

Human action recognition in still images via text analysis

Dieu-Thu Le

Email: dieuthu.le@unitn.it
Trento University

SEMINARS in SATO Laboratory
July 24, 2012

Outline

- 1 Introduction
- 2 Related work
- 3 Our system
- 4 Conclusion

University of Trento

- An Italian university located in Trento and Rovereto, achieve considerable results in didactics, research and international relations
- In 2009, it ranked first in the Italian national ranking (quality of the research and teaching activities, the success in attracting funds)(*)



Action recognition in still images

- Most action recognition systems are in the scope of analyzing video sequences
- However, many actions can be recognized from single images
- Studies have mainly focused on person-centric action recognition



Riding a bike



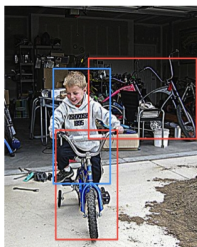
Riding a horse



Feeding a horse

How to recognize actions in images?

- Based on objects recognized in images
- Based on human poses [Lubomir Bourdev, Jitendra Malik, 2009]
- Based on scene background/type [Gupta et al. 2009]
- Based on clothing, camera viewpoint, and so on.



bike, person



horse, person



dog, person

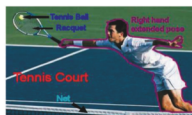
How to recognize actions in images?

- Based on objects recognized in images
- Based on human poses [Lubomir Bourdev, Jitendra Malik, 2009]
- Based on scene background/type [Gupta et al. 2009]
- Based on clothing, camera viewpoint, and so on.



How to recognize actions in images?

- Based on objects recognized in images
- Based on human poses [Lubomir Bourdev, Jitendra Malik, 2009]
- Based on scene background/type [Gupta et al. 2009]
- Based on clothing, camera viewpoint, and so on.



Tennis court vs. Baseball ground [Sport action dataset, Gupta et al., 2009]



Beach vs. Forrest

How to recognize actions in images?

- Based on objects recognized in images
- Based on human poses [Lubomir Bourdev, Jitendra Malik, 2009]
- Based on scene background/type [Gupta et al. 2009]
- Based on clothing, camera viewpoint, and so on.



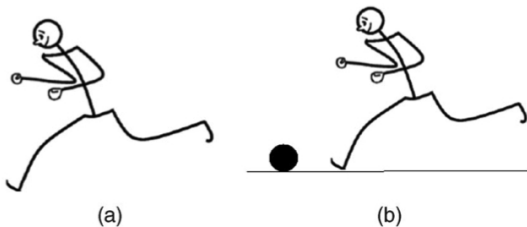
Wedding ceremony



Baseball player

Challenge: Interaction between human-object

[Gupta et al. 2009]

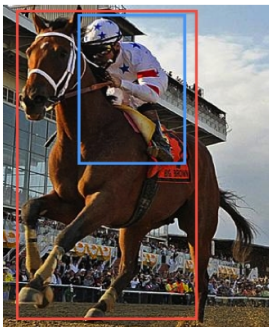


(a) running

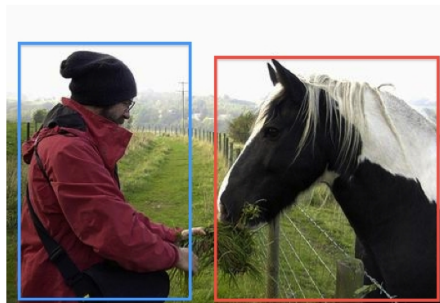
(b) kicking a ball

Similar human pose, different contexts (objects)

Challenge: Interaction between human-object



Riding a horse

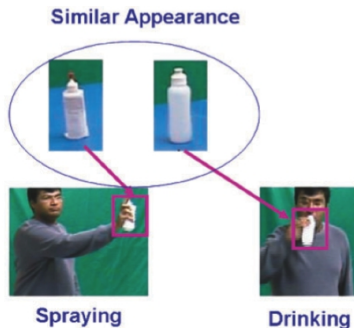


Feeding a horse

Same objects, different actions

Challenge: Interaction between human-object

[Gupta et al. 2009]

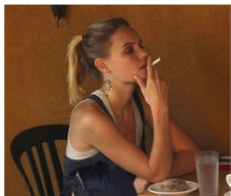


Contextual cues from human interactions aids in
object recognition

Similar shaped objects, different human poses

Challenges

- We cannot base solely on human and objects but the interaction between them
- Further information (such as human pose, scene background) is necessary to disambiguate actions in many cases
- False object recognition and inaccurate pose estimation can cause wrong action detection: background clutter, occlusions, similar shaped objects, etc.



Smoking cigarette
(cigarette is hard to recognize)

Action recognition in still images

- Gupta et al., 2009: sport action recognition using spatial and functional constraints for recognition
- B.Yao & Li Fei Fei, 2010: people playing musical instrument, image feature representation “grouplet”
- V.Delaitre, 2010: seven everyday action recognition, using bag-of-feature representation
- B.Yao et al., 2011: 40 action recognition, using “parts” and “attributes”

Action Dataset [B.Yao et al., 2011]

Dataset	No. of actions	No. of images	Clutter?	Poses vary?	Visibility varies?
Ikizler [11]	5	1727	Yes	Yes	Yes
Gupta [10]	6	300	Small	Small	No
PPMI [26]	24	4800	Yes	Yes	No
PASCAL [6]	9	1221	Yes	Yes	Yes
Stanford 40	40	9532	Yes	Yes	Yes

Problem statement

- These systems have mainly focused on extracting visual features from images, with the requirement of annotated dataset
- The actions recognized are limited to a small predefined set
- Object recognition systems on the other hand have been able to recognize more objects

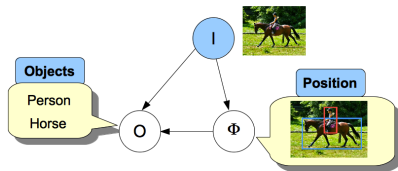
Our approach

- Based on objects recognized in images
- Take advantage of available textual datasets
- Automatically suggest the most/least plausible actions
- Does not require action annotated dataset
- Flexible, easy to extend

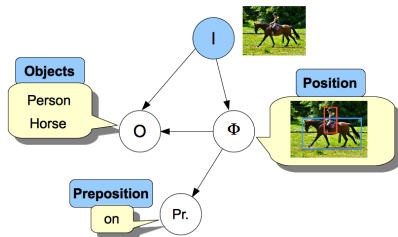
Action recognition in still images: A probabilistic model



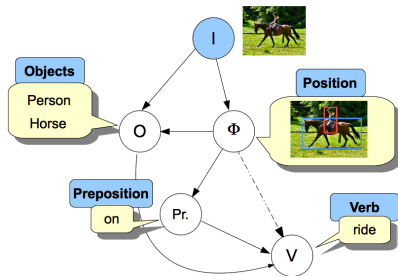
Action recognition in still images: A probabilistic model



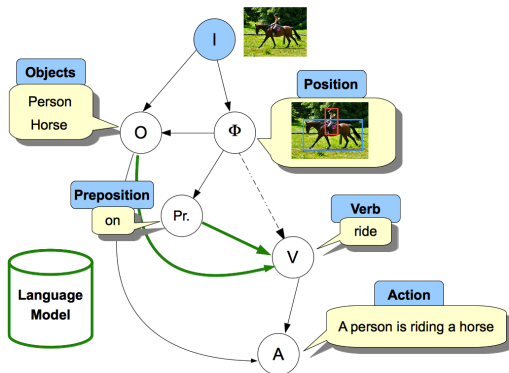
Action recognition in still images: A probabilistic model



Action recognition in still images: A probabilistic model



Action recognition in still images: A probabilistic model

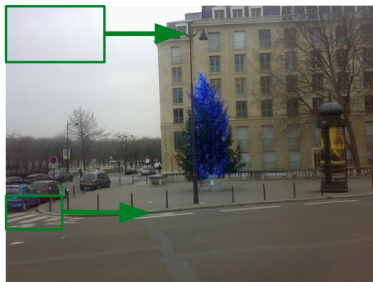


$$P(A|I) = P(O|I) \times P(\phi|I) \times P(\text{Pr.}|\phi) \times P(V|\text{Pr.}, O) \quad (1)$$

Object recognizer: The most telling window

Problem: There are many possible locations to search

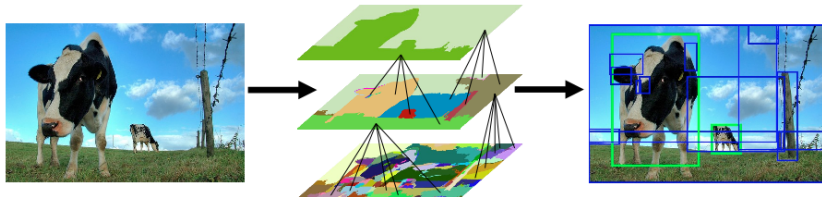
- Standard method is an exhaustive search, visiting all possible locations on a regular grid
- MST introduces Selective Search



Object recognizer: The most telling window

Problem: There are many possible locations to search

- Standard method is an exhaustive search, visiting all possible locations on a regular grid
- MST introduces Selective Search



Segmentation as Selective Search for Object Recognition, ICCV 2011,
K.E.A. van de Sande, J.R.R. Uijlings, T. Gevers, and A.W.M. Smeulders

How to learn from general textual corpora?

- We aim to discover the interaction between objects in images by exploiting general knowledge learning from textual corpora
- This problem is closely related to verbs' *selectional preferences*¹: the semantic preferences of verbs on their arguments (e.g., the verb “drink” prefers subjects that denotes human or animals, objects such as “water”, “milk”, etc.)
- We employ two different ways to extract this information:
 - Distributional semantic models
 - Topic models

¹alternative terms: selectional rules, selectional restrictions, sortal (in)correctness

Distributional Memory [Baroni & Lenci, 2010]²

- a state-of-the-art multi-purpose framework for semantic modeling
- extracts distributional information in the form of a set of weighted <word-link-word> tuples
- tuples are extracted from a dependency parse of a corpus

²<http://clic.cimec.unitn.it/dm/>

Distributional Memory [Baroni & Lenci, 2010]: TypeDM

- Training corpus: the concatenation of ukWaC corpus, English Wikipedia, British National corpus (≈ 2.8 billion tokens)
- contains 25,336 direct and inverse links that correspond to the patterns in the LexDM links, 130M tuples
- the top 20K most frequent nouns, 5K verb and 5K adjectives are selected

DM for action recognition in still images: Our experiment

- Test on the Stanford 40 action dataset
- We try the system over those 6 verbs shared by the PASCAL object and STANFORD 40 action data sets (riding, rowing, walking, watching, repairing, feeding)
- These verbs gave rise to 8 actions: Riding+horse, Rowing+boat, Riding+bike, Walking+dog, Watching+TV, Feeding+horse, Repairing+car, Repairing+bike

DM for action recognition in still images: Our experiment

Object recognizer:

- Training set: PASCAL object competition (20 objects)
- Testing set: Stanford 40 action testing data set (5,532 images)
- Evaluation: mAP, single average precision evaluated against all images in the test set:
 - 1 horse 54%
 - 2 TV: 33%
 - 3 Car: 14%
 - 4 Dog: 8%
 - 5 Bike: 54%
 - 6 Boat: 14%

DM for action recognition in still images: Our experiment

Action ranked list based on objects

Ride-v	obj-1	bike-n	3433
Ride-v	obj-1	bicycle-n	1378
Ride-v	on-1	bike-n	762
Ride-v	on-1	bicycle-n	278
Fix-v	obj-1	bike-n	182
Ride-v	sbj_intr-1	bike-n	155
Fix-v	obj-1	bicycle-n	23
Fix-v	sbj_intr-1	bike-n	19
Row-v	obj-1	bicycle-n	10
Row-v	obj-1	boat-n	520
Row-v	in-1	boat-n	224
Row-v	sbj_intr-1	boat-n	192
Watch-v	obj-1	boat-n	175
Watch-v	from-1	boat-n	59
Row-v	as-1	boat-n	39
Fix-v	obj-1	boat-n	35
Row-v	on-1	boat-n	22
Fix-v	on-1	boat-n	19
Row-v	for-1	boat-n	15
Row-v	sbj_tr-1	boat-n	15
Row-v	with-1	boat-n	15
Feed-v	around-1	boat-n	15
Feed-v	from-1	boat-n	15
Row-v	by-1	boat-n	12

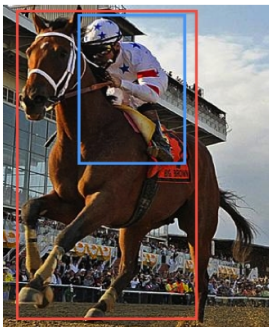
Ride-v	in-1	car-n	445
Fix-v	obj-1	car-n	305
Ride-v	obj-1	car-n	152
Watch-v	obj-1	car-n	142
Watch-v	from-1	car-n	72
Fix-v	in-1	car-n	31
Fix-v	sbj_intr-1	car-n	21
Fix-v	on-1	car-n	18
Ride-v	obj-1	horse-n	2465
Ride-v	on-1	horse-n	692
Feed-v	obj-1	horse-n	323
Ride-v	sbj_intr-1	horse-n	188
Watch-v	obj-1	horse-n	129
Feed-v	sbj_intr-1	horse-n	34
Feed-v	for-1	horse-n	12
Feed-v	sbj_tr-1	horse-n	12
Feed-v	obj-1	person-n	82
Watch-v	obj-1	person-n	66
Watch-v	obj-1	TV-n	1766
Watch-v	on-1	TV-n	1232
Watch-v	via-1	TV-n	56
Watch-v	sbj_intr-1	TV-n	52
Fix-v	obj-1	TV-n	42
Feed-v	obj-1	TV-n	32
Feed-v	into-1	TV-n	16
Feed-v	on-1	TV-n	15

DM for action recognition in still images: Our experiment

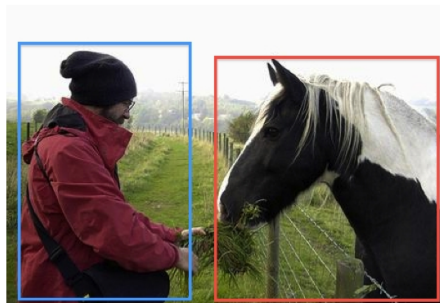
In many cases, objects themselves cannot decide which actions are correct

Object	Gold-Standard	DM
arrow	shoot	shoot
bike	fix/ride	ride
board	write	take
boat	row	take
book	write	read
bubble	blow	blow
car	fix/ride	use
cart	push	push
computer	use	use
dish	wash	cook
dog	walk	walk
floor	clean	clean
guitar	play	play
hand	wave	hold
horse	feed/ride	ride
liquid	pour	pour
message	text	read
microscope	look	use
photo	take	take
telescope	look	use
tooth	brush	brush
tree	cut	cut
TV	watch	watch
vegetable	cut	cook
violin	play	play

Person & Horse: “riding” or “feeding”?



Riding a horse



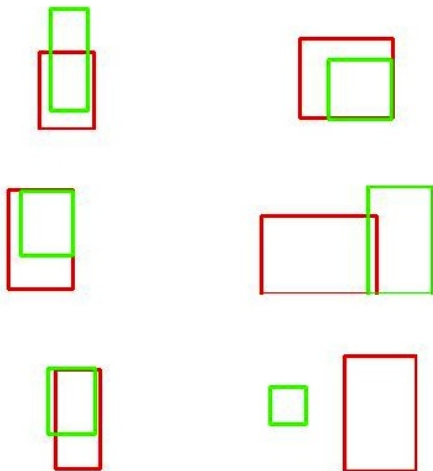
Feeding a horse

Same objects, different actions

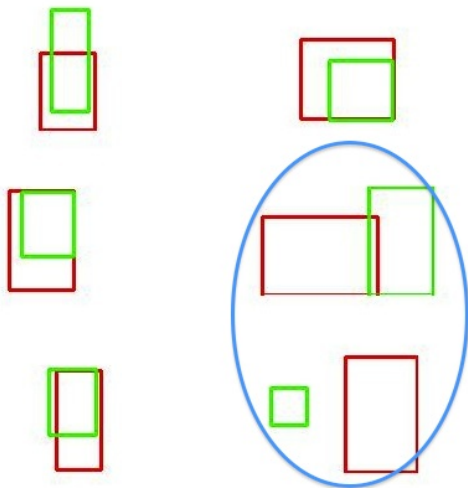
How to disambiguate actions in an image given its objects

- Human pose
- Object localization
 - Example:
 - Riding a horse: a person is on the top of the horse
 - Feeding a horse: a person is usually on the same level with a horse
 - Using preposition (i.e., link in the DM) to map with the localization of objects recognized in the images to automatically define the relative position between two objects (e.g., human - horse)

Experiment: Riding horse or feeding horse?



Experiment: Riding horse or feeding horse?



Experiment: Riding horse or feeding horse?



Relative position between person and other objects

Position between object and person vs. their possible preposition extracted from the distributional semantic model

Action	Upper	Below	Inside	Outside	Preposition from DM
ride a bike	172	12	10	190	on
fix a bike	67	53	13	128	-
ride a horse	184	6	33	196	on
feed a horse	50	124	21	183	for
row a boat	51	31	13	79	in, as, for, with, by
fix a car	115	128	22	275	in, on
walk a dog	118	49	5	187	with
watch TV	10	101	0	97	on, via

Disambiguating actions based on relative positions

Position between object and person vs. their possible preposition extracted from the distributional semantic model

Action	Upper	Below	Inside	Outside	Preposition from DM
ride a bike	172	17% 12	10	190	→ on
fix a bike	67	53	13	128	-
ride a horse	184	5% 6	33	196	→ on
feed a horse	50	124	21	183	for
row a boat	51	31	13	79	in, as, for, with, by
fix a car	115	128	22	275	in, on
walk a dog	118	49	5	187	with
watch TV	10	101	0	97	on, via

Disambiguating actions based on relative positions

- Based on Allens interval algebra
- Building a position-based SVM action classifier: use the coordinate of the center of each bounding box, height and weight ratio as features for the action classifier

Results:

Bike Training set: 200

Testing set: 321

Allen's interval: 68% **SVM:** 66% **Human:** 70%

Horse Training set: 200

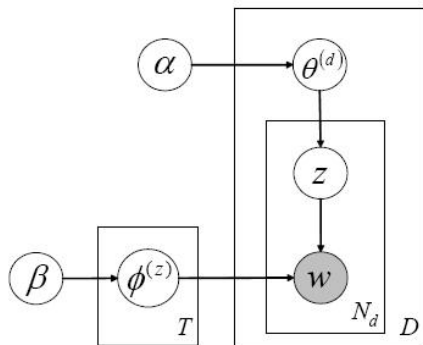
Testing set: 383

Allen's interval: 74% **SVM:** 72% **Human:** 75%

Topic Models

- Provide methods for statistical analysis of document collections & other discrete data
- Each document is viewed as a mixture of various topics
- Discover the abstract “topics” that occur in a collection of documents
- Use topic models for selectional preferences:
 - model the class-based nature of selectional preferences
 - do not take a pre-defined set of classes as input
 - naturally handle ambiguous arguments
 - scalable

Latent Dirichlet Allocation (LDA) [Blei et al., 2003]



- α, β : Dirichlet prior
- D : number of doc
- N_d : number of words in d
- z : latent topic
- w : observed word
- θ : distribution of topic in doc
- ϕ : distribution of words generated from topic z
- T : number of topics

Using plate notation:

- Sampling of distribution over topics for each document d
- Sampling of word distributions for each topic z until T topics have been generated

Topic models for action recognition in still images

- LDA:
 - trained on raw text
 - Extract triplets <subject, verb, object> before feeding to LDA
- Linked-LDA: Inspired by relevant work in selectional preferences [Alan Ritter et al., 10], [S.O, 10]
- Intuition: topic models
 - capture the “latent” relationship between words in corpora, hence can group together objects appearing in the same scene together
 - not only strictly focus on the relation between person-objects, but can be easily extended to more objects interacting with each other
 - can further be used to suggest possible scene, events for images

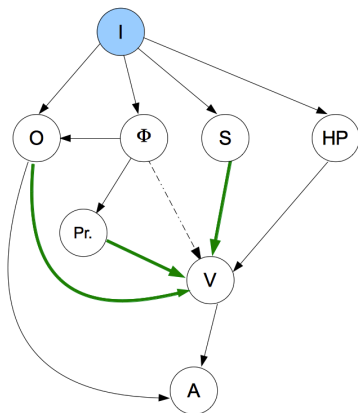
Topic models for action recognition in still images: Our experiment

- The LDA model is trained on the dataset containing 8,000 image descriptions collected from Flickr³
- Objects appearing together in an image in the PASCAL VOC gold standard
- Possible actions suggested by the LDA model

Objects	Topics	Actions
aeroplane-OBJ car-OBJ	Topic:70	drive pack pass
dog-OBJ sofa-OBJ	Topic:37 Topic:59	come break carry fetch
bottle-OBJ car-OBJ	Topic:70 Topic:84	drive drink
bus-OBJ car-OBJ motorbike-OBJ person-PER	Topic:35 Topic:63 Topic:70	kneel drive pack sit
bicycle-OBJ bottle-OBJ motorbike-OBJ	Topic:35 Topic:52 Topic:84	kneel ride
person-PER sofa-OBJ	Topic:35 Topic:37	come break

³<http://vision.cs.uiuc.edu/~pyoung2/8k-pictures.html>

Adding more features..



$$P(A|I) = P(O|I) \times P(\phi|I) \times P(Pr.|\phi) \times P(S|I) \times P(HP|I) \times P(V|Pr., O, S, HP) \quad (2)$$

Conclusion

- Action recognition in still images involves object, human pose, scene recognition and the interaction between them
- Most studies in action recognition have only focused on visual features without any help from general knowledge
- Learning from textual corpora can suggest plausible actions within any domain, not only limited to human actions
- Distributional memory and topic models are promising for learning general knowledge for this task
- This approach can be extended to recognize themes and events in images

Future work

- Train LDA-like model on the same corpora with the TypeDM model, compare these two models
- Exploit the possible mapping between prepositions in DM with the localization of objects in images
- Combine object recognition system with human pose classification to disambiguate actions
- Move to a broader domain with more interactions between objects in images, which is the main advantage of our approach

Bibliography

- [1] B. Yao, X. Jiang, A. Khosla, A.L. Lin, L.J. Guibas, and L. Fei-Fei. Human Action Recognition by Learning Bases of Action Attributes and Parts. International Conference on Computer Vision (ICCV), Barcelona, Spain. November 6-13, 2011.
- [2] Diarmuid O Seaghdha. 2010. Latent variable models of selectional preference. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 435-444.
- [3] Alan Ritter, Mausam, and Oren Etzioni. 2010. A latent dirichlet allocation method for selectional preferences. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 424-434.
- [4] M. Baroni and A. Lenci. Distributional Memory: A general framework for corpus-based semantics. 2010. Computational Linguistics 36 (4): 673-721.
- [5] J.R.R. Uijlings, A.W.M. Smeulders and R.J.H. Scha. Real-Time Visual Concept Classification. In IEEE Transactions on Multimedia, 99, 2010.

Thank you for your attention!