Social Media Annotation: from Images to Videos

Lamberto Ballan, PhD

www.micc.unifi.it/ballan

MICC
Media Integration and Communication Center

Stanford University - July 19, 2013

Joint work with:

Tiberio Uricchio

Marco Bertini, PhD

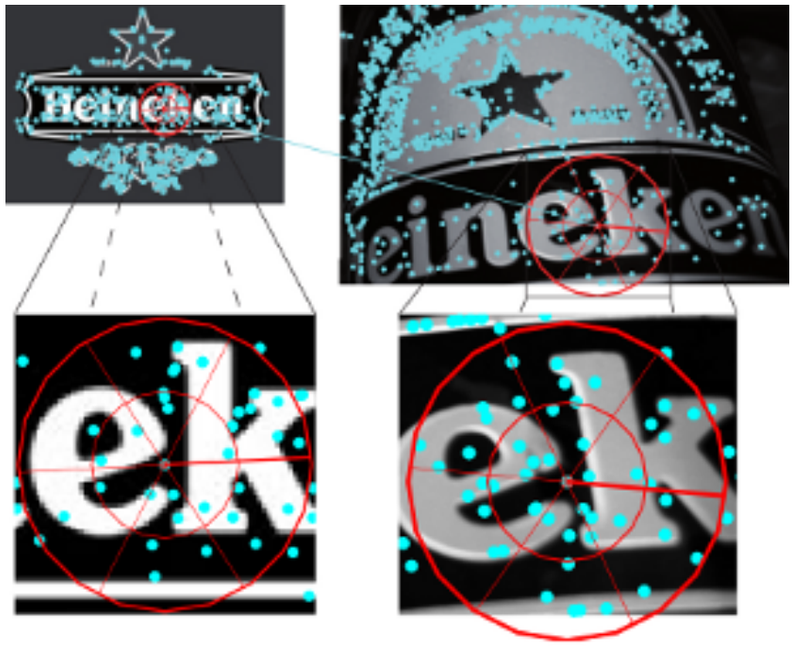
Prof. Alberto Del Bimbo

Univ. of Florence

Giuseppe Serra, PhD

Univ. of Modena and Reggio Emilia

Some of My Previous Works...

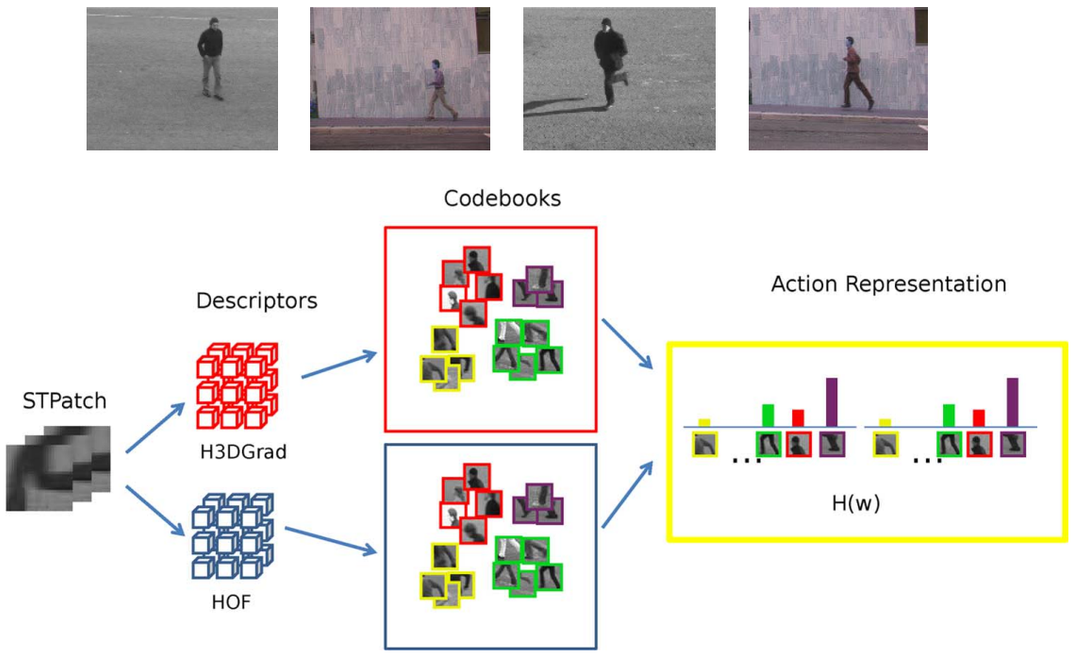


A Context-Dependent Kernel for Object Recognition (applied to logo recognition)

by Sahbi, Ballan, Serra, Del Bimbo, IEEE-TIP'13

Copy-Move Forgery Detection and Localization in Images

by Amerini, Ballan et al, ICASSP'10, IEEE-TIFS'11



Effective Codebooks for Human Action Categorization in Unconstrained Videos

by Ballan, Seidenari et al., ICCVW'09, IEEE-TMM'12

Web 2.0: Social Media

- Facebook
  - 964 million monthly active users on March 2013
  - an average user has 130 friends (Dunbar's number = 150)
  - more than 3.5 billion images/videos/etc. shared per week
- Twitter
  - 200 Millions of monthly active Twitter users
  - 175 Millions of tweets per day sent in 2012 (307 avg per user)
- Flickr
  - Flickr hosts more than 6.7 billion images
  - ~4 millions new uploads per day
- Youtube
  - ~4 billion views a day and 60-70 hours of videos uploaded per minute

It took to reach 50 million users:
- Telephone: 75 years
- Radio: 38 years
- TV: 13 years
- Internet: 4 years

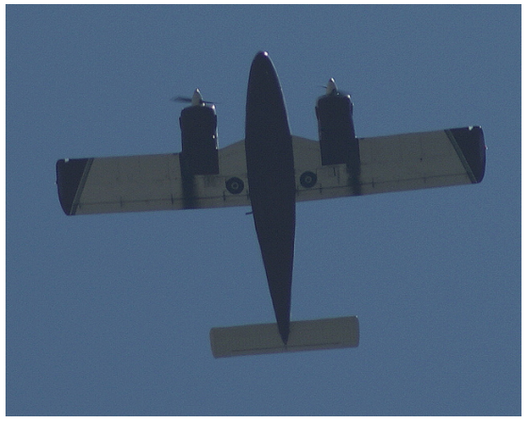
Source: Social Media Statistics (2012)

Tags and Folksonomies

- Tags imposed by social networking define soft organizations on data (*folksonomies*); they pose new opportunities of semantic extraction from visual data w.r.t. to fixed taxonomies that are rigid and centralized
- Main challenges:
  - tags are often imprecise and ambiguous; their order does not correspond to tag relevance and they are influenced by cultural aspects
  - tags are often irrelevant to the visual content and overly personalized
  - spontaneous choice of words with large variability among different users: *polysemy*, *synonymy*, ...
  - semantic loss in the textual descriptions: meaningful tags missing

Query: "airplane"

flickr

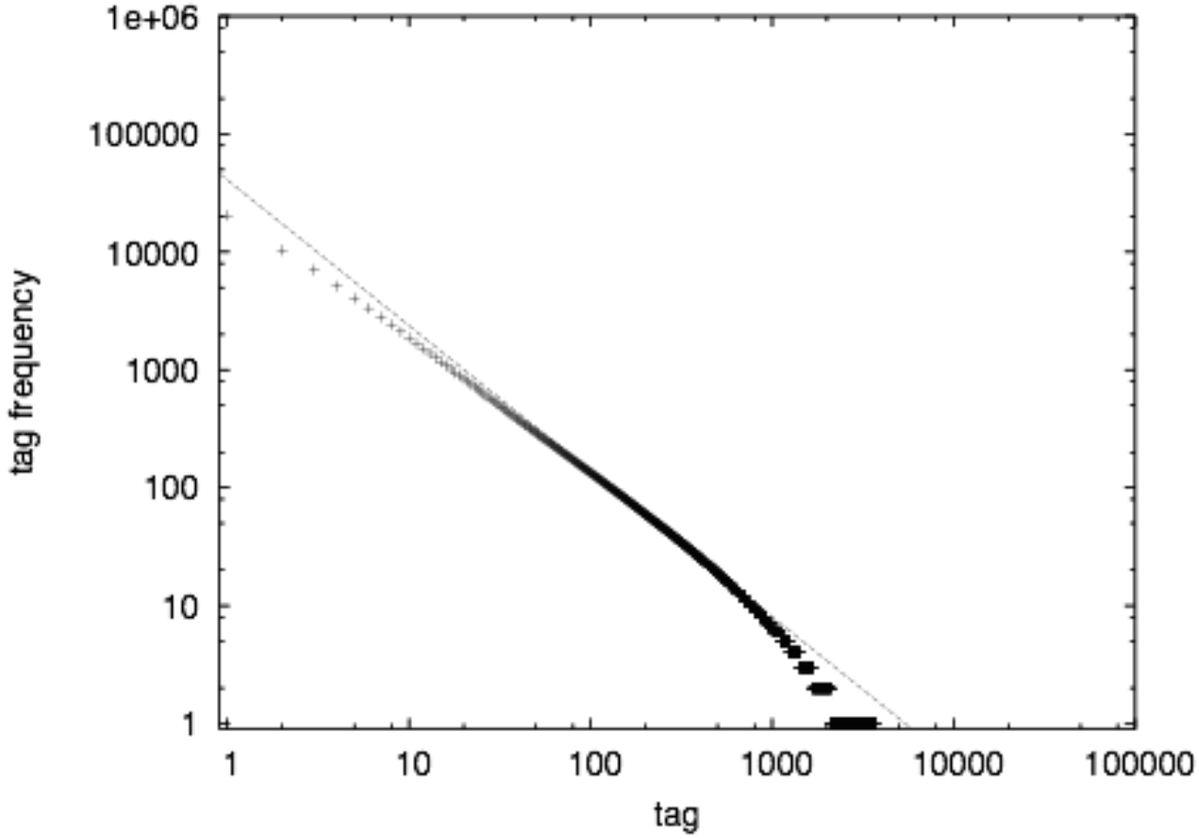


airplane
twin
engine
los angeles

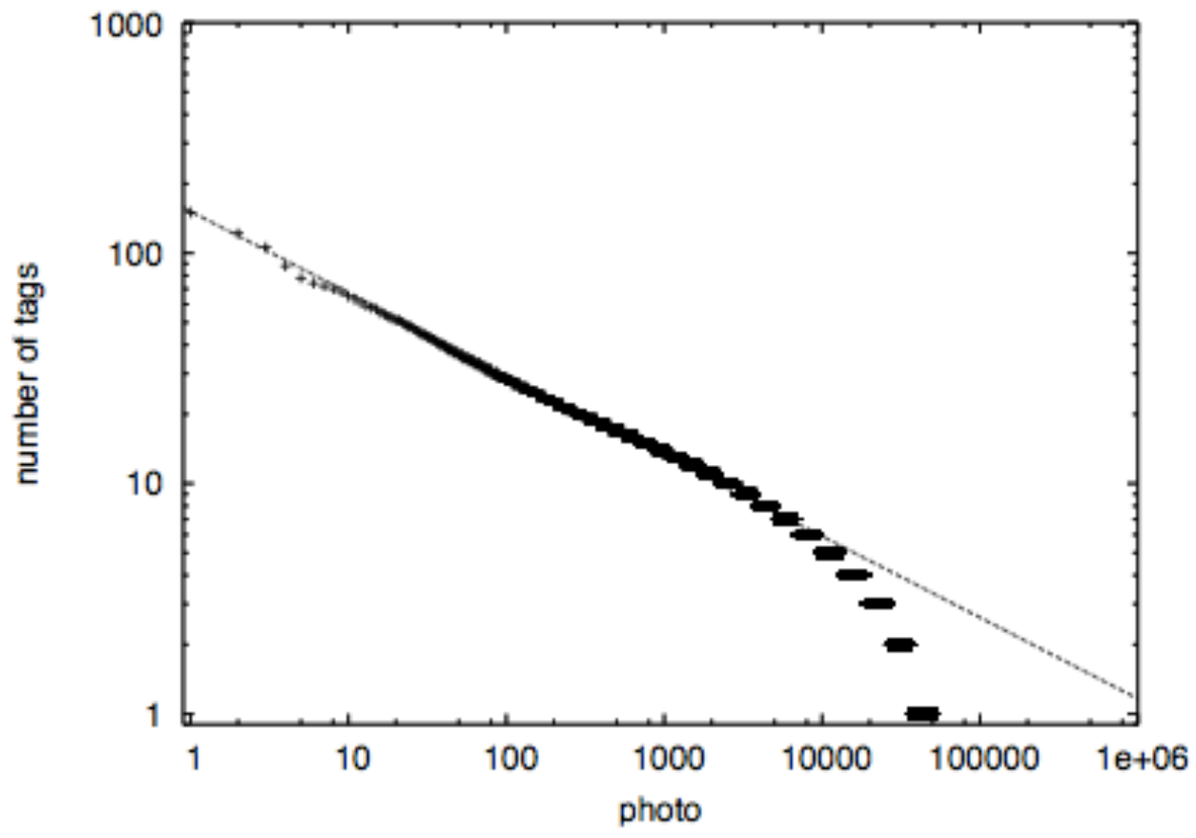
daytime
beach
airplane
ocean

Flickr Tags Distributions

Distribution of the tag frequency in Flickr



Distribution of the number of tags per image



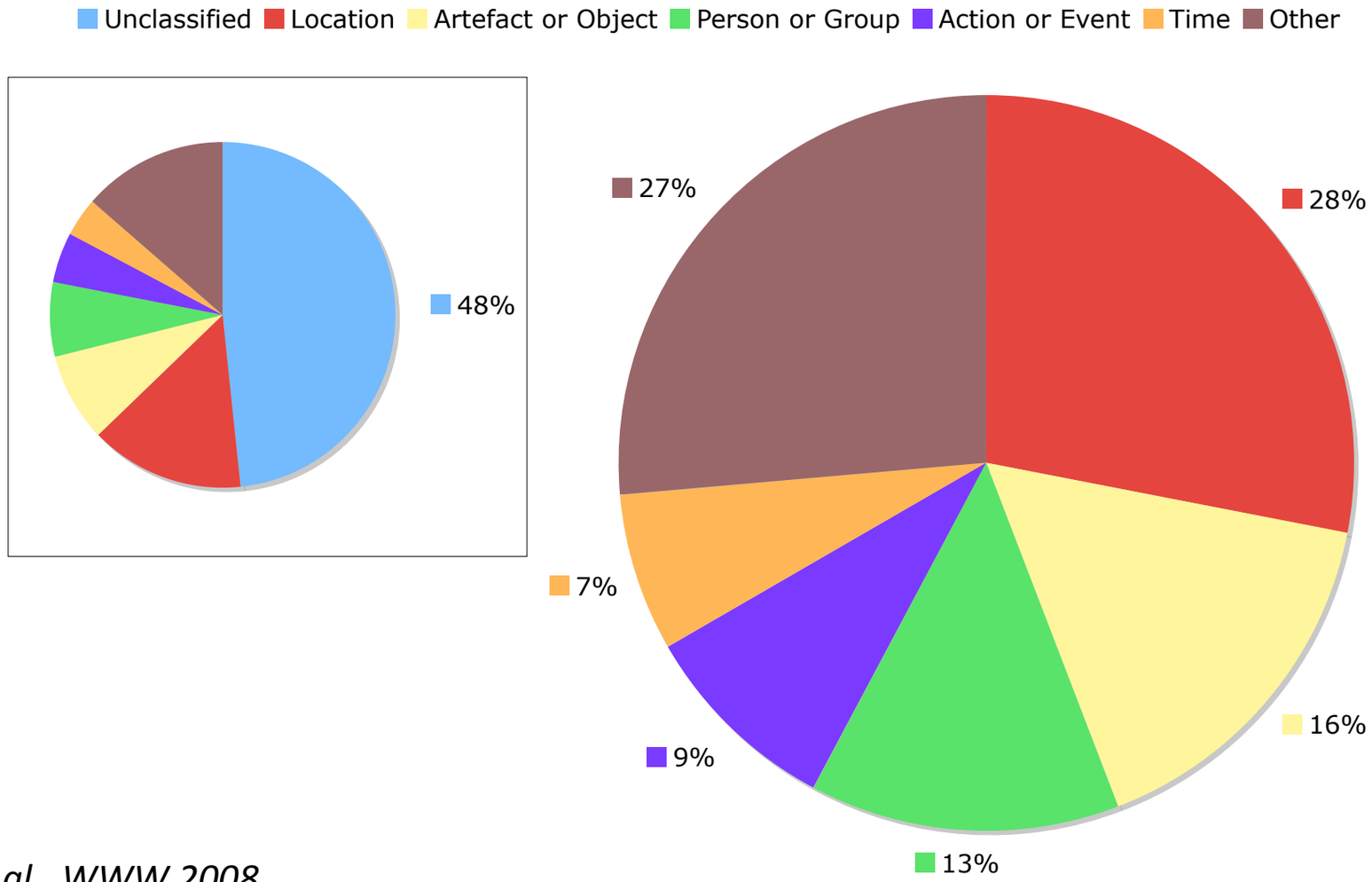
- Tag frequency:
  - the head of the distribution contains too generic tags (wedding, party,…)
  - the tail contains the infrequent tags with incidentally occurring terms such as misspellings and complex phrases
- Number of tags per image:
  - about 64% of images have only 1-3 tags

Source data:
- 52 Million Flickr photos
- 3.7 million unique tags

Source: Sigurbjörnsson et al., WWW 2008

WordNet Categories for Flickr Tags

- The distribution of Flickr tags over the most common WordNet categories
  - 52% of the tags is correctly classified
  - 48% of the tags is left unclassified
- Nearly one half of tags are irrelevant for general audience

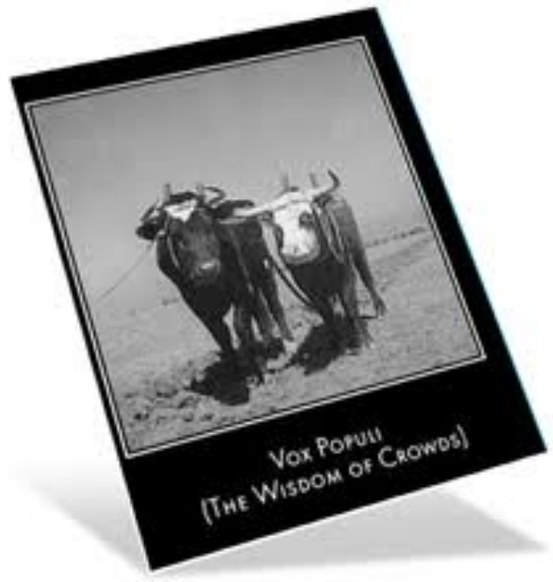


Source: Sigurbjörnsson et al., WWW 2008

The Wisdom of Crowds

- *The Wisdom of Crowds: "the verdict of a group of people is closer to the truth than that of any individual in the group"* [Galton 1906]

- The crowd could contribute to reach a "statistical regularity" in the tag vocabulary
- Mechanisms to convert opinions into an aggregated verdict:
  - *tag co-occurrence*: the number of images where several tags are used in the same annotation is the key to tag recommendation
  - *visual content-tag association*: if different persons label visually similar images with the same tags, these tags are likely to reflect objective aspects of the visual content
  - *consider the complex relationships* of tags in a folksonomy



How to Improve Image Tags?

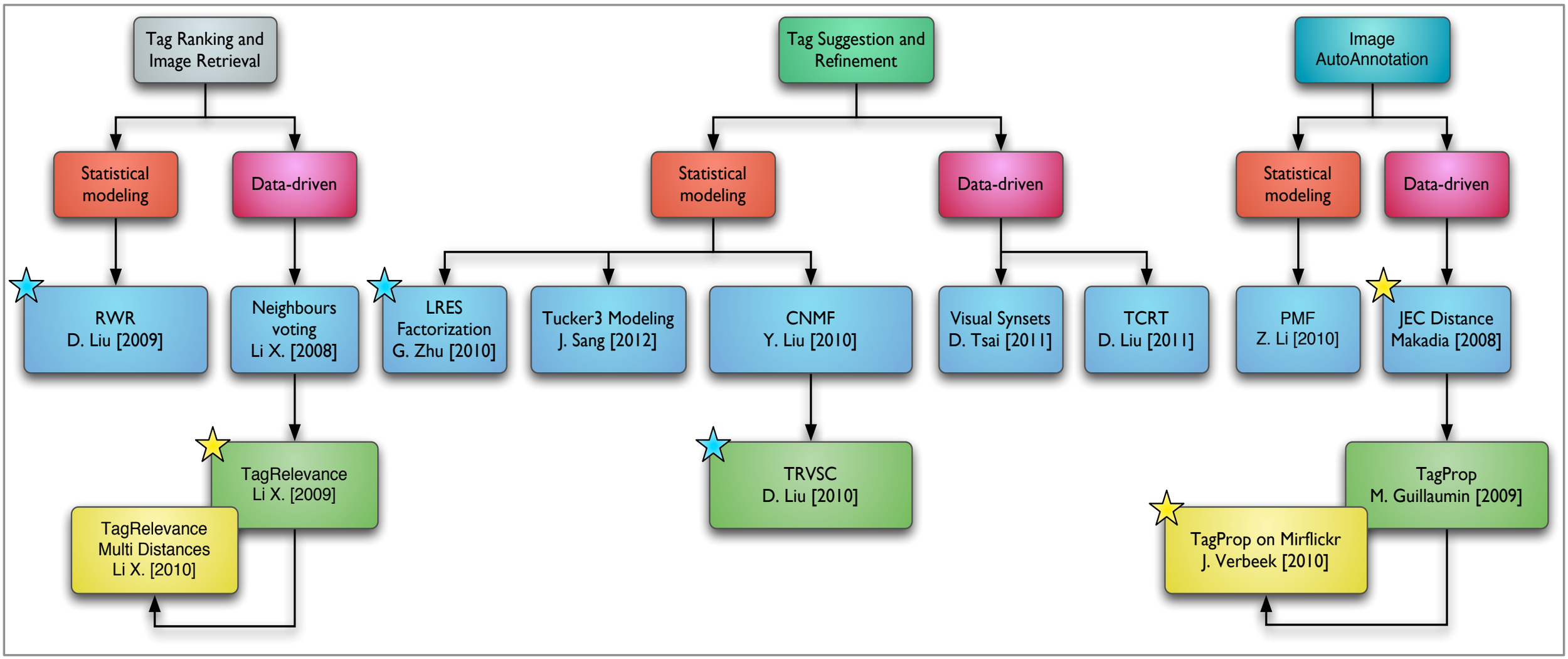
- ***Tag Refinement***: the goal is removing noisy tags, disambiguating tags and recommending new tags that are relevant to the visual content and the other tags
- Related tasks are: *Tag Suggestion/Recommendation*, *Tag Re-ranking/Relevance*

Luigi Torreggiani, CC BY 2.0 license.

~~*child*~~	
~~*girl child*~~	
party	***context***
birthday	***context***
~~*nikon*~~	
~~*d40*~~	
candle	***content***
pie	***content***
~~*apple*~~	
berries	*content*
hand	*content*

Taxonomy of Main Research Contributions

- Previous works may be divided into two main categories of approaches:
  - Based on **statistical modeling** (e.g. matrix factorization)
  - Based on **data-driven** techniques (e.g. NN voting)



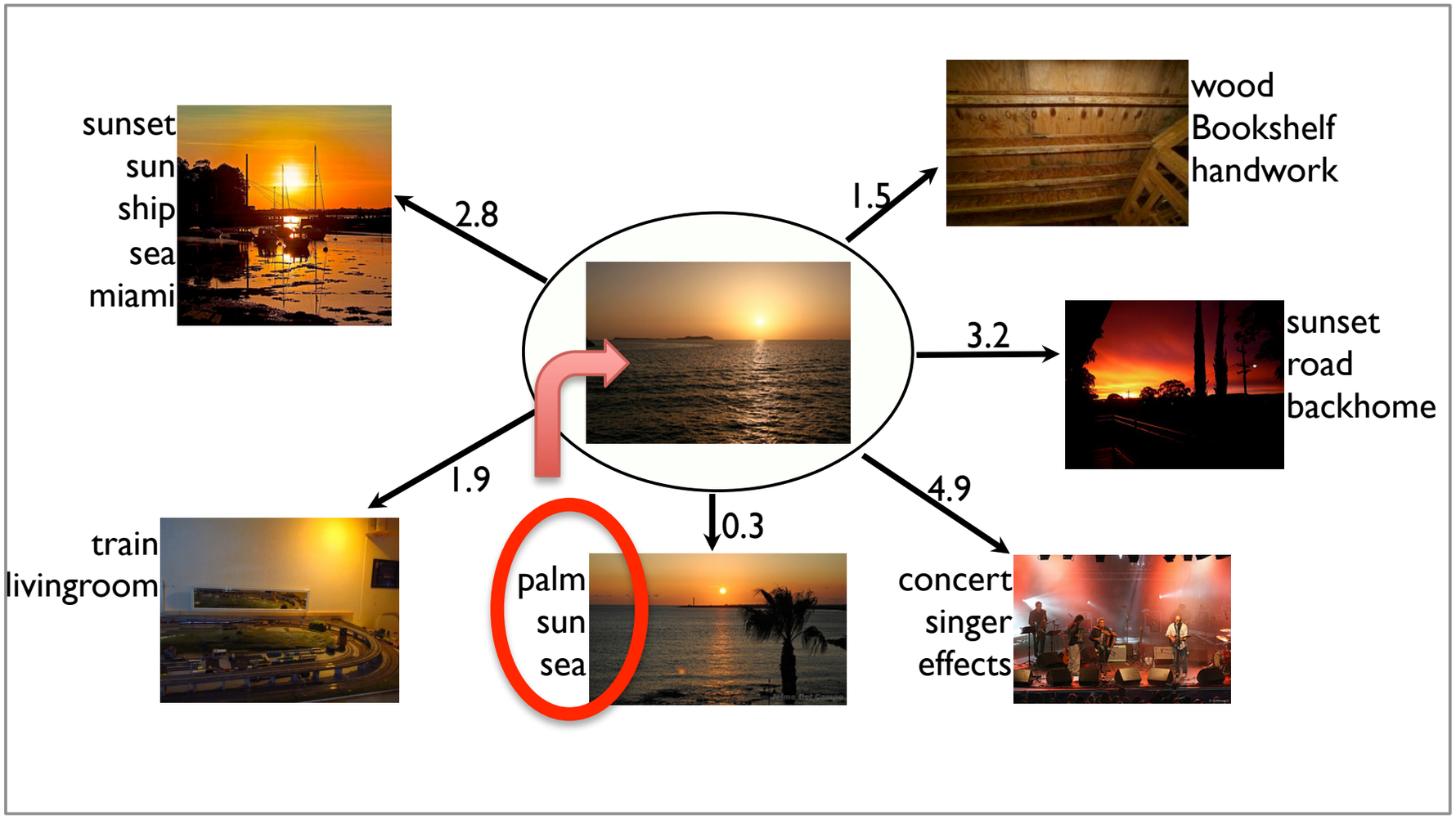
Tag Refinement: Data-driven Approach

- Data driven methods exploit binary image-tag relations; they assume there exist large well labeled dataset where one can find visual "near-duplicates" of the image
- Ground on the idea of selecting a set of visually similar images and then extract a set of relevant tags using a tag transfer procedure (usually a Nearest-Neighbor voting scheme)

- Usually applied for Image Annotation or Retrieval:
  - *Simple Label Transfer (SLT) / JEC:* Makadia et al. ECCV'08, IJCV'10
  - *Tag Relevance Learning (TR):* X. Li & Snoek, IEEE-TMM'09, CIVR'10
  - *Tag Propagation (TagProp):* Guillaumin et al. ICCV'09, MIR'10

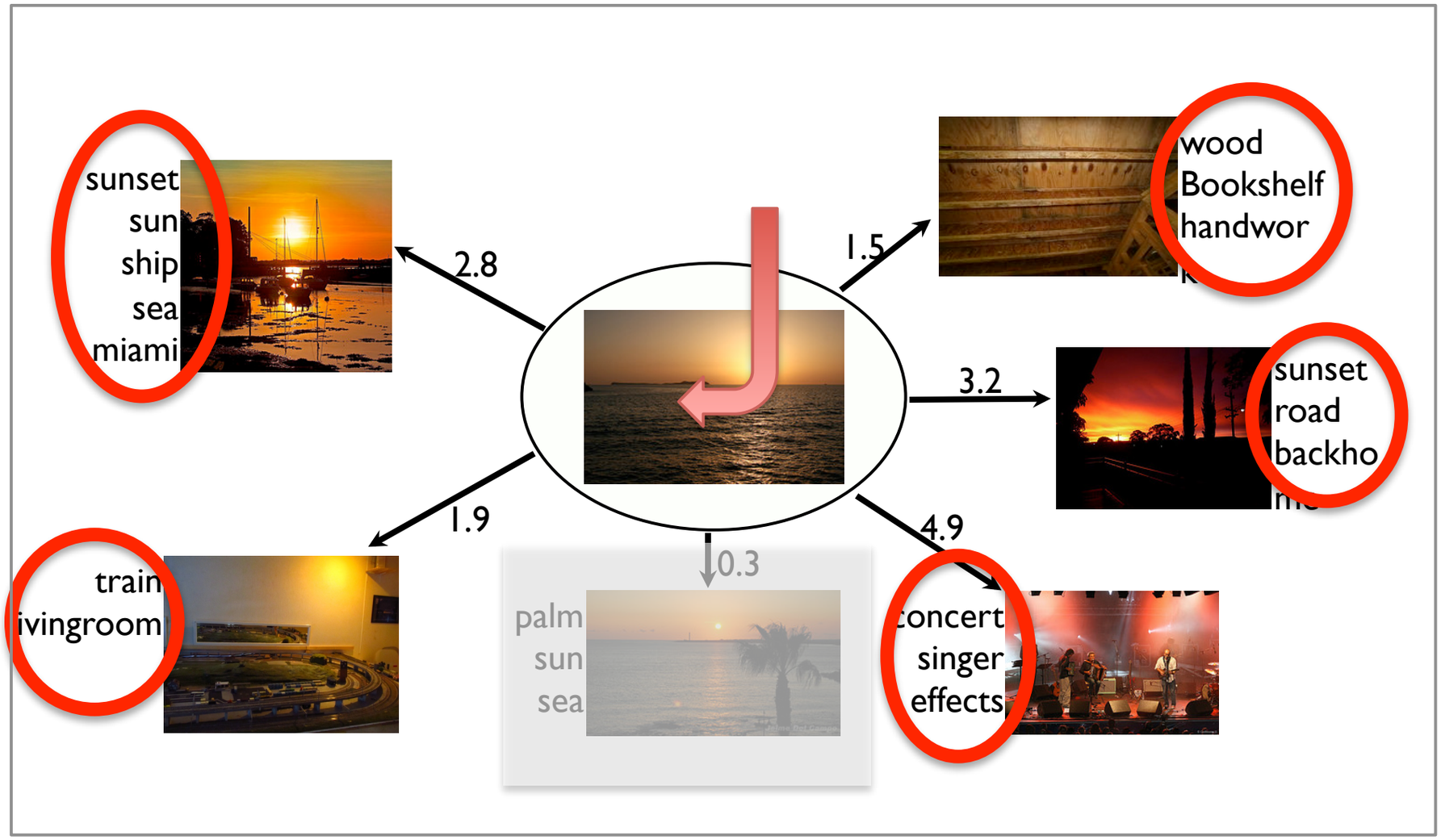
Simple Label Transfer (SLT)

- Images are ranked according to content similarity distances (using multiple visual features)
- Two strategies for fusion: Joint Equal Contribution (JEC) between distances or Lasso
- The most similar image is selected and its tags are applied



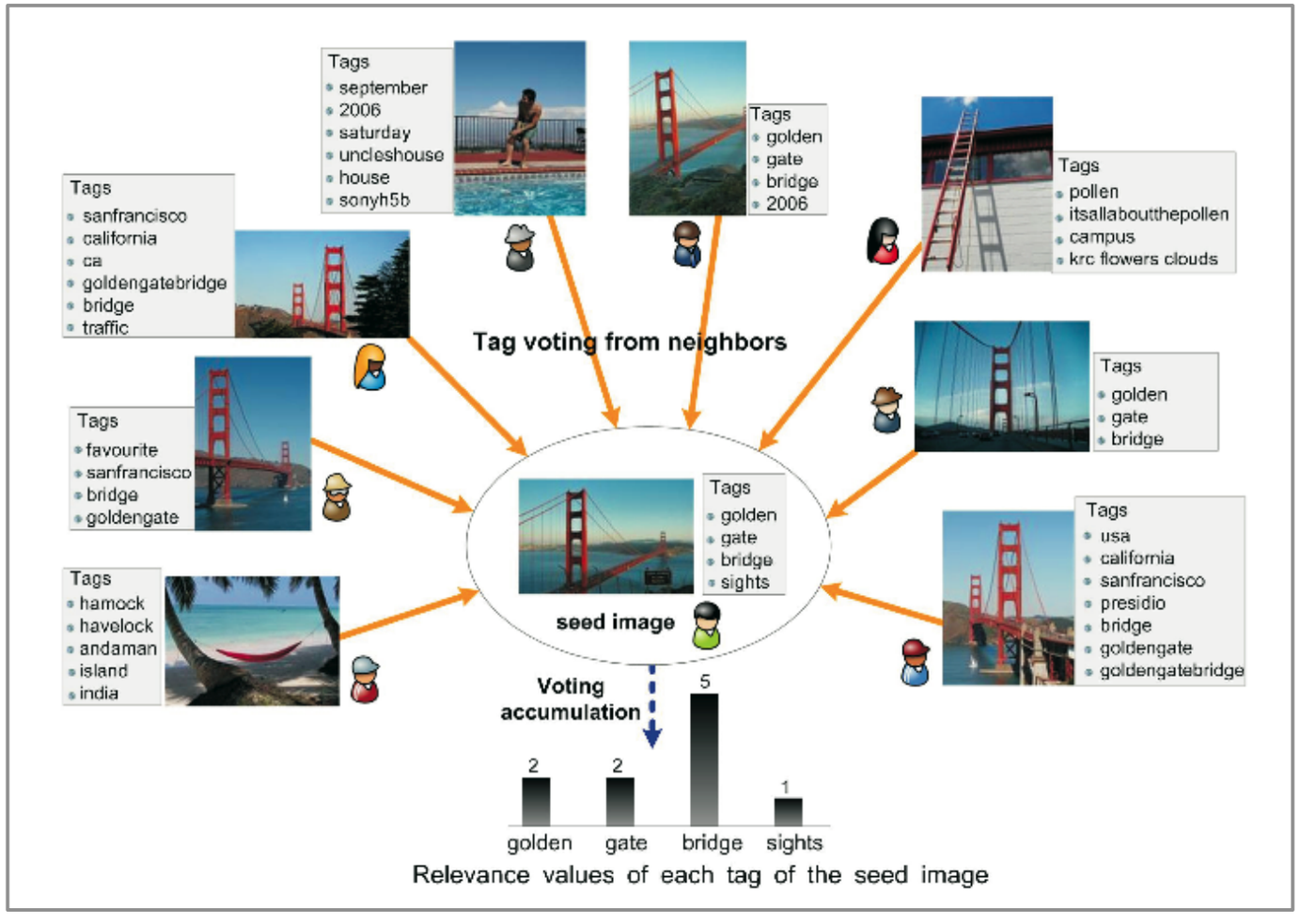
References: *Makadia et al., "A new baseline for image annotation", ECCV 2008, IJCV 2010*

- If additional tags are required, the closest images are selected and their tags applied, according to their co-occurrence with the keywords transferred and their frequency



Tag Relevance

- Key assumption: "If different persons label visually similar images using the same tags, then these tags are more likely to reflect objective aspects of the visual content"
  - define a tag relevance measure by considering the distribution of the tag in the neighbor set of the image and in the entire collection
  - the more frequent a tag is in the neighbor set the more relevant it is

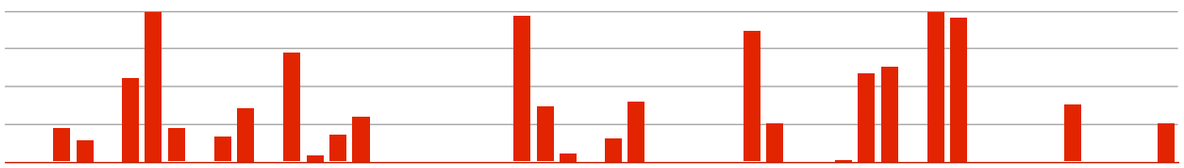


References: *X. Li and C. Snoek, "Learning social tag relevance by neighbor voting", IEEE-TMM 2009, CIVR 2010*

$$tagRelevance(w, I, k) := n_w[N_f(I, k)] - Prior(w, k)$$

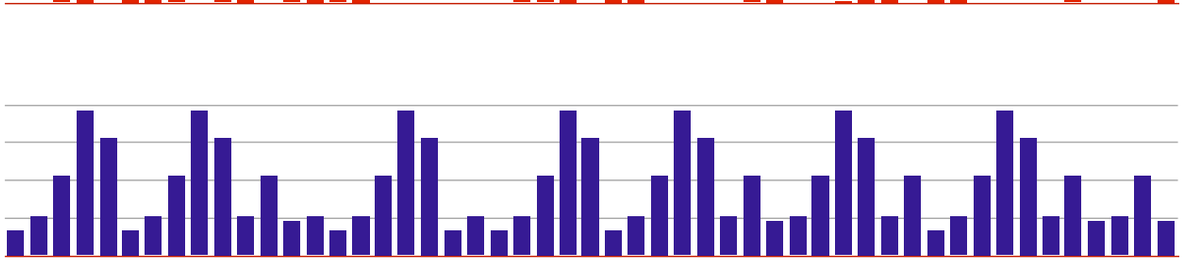
n_w counts the occurrences of w in the neighborhood $N_f(I, k)$ of k similar images

$Prior(w, k)$ is the frequency of occurrence of w in the collection

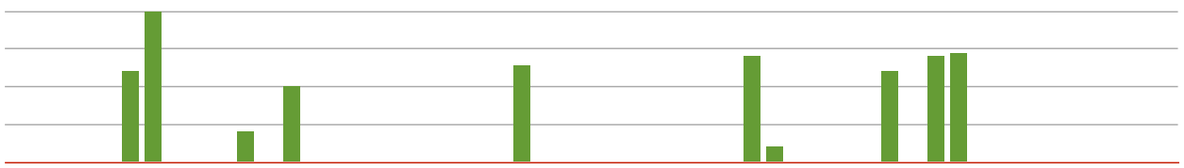


distribution of each tag in $N_f(I, k)$

-



prior distribution of each tag in the collection



final tag relevance

TagProp: Weighted NN Image Annotation

- Learns a weighted nearest neighbor model to find the optimal combination of feature distances
  - the model is defined using a probabilistic framework:

$$p(y_{iw} = +1) = \sum_j \pi_{ij} p(y_{iw} = +1|j) \qquad \pi_{ij} \geq 0 \wedge \textstyle\sum_j \pi_{ij} = 1$$

$$p(y_{iw} = +1|j) = \begin{cases} 1 - \epsilon & \text{for} \quad y_{jw} = +1 \\ \epsilon & \text{otherwise.} \end{cases}$$

where $y_{iw} \in \{+1,-1\}$ indicates whether tag w is relevant or not for image i

and π_{ij} is the weight of image j (from the visual neighbors) in respect to image i to be learned

- The objective is to maximize the log-likelihood by using EM

$$\mathcal{L} = \sum_{i,w} \ln p(y_{iw}) = \sum_{i,w} \ln \sum_j \pi_{ij}\, p(y_{iw}|j)$$

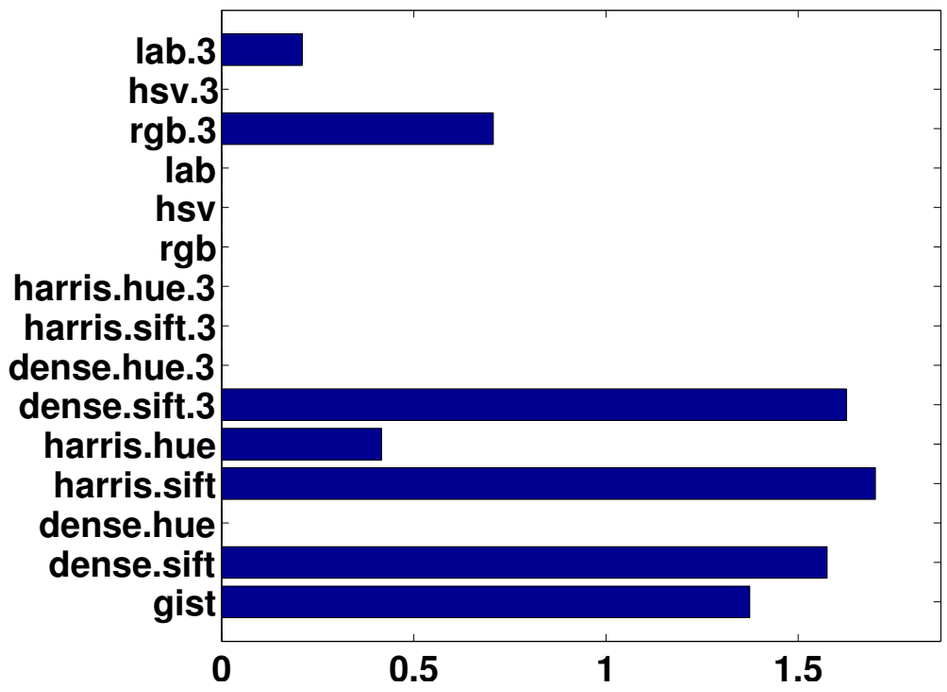
References: *Guillaumin et al. "TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation", ICCV 2009*

- weights can be defined as a function of distance of neighbors images:

$$\pi_{ij} = \frac{\exp(-d_\theta(i,j))}{\sum_{j'} \exp(-d_\theta(i,j'))}$$

- due to the unbalanced tags frequency, a *word-specific logistic discriminant* is used to boost the probability for rare terms and decrease it for the most frequent ones:

$$p(y_{iw} = +1) = \sigma(\alpha_w x_{iw} + \beta_w)$$



clouds
male
people
sea
sky
water

sky (0.99)
clouds (0.99)
water (0.69)
structures (0.64)
sea (0.32)
tree (0.32)

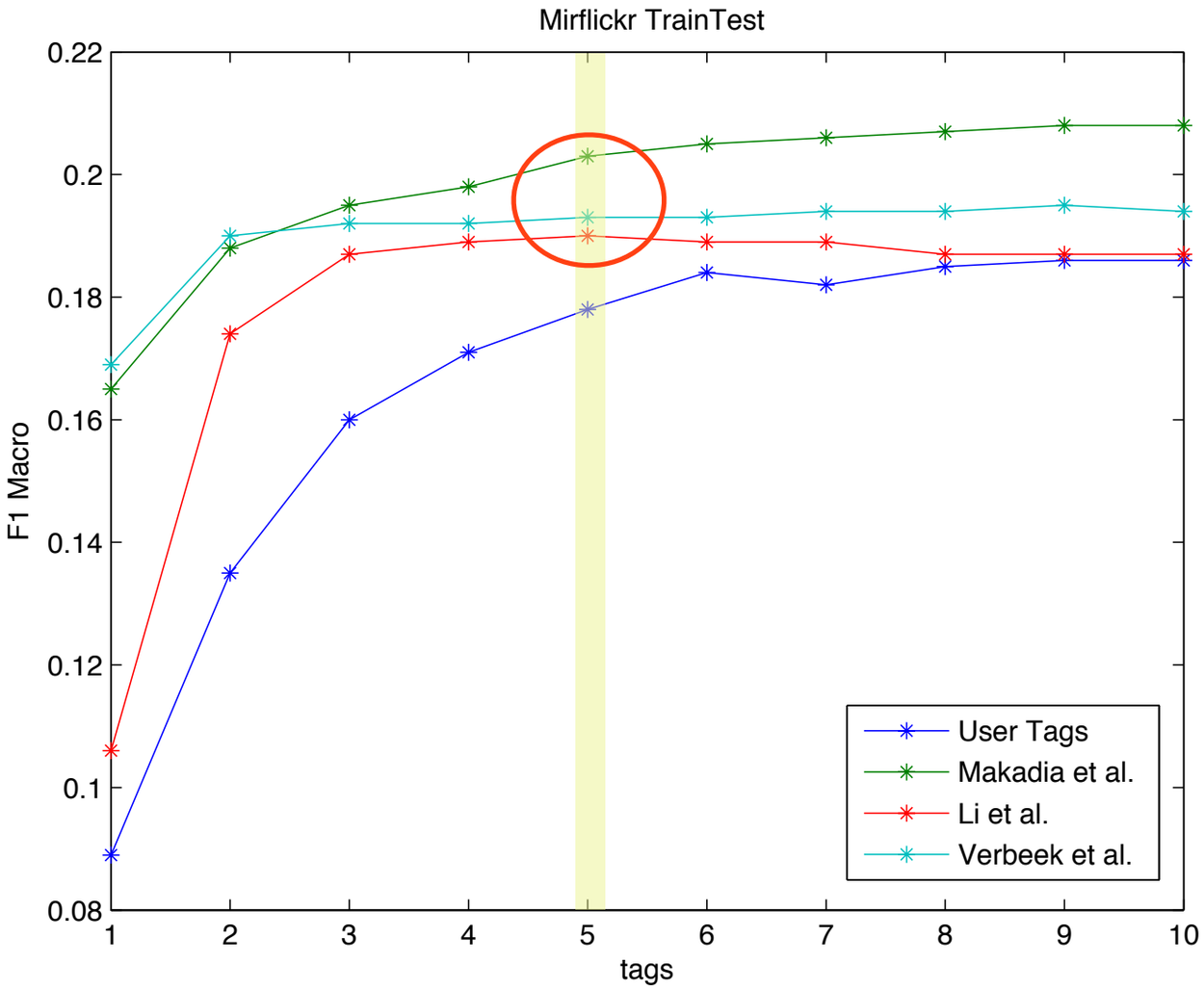
An Evaluation of NN Methods for Tag Refinement

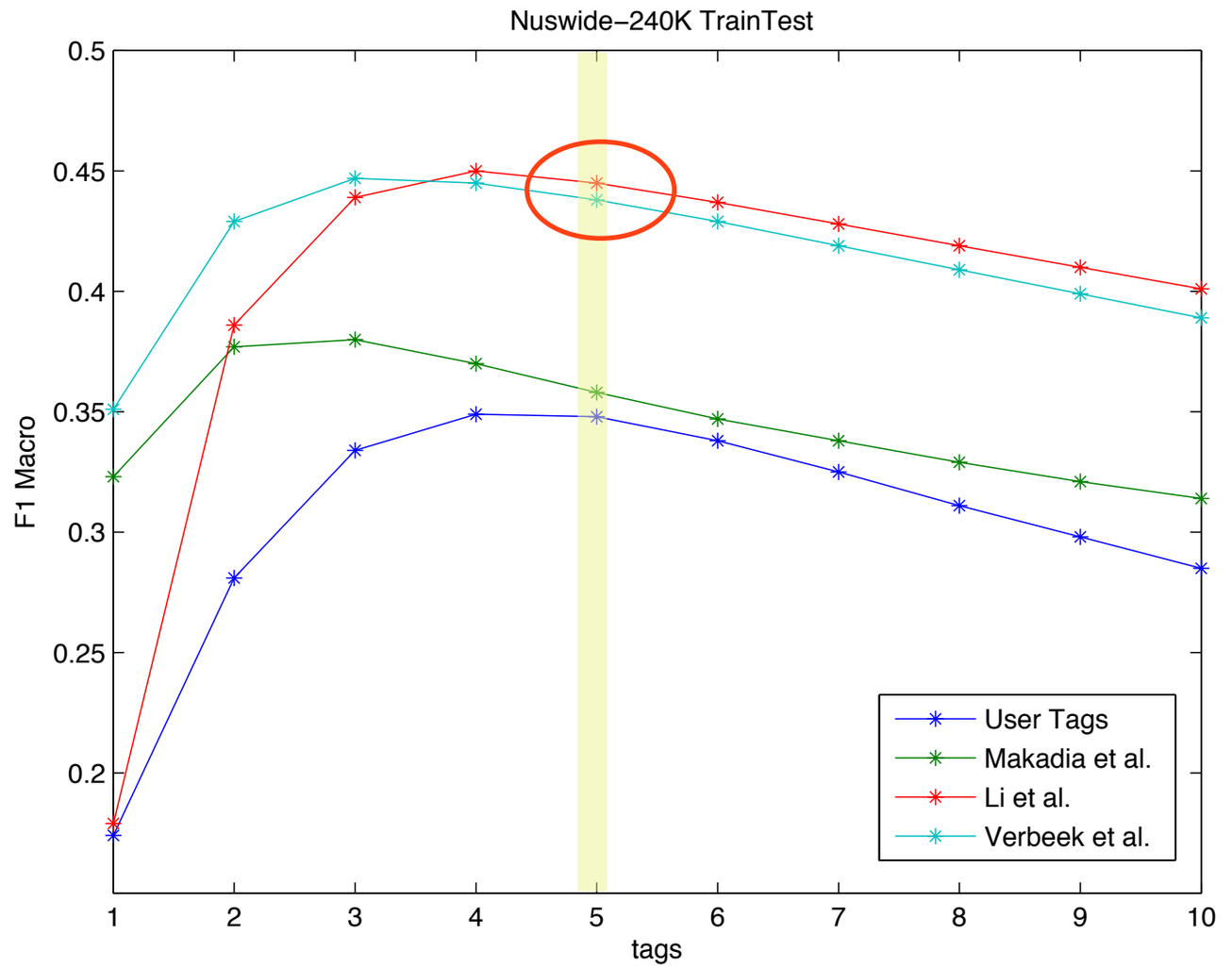
- MIRFlickr dataset
  - 16 global and local features
  - distances: combination of L2 and e KL-divergence
  - performance: macro and micro-average
- NUSWIDE dataset
  - 428-d global features (color, wavelet, edge histograms)
  - distance: L2
  - performance: macro and micro-average

	NUSWIDE 270K	NUSWIDE 240K	MIRFlickr
Images	269,648	238,251	25,000
Train Set	161,789	158,834	10,000
Test Set	107,859	79,417	15,000
Ground-Truth Tags	81	81	27
Users	-	24,625	9,862
Original Tags	5,018	5,018	1,386
Filtered Tags (Wikipedia)	521	-	-
Filtered Tags (WordNet)	-	684	219

References: *Uricchio et al. "An evaluation of nearest-neighbor methods for tag refinement", ICME 2013*

Performance Evaluation





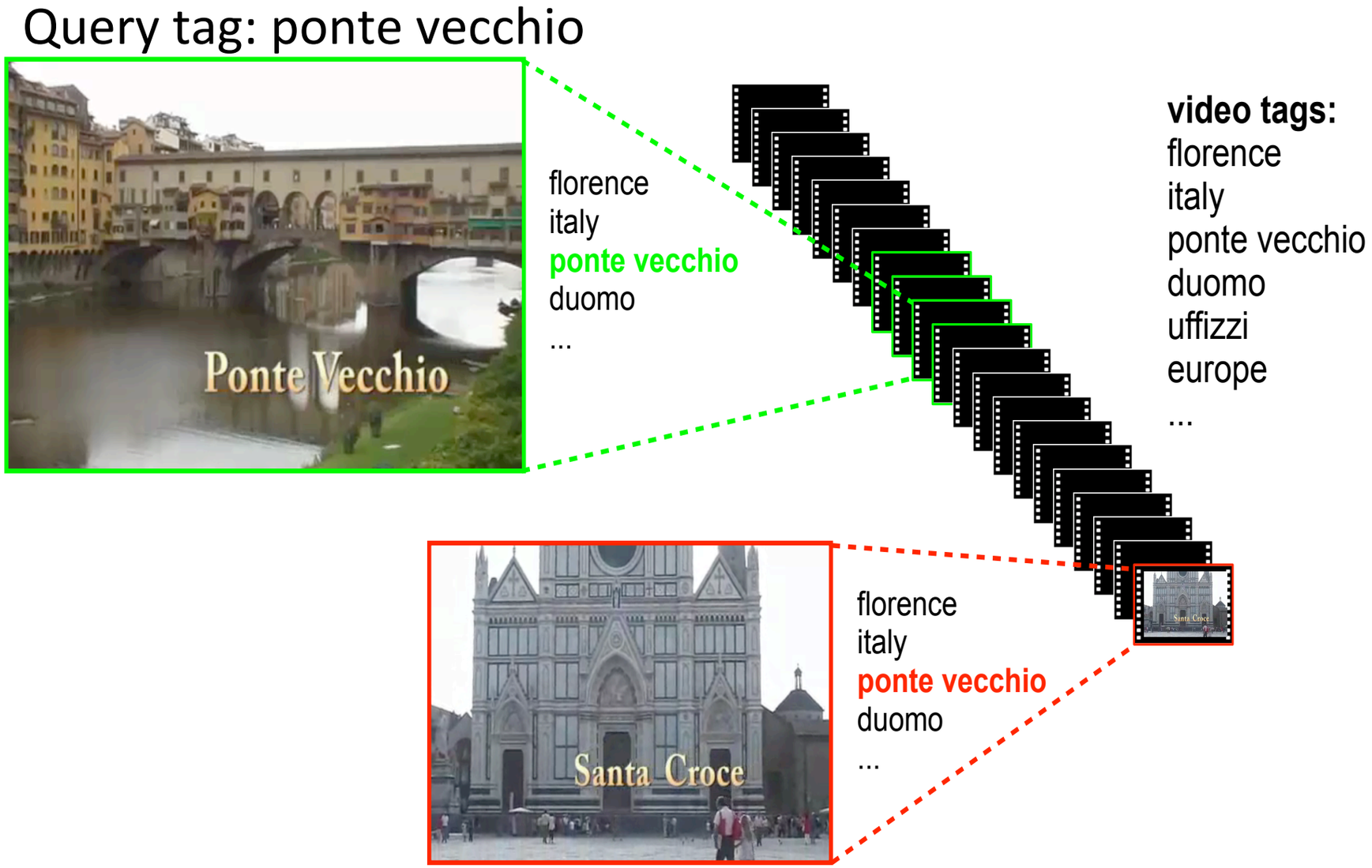
NN methods give comparable results to more complex state-of-the-art approaches, despite their simplicity

Nuswide−240K TrainTest

	UT	RWR [wang07]	TRVSC [liu10]	LR [zhu10]
Zhu *et al.* ACMMM2010	0.27	-	-	0.35
Liu *et al.* TMM2011	0.45	-	0.55	-
Xu *et al.* TMM2012	0.48	0.48	0.49	0.52

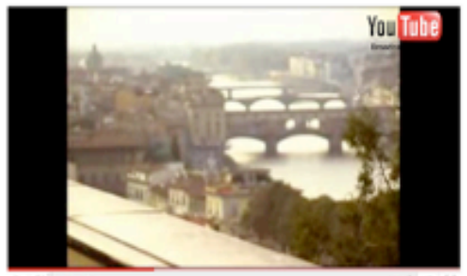
Extending Data-driven Methods to Video

- We have the same problems/challenges of images
- Moreover tags are not "localized" at the frame (shot) level



References: *Ballan et al. "Enriching and Localizing Semantic Tags in Internet Videos", ACM-MM 2011*

Our Framework



Shot segmentation and Keyframe extraction

Keyframes

K

Video tags: firenze florence tuscany italy culture tribute most beautiful town travel love art

V

Identification of the nearest cluster and Tag localization

flickr

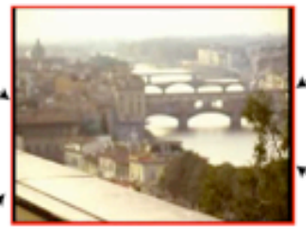
Retrieved Flickr images and Image Clustering

I

T

Image tags:
Italy
old
bridge
culture
art
Arno

Image tags:
Firenze
bridge
vecchio
tuscany
Italy



Suggested Tags:
FIRENZE
ART
FLORENCE
ITALY
bridge

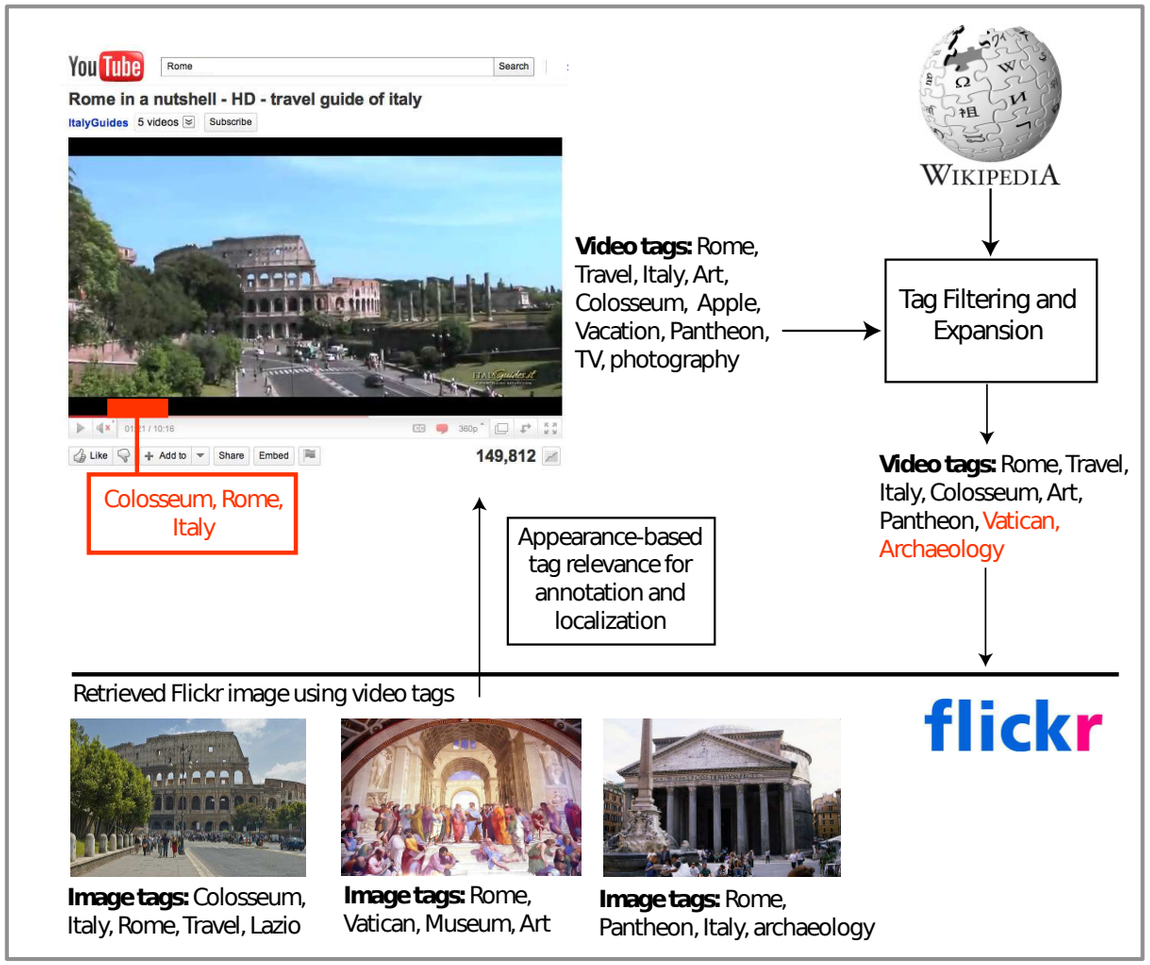
Image tags:
Firenze
Florence
river
bridge
Ponte
Vecchio
Canon 300D

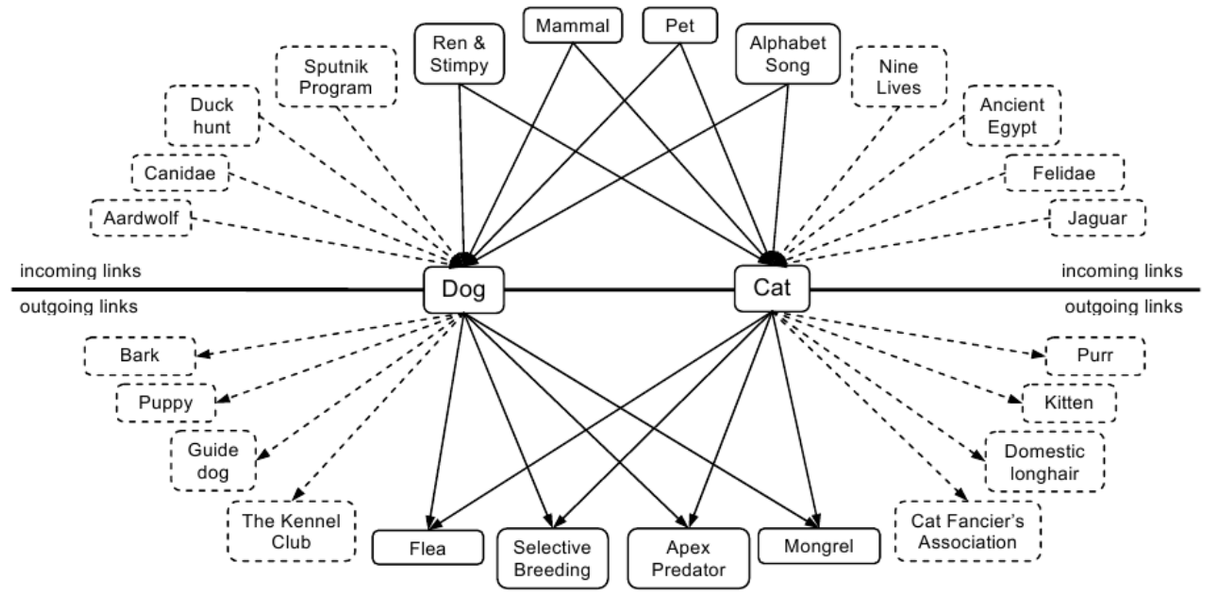
Image tags:
Firenze
old bridge
Florence
ponte
love
art

- Images in $\mathbf{I}$ are retrieved using the tags in $\mathbf{V}$; for each I_i in $\mathbf{I}$ and K_j in $\mathbf{K}$ we compute a global feature vector (GIST, HSV and edge hist)

- All the tags associated to images in the most similar cluster to K_j are retained in $\mathbf{T_j}$

- The original tags in $\mathbf{V}$ are assumed as valid only if they are also in $\mathbf{T_j}$
- The lis of tags is refined by analyzing Wikipedia (e.g. synonyms)

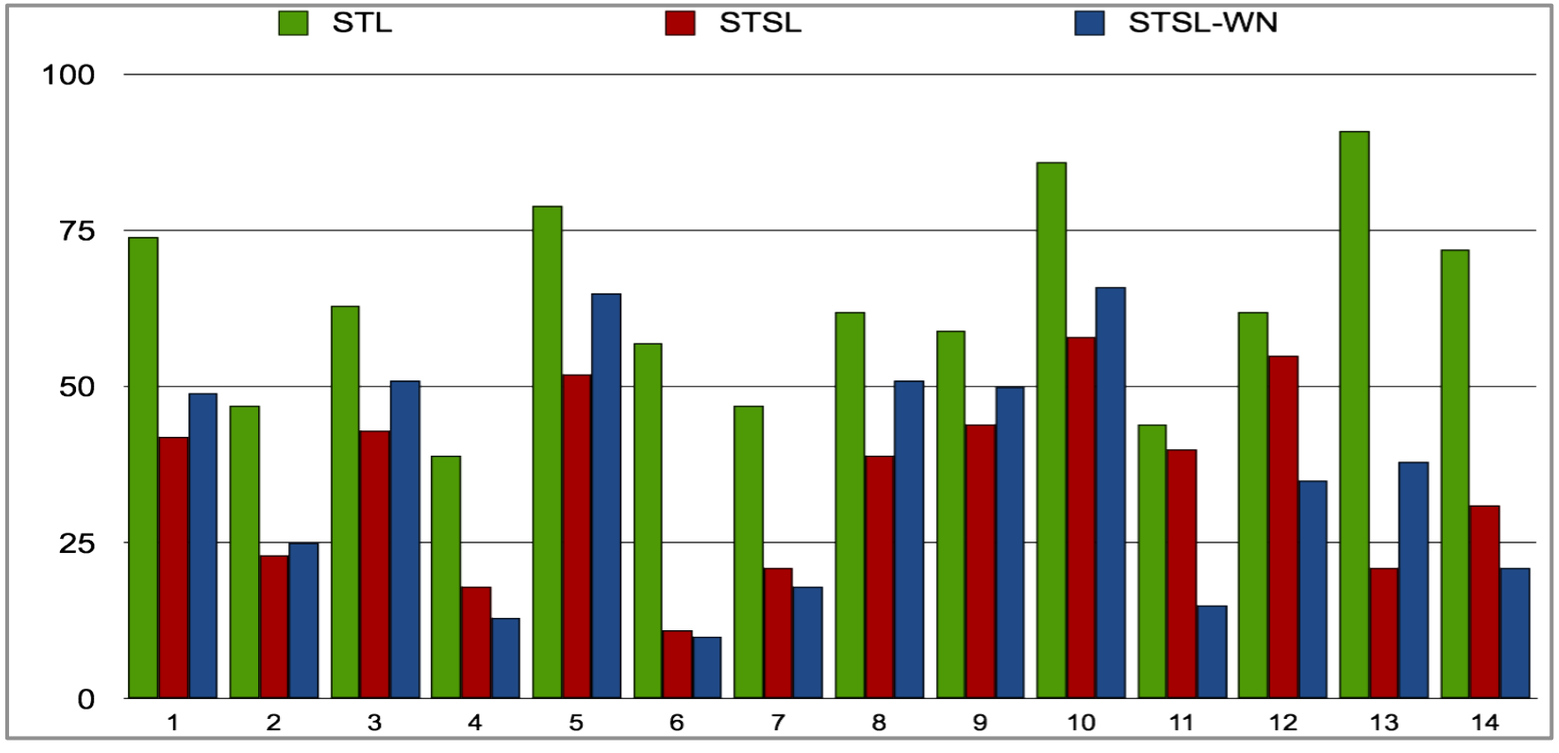




- Tag Relevance is computer for each t in $\mathbf{T_j}$ as previously reported [Li and Snoek, TMM'09]
- The five most relevant tags are added at the shot level
- The union of all tags that have been added at the shot level are used for video annotation

Experiments

- YouTube60 dataset
  - 1,135 shots - 3,405 keyframes annotated
  - all the original tags are provided (min 3, max 26 per video)
  - for each tag 15 Flickr images have been downloaded
  - 5 additional Flickr images for each synonym



Shot level Tag Localization (STL)

Accuracy of localization of the YouTube tags within the correct shots

Shot level Tag Suggestion and Localization (STSL)

Accuracy of localization at the shot level for both YouTube and suggested tags

STSL + WordNet expansion (STSL-WN)

Accuracy of STSL after WordNet expansion of the YouTube tags

1. Cars & Vehicles
2. Comedy
3. Education
4. Entertainment
5. Film & Animation
6. Gaming
7. Howto & Style
8. Music
9. News & Politics
10. People & Blogs
11. Pets & Animals
12. Science & Technology
13. Sport
14. Travel & Events


Scene 14: PARK, TERRAIN, LAND,
landscape, sky, mountain,
scenery, colors


Scene 1: MAID, MIST, NIAGRA, FALLS,
scotland, waterfall, trees,
crossdresser, tablier


Scene 1: VOLCANO, ERUPTION,
EYJAFJALLAJÖKULL, ICELAND,
glacier, landscape,
volcaniceruption, eldgos, nature

Ongoing Works

- Use multiple semantic taxonomies for image/video annotation and tag refinement
  - WordNet and ImageNet
  - a folksonomy learned from user tags

- Define a unified model for image and tags based on an intermediate semantic representation (attributes?)

Thank you!