Visual Sense

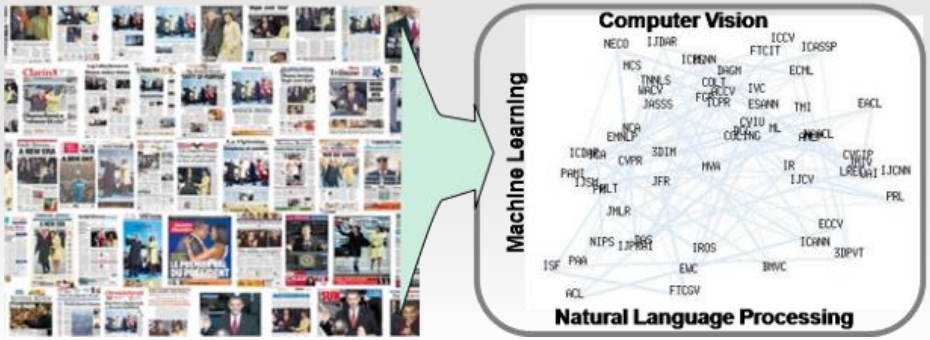
Tagging visual data with semantic descriptions

University of Surrey
Institut de Robòtica i Informàtica Industrial
Ecole Central de Lyon
University of Sheffield

Start: Jan 2013

Tagging visual data with semantic descriptions

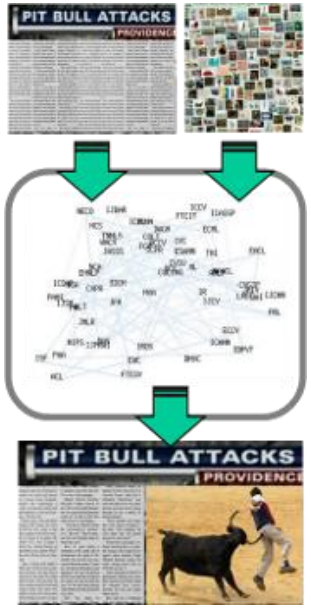
- Extract a semantic representation of visual content.
  - beyond the detection of objects and scenes

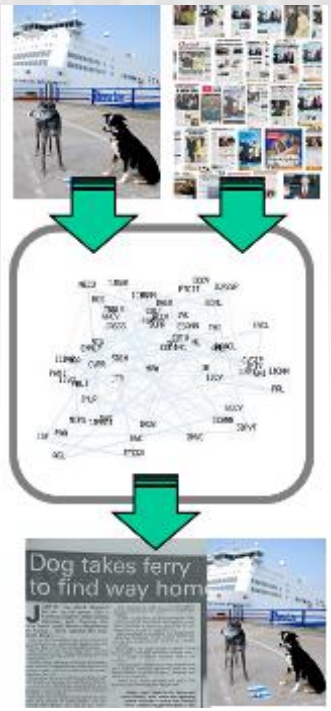


- Generate image captions using both
  - semantic representations of visual content
  - object/scene models derived from semantic representations of the multi-modal documents.
- Exploit available multi-modal data to discover mappings between visual and textual content.

Application scenarios

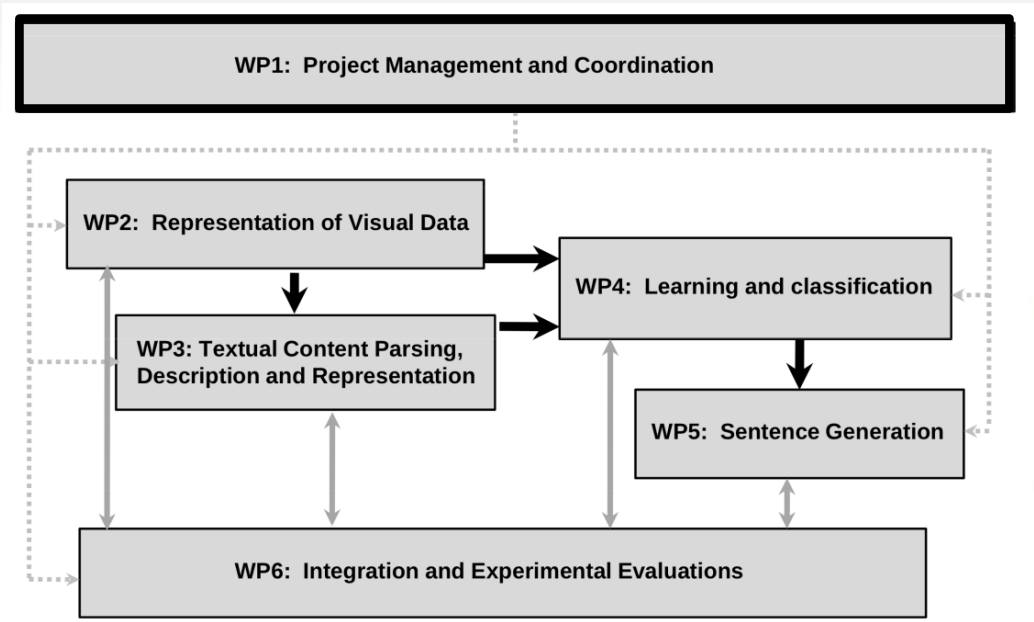
- Annotating visual documents with a rich semantic description of images
- Re-ranking the output of a baseline image search engine
- Finding suitable illustrations for text documents





ViSen research activities

- WP1:
  - Webpage: https://sites.google.com/site/visenproject/
  - Dissemination
    - 15 Publications in 2013
    - Collaboration with the BBC



ÉCOLE CENTRALE LYON

Institut de Robòtica i Informàtica Industrial

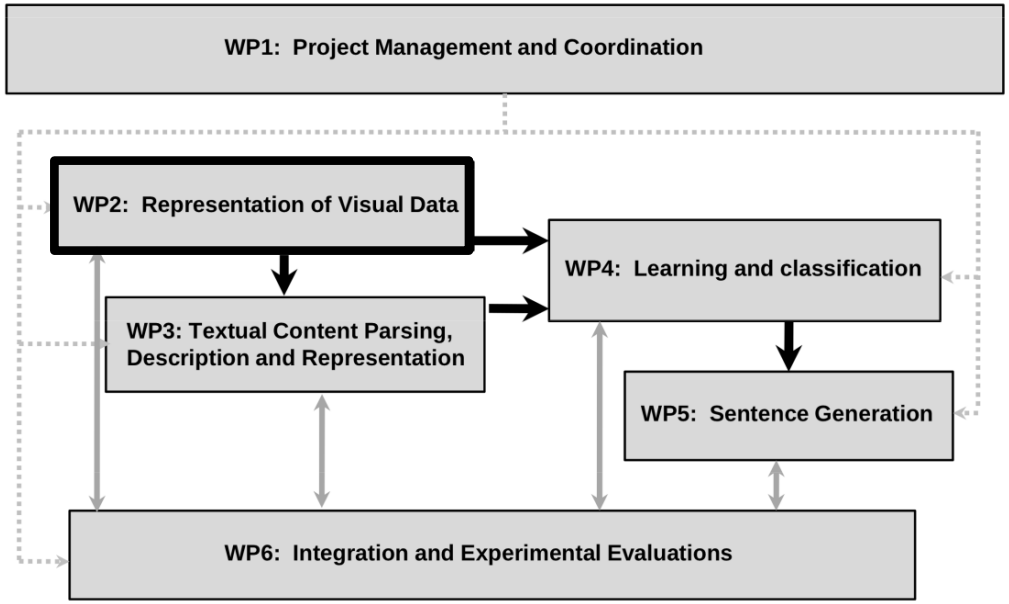
UNIVERSITY OF SURREY

The University Of Sheffield.

UNIVERSITY OF SURREY
Institut de Robòtica i Informàtica Industrial
ÉCOLE CENTRALE LYON
The University Of Sheffield.

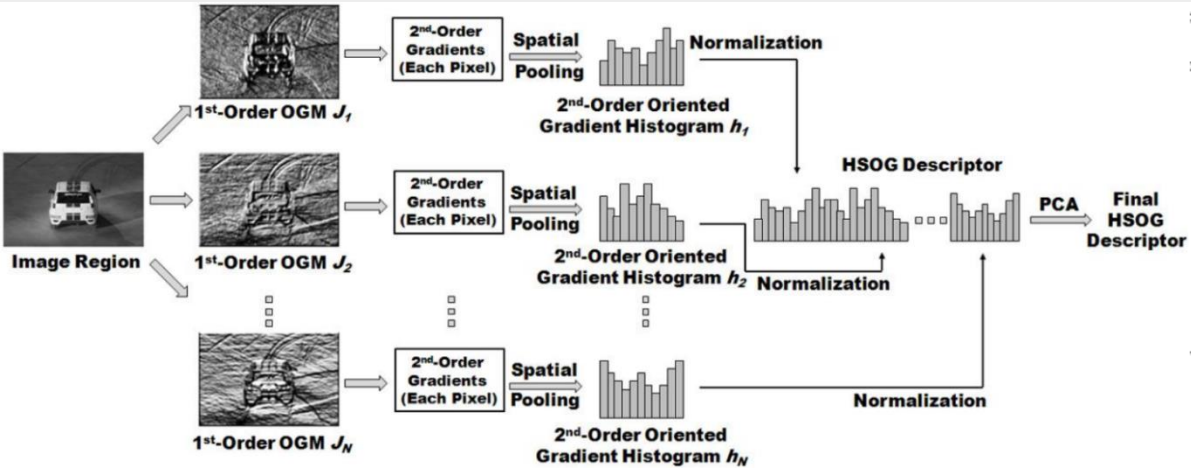
WP2: Representation of visual data

- Low-level features [ICPR2012, CPRAM2013]
  - Novel image descriptors [ICMR2013]
- Mid-level features [CVIU2013]
- Object detection and segmentation



WP2: Low-level image representation

- HSOG: Histogram of Second Order Gradients

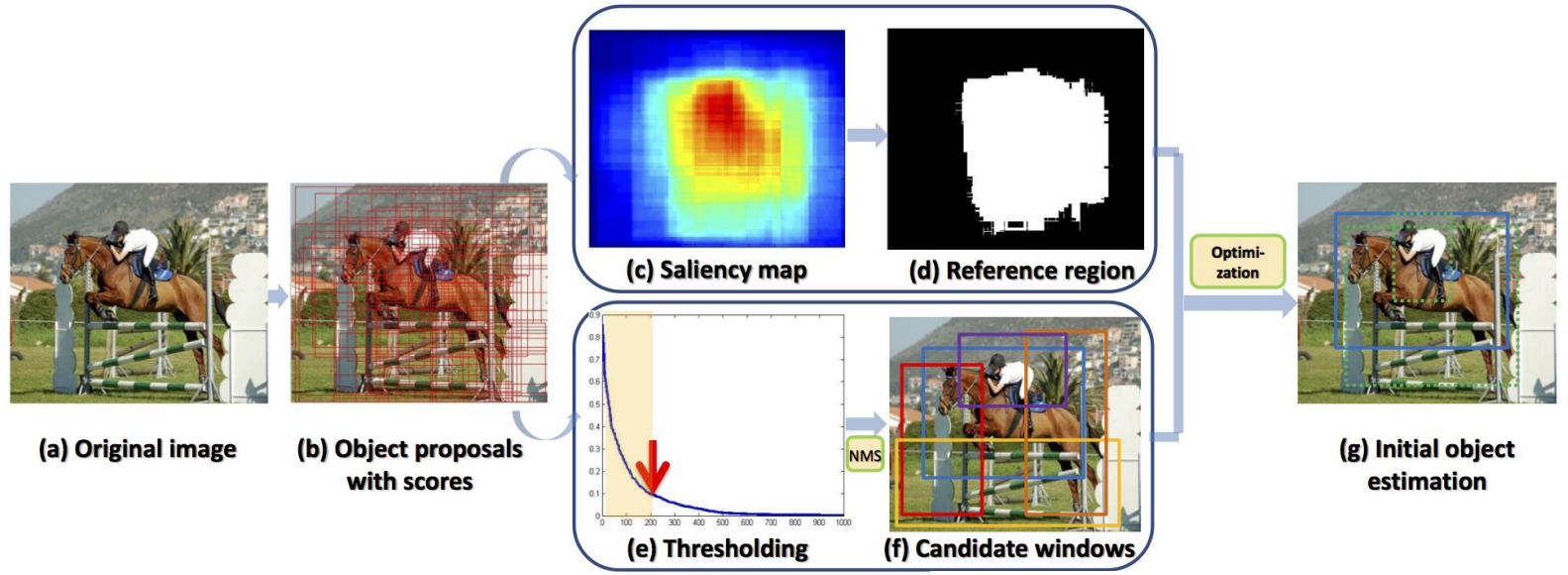


D.Huang, et al. ICMR 2013

Caltech 256

Descriptor	Results (%)
SIFT	20.68
HOG	19.55
DAISY	21.61
CS-LBP	20.39
T3i-S4-13	20.28
HSOG (Ss)	**22.36**
HSOG (Ms)	**27.92** /25.52
HSOG (Ss) + SIFT	**30.45** /28.21
HSOG (Ss) + HOG	28.70 /26.39
HSOG (Ss) + DAISY	28.83 /26.61
HSOG (Ss) + CS-LBP	29.67 /27.55
HSOG (Ms) + SIFT	**32.47** /29.68
HSOG (Ms) + HOG	30.52 /28.05
HSOG (Ms) + DAISY	30.33 /27.83
HSOG (Ms) + CS-LBP	32.05 /29.33

WP2: **Improving DPM with Objectness for Object Detection**



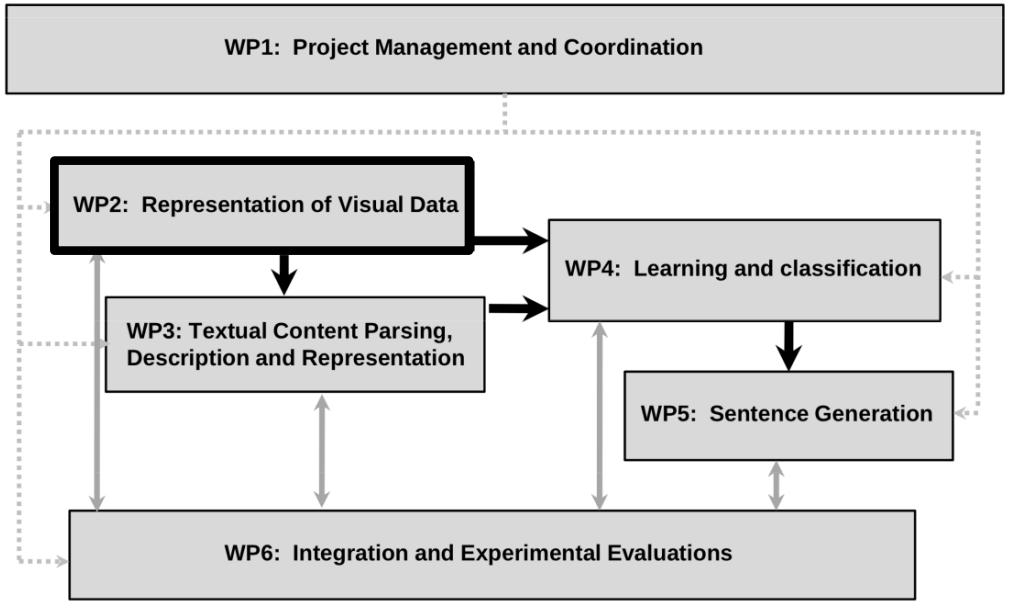
Average detection (%)	*VOC07-14*					*VOC07-6×2*				
	no post-processing		with post-processing			no post-processing		with post-processing		
	[5]	ours	[5]	ours-[5]	ours-ES	[5]	ours	[5]	ours-[5]	ours-ES
Initialization	19.98	**21.73**	23.00	24.20	**25.12**	37.22	**38.74**	44.62	47.85	**48.59**
Refinement 1	25.11	27.46	26.38	28.21	28.94	51.63	55.85	53.11	56.78	58.02
Refinement 2	27.69	28.95	29.39	**32.87**	32.82	56.99	**59.82**	59.31	63.31	**63.91**
Refinement 3	**28.98**	—	30.31	—	—	59.32	—	61.05	—	—
Result from [4]	26.00					50.00				

[4] T. Deselaers, B. Alexe, and V. Ferrari, "Localizing objects …", ECCV, 2010

[5] M. Pandey and S. Lazebnik, "Scene recognition and weakly …", ICCV, 2011

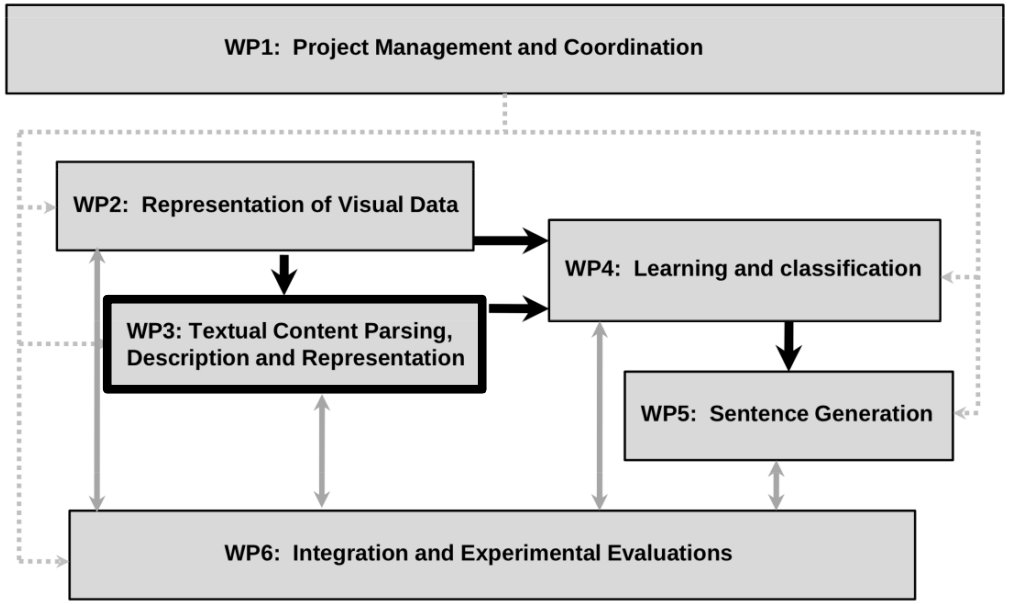
WP2: Workplan

- Investigate deep learning for improved object detection
- Conceptual graphs for representing spatial relationships of objects in images



WP3: Textual Content Parsing, Description and Representation

- Shallow linguistic processing
- Deep linguistic processing
- Extraction of <predicate, agent, patient, locative>



WP3: Representing Sentential Image Annotations : Beyond Bag of Words

What is relevant in a sentential image description?

- We need to extract information about the events described in the annotation:
  - What is the action or event in the image? (i.e. predicate).
  - What are the actors and objects involved in the event? (i.e. agents and patients of the action).
  - Where is the event happening? (i.e. locative).

A man tries to open a small door next to a large window.

Semantic Tuple
Predicate: open
Agent: man
Patient: door
Locative: window [next to]

A man is opening a door in front of a large window which has white curtains.

Semantic Tuple
Predicate: open
Agent: man
Patient: door
Locative: window [in front of]

WP3: The Challenges of Semantic Tuple Extraction

Multiple tuples per annotation

- A little boy is standing on the street while a man in overalls is working on a stone wall.

- Seven climbers are ascending a rock face whilst another man stands holding the rope.

Multiple prepositional phrases (hard to find the correct locative):

- A little girl [in a pink dress] climbs a rope bridge [at the park].
Locative (of climb): park [at]

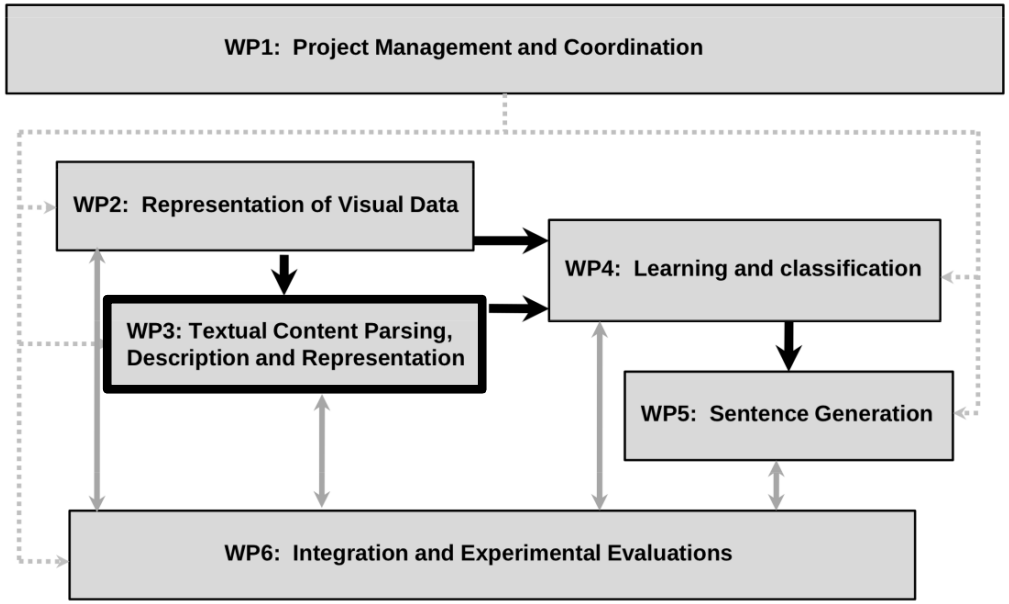
- A man [in a hat] is displaying pictures [next to a skier] [in a blue hat].
Locative (of displaying): skier [next to]

Co-reference for arguments

- A [dog] shakes its head near the shore, a red ball is next to [it].
Locative (of is): dog [next to]

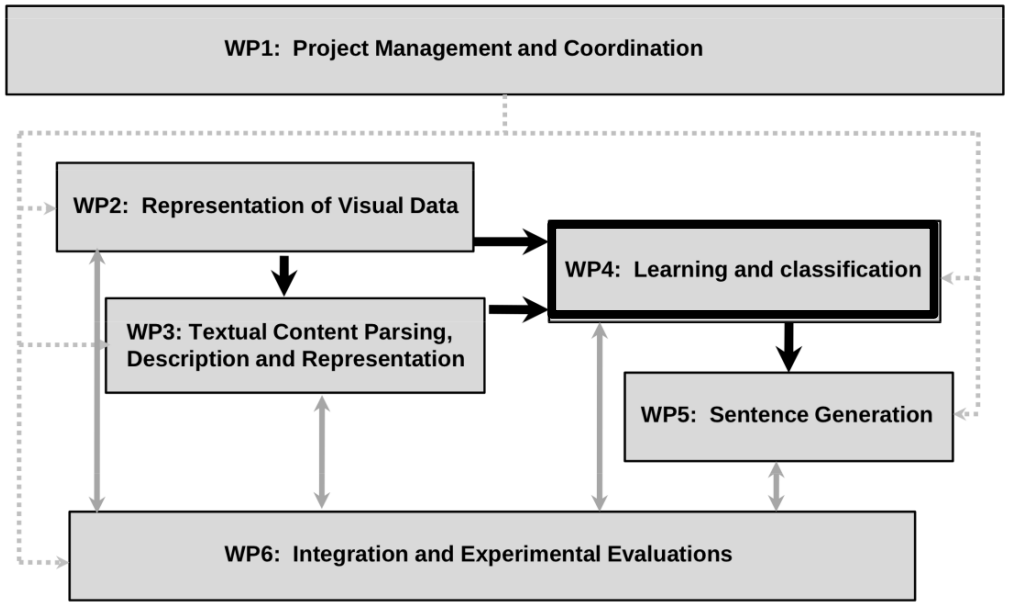
WP3: Workplan

- Create ground-truth annotation of sentences paired with semantic tuples for learning and evaluation
- Handle synonyms, e.g. child and boy refer to the same world entity.
- Develop stronger text kernels
  - To be combined with visual kernels



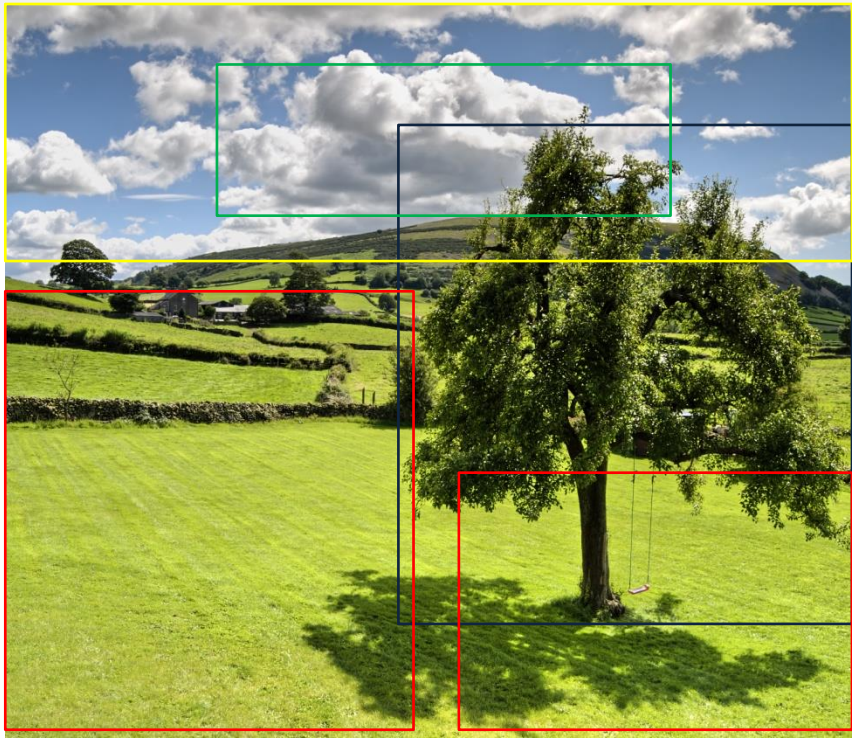
WP4: Visual Learning

- Large scale object detection and image classification
- Kernel based approach to combine image and text representation



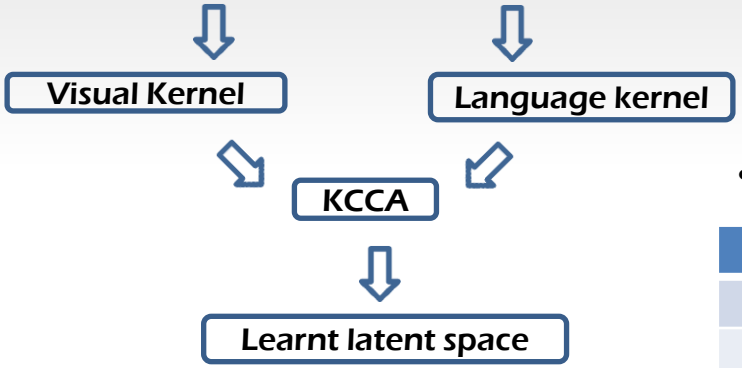
WP4: Large scale classification

- Scale is the challenge:
  - 1372 classifiers, 572 detectors, 25 attribute classifiers
  - 1.5M+ images in ImageNet for training
  - Efficient techniques investigated
  - Processing heavily parallelised
  - Both learning and testing can still take weeks
- Results of several versions of visual learning pipelines



WP4: Kernel based approach

- The skiers are in front of the lodge.
- A brown dog is running on a rock.
- A group of people are crossing a river.

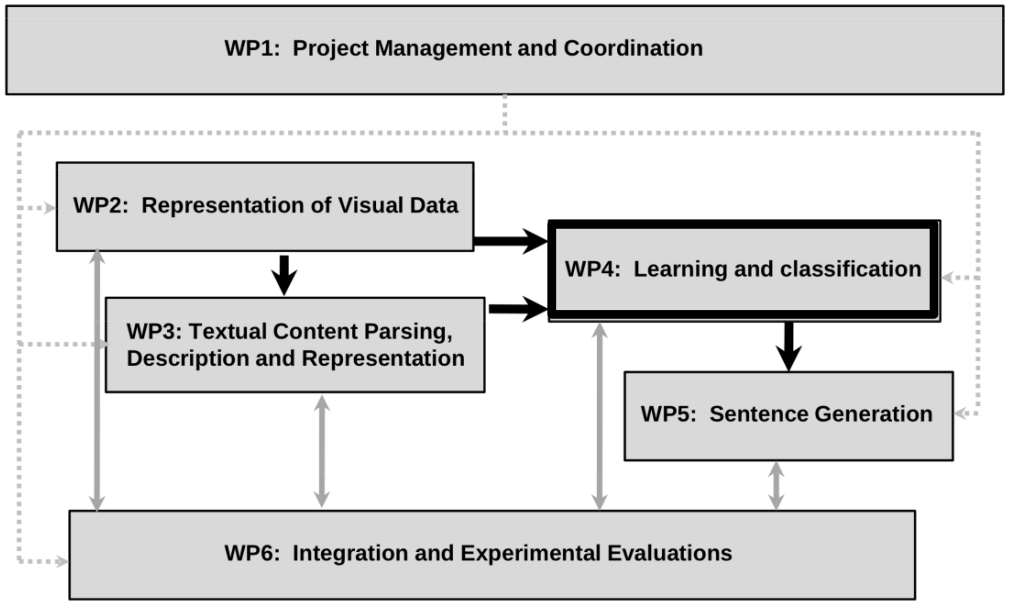


- Mid-level visual kernel better than low-level

Median rank on test set					
Hodosh2013JAIR			ours		
V	L		V	L	
SPM	BoW1	64	Mid-level	BoW1	36
SPM	BoW5	58	Mid-level	BoW5	32
SPM	TagRank	56			
SPM	Tri5	53			
SPM	Tri5Sem	34	Mid-level	Tri5Sem	?

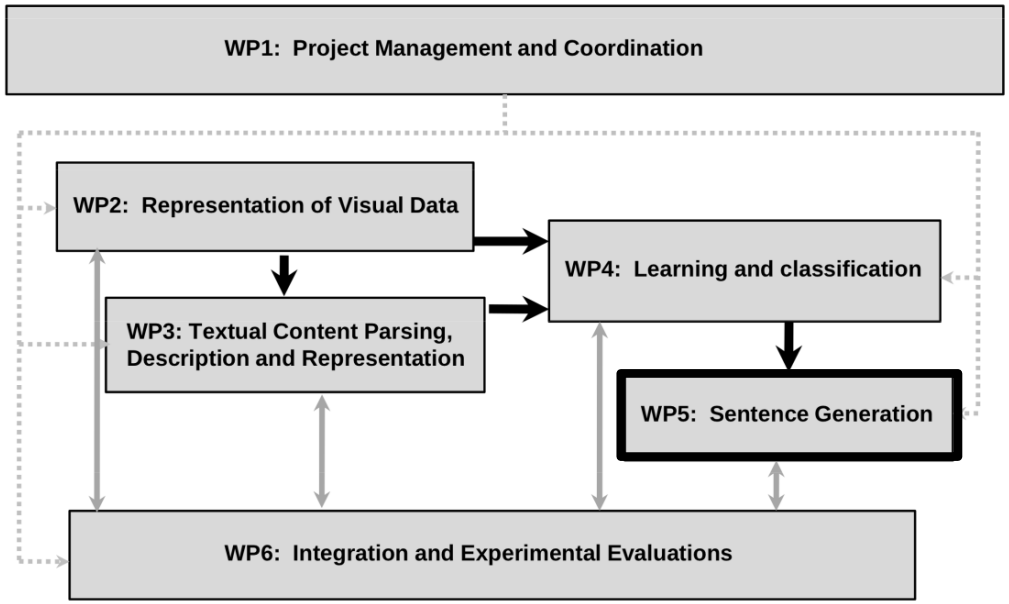
WP4: Workplan

- High-level features
  - Vision: combining classification and detection, deep learning
  - NLP: deep linguistic analysis
- Machine learning
  - Parameterise kernels, learn embedding and projection jointly
  - Other multi-view learning methods, e.g. co-training



WP5: Sentence Generation

- Identifying and Gathering Corpora for Object/Scene Modelling
- Deriving Object/Scene models from corpora



WP5: Gathering Prior Knowledge from Corpora

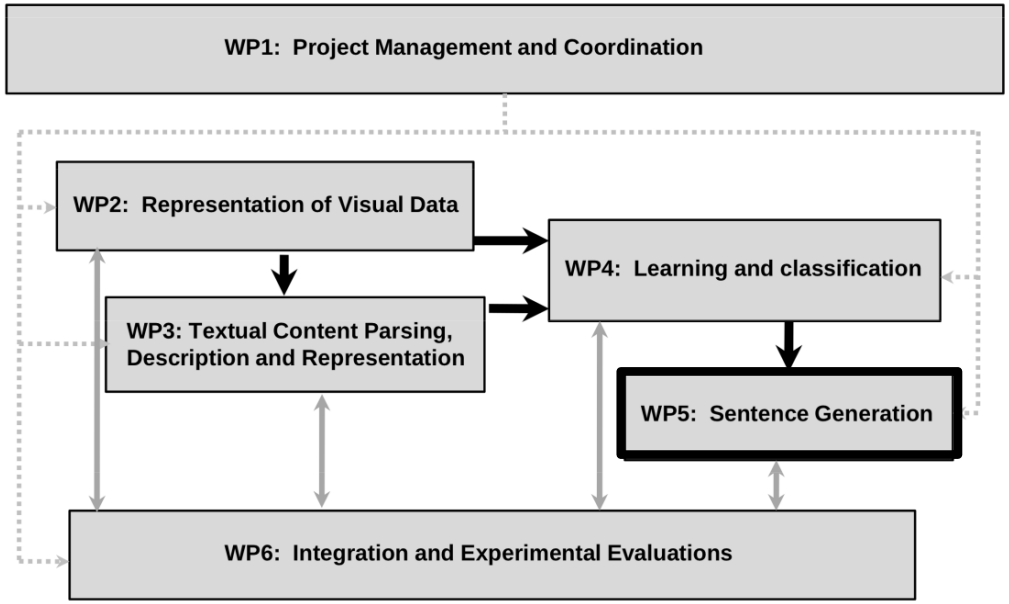
- Identifying and Gathering Corpora for Object/Scene Modelling
  - Gathered prior knowledge from textual sources
    - reduce noise from visual systems
  - Identified **objects**, **subjects** and **attributes** to be used by all partners
    - From image descriptions

      `The black dog is running through the snow.`
  - Developed tools for data gathering
    - Gathering web-documents related to objects using ChatNoir search engine
    - Gathering image captions from Flickr

- Deriving Object/Scene Models from Corpora
  - Derive inference models to improve vision outputs:
    - Inferring object-object relations
    - Inferring object-attribute relations
    - Inferring object-scene relations

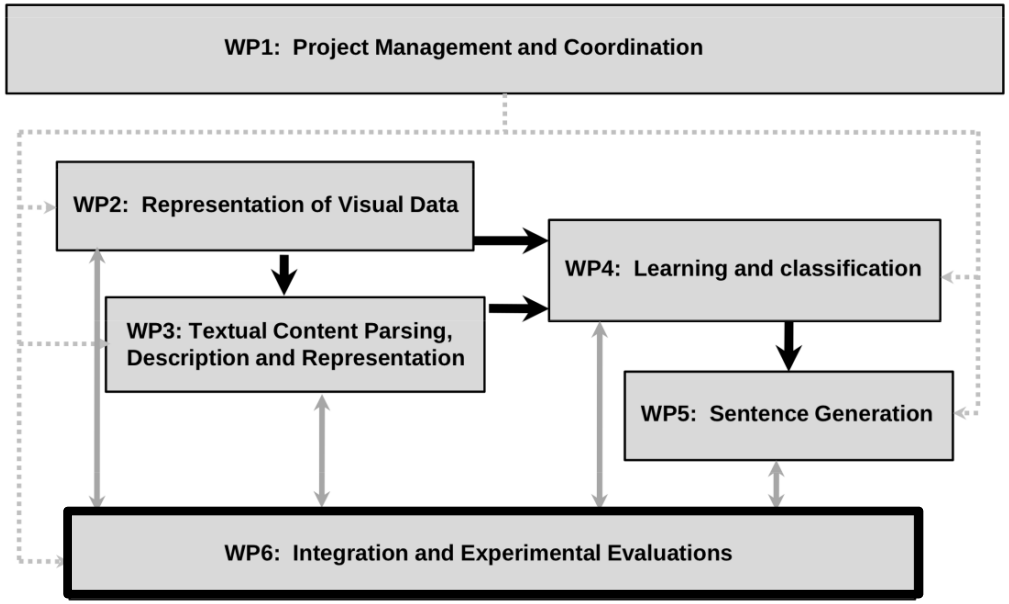
WP5: Workplan

- Continue gathering corpora for object/scene modelling
- Continue investigating entity or object type models derived from corpora and other textual sources
- Sentence generation from simulated and real data



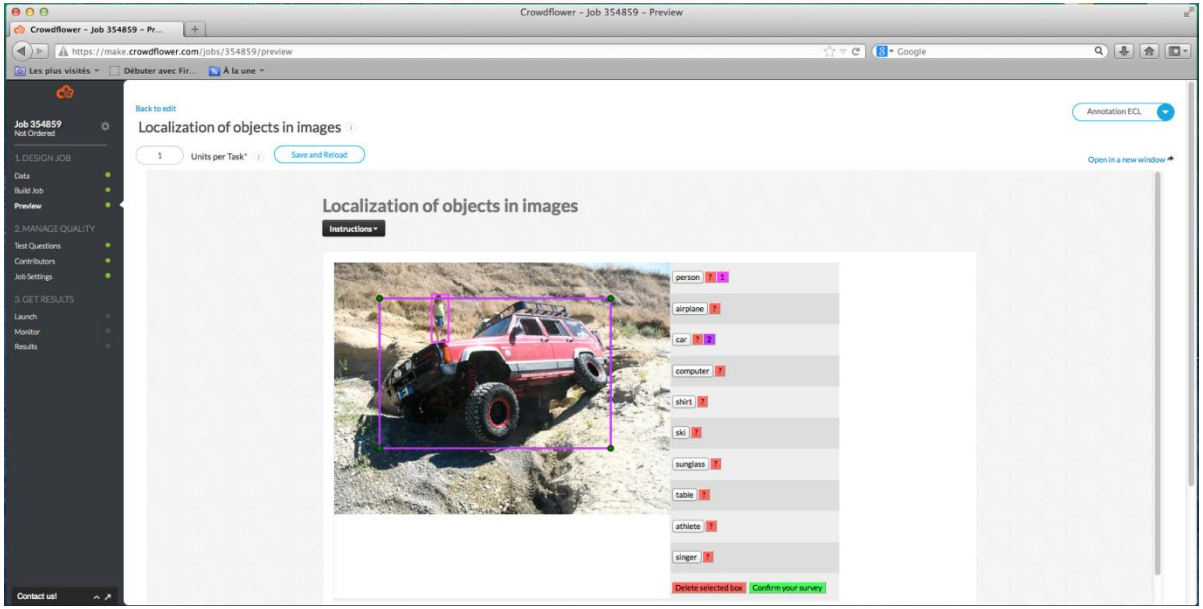
WP6: Integration and Experimental Evaluations

- Definition of common evaluation frameworks
  - Evaluation of visual classification systems
  - Existing datasets not suitable for testing both vision and text
  - Existing evaluation measures for sentence generation are not established or very limited
- Design of the architecture and Integration of software tools
  - Definition of building blocks with IO formats.



WP6: Evaluation benchmark

- Adopting 8k Flickr data for our benchmark
  - 8k images (common outdoor & indoor activities)
- Enriching annotation of the evaluation dataset
- Using Crowdflower environment to crowdsource the annotations



UNIVERSITY OF SURREY

Institut de Robòtica i Informàtica Industrial

ÉCOLE CENTRALE LYON

The University Of Sheffield.

WP6: Evaluating Image Descriptions

A person hits a ball with a tennis racket.
A person swings at a tennis ball.
A tennis player wearing a green shirt about to hit a ball with his racquet.
A woman in a green shirt and blue hat is playing tennis.
Miami tennis player hits the ball with a forehand.

WP6: Evaluating Image Descriptions

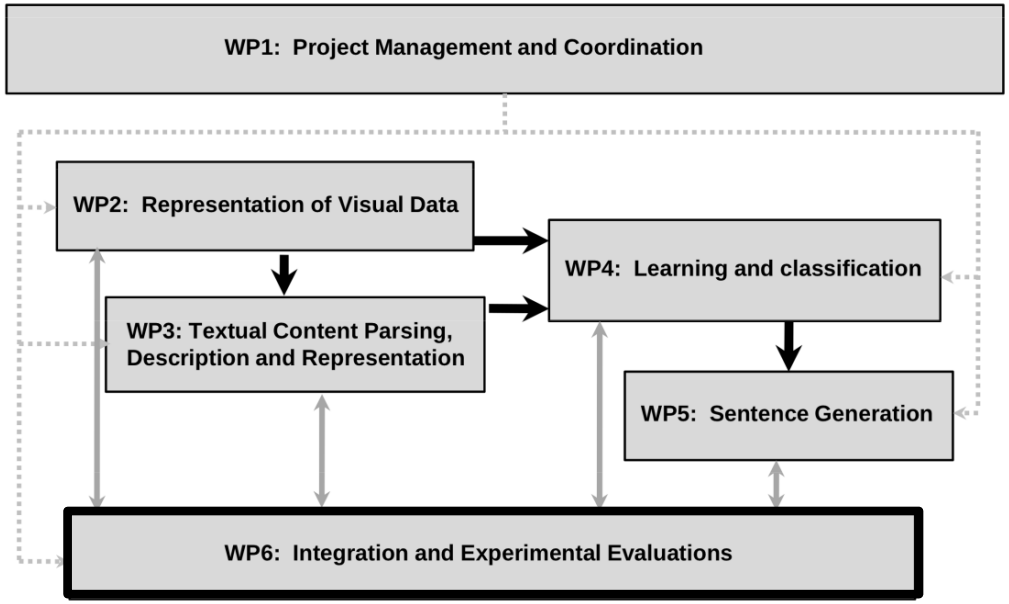
- Investigated methods for evaluating image descriptions against a reference set of human-authored descriptions
  - Important for assessing the results of automatic caption generation
- Challenges:
  - Variations in *what* is described
  - Variations in *how* the same things are described
- Existing methods (rouge, bleu) don't work well on short texts

WP6: Evaluating Visual Classifiers

- Large number of categories to detect (>1k)
- Do visual systems agree with how humans describe an image?
  - Gives idea of how difficult it will be to generate image descriptions
- Results:
  - For 80% of images, at least one category in the top-40 visual classifier output is correct

WP6: Workplan

- Completing annotation with the crowd sourcing task
- Continue investigating methods for evaluating image descriptions for sentence generation
- Investigate methods of evaluating text kernels independent of vision kernels
  - And whether better text kernels necessarily lead to better image annotation/retrieval



Summary

- Intensive collaboration between all partners
  - Exchange of expertise
  - Data processing/generation services
- Good and timely progress towards the objectives
- Main challenges so far
  - Establishing common test data with evaluation measures for all application scenarios
  - Bridging the gap between state-of-the-art vision output and input required for language generation
- Inter-project collaborations
  - Visen, Mucke, Camomile

Visual Sense

Tagging visual data with semantic descriptions

University of Surrey
Institut de Robòtica i Informàtica Industrial
Ecole Central de Lyon
University of Sheffield

Start: Jan 2013