participants on both validation and test sets were made public.
• In the end, **the final error rate on the test data set was around 12%.**

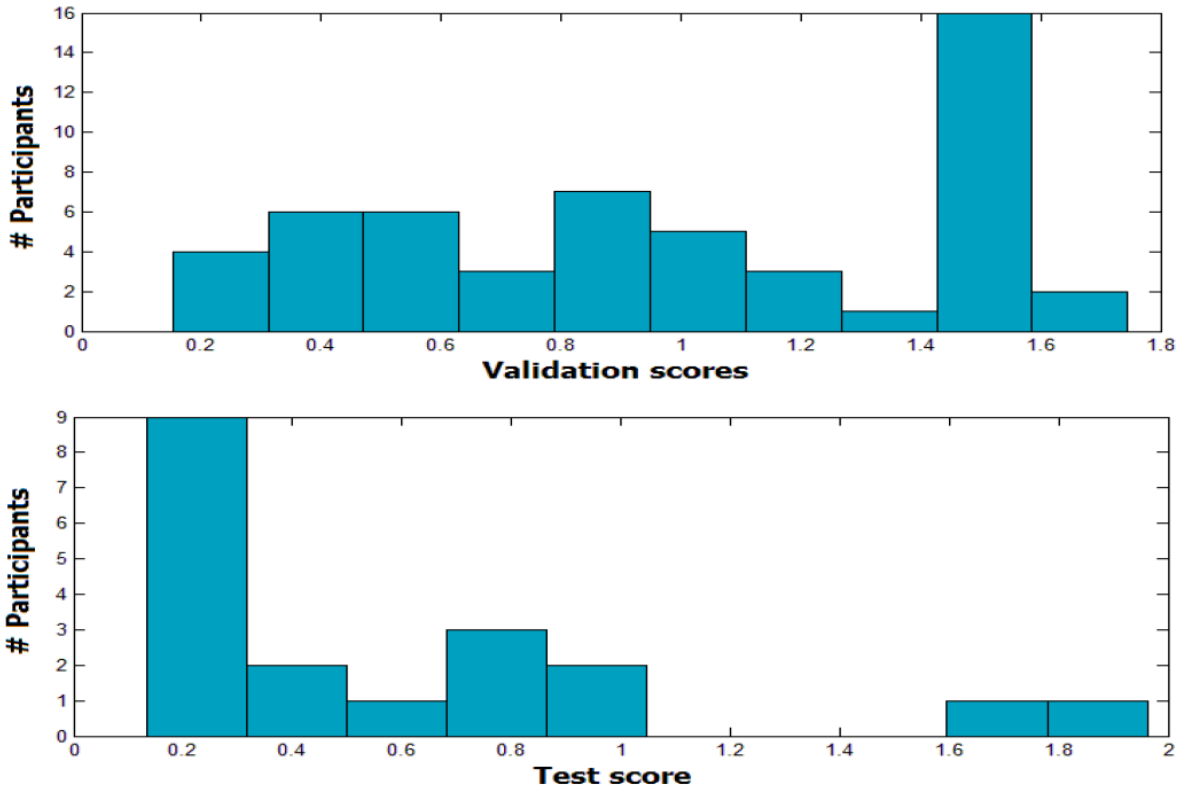
Top rank results on validation and test sets.

TEAM	Validation score	Test score
IVA MM	0.20137	**0.12756**
WWEIGHT	0.46163	**0.15387**
ET	0.33611	**0.16813**
MmM	0.25996	0.17215
PPTK	0.15199	0.17325
LRS	0.18114	0.17727
MMDL	0.43992	0.24452
TELEPOINTS	0.48543	0.25841
CSI MM	0.32124	0.28911
SUMO	0.49137	0.31652
GURU	0.51844	0.37281
AURINKO	0.31529	0.63304
STEVENWUDI	1.43427	0.74415
JACKSPARROW	0.86050	0.79313
JOEWAN	0.13653	0.83772
MILAN KOVAC	0.87835	0.87463
IAMKHADER	0.93397	0.92069

ChaLearn Looking at People Challenges and Workshops

ICMI 2013 Workshop and Challenge on Multi-modal gesture recognition





Validation and test scores histograms.

ChaLearn Looking at People Challenges and Workshops

ICMI 2013 Workshop and Challenge on Multi-modal gesture recognition

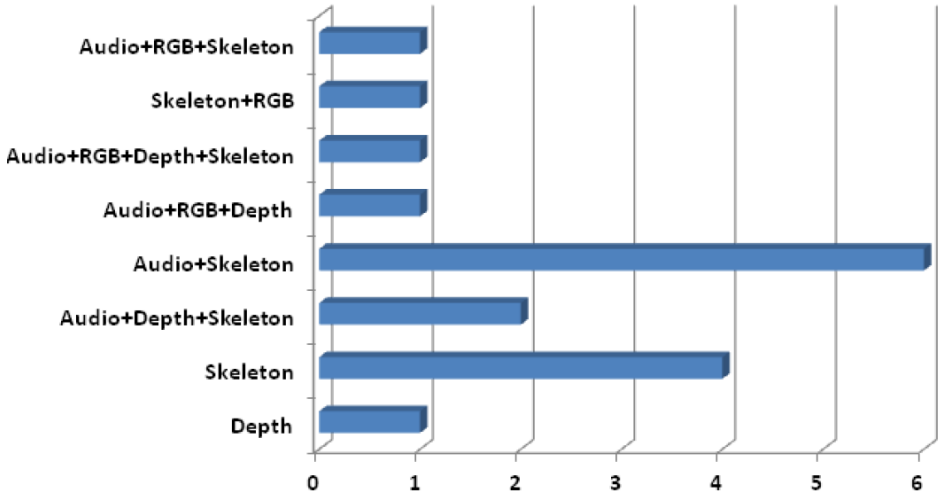
Team methods and results. Early and late refer to early and late fusion of features/classifier outputs. HMM: Hidden Markov Models. KNN: Nearest Neighbor. RF: Random Forest. Tree: Decision Trees. ADA: Adaboost variants. SVM: Support Vector Machines. Fisher: Fisher Linear Discriminant Analysis. GMM: Gaussian Mixture Models. NN: Neural Networks. DGM: Deep Boltzmann Machines. LR: Logistic Regression. DP: Dynamic Programming. ELM: Extreme Learning Machines.

TEAM	Test score	Rank position	Modalities	Segmentation	Fusion	Classifier
IVA MM	0.12756	1	Audio,Skeleton	Audio	None	HMM,DP,KNN
WWEIGHT	0.15387	2	Audio,Skeleton	Audio	Late	RF,KNN
ET	0.16813	3	Audio,Skeleton	Audio	Late	Tree,RF,ADA
MmM	0.17215	4	Audio,RGB+Depth	Audio	Late	SVM,Fisher,GMM,KNN
PPTK	0.17325	5	Skeleton,RGB,Depth	Sliding windows	Late	GMM,HMM
LRS	0.17727	6	Audio,Skeleton,Depth	Sliding windows	Early	NN
MMDL	0.24452	7	Audio,Skeleton	Sliding windows	Late	DGM+LR
TELEPOINTS	0.25841	8	Audio,Skeleton,RGB	Audio,Skeleton	Late	HMM,SVM
CSI MM	0.28911	9	Audio,Skeleton	Audio	Early	HMM
SUMO	0.31652	10	Skeleton	Sliding windows	None	RF
GURU	0.37281	11	Audio,Skeleton,Depth	DP	Late	DP,RF,HMM
AURINKO	0.63304	12	Skeleton,RGB	Skeleton	Late	ELM
STEVENWUDI	0.74415	13	Audio,Skeleton	Sliding windows	Early	DNN,HMM
JACKSPARROW	0.79313	14	Skeleton	Sliding windows	None	NN
JOEWAN	0.83772	15	Skeleton	Sliding windows	None	KNN
MILAN KOVAC	0.87463	16	Skeleton	Sliding windows	None	NN
IAMKHADER	0.92069	17	Depth	Sliding windows	None	RF

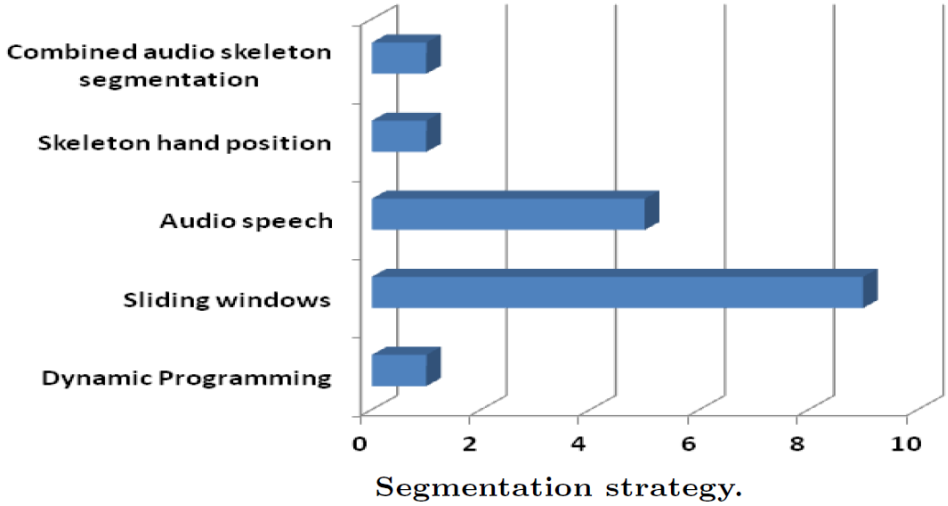
Data structure information: *S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athistos, H.J. Escalante, "Multi-modal Gesture Recognition Challenge 2013: Dataset and Results", ICMI 2013*.

ChaLearn Looking at People Challenges and Workshops

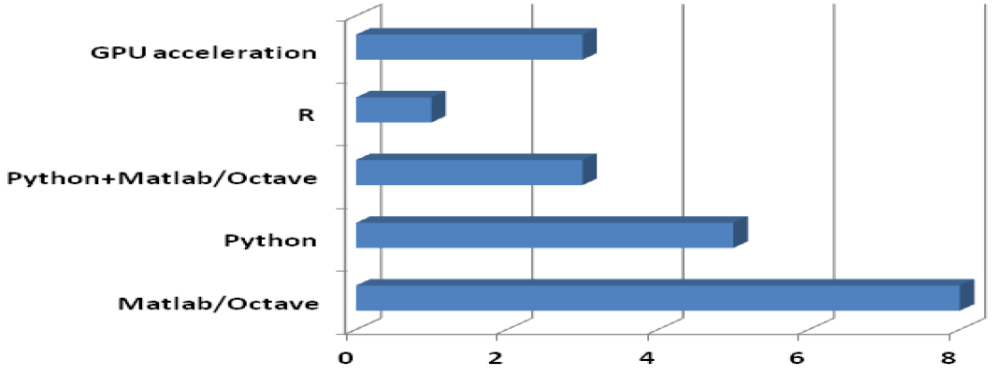
ICMI 2013 Workshop and Challenge on Multi-modal gesture recognition



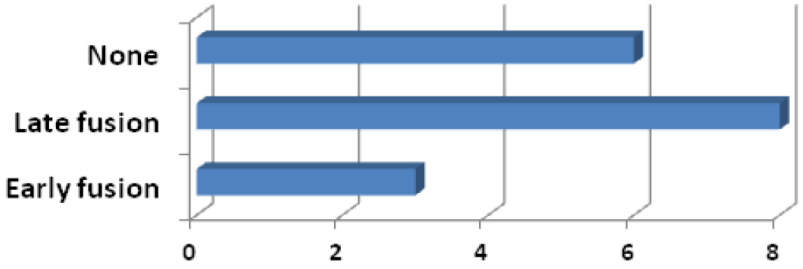
Modalities considered.



Segmentation strategy.



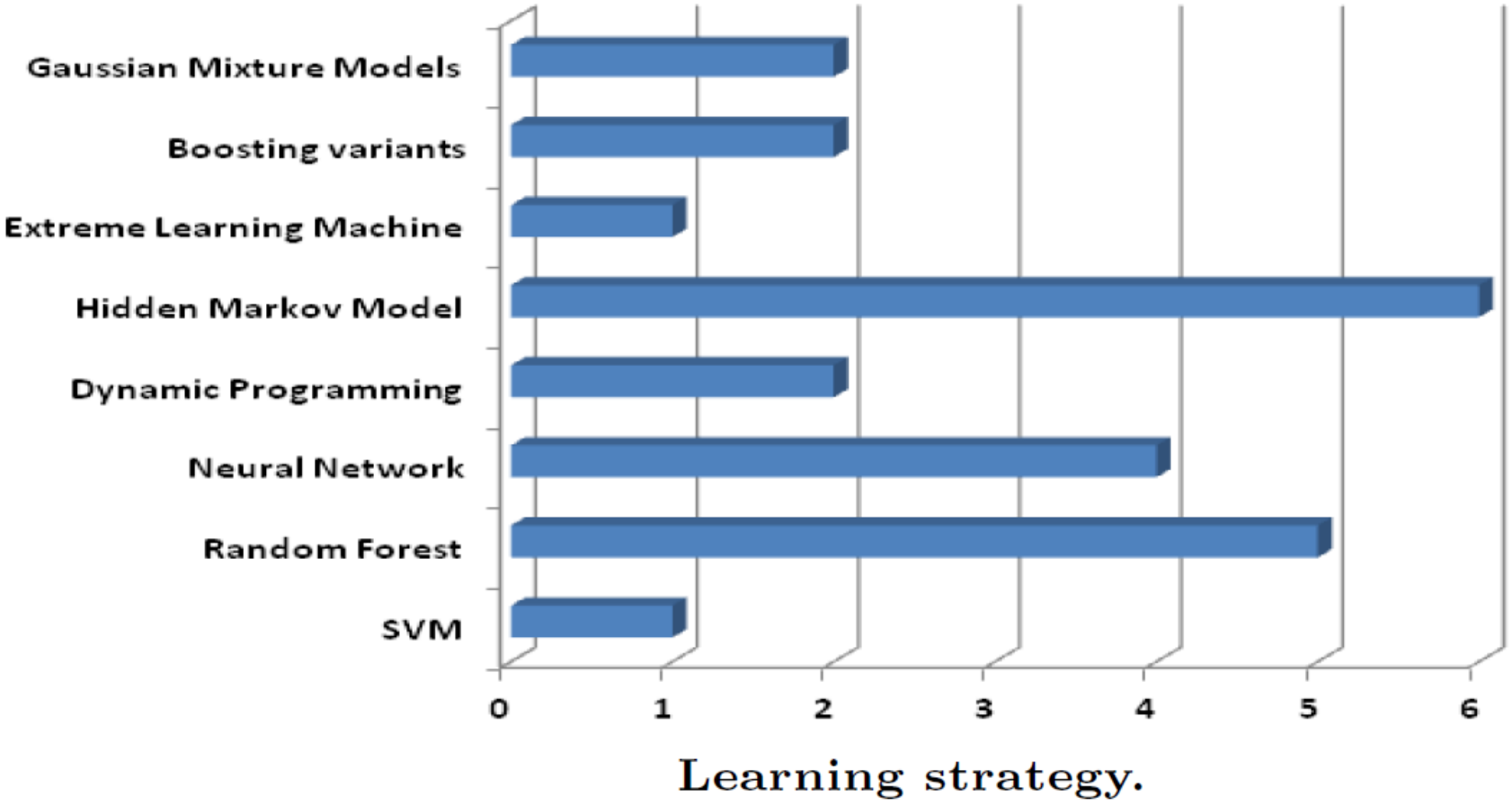
Programming language.



Fusion strategy.

ChaLearn Looking at People Challenges and Workshops

ICMI 2013 Workshop and Challenge on Multi-modal gesture recognition



ChaLearn Looking at People Challenges and Workshops

ECCV 2014 Workshop and Challenge on multi-modal gesture spotting, human pose recovery, action/interaction spotting

•**Track 1: Human Pose Recovery:** More than 8,000 frames of continuous RGB sequences are recorded and labeled with the objective of performing human pose recovery by means of recognizing more than 120,000 human limbs of different people.

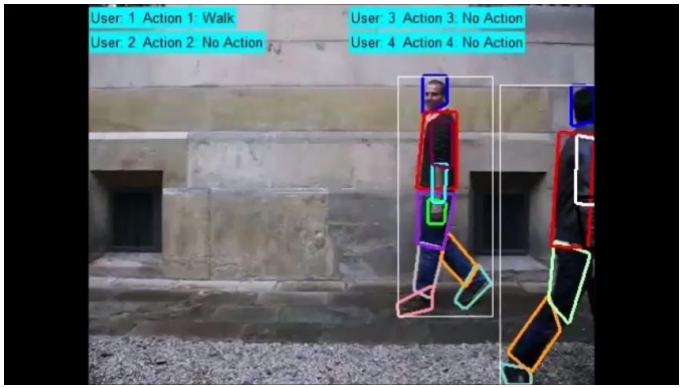
•**Track 2: Action/Interaction Recognition:** 235 performances of 11 action/interaction categories are recorded and manually labeled in continuous RGB sequences of different people performing natural isolated and collaborative behaviors.

•**Track 3: Gesture Recognition:** The gestures are drawn from a vocabulary of Italian sign gesture categories. The emphasis of this third track is on multi-modal automatic learning of a set of 20 gestures performed by several different users, with the aim of performing user independent continuous gesture spotting.

ChaLearn Looking at People Challenges and Workshops

ECCV 2014 Workshop and Challenge on multi-modal gesture spotting, human pose recovery, action/interaction spotting

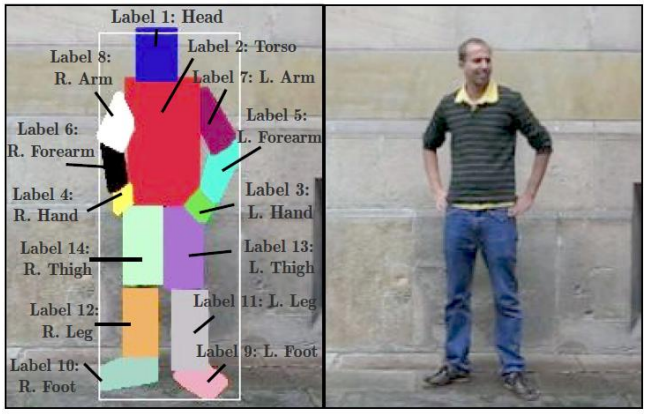
•**Track 1: Human Pose Recovery:** More than 8,000 frames of continuous RGB sequences are recorded and labeled with the objective of performing human pose recovery by means of recognizing more than 120,000 human limbs of different people.



Training frames	Validation frames	Test frames	Sequence duration	FPS
4,000	2,000	2,236	1-2 min	15

Modalities	Num. of users	Limbs per body	Labeled frames	Labeled limbs
RGB	14	14	8,234	124,761

Human pose recovery data characteristics.



•**9 videos** (RGB sequences) and a total of **14** different **actors**. Stationary camera with the same static background.

•**15 fps** rate, resolution **480x360** in BMP file format.

• For each actor **14 limbs** (if not occluded) were manually tagged: Head, Torso, R-L Upper-arm, R-L Lower-arm, R-L Hand, R-L Upper-leg, R-L Lower-leg, and R-L Foot.

• **Limbs are manually labeled** using binary masks and the minimum bounding box containing each subject is defined.

• The actors appear in a **wide range of different poses** and **performing different actions/gestures** which vary the visual appearance of human limbs. So there is a large variability of human poses, self-occlusions and many variations in clothing and skin color.

ChaLearn Looking at People Challenges and Workshops

ECCV 2014 Workshop and Challenge on multi-modal gesture spotting, human pose recovery, action/interaction spotting

•**Track 1: Human Pose Recovery:** More than 8,000 frames of continuous RGB sequences are recorded and labeled with the objective of performing human pose recovery by means of recognizing more than 120,000 human limbs of different people.

Overlap evaluation

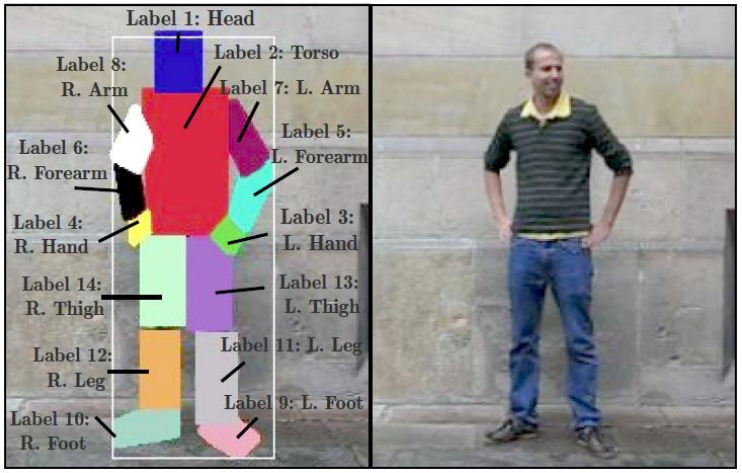
$$J_{i,n} = \frac{A_{i,n} \bigcap B_{i,n}}{A_{i,n} \bigcup B_{i,n}},$$

$$H_{i,n} = \begin{cases} 1 & \text{if } \frac{A_n \bigcap B_n}{A_n \bigcup B_n} \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

$A_{i,head}$ $B_{i,head}$

$$J_{i,head} = \frac{A_{i,head} \cap B_{i,head}}{A_{i,head} \cup B_{i,head}} = 0.82$$

$$J_{i,head} > 0.5 \longrightarrow HR_{i,head} = 1$$



ChaLearn Looking at People Challenges and Workshops

ECCV 2014 Workshop and Challenge on multi-modal gesture spotting, human pose recovery, action/interaction spotting

•**Track 2: Action/Interaction Recognition:** 235 performances of 11 action/interaction categories are recorded and manually labeled in continuous RGB sequences of different people performing natural isolated and collaborative behaviors.

Training actions	Validation actions	Test actions	Sequence duration	FPS
150	90	95	9× 1-2 min	15

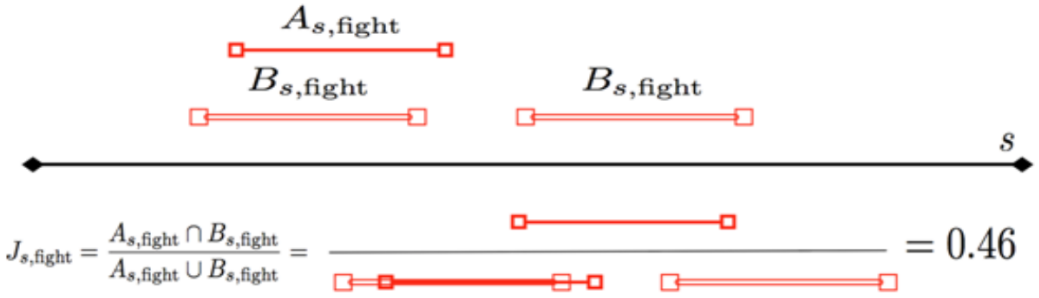
Modalities	Num. of users	Action categories	interaction categories	Labeled sequences
RGB	14	7	4	235

Action and interaction data characteristics.

- **235 action/interaction** samples performed by **14 actors**.
- Large **difference in length** about the performed actions and interactions.
- Several **distracter actions** out of the 11 categories are also present.
- **11 action categories, containing isolated and collaborative actions**: Wave, Point, Clap, Crouch, Jump, Walk, Run, Shake Hands, Hug, Kiss, Fight. There is a high intra-class variability among action samples.

Overlap evaluation

$$J_{s,n} = \frac{A_{s,n} \bigcap B_{s,n}}{A_{s,n} \bigcup B_{s,n}},$$



ChaLearn Looking at People Challenges and Workshops

ECCV 2014 Workshop and Challenge on multi-modal gesture spotting, human pose recovery, action/interaction spotting

Track 2: Action/Interaction Recognition

Action categories


Wave


Point


Clap


Crouch


Jump


Walk

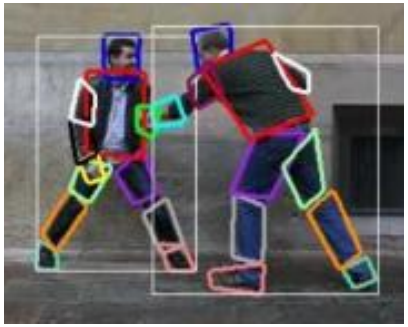

Run

Interaction categories


Shake Hands


Hug


Kiss


Fight

ChaLearn Looking at People Challenges and Workshops

ECCV 2014 Workshop and Challenge on multi-modal gesture spotting, human pose recovery, action/interaction spotting

Track 3: Gesture Recognition: The gestures are drawn from a vocabulary of Italian sign gesture categories. The emphasis of this third track is on multi-modal automatic learning of a set of 20 gestures performed by several different users, with the aim of performing user independent continuous gesture spotting.

Training seq.	Validation seq.	Test seq.	Sequence duration	FPS
393 (7,754 gestures)	287 (3,362 gestures)	276 (2,742 gestures)	1-2 min	20

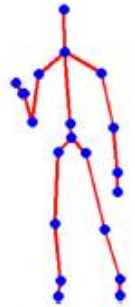
Modalities	Num. of users	Gesture categories	Labeled sequences	Labeled frames
RGB, Depth, User mask, Skeleton	27	20	13,858	1,720,800

Main characteristics of the *Montalbano* gesture dataset.

RGB

Depth

User mask



Skeletal model

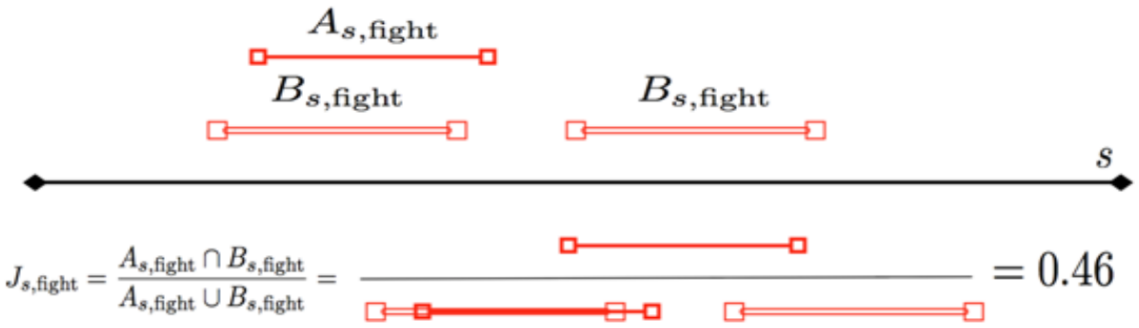
ChaLearn Looking at People Challenges and Workshops

ECCV 2014 Workshop and Challenge on multi-modal gesture spotting, human pose recovery, action/interaction spotting

- **Largest dataset** in the literature with a large duration of each individual performance showing **no resting poses and self-occlusions**.
- There is **no information about the number of gestures to spot** within each sequence, and **several distracter gestures** (out of the vocabulary) are present.
- **High intra-class variability** of gesture samples and **low inter-class variability** for some gesture categories.

Overlap evaluation

$$J_{s,n} = \frac{A_{s,n} \bigcap B_{s,n}}{A_{s,n} \bigcup B_{s,n}},$$



ChaLearn Looking at People Challenges and Workshops

ECCV 2014 Workshop and Challenge on multi-modal gesture spotting, human pose recovery, action/interaction spotting

- State of the art comparison

	Labeling at pixel precision	Number of limbs	Number of labeled limbs	Number of frames	Full body	Limb annotation	Gesture-action annotation	Number of gestures-actions	Number of gest-act. samples
Montalbano[8]	No	16	27 532 800	1 720 800	Yes	Yes	Yes	20	13 858
HuPBA 8K+ [7]	Yes	14	124 761	8 234	Yes	Yes	Yes	11	235
LEEDS SPORTS[4]	No	14	28 000	2 000	Yes	Yes	No	-	-
UIUC people[10]	No	14	18 186	1 299	Yes	Yes	No	-	-
Pascal VOC[2]	Yes	5	8 500	1 218	Yes	Yes	No	-	-
BUFFY[3]	No	6	4 488	748	No	Yes	No	-	-
PARSE[11]	No	10	3 050	305	Yes	Yes	No	-	-
MPII Pose[12]	Yes	14	-	40 522	Yes	Yes	Yes	20	491
FLIC[13]	No	29	-	5 003	No	No	No	-	-
H3D[14]	No	19	-	2 000	No	No	No	-	-
Actions[15]	No	-	-	-	Yes	No	Yes	6	600
HW[5]	-	-	-	-	-	No	Yes	8	430

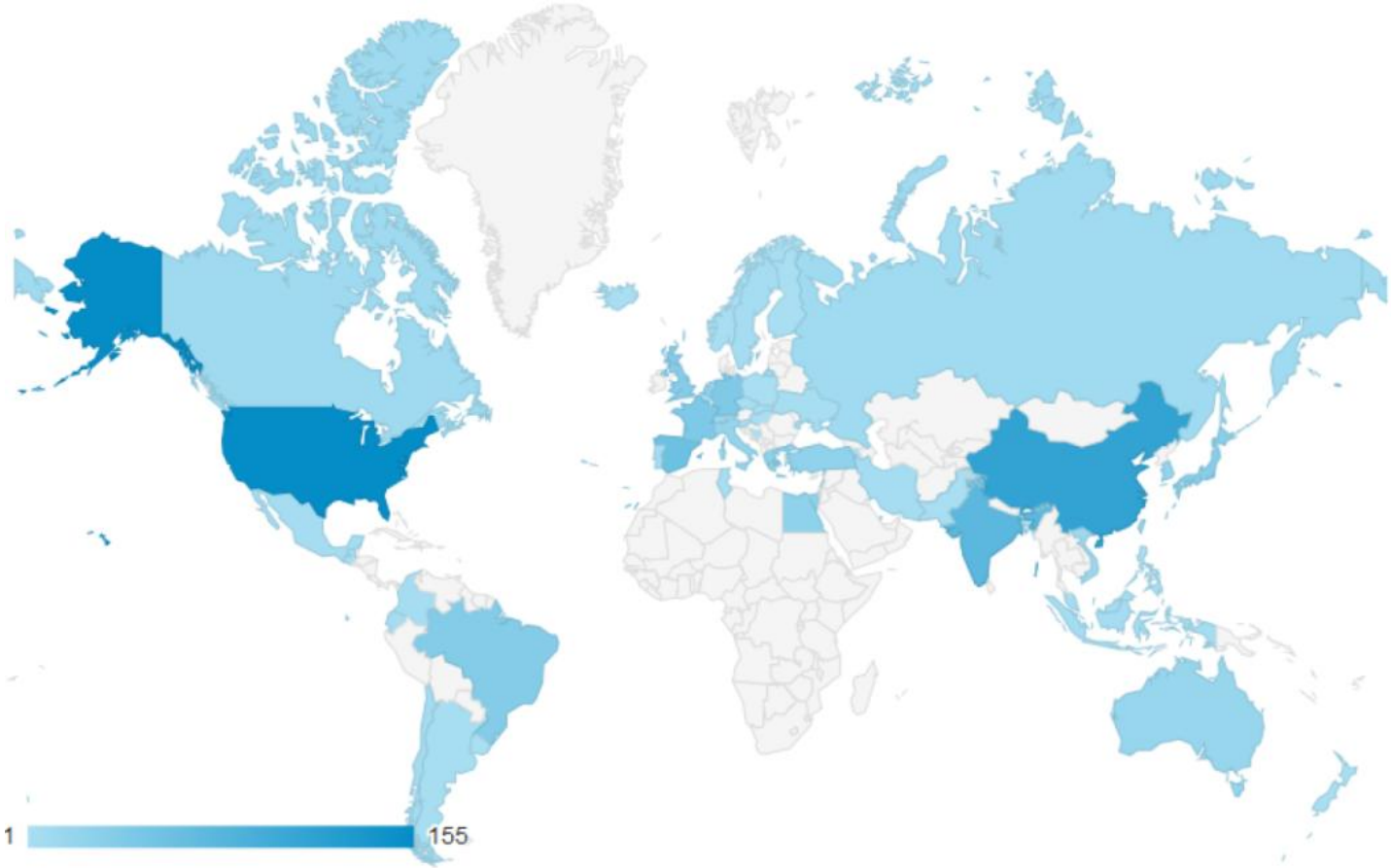
Comparison of public dataset characteristics.

ChaLearn LAP data sets, public available at:
http://sunai.uoc.edu/chalearnLAP/
ChaLearn LAP challenges and news:
http://gesture.chalearn.org/

ChaLearn Looking at People Challenges and Workshops

ECCV 2014 Workshop and Challenge on multi-modal gesture spotting, human pose recovery, action/interaction spotting

- **Connectivity:** During the Challenge period, the download page had a total of 2.895 visits from 920 different users of 59 countries.

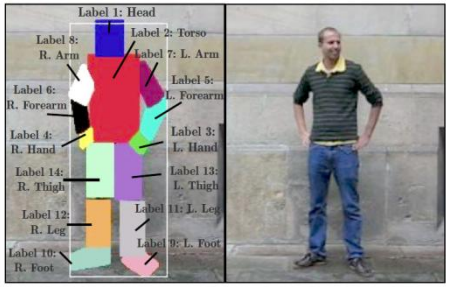


ChaLearn Looking at People Challenges and Workshops

ECCV 2014 Workshop and Challenge on multi-modal gesture spotting, human pose recovery, action/interaction spotting

- **Track1 results**

CodaLab



Team	Accuracy	Rank position	Features	Pose model
ZJU	**0.194144**	1	HOG	tree structure
Seawolf Vision	**0.182097**	2	HOG	tree structure

Track 1 Pose Recovery results.

Both winner participants applied a similar approach based on [*].

- Mixture of templates for each part. This method incorporates the co-occurrence relations, appearance and deformation into a model represented by an objective function of pose configurations. Model is tree-structured, and optimization is conducted via dynamic programming.

[*] Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE TPAMI (2013)

ChaLearn Looking at People Challenges and Workshops

ECCV 2014 Workshop and Challenge on multi-modal gesture spotting, human pose recovery, action/interaction spotting

- **Track2 results**

Team name	Accuracy	Rank	Features
CUHK-SWJTU	**0.507173**	1	Improved trajectories [*]
ADSC	**0.501164**	2	Improved trajectories [*]
SBUVIS	**0.441405**	3	Improved trajectories [*]
DonkeyBurger	0.342192	4	MHI, STIP
UC-T2	0.121565	5	Improved trajectories [*]
MindLAB	0.008383	6	MBF

Team name	Dimension reduction	Clustering	Classifier	Temporal coherence	Gesture representation
CUHK-SWJTU	PCA	-	SVM	Sliding windows	Fisher Vector
ADSC	-	-	SVM	Sliding windows	-
SBUVIS	-	-	SVM	Sliding windows	-
DonkeyBurger	-	Kmeans	Sparse code	Sliding windows	-
UC-T2	PCA	-	Kmeans	Sliding windows	Fisher Vector
MindLAB	-	Kmeans	RF	Sliding windows	BoW

* Wang, H., Schmid, C.: Action recognition with improved trajectories. ICCV (2013)

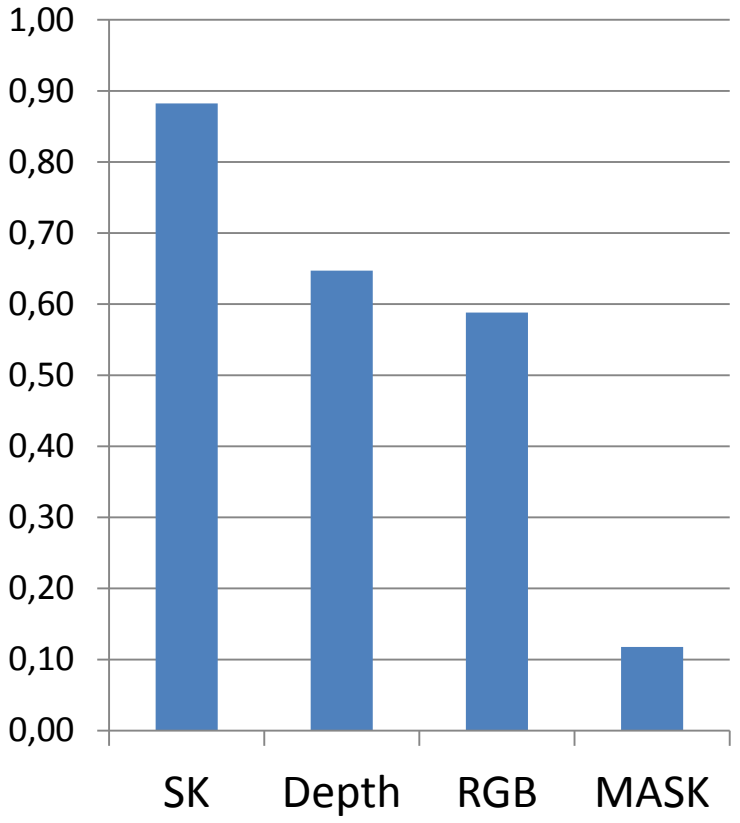
ChaLearn Looking at People Challenges and Workshops

ECCV 2014 Workshop and Challenge on multi-modal gesture spotting, human pose recovery, action/interaction spotting

- **Track3 results**

Team	Accuracy	Rank	Modalities
LIRIS	**0.849987**	1	SK, Depth, RGB
CraSPN	**0.833904**	2	SK, Depth, RGB
JY	**0.826799**	3	SK, RGB
CUHK-SWJTU	0.791933	4	RGB
Lpigou	0.788804	5	Depth, RGB
stevenwudi	0.787310	6	SK, depth
Ismar	0.746632	7	SK
Quads	0.745449	8	SK
Telepoints	0.688778	9	SK, Depth, RGB
TUM-fortiss	0.648979	10	SK, Depth, RGB
CSU-SCM	0.597177	11	Skeleton, Depth, mask
iva.mm	0.556251	12	Skeleton, RGB, depth
Terrier	0.539025	13	Skeleton
Team Netherlands	0.430709	14	Skeleton, Depth, RGB
VecsRel	0.408012	15	Skeleton, Depth, RGB
Samgest	0.391613	16	Skeleton, Depth, RGB, mask
YNL	0.270600	17	Skeleton

Percentage of methods using each independent modality

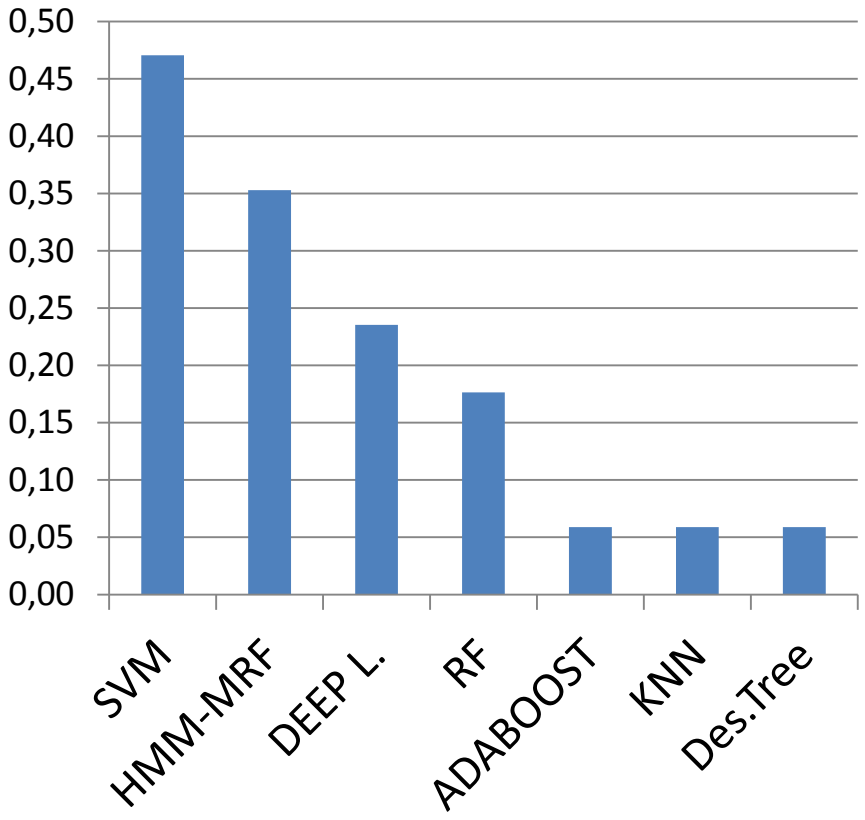


ChaLearn Looking at People Challenges and Workshops

ECCV 2014 Workshop and Challenge on multi-modal gesture spotting, human pose recovery, action/interaction spotting

- **Track3 results** **Percentage of methods using each gesture classification strategy**

Team	Gesture representation	Classifier
LIRIS	-	DNN
CraSPN	BoW	Adaboost
JY	-	MRF, KNN
CUHK-SWJTU	Fisher Vector, VLAD	SVM
Lpigou	-	CNN
stevenwudi	-	HMM, DNN
Ismar	-	RF
Quads	Fisher Vector	SVM
Telepoints	-	SVM
TUM-fortiss	-	RF, SVM
CSU-SCM	2DMTM	SVM, HMM
iva.mm	BoW	SVM, HMM
Terrier	-	RF
Team Netherlands	-	SVM, RT
VecsRel	-	DNN
Samgest	-	HMM
YNL	Fisher Vector	HMM, SVM



ChaLearn Looking at People Challenges and Workshops

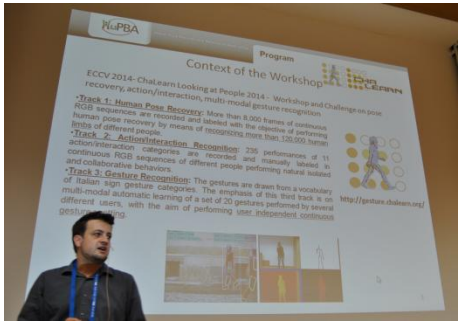
ECCV 2014 Workshop and Challenge on multi-modal gesture spotting, human pose recovery, action/interaction spotting

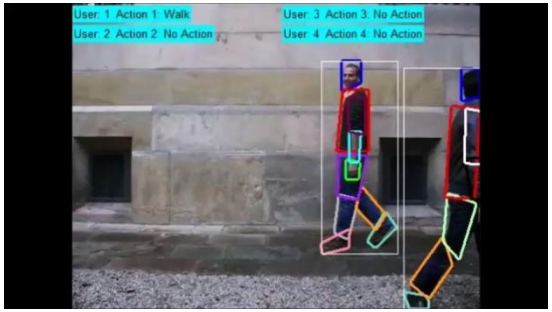
Team	Features	Fusion	Temp. segmentation	Dimension reduction
LIRIS	RAW, SK joints	Early	Joints motion	-
CraSPN	HOG, SK	Early	Sliding windows	-
JY	SK, HOG	Late	MRF	PCA
CUHK-SWJTU	Improved trajectories	-	Joints motion	PCA
Lpigou	RAW, SK joints	Early	Sliding windows	Max-pooling CNN
stevenwudi	RAW	Late	Sliding windows	-
Ismar	SK	-	Sliding windows	-
Quads	SK quads	-	Sliding windows	-
Telepoints	STIPS, SK	Late	Joints motion	-
TUM-fortiss	STIPS	Late	Joints motion	-
CSU-SCM	HOG, Skeleton	Late	Sliding windows	-
iva.mm	Skeleton, HOG	Late	Sliding windows	-
Terrier	Skeleton	-	Sliding windows	-
Team Netherlands	MHI	Early	DTW	Preserving projections
VecsRel	RAW, skeleton joints	Late	DTW	-
Samgest	Skeleton, blobs, moments	Late	Sliding windows	-
YNL	Skeleton	-	Sliding windows	-

For more details of the challenge and the results: *Sergio Escalera, Xavier Baró, Jordi Gonzàlez, Miguel Ángel Bautista, Meysam Madadi, Miguel Reyes, Víctor Ponce-López, Hugo J. Escalante, Jamie Shotton, Isabelle Guyon, ChaLearn Looking at People Challenge 2014: Dataset and Results, ChaLearn Looking at People, European Conference on Computer Vision, 2014.*

ChaLearn Looking at People Challenges and Workshops

ECCV 2014 Workshop and Challenge on multi-modal gesture spotting, human pose recovery, action/interaction spotting





ChaLearn Looking at People Challenges and Workshops

CVPR 2015 Workshop and Challenge on **action recognition** and cultural event recognition

Action/Interaction Recognition (second round)

Action categories


Wave


Point


Clap


Crouch


Jump


Walk

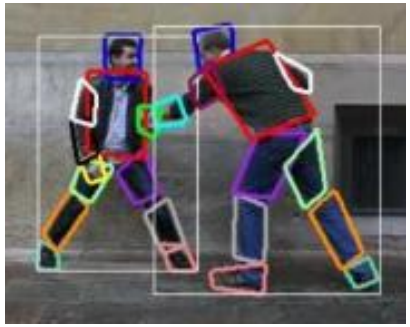

Run

Interaction categories


Shake Hands


Hug


Kiss


Fight

ChaLearn Looking at People Challenges and Workshops

CVPR 2015 Workshop and Challenge on action recognition and **cultural event recognition**

Dataset	#Images	#Classes	Year
Action Classification Dataset [8]	5,023	10	2010
Social Event Dataset [11]	160,000	149	2012
Event Identification Dataset [1]	594,000	24,900	2010
Cultural Event Dataset	11,776	50	2015

First state of the art data set for cultural event recognition in still images

ChaLearn Looking at People Challenges and Workshops

CVPR 2015 Workshop and Challenge on action recognition and cultural event recognition

Team name	Accuracy	Rank	Features
CUHK-SWJTU	**0.507173**	1	Improved trajectories [*]
ADSC	**0.501164**	2	Improved trajectories [*]
SBUVIS	**0.441405**	3	Improved trajectories [*]
DonkeyBurger	0.342192	4	MHI, STIP
UC-T2	0.121565	5	Improved trajectories [*]
MindLAB	0.008383	6	MBF

2014

Action/Interaction Track								
Rank	Team name	Score	Features	Dimension reduction	Clustering	Classification	Temporal coherence	Action representation
1	MMLAB	**0.5385**	IDT [19]	PCA	-	SVM	-	Fisher Vector
2	FIKIE	**0.5239**	IDT	PCA	-	HMM	Appearance+Kalman filter	-

Cultural Event Track				
Rank	Team name	Score	Features	Classification
1	MMLAB	**0.855**	Multiple CNN	Late weighted fusion of CNNs predictions.
2	UPC-ST	**0.767**	Multiple CNN	SVM and late weighted fusion.
3	MIPAL_SNU	**0.735**	Discriminant regions [18] + CNNs	Entropy + Mean Probabilities of all patches
4	SBU_CS	**0.610**	CNN-M [2]	SPM [10] based on LSSVM [16]
5	MasterBlaster	**0.58**	CNN	SVM, KNN, LR and One Vs Rest
6	Nyx	**0.319**	Selective-search approach [17] + CNN	Late fusion AdaBoost

2015

ChaLearn Looking at People Challenges and Workshops

ICCV 2015 Workshop and Challenge on **action recognition**, cultural event recognition, and apparent age recognition

ACTION RECOGNITION

- Large video collection (25 hours for testing) in which actions to detect are rare.
- Many examples for training (900 clips for 'Bull Charge Cape' and 500 for 'Horse Riding').
- High intra-class variability: different points of view, zoom level, action direction, color, occlusions.
- Actions are not related only to the human pose but with scene understanding.
- Videos produced during a 60 year period (1945 to 2012).

ChaLearn Looking at People Challenges and Workshops

ICCV 2015 Workshop and Challenge on action recognition, **cultural event recognition**, and apparent age recognition

Total number of images	Number of countries for all the events	Number of images per category	Number of categories	Number of test images	Number of validation images	Number of training images
25000	45	>200	90	5000	5000	15000

CULTURAL EVENT RECOGNITION

- First database on cultural events.
- More than 25,000 images representing 90 different categories.
- High intra- and inter-class variability.
- For this type of images, different cues can be exploited like garments, human poses, crowds analysis, objects and background scene.
- The evaluation metric will be the recognition accuracy.

ChaLearn Looking at People Challenges and Workshops

ICCV 2015 Workshop and Challenge on action recognition, cultural event recognition, and **apparent age recognition**

Range of labeled ages	Information from the labelers	Contains real age	Contains estimated age by the labelers	Number of labelers	Number of actors	Number of images
0-85	Nationality, age, and gender of the labelers	YES	YES	> 3600	>2000	5000

AGE ESTIMATION

- More than 5,000 faces from more than 2000 different people.
- Images with background.
- Non-controlled environments.
- Non-labeled faces neither landmarks, making the estimation problem even harder.
- One of the first datasets in the literature including estimated age labeled by many users to define the ground truth with the objective of estimating the age.
- The evaluation metric will be pondered by the mean and the variance of the labeling by the participants.
- The dataset also provides for each image the real age although not used for recognition (just for analysis purposes). In the same way for all the labelers we have their nationality, age, and gender, which will allow analyzing demographic and other interesting studies among the correlation of labelers.

INDEX

Other CV datasets

Microsoft COCO
Common Objects in Context

cocodataset@outlook.com

Home | People | Explore | Dataset

News

MS COCO Captions Challenge

Participate to the MS COCO Captioning Challenge organized with the LSUN Challenge at CVPR 2015

Click here for info

- a woman is playing a frisbee with a dog.
- a woman is playing frisbee with her large dog.
- a girl holding a frisbee with a dog coming at her.
- a woman kneeling down holding a frisbee in front of a white dog.
- a young lady is playing frisbee with her dog.

What is Microsoft COCO?

Microsoft COCO is a new image recognition, segmentation, and captioning dataset. Microsoft COCO has several features:

- **Object segmentation**
- **Recognition in Context**
- **Multiple objects per image**
- **More than 300,000 images**
- **More than 2 Million instances**
- **80 object categories**
- **5 captions per image**

Collaborators

Tsung-Yi Lin Cornell Tech

Michael Maire TTI Chicago

Serge Belongie Cornell Tech

Lubomir Bourdev Facebook AI

Ross Girshick Microsoft Research

James Hays Brown University

Pietro Perona Caltech

Deva Ramanan UC Irvine

Larry Zitnick Microsoft Research

Piotr Dollár Facebook AI

CORNELL NYC TECH

Caltech

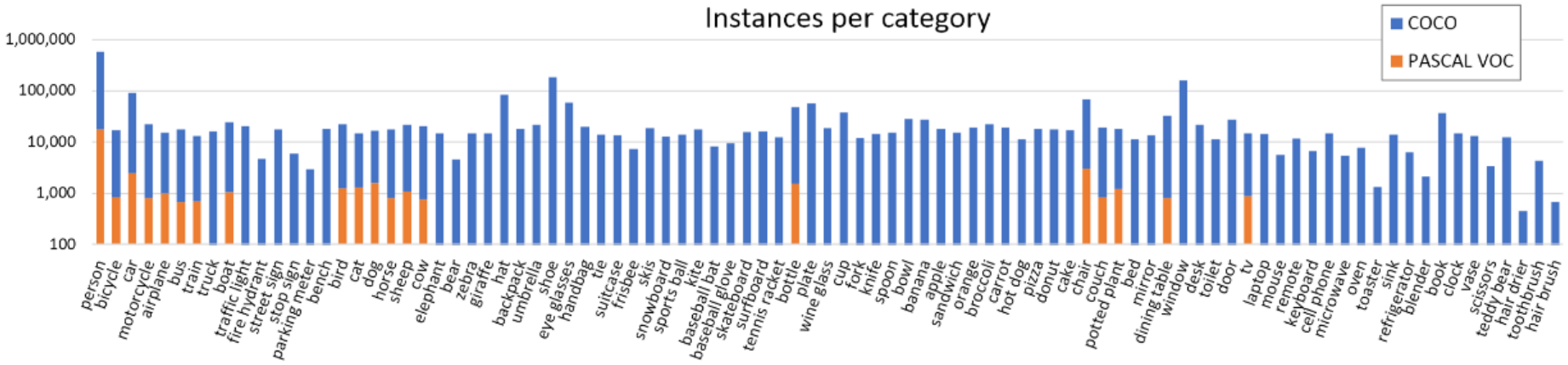
facebook

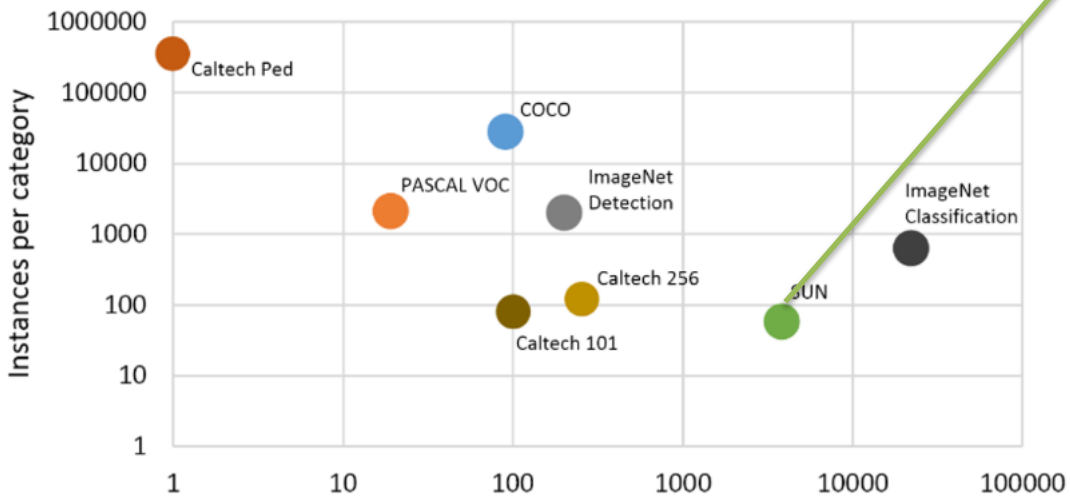
Brown University

UCIrvine
University of California, Irvine

Microsoft Research

Other CV datasets





http://groups.csail.mit.edu/vision/SUN/

Other CV datasets

Microsoft COCO
Common Objects in Context

cocodataset@outlook.com

Home People Explore Dataset

Overview Download Evaluate Leaderboard Challenges

Welcome to the MS COCO Captioning Challenge 2015!

The man at bat readies to swing at the pitch while the umpire looks on.

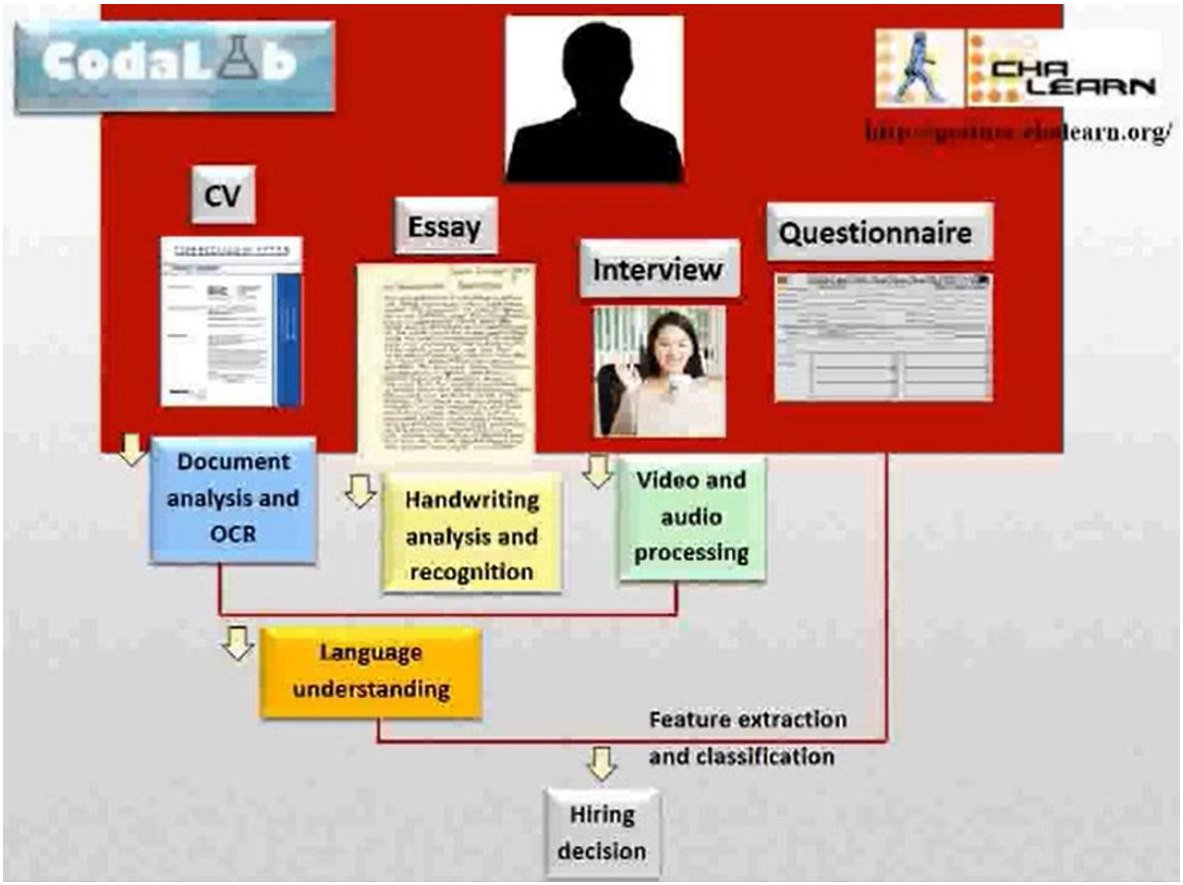
A large bus sitting next to a very tall building.

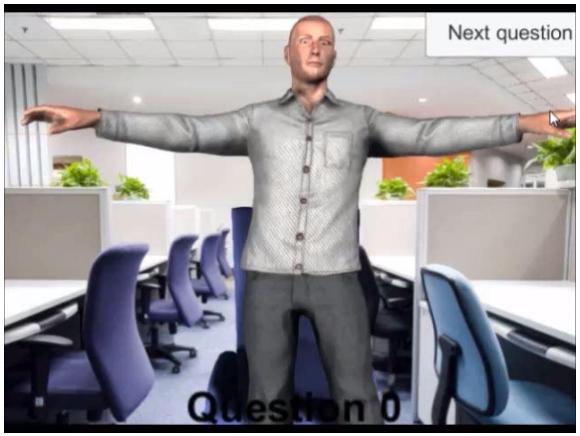
1. Introduction

The MS COCO Captioning Challenge is designed to spur the development of algorithms producing image captions that are informative and accurate. Teams will be competing against each other by training their algorithms on the MS COCO 2014 dataset and their results will be scored by **human judges**.

Other CV datasets

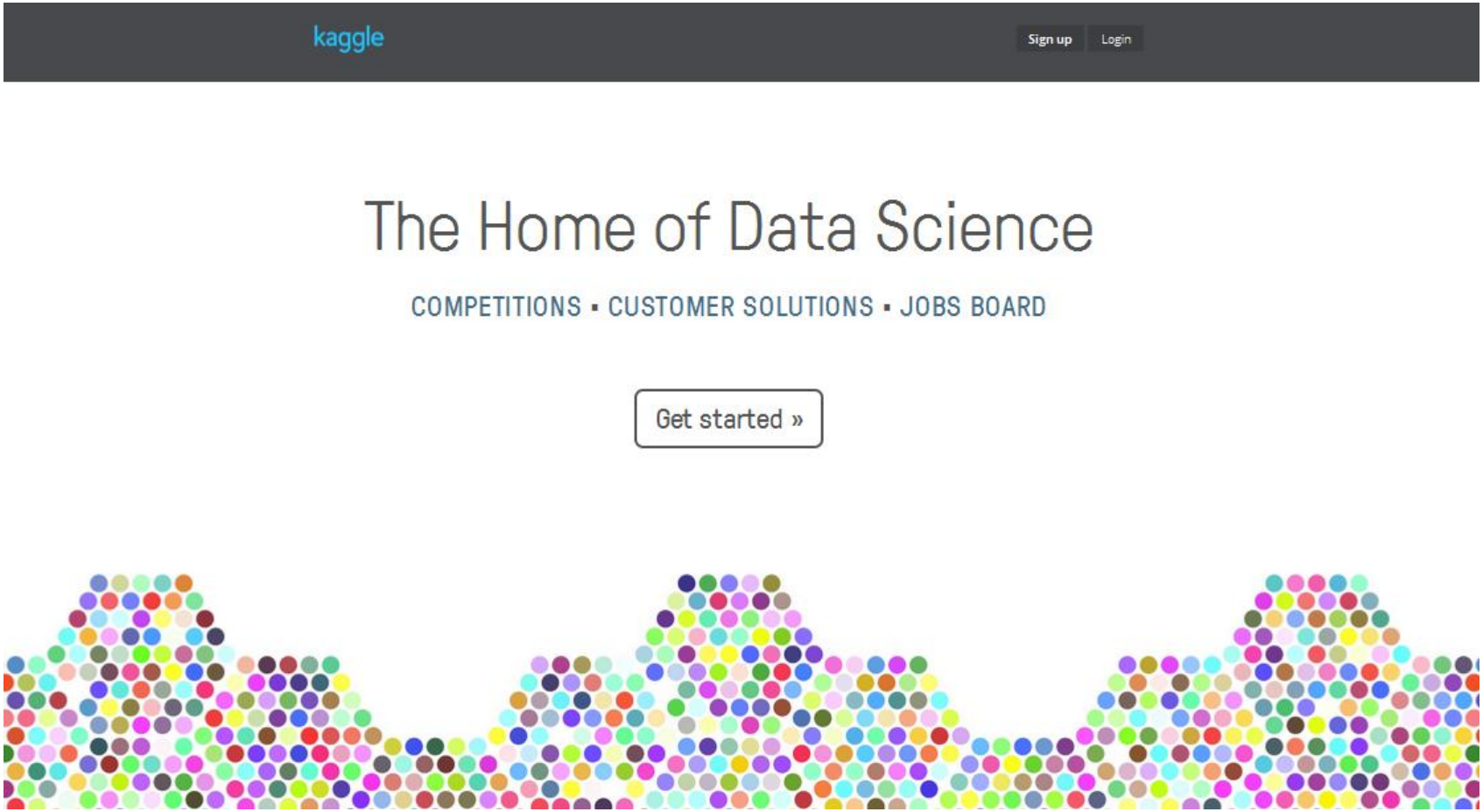
- CHALEARN SPEED INTERVIEWS WORLD CHALLENGE (2016?)





http://gesture.chalearn.org/speed-interviews

Platforms for CV competitions: KAGGLE



Platforms for CV competitions: CODALAB

CodaLab is an ecosystem for conducting computational research in a more efficient, reproducible, and collaborative manner.

Worksheets allow you to capture complex research pipelines in a reproducible way and create "executable papers". Use any data format or programming language — great for the power user!

» Explore worksheets

Competitions bring together the entire community to tackle the most challenging data and computational problems today. You can win prizes and also create your own competition.

» Explore competitions

Contribute your skills to help develop the CodaLab platform!

ChaLearn Looking at people news

ICCV 2015 COMPETITIONS AND WORKSHOP!! ALREADY STARTED!

WANT TO COLLABORATE IN CV CHALLENGE ORGANIZATIONS? MAIL US. There are many tasks to do.

ChaLearn LAP challenges and news:

http://gesture.chalearn.org/

Organization of ChaLean Looking at People requires:

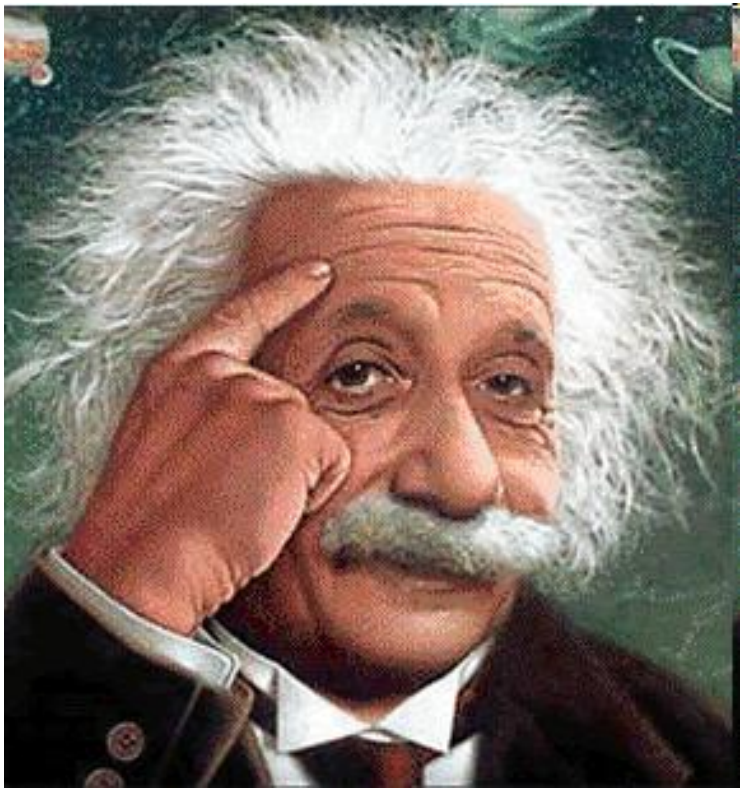
-Good ideas to solve real problems focused on humans
-Collecting data
-Labeling tools
-Dissemination and repositories
-Baseline designs based on state of the art approaches
-Online platform for the competition
-Sponsoring
-Presentation of the results in a relevant events
-Organization of special issues and challenge report documents, making competition data public for the scientific community

For each competition many organizers contribute. Our plan is to perform yearly challenges.

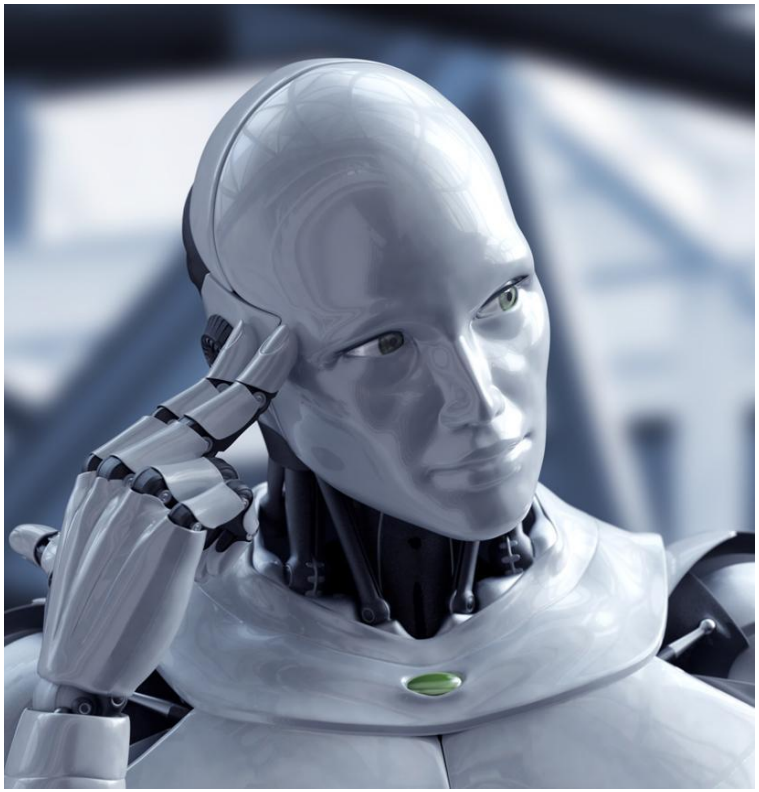
Feel free to contact us if you want to be included in our ChaLearn LAP mailing list or collaborate in some aspect propose ideas related to ChaLearn Looking at People competitions:

sergio@maia.ub.es

Thank you!



XX CENTURY



XXI CENTURY ?