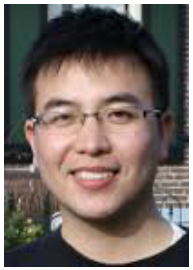
IMAGE TAG ASSIGNMENT, REFINEMENT AND RETRIEVAL

ACM Multimedia 2015 Tutorial

October 26, 2015



Xirong Li
Renmin University of China

Tiberio Uricchio
University of Florence

Lamberto Ballan
University of Florence &
Stanford University

Marco Bertini
University of Florence

Cees Snoek
University of Amsterdam &
Qualcomm Research
Netherlands

Alberto Del Bimbo
University of Florence

PART I
INTRODUCTION

Cees Snoek
University of Amsterdam &
Qualcomm Research Netherlands

cgmsnoek@uva.nl

Marco Bertini
University of Florence

- Problem statement
- Course organization

ABOUT THIS TUTORIAL

- This tutorial focuses on challenges and solutions for content-based image retrieval in the context of online image sharing and tagging.
- We present a unified review on three closely linked problems, i.e., tag assignment, tag refinement, and tag-based image retrieval.
- We introduce a taxonomy to structure the growing literature, understand the ingredients of the main works, clarify their connections and difference, and recognize their merits and limitations.
- We present an open-source testbed, with training sets of varying sizes and three test datasets, to evaluate methods of varied learning complexity.
  11 representative works have been implemented and evaluated.

http://www.micc.unifi.it/tagsurvey/

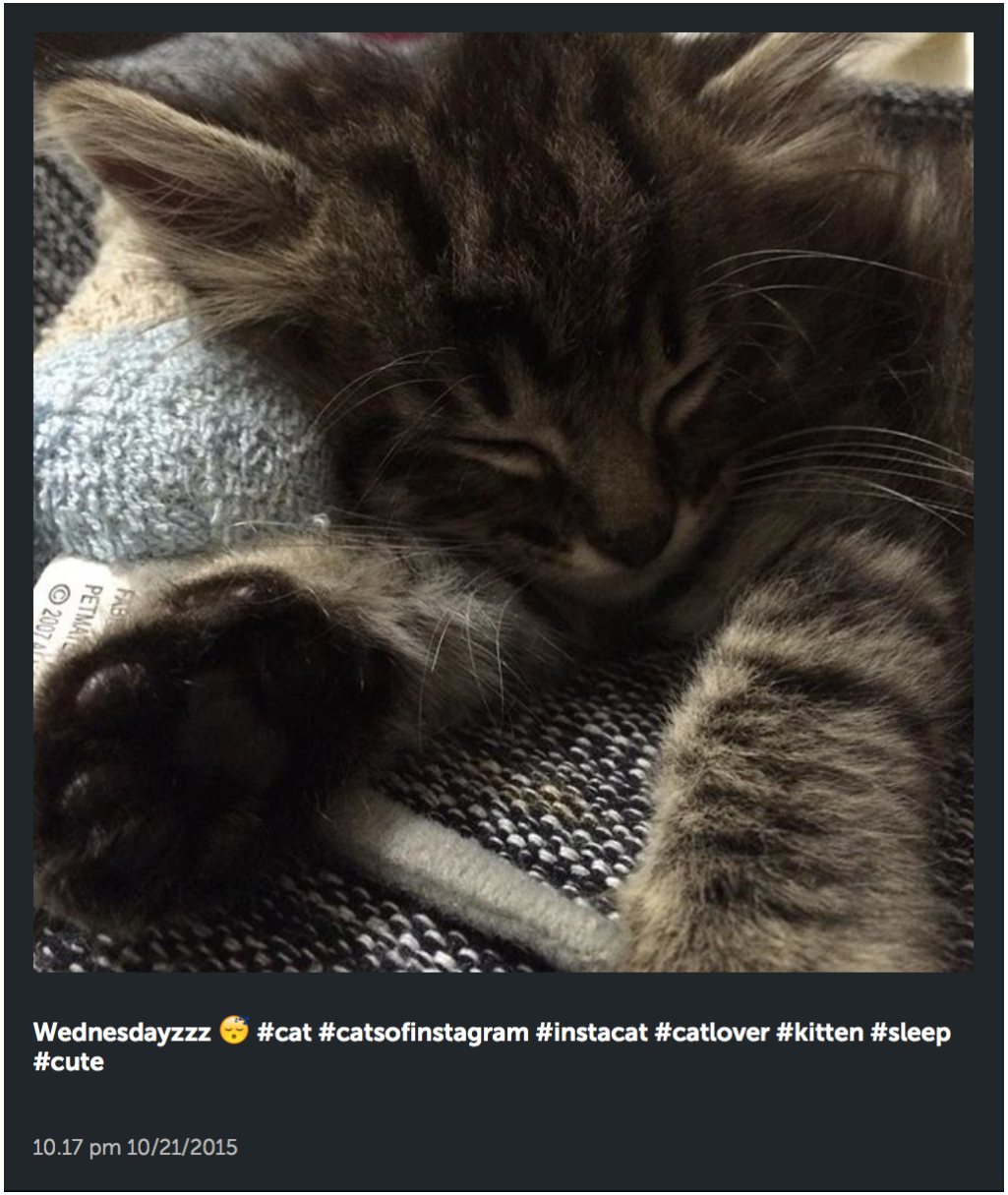
INTRODUCTION

- People want to share photos, and the process that goes from image capture to uploading to internet has become so smooth that even the least sophisticated user can do it.

- According to several estimations, every day hundreds of millions of photos are shared:
  - 50 millions photos are uploaded on Flickr
  - 80 millions on Instagram
  - 350 millions on Facebook

- All these services allow users to tag photos.
  Tagging, commenting and rating (or liking) is now a common practice.

EXAMPLES


Wednesdayzzz 😴 #cat #catsofinstagram #instacat #catlover #kitten #sleep #cute
10.17 pm 10/21/2015

EXAMPLES

Tags BETA Add tags

pentaxk10d 31mm

Beijing sonnet 116

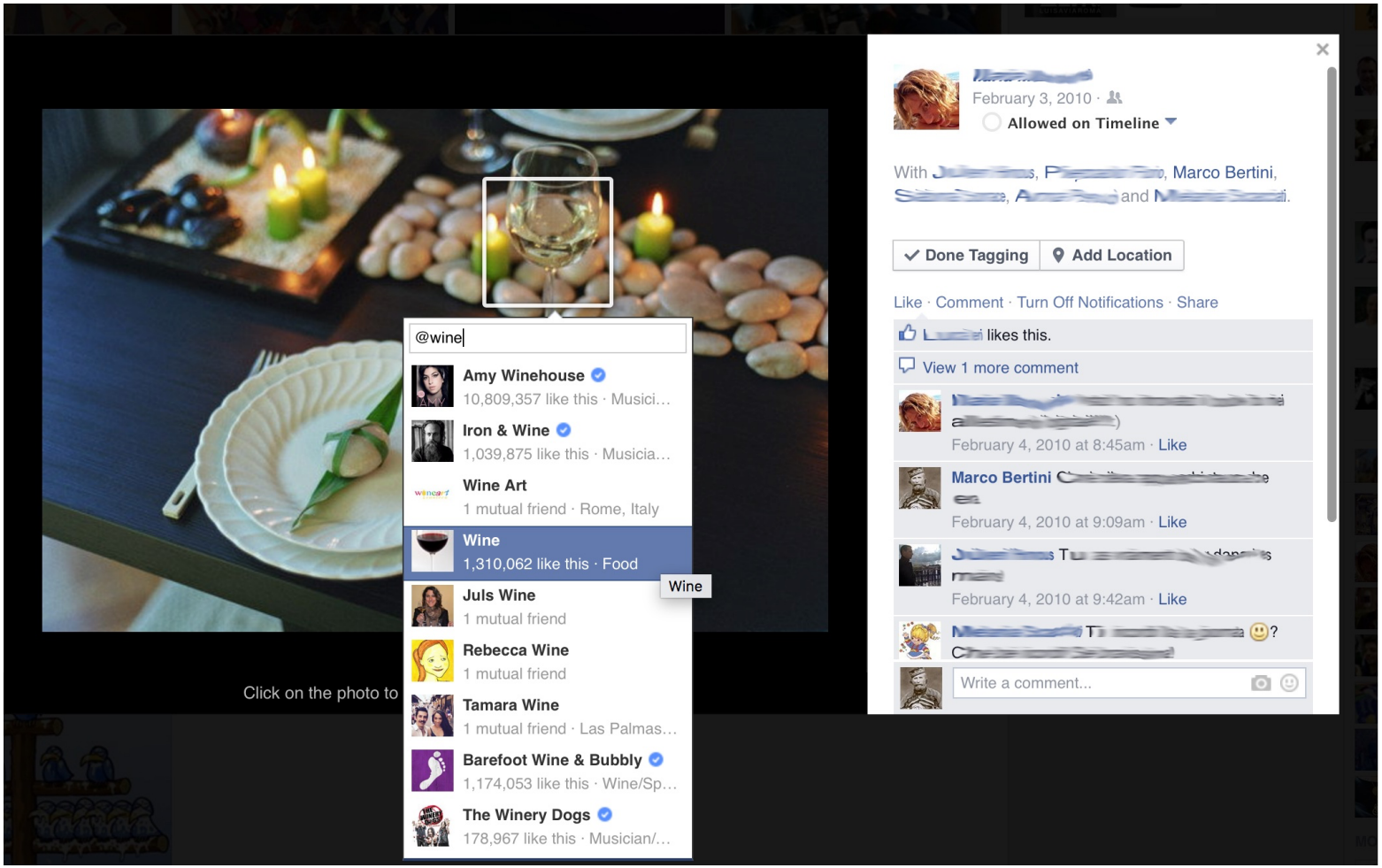
stealing photoshooting

chinese bride

Remember That Moment ...

People in photo Add people

EXAMPLES

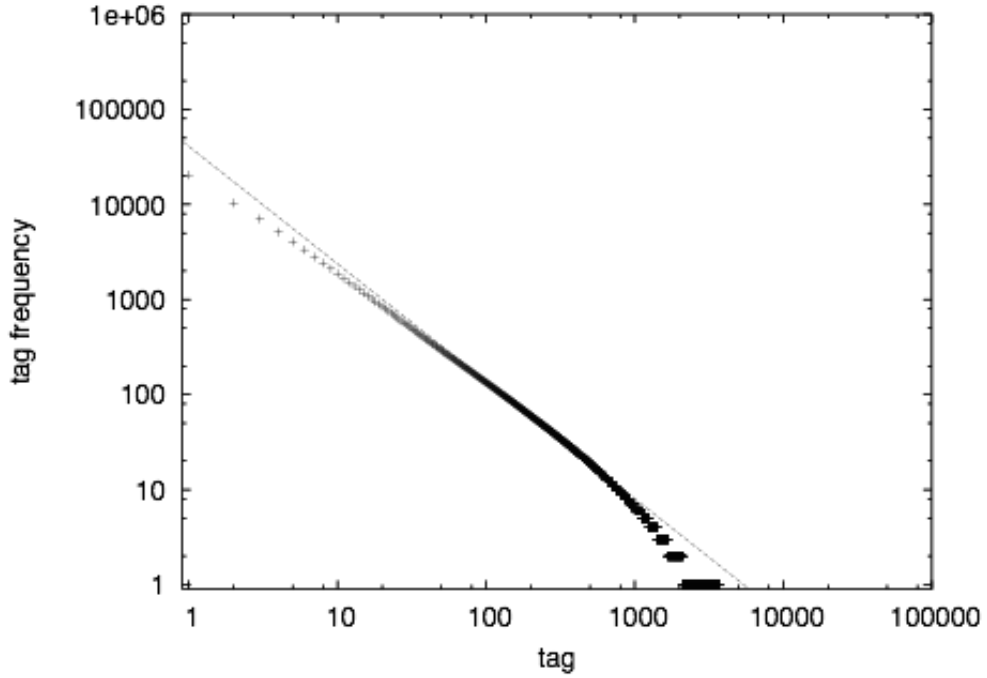

@wine
Amy Winehouse
10,809,357 like this · Musici…
Iron & Wine
1,039,875 like this · Musicia…
Wine Art
1 mutual friend · Rome, Italy
Wine
1,310,062 like this · Food
Wine
Juls Wine
1 mutual friend
Rebecca Wine
1 mutual friend
Tamara Wine
1 mutual friend · Las Palmas…
Barefoot Wine & Bubbly
1,174,053 like this · Wine/Sp…
The Winery Dogs
178,967 like this · Musician/…
Click on the photo to
February 3, 2010 ·
Allowed on Timeline
With Marco Bertini,
and
Done Tagging
Add Location
Like · Comment · Turn Off Notifications · Share
likes this.
View 1 more comment
February 4, 2010 at 8:45am · Like
Marco Bertini
February 4, 2010 at 9:09am · Like
February 4, 2010 at 9:42am · Like
Write a comment...

USER-GENERATED META-DATA

- The success of online social platforms and the availability of huge quantities of user-generated information motivates social image analysis, annotation and retrieval as important research topics for the multimedia community.

- Multimedia content and descriptions, location and comments in various forms (ranking, votes, likes) and associated metadata, social connections … are valuable resources for improving the results of tasks such as semantic indexing and retrieval.

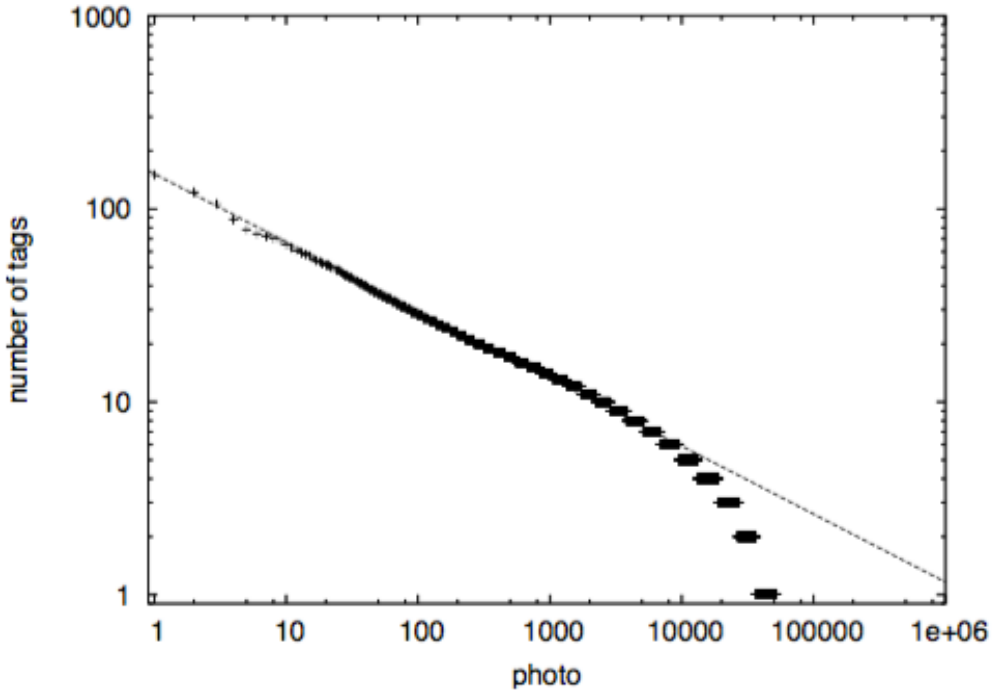
TAGGING BEHAVIOR



From: [Sigurbjörnsson and van Zwol 2008]

- Tag distribution in Flickr:
  - x-axis: the 3.7 million unique tags, ordered by descending tag frequency
  - y-axis: the tag frequency.
- The head of the distribution contains too generic tags to be useful (the top 5 most frequent: 2006, 2005, wedding, party, and 2004).
- The tail contains the infrequent tags with incidentally occurring terms such as misspellings and complex phrases.

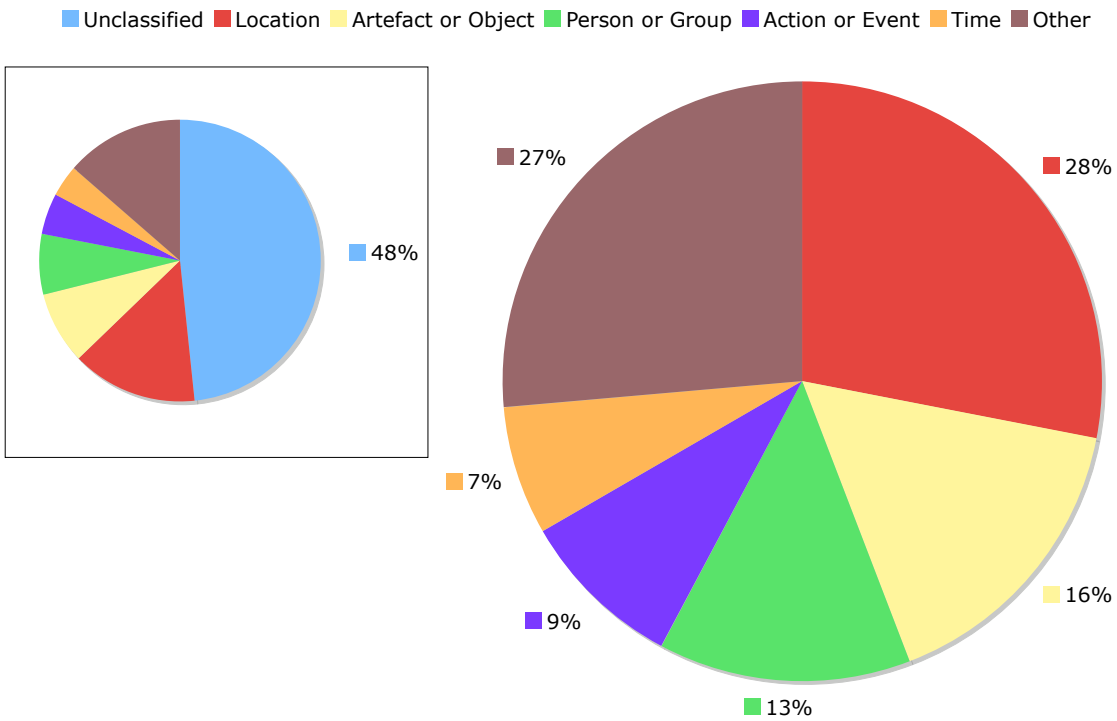
TAGGING BEHAVIOR



From: [Sigurbjörnsson and van Zwol 2008]

- distribution of the number of tags per photo in Flickr:
  - x-axis: 52 million photos
  - y-axis: number of tags per photo.
- A few photos are exceptionally tagged
- 64% of photos have 1, 2 or 3 tags only.

CATEGORIES OF TAGS



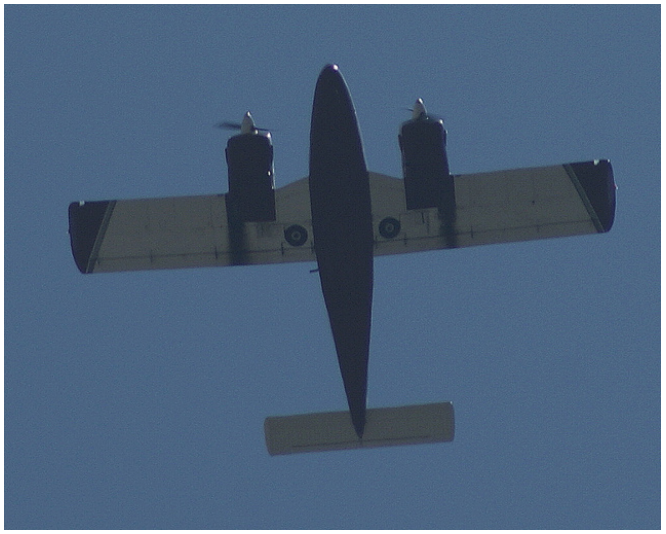
From: [Sigurbjörnsson and van Zwol 2008]

- The distribution of Flickr tags over the most common WordNet categories: selecting the highest ranked category, 52% of the tags is correctly classified, and 48% of the tags is left unclassified, of a 3.7M collection.

PROBLEMS

- Tags are few, imprecise, ambiguous and overly personalized [Golder and Huberman 2006; Sen et al. 2006; Sigurbjörnsson and van Zwol 2008; Kennedy et al. 2006].
- Tags might be irrelevant to the visual content.
- In a social network, users continuously add images and create new terms given the freedom of tagging.
- Web-scale quantity of media.

Query tag: airplane



airplane
twin
engine
los angeles
…

daytime
beach
airplane
ocean
…

TASK: TAG ASSIGNMENT

- Given an unlabeled image, tag assignment strives to assign a number of tags related to the image content
  - How many tags ? Fixed or variable number ?

Photo courtesy of Nicola Bertini (Flickr member: niK10d).

bride
bridegroom
wedding

TASK: TAG REFINEMENT

- Given an image associated with some initial tags, tag refinement aims to remove irrelevant tags from the initial tag list and enrich it with novel, yet relevant, tags.

Photo courtesy of Nicola Bertini (Flickr member: niK10d).

~~*stealing*~~
~~*sonnet*~~
photoshooting
~~*pentaxk10d*~~
~~*31mm*~~
bride
Chinese
bridegroom
photographer
wedding

TASK: TAG RETRIEVAL

- Given a tag and a collection of images labeled with the tag (and possibly other tags), the goal of tag retrieval is to retrieve images relevant with respect to the tag of interest.

Query: **bride**

Photo courtesy of Nicola Bertini (Flickr member: niK10d).

stealing
sonnet
photoshooting
pentaxk10d
31mm
bride
Chinese

wedding
father of the bride
bride
puglia
italianwedding
romance
romantic
bridegroom

ORGANIZATION OF THE TUTORIAL

- The tutorial is divided in 4 slots:

- Morning:
  - Introduction and overview of methods

  - Description of experimental setup and of the implemented methods

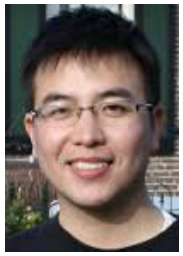
- Evening:
  - Practical session using open source implementations of selected methods

  - Final comments and related works; recap of hands-on session

IMAGE TAG ASSIGNMENT, REFINEMENT AND RETRIEVAL

ACM Multimedia 2015 Tutorial

October 26, 2015



Xirong Li
Renmin University of China

Tiberio Uricchio
University of Florence

Lamberto Ballan
University of Florence &
Stanford University

Marco Bertini
University of Florence

Cees Snoek
University of Amsterdam &
Qualcomm Research
Netherlands

Alberto Del Bimbo
University of Florence

PART 2
TAXONOMY

Lamberto Ballan
University of Florence &
Stanford University

Marco Bertini
University of Florence

- Foundations
  - tag relevance
- A two-dimensional taxonomy
  - Media for tag relevance
  - Learning for tag relevance

FOUNDATIONS

- The basic elements to be considered when developing methods for tag assignment, refinement and retrieval are:

- An image ***x***
- A tag ***t***
- A user ***u***

- A user ***u*** can share an image ***x***, assigning tag ***t*** to it
- A set of users ***U*** contributes a set of *n* socially tagged images ***X***, with $\boldsymbol{X_t}$ the set of images tagged with ***t***. All the tags used to describe ***X*** form a vocabulary ***V*** composed by *m* tags.
- Depending on the social network we can assume the availability of a set of user information $\boldsymbol{\theta}$ (e.g. user contacts, geo-localization, etc.)

TAG RELEVANCE

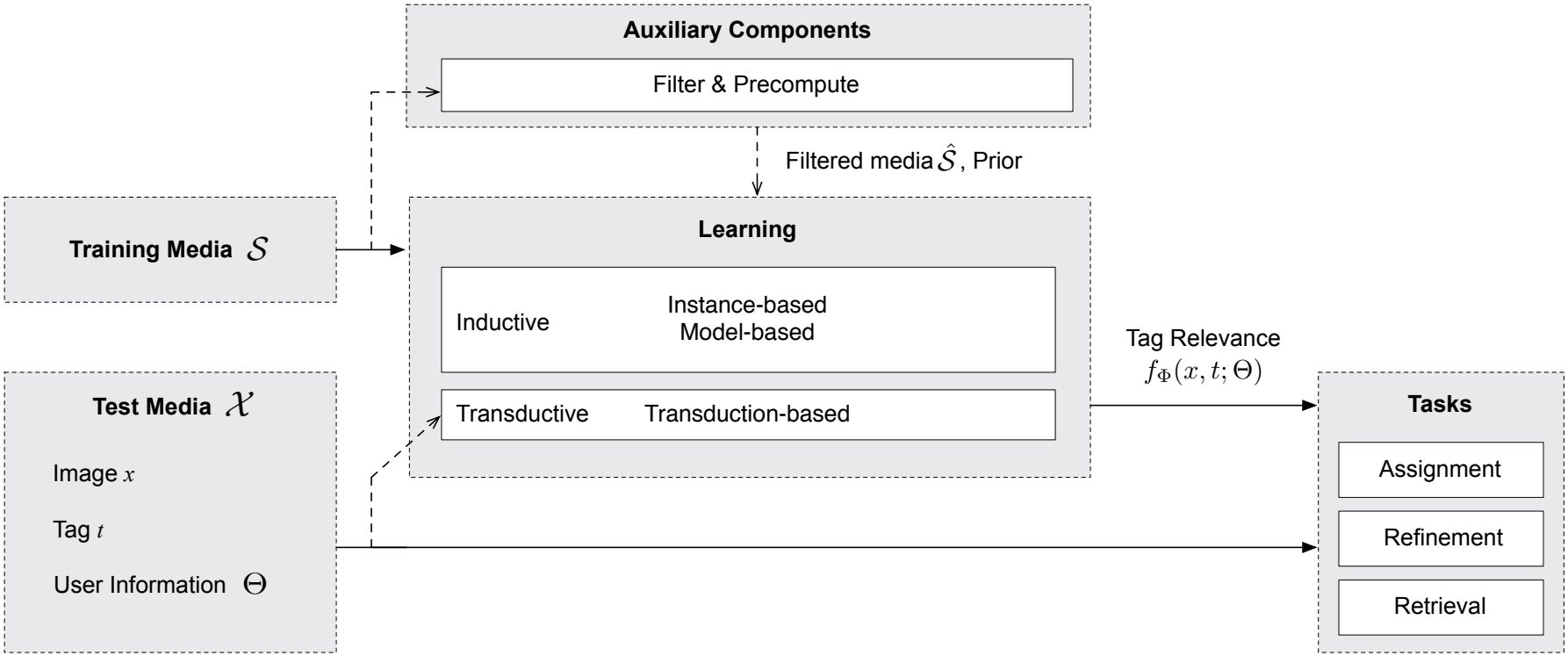
- Tag assignment, refinement and retrieval share a common essential component: a way to measure the relevance between a tag and a given image

- This function considers the image $\boldsymbol{x}$, tag $\boldsymbol{t}$ and user information $\boldsymbol{\theta}$:

$$f_{\phi}(x, t; \theta)$$

- Sorting $\boldsymbol{V}$ in descending order by $f_{\phi}(x, t; \theta)$ implements tag assignment and refinement
- Sorting $\boldsymbol{X}_t$ in descending order in terms of $f_{\phi}(x, t; \theta)$ implements retrieval
- **Note:** this formalization does not necessarily imply that the same implementation of tag relevance is applied for all the three tasks.

UNIFIED FRAMEWORK



- ***S*** is a set of training media obtained from social networks, i.e. with unreliable user-generated annotations. It can be optionally filtered to remove unwanted tags or images, obtaining ***Š***.

TAXONOMY

Media	Learning: *Instance-based*	Learning: *Model-based*	Learning: *Transduction-based*
tag	**[Sigurbjörnsson and van Zwol 2008]** [Sun et al. 2011] **[Zhu et al. 2012]**	[Xu et al. 2009]	–
tag + image	**[Liu et al. 2009]** **[Makadia et al. 2010]** [Tang et al. 2011] [Wu et al. 2011] [Yang et al. 2011] [Truong et al. 2012] [Qi et al. 2012] [Lin et al. 2013] [Lee et al. 2013] [Uricchio et al. 2013] [Zhu et al. 2014] [Ballan et al. 2014] [Pereira et al. 2014]	[Wu et al. 2009] **[Guillaumin et al. 2009]** [Verbeek et al. 2010] [Liu et al. 2010] [Ma et al. 2010] [Liu et al. 2011b] [Duan et al. 2011] [Feng et al. 2012] [Srivastava and Sala **[Chen et al. 2012]** [Lan and Mori 2013] **[Li and Snoek 2013]** [Li et al. 2013] [Wang et al. 2014] [Niu et al. 2014]	**[Zhu et al. 2010]** [Wang et al. 2010] [Li et al. 2010] [Zhuang and Hoi 2011] [Richter et al. 2012] [Kuo et al. 2012] [Liu et al. 2013] [Gao et al. 2013] [Wu et al. 2013] [ang et al. 2014] eng et al. 2014] [Xu et al. 2014]
tag + image + user	**[Li et al. 2009b]** [Kennedy et al. 20([Li et al. 2010] [Znaidia et al. 2013] [Liu et al. 2014]	[Sawant et al. 2010] [Li et al. 2011b] [McAuley and Leskovec 2012] [Kim and Xing 2013] [McParlane et al. 2013b] [Ginsca et al. 2014] [Johnson et al. 2015]	**[Sang et al. 2012a]** [Sang et al. 2012b] [Qian et al. 2015]

TagCooccur
SemanticField
TagRanking
KNN
TagProp
RobustPCA
TagFeature
RelExample
TagVote
TagCooccur+
TensorAnalysis

MEDIA FOR TAG RELEVANCE

- Depending on the modalities exploited we can divide the methods between those that use:

- Tags – e.g. considering ranking of tags as a proxy of user's priorities

- Tags and images – e.g. considering the set of tags assigned to an image

- Tags, images and user information – e.g. considering the behaviors of different users tagging similar images

TAG BASED

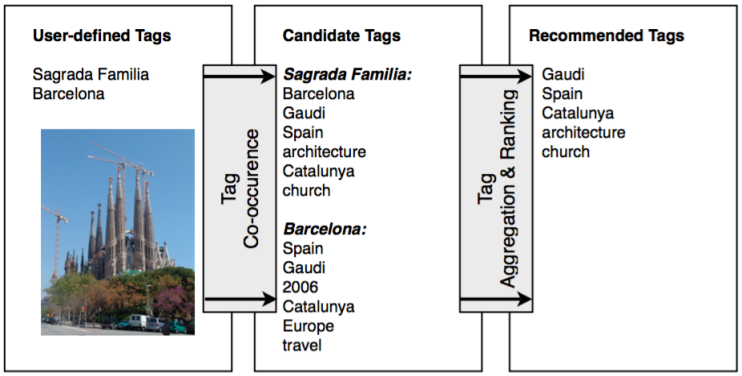
- These methods consider the original ranking of tags provided by users [Sun et al. 2011], tag co-occurrence [Sigurbjönsson and van Zwol 2008; Zhu et al. 2012] or topic modelling [Xu et al. 2009] to find semantically similar tags.

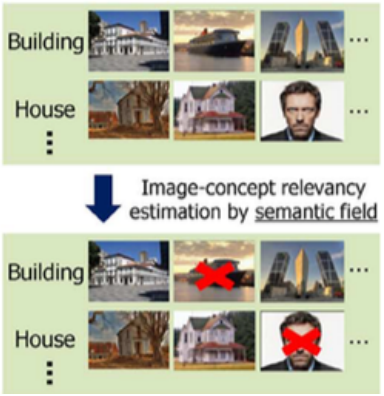
- These methods consider that the test image has already been labelled by the user so they can not be employed for **tag suggestion.**

TAG BASED

- We will review in detail:
- TagCooccur [Sigurbjörnsson and van Zwol 2008] that uses tag co-occurrence to create a list of candidate tags, aggregating them with through voting, and then weights the votes with a promotion function that accounts for characteristics like descriptiveness and statistical stability of tags.

- SemanticField [Zhu et al. 2012]: measures tag relevance in terms of an averaged semantic similarity between the tag and the other tags assigned to the image





TAG + IMAGE BASED

- These methods are the vast majority of those that we have analyzed in the review, and of those that have been re-implemented.
- The main idea of these works is to exploit visual consistency, i.e. the fact that visually similar images should have similar tags.
- Unlike previous methods they can be applied to tag suggestion.

- Three main approaches:
  - Use visual similarity between test image and database (e.g. [Li et al. 2009b; 2010; Verbeek et al. 2010; Ma et al. 2010; Wu et al. 2011; Feng et al. 2012])
  - Use similarity between images with same tags[Liu et al. 2009; Richter et al. 2012; Liu et al. 2011b; Kuo et al. 2012; Gao et al. 2013]
  - Learn classifiers from social images + tags[Wang et al. 2009a; Chen et al. 2012; Li and Snoek 2013; Yang et al. 2014]

TAG + IMAGE BASED

- The previously mentioned methods exploit image modality to compute the visual similarity, then use the tag modality in a subsequent step.

- A few methods use both modalities at the same time, creating a common latent-space, e.g. with Canonical Correlation Analysis[Pereira et al. 2014], building a unified graph composed by the fusion of a visual similarity graph with a image-tag connection graph[Ma et al. 2010], or using tag and image similarities as constraints to reconstruct a image-tag association matrix[Wu et al. 2013; Xu et al. 2014; Zhu et al. 2010].

TAG + IMAGE BASED

- These methods can be considered mainstream, and on the following will be reviewed in detail:

- TagRanking [Liu et al. 2009]
- KNN [Makadia et al. 2010]
- TagProp [Guillaumin et al. 2009; Verbeek et al. 2010]
- TagFeature [Chen et al. 2012]
- RelExample [Li and Snoek 2013]
- RobustPCA [Zhu et al. 2010]

TAG + IMAGE + USER INFORMATION

- Personal tagging behavior can be used in the form of tag statistics computed from images a user has uploaded in the past[Sawant et al. 2010; Li et al. 2011b], or learning a specific user embedding[Liu et al. 2014] .
- Another approach has been to combine tagging behavior of different users, e.g. to use more varied learning examples of different users[Li et al. 2009b] or keeping more robust tags that are used by different users for similar images[Kennedy et al. 2009].

- To discover latent relations between images, tags and user information several approaches use tensor analysis [Sang et al. 2012a, Qian et al. 2015].

TAG + IMAGE + USER INFORMATION

- Among the different types of metadata generated by users that have been exploited so fare we have:

- Photo time stamps [Kim and Xing 2013, McParlane et al. 2013a]

- Geo-localization[McParlane et al. 2013b]

- User interaction (e.g. comments)[Sawant et al. 2010] and group memberships[Wang et al. 2009b; McAuley and Leskovec 2012; Johnson et al. 2015]

TAG + IMAGE + USER INFORMATION

- The methods reviewed in detail in the following are:

- TagVote [Li et al. 2009b] + TagCooccur+ [Li et al. 2009b]: that use a unique-user constraint to create the visual neighborhood used in the voting algorithm, so to have a more objective voting and reduce the effect of batch tagging
- TensorAnalysis [Sang et al. 2012a]: that explicitly models the relation between users, tags and images

LEARNING FOR TAG RELEVANCE

- We can divide the learning methods in transductive and inductive. The former do not make a distinction between learning and test dataset, the latter may be further divided in methods that produce an explicit model and those that are instance based.

- We therefore divide the methods in instance-based, model-based and transduction-based.

- Typically inductive methods have better computational scalability than transductive ones.

INSTANCE BASED

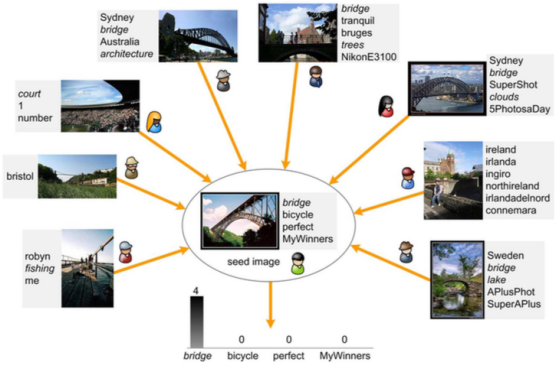
- This class of methods does not perform explicit generalization but, instead, compares new test images with training instances. There are no parameters and the complexity grows with the number of instances.

- In a neighbor voting approach[Li et al. 2009b, Li et al. 2010, Kennedy et al. 2009] it is estimated the relevance of tag ***t*** w.r.t. image ***x*** by counting the occurrence of ***t*** in the image's visual neighborhood.

- Weighted voting, e.g. using visual similarity, provides limited increases in performance.
- Improving the quality of the visual neighborhood improves the performance[Ballan et al. 2014]

INSTANCE BASED

- Of the implemented methods those following this approach are:



- KNN [Makadia et al. 2010]
- TagVote [Li et al. 2009b]
- TagCooccur+ [Li et al. 2009b]
- TagRanking [Liu et al. 2009]: that all build a visual neighborhood, to compute tag relevance

- Also the methods based on tags only (TagCooccur [Sigurbjörnsson and van Zwol 2008] and SemanticField [Zhu et al. 2012]) evaluate tag co-occurrence and similarity without building a model.

MODEL BASED

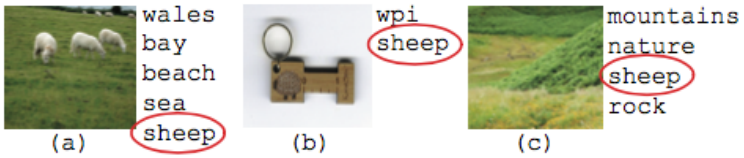
- This class of methods learns its parameters from a training set. A model can be tag-specific or holistic, i.e. for all tags.

- Methods of the first type are those of [Chen et al. 2012], that use linear SVMs trained on features augmented by pre-trained classifiers of popular tags, and [Li and Snoek 2013] that uses intersection kernel SVMs trained on relevant positive and negative examples and [Zhou et al. 2015] that treats tagged images as positive training examples and untagged images as candidate negative training examples.
- Examples of the second type use topic modelling[Wang et al. 2014], where relevance is computed using a topic vector of the image and a topic vector of the tag.

MODEL BASED

- The methods of this class is going to be analyzed in depth are:

- TagProp [Guillaumin et al. 2009; Verbeek et al. 2010]: that uses distance metric learning and a logistic model per tag to penalize frequent tags and promote rare ones.
- TagFeature [Chen et al. 2012]: that builds a two-class linear SVM for each tag from web images, extending pre-trained SVMs.
- RelExample [Li and Snoek 2013]: that proposes a system that selects positive and negative examples, deemed most relevant with respect to the given tag from crowd-annotated images, to train an ensemble of discriminative classifiers.



TRANSDUCTION BASED

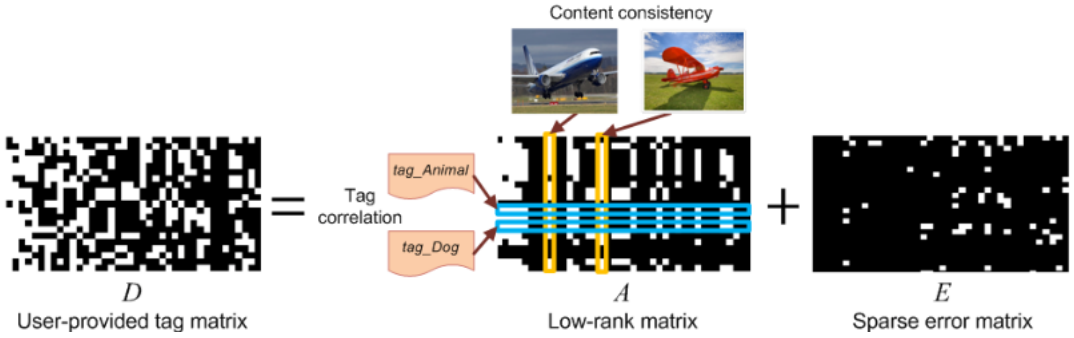
- This class of methods consists in procedures that evaluate tag relevance for a given image-tag pair of a set of images by minimizing some specific cost function.
- There's no separation between training and testing: a matrix ***D*** that associates all the images and tags of the dataset is the input of the method, while the output is anew matrix ***D**** whose elements are considered relevance scores.
- The majority of these methods is based on matrix factorization[Zhu et al. 2010, Sang et al. 2012a, Xu et al. 2014, Kalayeh et al. 2014].
- Graph-based label propagation is also used[Richter et al. 2012, Kuo et al. 2012], where image-tag pairs are represented as a graph in which each node corresponds to a specific image and the edges are weighted according to a multi-modal similarity measure.

TRANSDUCTION BASED

- The methods of this class is going to be analyzed in depth are:

- TensorAnalysis [Sang et al. 2012a]: that extend the ***D*** matrix to a tensor that comprises users.

- RobustPCA [Zhu et al. 2010]: that factorizes ***D*** by a low rank decomposition taking into account image and tag affinities.



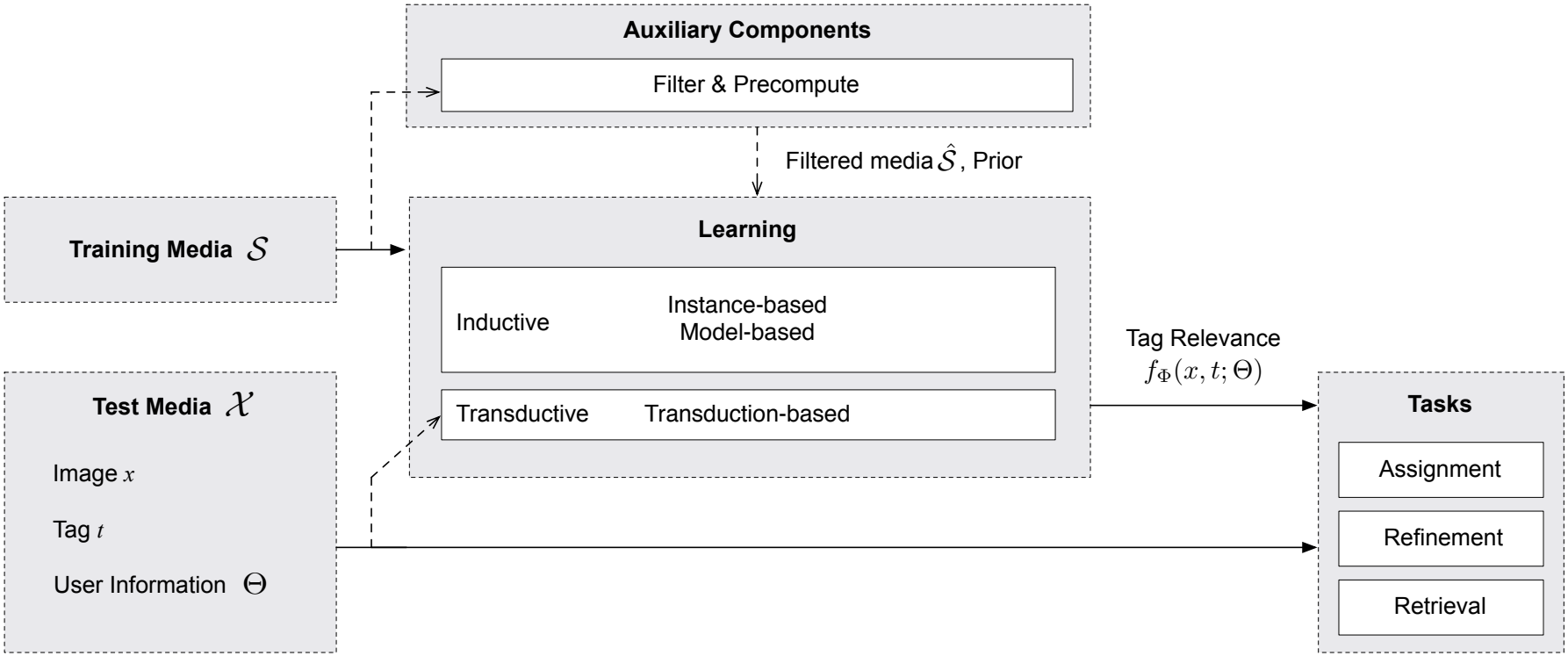
PROS AND CONS

- Instance-based methods:
  - Pro: flexible and adaptable to manage new images and tags.
  - Con: require to manage ***S***, a task that may become complex with increasing amount of data.

- Model-based methods:
  - Pro: training data is represented compactly, leading to swift computations, especially when using linear classifiers.
  - Con: need to retrain to cope with new imagery of a tag or when expanding the vocabulary ***V***.

- Transduction-based methods:
  - Pro: exploit better inter-tag and inter-image relationships, through matrix factorization.
  - Con: difficult to manage large datasets, because of memory or computational complexity.

UNIFIED FRAMEWORK



- ***S*** is a set of training media obtained from social networks, i.e. with unreliable user-generated annotations. It can be optionally filtered to remove unwanted tags or images, obtaining ***Š***.

AUXILIARY COMPONENTS: FILTER

- A common practice is to eliminate overly personalized tags (e.g. hadtopostsomething), e.g. excluding tags that are not part of WordNet or Wikipedia

- Often tags that do not appear enough times in the collection are eliminated.
- Reduction of vocabulary ***V*** size is also important for several methods that use image-tag association matrix, like the transductive methods(e.g. [Zhu et al. 2010; Sang et al. 2012a; Wu et al. 2013])
- Since batch tagging tends to reduce the quality of tags, these types of images can be excluded[Li et al. 2012]

AUXILIARY COMPONENTS: PRECOMPUTE

- It is practical to precompute information from ***S***, and use this information along with the refined media ***Š*** in the learning.

- The most common precomputation is tag occurrence and co-occurrence.

  Occurrence can be used to penalize excessively frequent tags[Li et al. 2009b].

  Co-occurrence is used to capture semantic similarity of tags directly from users' behavior

  A common method(e.g. in [Liu et al. 2009, Zhu et al. 2010, Zhu et al. 2012]) to obtain semantic similarity is to use Flickr context distance[Jiang et al. 2009], i.e. Normalized Google Distance computed on Flickr image collections.

REFERENCES

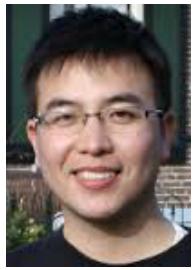
- All the references presented in these slides are available in:

X. Li, T. Uricchio, L. Ballan, M. Bertini, C.G.M. Snoek, A. Del Bimbo, “Socializing the Semantic Gap: A Comparative Survey on Image Tag Assignment, Refinement and Retrieval”, arXiv:1503.08248

IMAGE TAG ASSIGNMENT, REFINEMENT AND RETRIEVAL

ACM Multimedia 2015 Tutorial

October 26, 2015



Xirong Li
Renmin University of China

Tiberio Uricchio
University of Florence

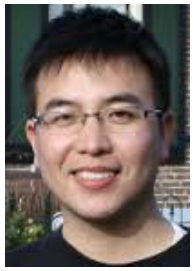
Lamberto Ballan
University of Florence &
Stanford University

Marco Bertini
University of Florence

Cees Snoek
University of Amsterdam &
Qualcomm Research
Netherlands

Alberto Del Bimbo
University of Florence

PART 3
A NEW EXPERIMENTAL PROTOCOL



Xirong Li
Renmin University of China

xirong@ruc.edu.cn

- Limitations in current evaluation
- Training and test data
- Evaluation setup

LIMITATIONS IN CURRENT EVALUATION

- Results are not directly comparable
  - homemade datasets
  - selected subsets of a benchmark set
  - varied implementation
    - preprocessing, parameters, features, …

- Results are not easily reproducible
  - For many methods, no source code or executable is provided.

- Single-set evaluation
  - Split a dataset into training/testing, at risk of overfitting

PROPOSED PROTOCOL

- Results are comparable
  - use full-size test datasets
  - same implemenation whenever applicable

- Results are reproducible
  - open-source

- Cross-set evaluation
  - Training and test datasets are constructed independently

SOCIALLY-TAGGED TRAINING DATA

- Data gathering procedure[Li et al. 2012]
  - using WordNet nouns as querie to uniformly sample Flickr images uploaded between 2006 and 2010
  - remove batch-tagged images (simple yet effective trick to improve data quality)

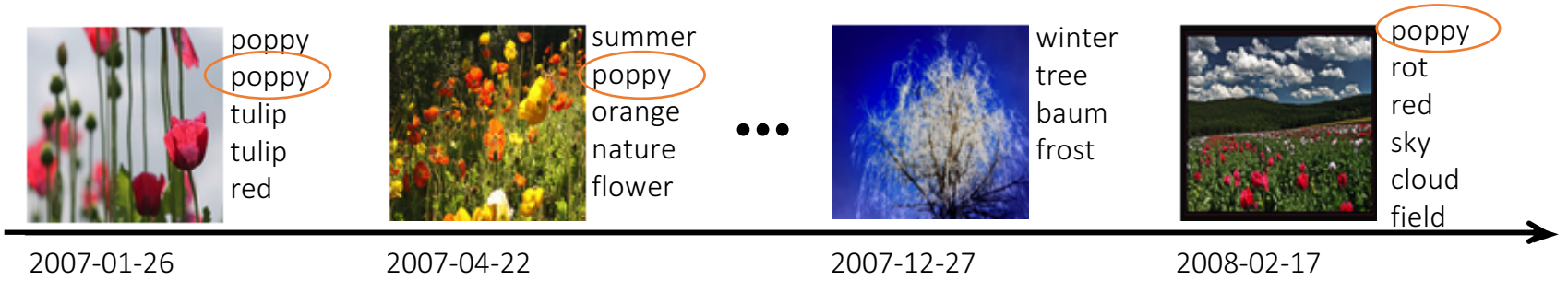
- Training sets of varied size
  - Train1M (a random subset of the collected Flickr images)
  - Train100k (a random subset of Train1m)
  - Train10k (a random subset of Train1m)

ImageNet already provides labeled examples for over 20k categories. Is it necessary to learn from socially tagged data?

SOCIAL TAGS VERUS IMAGENET ANNOTATIONS

- ImageNet annotations
  - computer vision oriented, focusing on fine-grained visual objects
  - single label per image

- Social tags
  - follow context, trends and events in the real world
  - describe both the situation and the entity presented in the visual content



A Flickr user's album

Credits: http://www.flickr.com/people/regina_austria

IMAGENET EXAMPLES ARE BIASED

- By web image search engines

(a) vehicles (b) carnivores

Figure from [Vreeswijk et al. 2012]

D. Vreeswijk, K. van de Sande, C. Snoek, A. Smeulders, All Vehicles are Cars: Subclass Preferences in Container Concepts, ICMR 2012

TEST DATA

- Three test datasets
  - contributed by distinct research groups

Test dataset	Contributors
MIRFlickr[Huiskes 2010]	LIACS Medialab, Leiden University
NUS-WIDE[Chua 2009]	LMS, National University of Sigapore
Flickr51[Wang 2010]	Microsoft Research Asia

MIRFLICKR

- Image collection
  - 25,000 high-quality photographic images from Flickr
- Labeling criteria
  - Potential labels: visibile to some extent
  - Relevant labels: saliently present
- Test tag set
  - 14 relevant labels: *baby bird car cloud dog flower girl man night people portrait river sea tree*
- Applicability
  - Tag assignment
  - Tag refinement

http://press.liacs.nl/mirflickr/

M. Huiskes, B. Thomee, M Lew, New trends and ideas in visual concept detection: the MIR Flickr retrieval evaluation initiative, MIR 2010

NUS-WIDE

- Image collection
  - 260K images randomly crawled from Flickr
- Labeling criteria
  - An active learning strategy to reduce the amount of manual labeling
- Test tag set
  - 81 tags containing objects (*car, dog*), people (*police, military*), scene (*airport, beach*), and events (*swimming, wedding*).
- Applicability
  - tag assignment
  - tag refinement
  - tag retrieval

T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y.-T. Zheng. NUS-WIDE: A Real-World Web Image Database from {National University of Singapore, CIVR 2009

http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

FLICKR51

- Image collection
  - 80k images collected from Flickr using a predefined set of tags as queries

- Labeling criteria
  - Given a tag, manually check the relevance of images labelled with the tag
  - Three relevance levels: very relevant, relevant, and irrelevant

- Test tag set
  - 51 tags, and some are ambiguous, e.g, apple, jaguar

- Applicability
  - Tag retrieval

[1] M. Wang, X.-S. Hua, H.-J. Zhang, Towards a relevant and diverse search of social images, IEEE Transactions on Multimedia 2010
[2] Y. Gao, M. Wang , Z.-J. Zha, J. Sheng, X. Li, X. Wu, Visual-Textual Joint Relevance Learning for Tag-Based Social Image Search, IEEE Transactions on Image Processing, 2013

VISUAL FEATURES

- Traditional bag of visual words[van de Sande 2010]
  - SIFT points quantized by a codebook of size 1,024
  - Plus a compact 64-d color feature vector[Li 2007]

- DeepNet feature
  - A 4,096-d FC7 vector after ReLU activation, extracted by the pre-trained 16-layer VGGNet[Simonyan 2015]

EVALUATION

Three tasks as introduced in Part 1

- Tag assignment
- Tag refinement
- Tag retrieval

EVALUATING TAG ASSIGNMENT

- A good method for tag assignment shall
  - rank relevant tags before irrelevant tags for a given image
  - rank relevant images before irrelevant images for a given tag

- Two criteria
  - Image-centric: Mean image Average Precision (MiAP)

$$iAP(x) := \frac{1}{R} \sum_{j=1}^{m_{gt}} \frac{r_j}{j} \delta(x, t_j)$$

  - Tag-centric: Mean Average Precision (MAP)

$$AP(t) := \frac{1}{R} \sum_{i=1}^{n} \frac{r_i}{i} \delta(x_i, t)$$

MiAP is biased towards frequent tags
MAP is affected by rare tags

EVALUATING TAG REFINEMENT

- Similar to tag assignment

EVALUATING TAG RETRIEVAL

- A good method for tag retrieval shall
  - rank relevant images before irrelevant images for a given tag

- Two criteria
  - Mean Average Precision (MAP) to measure the overall ranks

$$AP(t) := \frac{1}{R}\sum_{i=1}^{n}\frac{r_i}{i}\delta(x_i,t),$$

  - Normalized Discounted Cumulative Gain (NDCG) to measure the top ranks

$$NDCG_h(t) := \frac{DCG_h(t)}{IDCG_h(t)}, \qquad DCG_h(t) = \sum_{i=1}^{h}\frac{2^{rel_i}-1}{\log_2(i+1)}$$

SUMMARY

	Media characteristics				Tasks		
Media	**# images**	**# tags**	**# users**	**# test tags**	**assignment**	**refinement**	**retrieval**
Training media $\mathcal{S}$:							
Train10k	10,000	41,253	9,249	–	✓	✓	✓
Train100k	100,000	214,666	68,215	–	✓	✓	✓
Train1m [Li et al. 2012]	1,198,818	1,127,139	347,369	–	✓	✓	✓
Test media $\mathcal{X}$:							
MIRFlickr [Huiskes et al. 2010]	25,000	67,389	9,862	14	✓	✓	–
Flickr51 [Wang et al. 2010]	81,541	66,900	20,886	51	–	–	✓
NUS-WIDE [Chua et al. 2009]	259,233	355,913	51,645	81	✓	✓	✓

Data servers

[1] http://www.micc.unifi.it/tagsurvey
[2] http://www.mmc.ruc.edu.cn/research/tagsurvey/data.html

LIMITATIONS IN OUR PROTOCOL

- Tag informativeness in tag assignment

dog
pet

versus

dog
beach

X. Qian, X.-S. Hua, Y. Tang, T. Mei, Social Image Tagging With Diverse Semantics, IEEE Transactions on Cybernetics 2014

How to assess informativeness?

LIMITATIONS IN OUR PROTOCOL

- Image diversity in tag retrieval

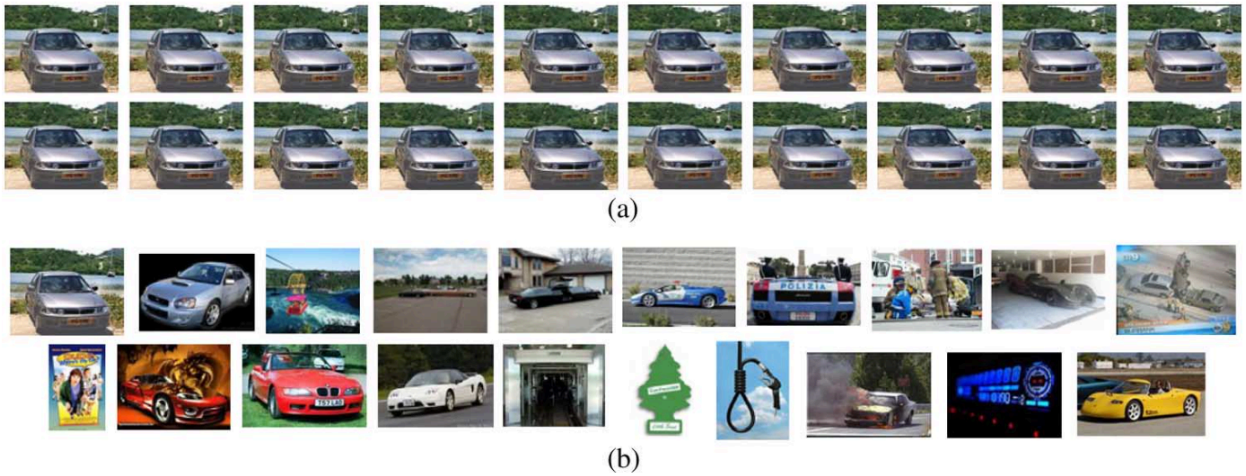


Figure from [Wang et al. 2010]

How to measure diversity?

M. Wang, X.-S. Hua, H.-J. Zhang, Towards a relevant and diverse search of social images, IEEE Transactions on Multimedia 2010

LIMITATIONS IN OUR PROTOCOL

- Semantic ambiguity
  - E.g., search for *jaguar* in Flickr51

SemanticField

RelExamples

Need fine-grained annotation

X. Li, S. Liao, W. Lan, X. Du, G. Yang, Zero-shot image tagging by hierarchical semantic embedding, SIGIR 2015

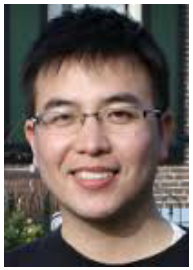
REFERENCES

- [Chua 2009] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y.-T. Zheng. NUS-WIDE: A Real-World Web Image Database from National University of Singapore, CIVR 2009
- [Huiskes 2010] M. Huiskes, B. Thomee, M Lew, New trends and ideas in visual concept detection: the MIR Flickr retrieval evaluation initiative, MIR 2010.
- [Li 2007] M. Li, Texture Moment for Content-Based Image Retrieval, ICME 2007
- [Li 2012] X. Li, C. Snoek, M. Worring, A. Smeulders, Harvesting social images for bi-concept search, IEEE Transactions on Multimedia 2012
- [Li 2015] X. Li, S. Liao, W. Lan, X. Du, G. Yang, Zero-shot image tagging by hierarchical semantic embedding, SIGIR 2015
- [Simonyan 2015] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR 2015
- [Qian 2014] X. Qian, X.-S. Hua, Y. Tang, T. Mei, Social Image Tagging With Diverse Semantics, IEEE Transactions on Cybernetics 2014
- [van de Sande 2010] K. van de Sande, T. Gevers, C. Snoek, Evaluating Color Descriptors for Object and Scene Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010
- [Vreeswijk 2012] D. Vreeswijk, K. van de Sande, C. Snoek, A. Smeulders, All Vehicles are Cars: Subclass Preferences in Container Concepts, ICMR 2012
- [Wang 2010] M. Wang, X.-S. Hua, H.-J. Zhang, Towards a relevant and diverse search of social images, IEEE Transactions on Multimedia 2010

IMAGE TAG ASSIGNMENT, REFINEMENT AND RETRIEVAL

ACM Multimedia 2015 Tutorial

October 26, 2015



Xirong Li
Renmin University of China

Tiberio Uricchio
University of Florence

Lamberto Ballan
University of Florence &
Stanford University

Marco Bertini
University of Florence

Cees Snoek
University of Amsterdam &
Qualcomm Research
Netherlands

Alberto Del Bimbo
University of Florence

PART 4
ELEVEN KEY METHODS

Tiberio Uricchio
University of Florence

tiberio.uricchio@unifi.it

- **Goal:** see several key methods of various Media and Learning
- Q: What are their key ingredients ?
- Q: How much do they cost computationally ?

KEY METHODS

- Covering all published methods is obviously impractical.

- We have to leave out methods that:
  - Do not show significant improvements or novelties w.r.t. the seminal papers in the field.
  - Methods that are difficult to replicate with the same mathematical preciseness as intended by their developers.

- We drive our choice by the intention to cover methods that aim for each of the three tasks, exploiting varied modalities by distinct learning mechanisms.

- 11 representative methods.

KEY METHODS

- Each method is required to output tag relevance of each test image and each test tag.

$$\begin{matrix} f(x_1,t_1) & f(x_1,t_2) & \dots & f(x_1,t_m) \\ f(x_2,t_1) & f(x_2,t_2) & \dots & f(x_2,t_m) \\ \vdots & \vdots & \ddots & \vdots \\ f(x_n,t_1) & f(x_n,t_2) & \dots & f(x_n,t_m) \end{matrix}$$

n images

m tags

KEY METHODS

Media \ Learning	Instance Based	Model Based	Transductive Based
Tag	**SemanticField** [Zhu et al. 2012] **TagCooccur** [Sigurbjörnsson and van Zwol 2008]		
Tag + Image	**TagRanking** [Liu et al. 2009] **KNN** [Makadia et al. 2010]	**TagProp** [Guillaumin et al. 2009] **TagFeature** [Chen et al. 2012] **RelExample** [Li and Snoek 2013]	**RobustPCA** [Zhu et al. 2010]
Tag + Image + User	**TagVote** **TagCooccur+** [Li et al. 2009b]		**TensorAnalysis** [Sang et al. 2012a]

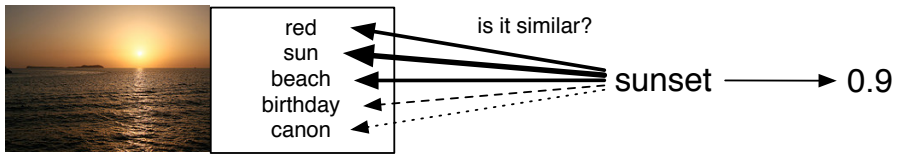
KEY METHODS

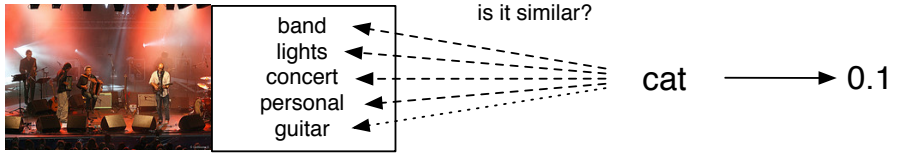
Media \ Learning	Instance Based	Model Based	Transductive Based
Tag	**SemanticField** [Zhu et al. 2012] **TagCooccur** [Sigurbjörnsson and van Zwol 2008]		
Tag + Image	**TagRanking** [Liu et al. 2009] **KNN** [Makadia et al. 2010]	**TagProp** [Guillaumin et al. 2009] **TagFeature** [Chen et al. 2012] **RelExample** [Li and Snoek 2013]	**RobustPCA** [Zhu et al. 2010]
Tag + Image + User	**TagVote** **TagCooccur+** [Li et al. 2009b]		**TensorAnalysis** [Sang et al. 2012a]

SEMANTICFIELD

[Zhu et al. 2012] Instance-Based Tag

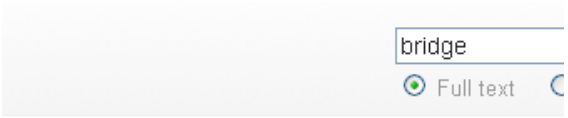
- Tags of similar semantics usually co-occur in user images.
- SemanticField measures an averaged similarity between a tag and the user tags already assigned to the image.
- Two similarity measures between words:
  - Flickr context similarity
  - Wu-Palmer similarity on WordNet



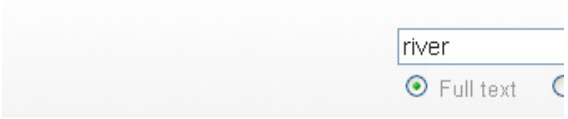


FLICKR CONTEXT SIMILARITY

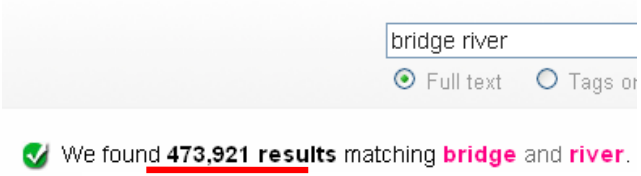
h(x)



h(y)



h(x,y)



FCS (bridge, river) = 0.65

- Based on the Normalized Google Distance.
- Measures the co-occurence of two tags with respect to the two single tag occurrencies.
- No semantics is involved, works for any tag.

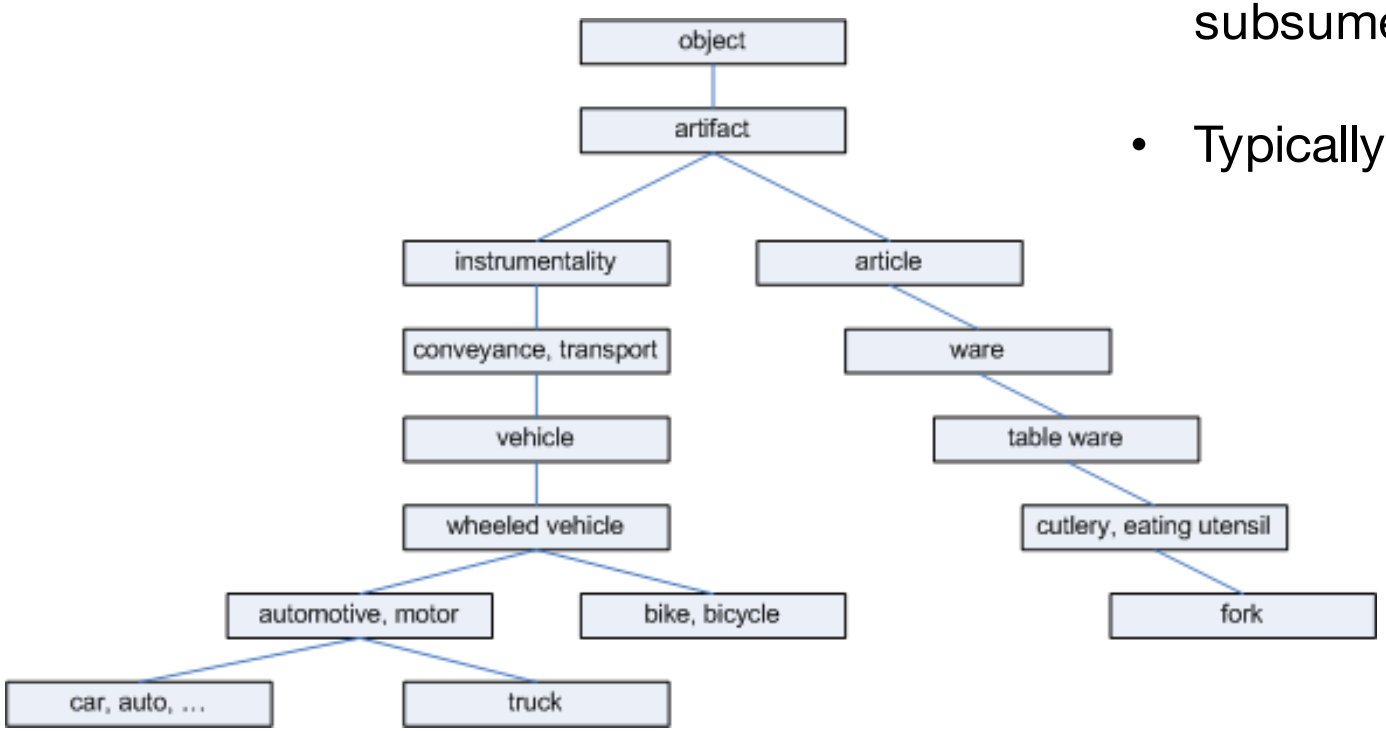
$$\mathrm{NGD}(x,y) = \frac{\max\{\log h(x), \log h(y)\} - \log h(x,y)}{\log N - \min\{\log h(x), \log h(y)\}},$$

$$\mathrm{FCS}(x,y) = e^{-\mathrm{NGD}(x,y)/\sigma}$$

[Jiang et al. 2009]

WU-PALMER SIMILARITY

$$\mathrm{Sim}(w_1, w_2) = \max \left[\frac{2 * \mathrm{depth}(\mathrm{LCS}(w_1, w_2))}{\mathrm{length}(w_1, w_2) + 2 * \mathrm{depth}(\mathrm{LCS}(w_1, w_2))} \right]$$



- It is a measure between concepts in an ontology restricted to taxonomic links.
- Considers the depth of x, y and their least common subsumer (LCS).
- Typically used with WordNet.

[Wu and Palmer 1994]

SEMANTICFIELD

[Zhu et al. 2012] | Instance-Based | Tag

$$f_{SemField}(x,t) := \frac{1}{l_x}\sum_{i=1}^{l_x} sim(t,t_i),$$

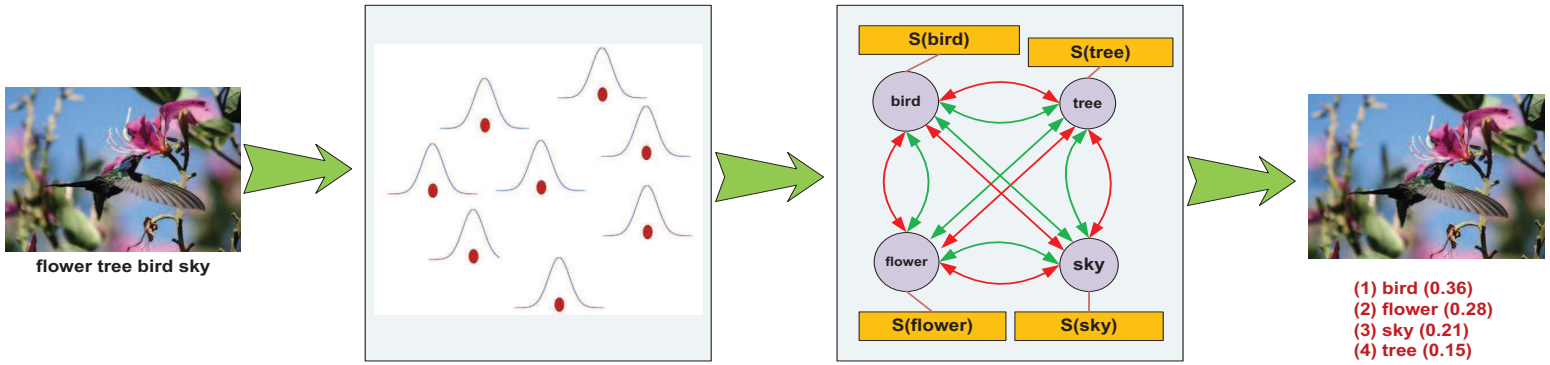
- *Sim* is the similarity between t and the other image tags.
- Needs some user tags. Not applicable to Tag Assignment.
- Complexity $O(m \cdot l_x)$ the number of image tags l_x times m tags
- Memory $O(m^2)$ quadratic in terms of vocabulary m tags

KEY METHODS

Media \ Learning	Instance Based	Model Based	Transductive Based
Tag	**SemanticField** [Zhu et al. 2012] **TagCooccur** [Sigurbjörnsson and van Zwol 2008]		
Tag + Image	**TagRanking** [Liu et al. 2009] **KNN** [Makadia et al. 2010]	**TagProp** [Guillaumin et al. 2009] **TagFeature** [Chen et al. 2012] **RelExample** [Li and Snoek 2013]	**RobustPCA** [Zhu et al. 2010]
Tag + Image + User	**TagVote** **TagCooccur+** [Li et al. 2009b]		**TensorAnalysis** [Sang et al. 2012a]

TAGRANKING

[Liu et al. 2009] | Instance-Based | Tag + Image



$p(t|x)$ Gaussian Kernel Density Estimation

Random walk on Tag graph

- TagRanking assigns a rank to each user tag, based on their relevance to the image content.
- Tag probabilities are first estimated in the KDE phase.
- Then a random walk is performed on a tag graph, built from visual exemplar similarity and tags semantic similarity.

TAGRANKING

[Liu et al. 2009] | Instance-Based | Tag + Image

- Suitable only for Tag Retrieval: it doesn't add or remove user tags.

$$f_{TagRanking}(x,t) = -rank(t) + \frac{1}{l_x},$$

- l_x is a tie-breaker when two images have the same tag rank.

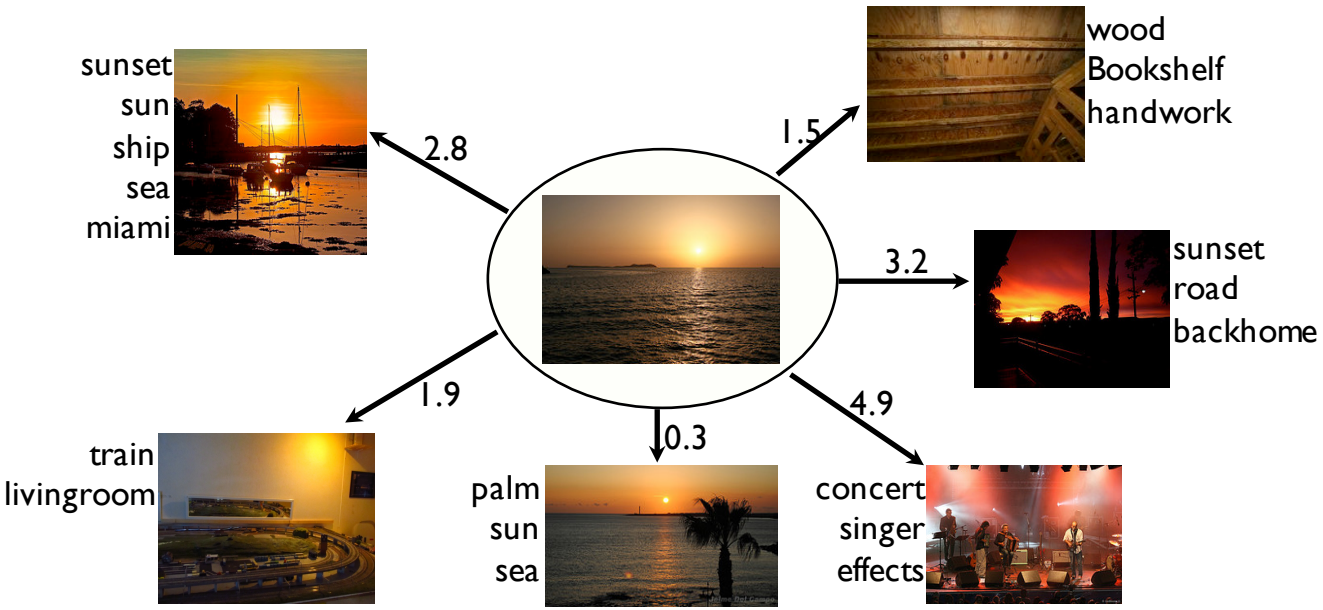
- Complexity $O(m \cdot d \cdot n + L \cdot m^2)$ – KDE on n images + L iter random walk
- Memory $O(\max(d \cdot n, m^2))$ – max of the two steps

KEY METHODS

Media \ Learning	Instance Based	Model Based	Transductive Based
Tag	**SemanticField** [Zhu et al. 2012] **TagCooccur** [Sigurbjörnsson and van Zwol 2008]		
Tag + Image	**TagRanking** [Liu et al. 2009] **KNN** [Makadia et al. 2010]	**TagProp** [Guillaumin et al. 2009] **TagFeature** [Chen et al. 2012] **RelExample** [Li and Snoek 2013]	**RobustPCA** [Zhu et al. 2010]
Tag + Image + Use	**TagVote** [Li et al. 2009b]		**TensorAnalysis** [Sang et al. 2012a]

KNN

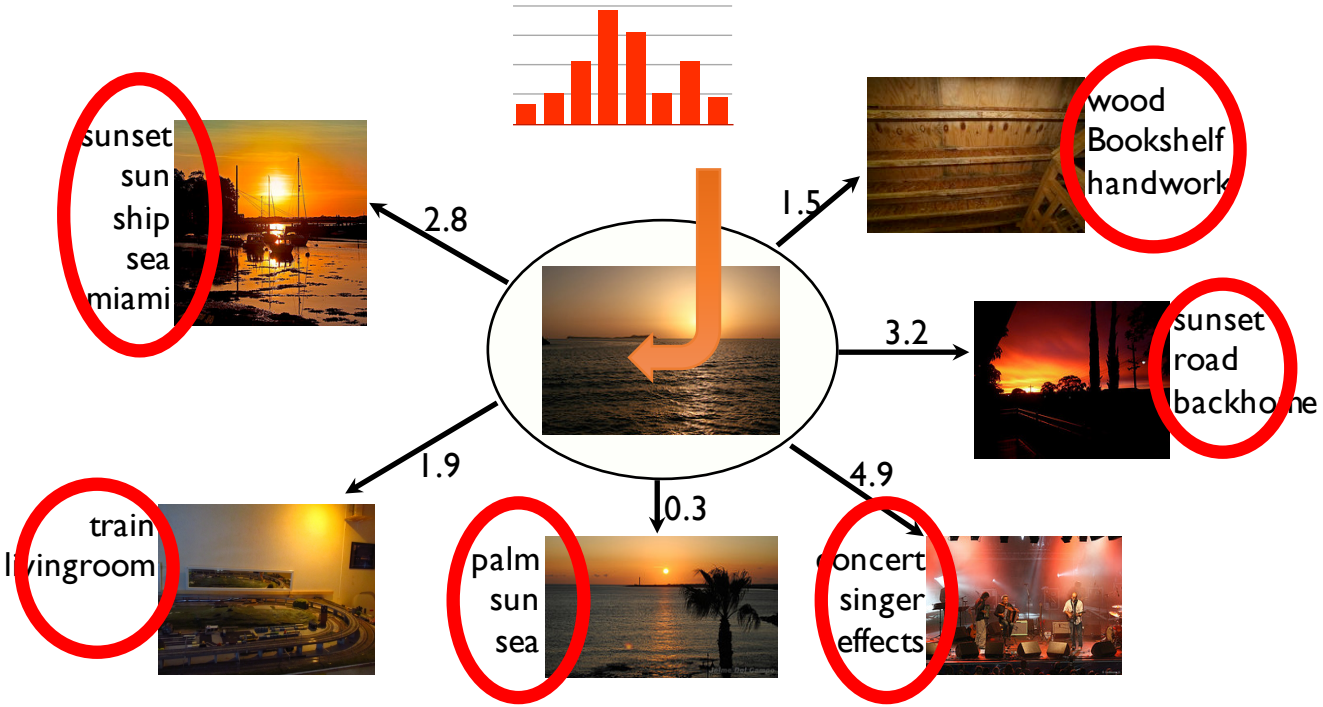
[Makadia et al. 2010] Instance-Based Tag + Image



- Similar images share similar tags.
- Finds k nearest images with a distance d.
- Counts the frequency of tags in the neighborhood.
- Assign the top ranked tags to the test image.

KNN

[Makadia et al. 2010] Instance-Based Tag + Image



- Similar images share similar tags.
- Finds k nearest images with a distance d.
- Counts the frequency of tags in the neighborhood.
- Assign the top ranked tags to the test image.

KNN

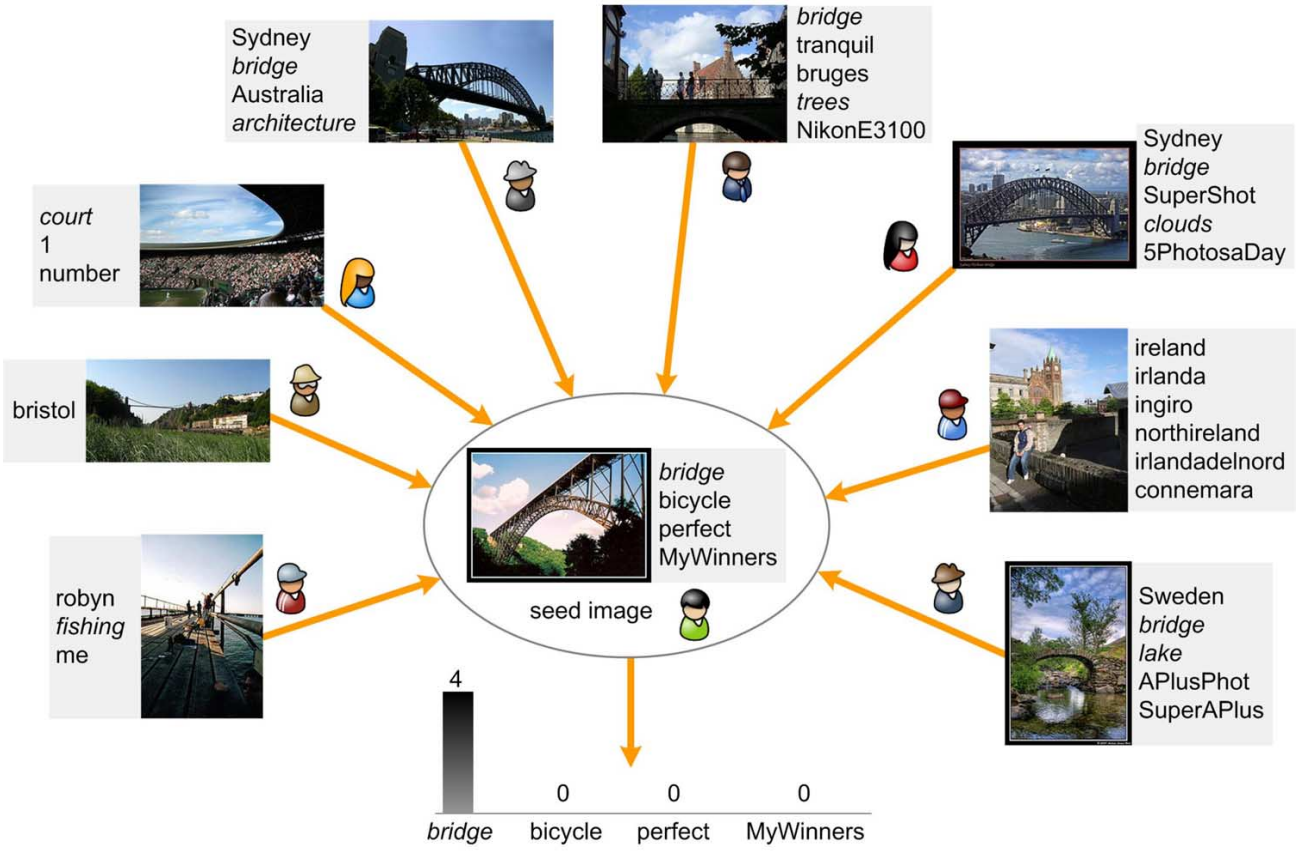
[Makadia et al. 2010] Instance-Based Tag + Image

$$f_{KNN}(x,t) := k_t,$$

- k_t is the number of images with t in the visual neighborhood of x.
- User tags on test image are not used. Not applicable to Tag Refinement.
- Complexity $O(d \cdot |S| + k \cdot \log|S|)$ – proportional to d feature dimensionality and k nearest neighbors.
- Memory $O(d \cdot |S|)$ – d-dimensional features.

TAGVOTE

[Li et al. 2009b] | Instance-Based | Tag + Image + User



- Adds two improvements w.r.t KNN:
  - Unique-user constraint
  - Tag prior frequency

TAGVOTE

[Li et al. 2009b] | Instance-Based | Tag + Image

$$f_{TagVote}(x,t) := k_t - k\frac{n_t}{|\mathcal{S}|},$$

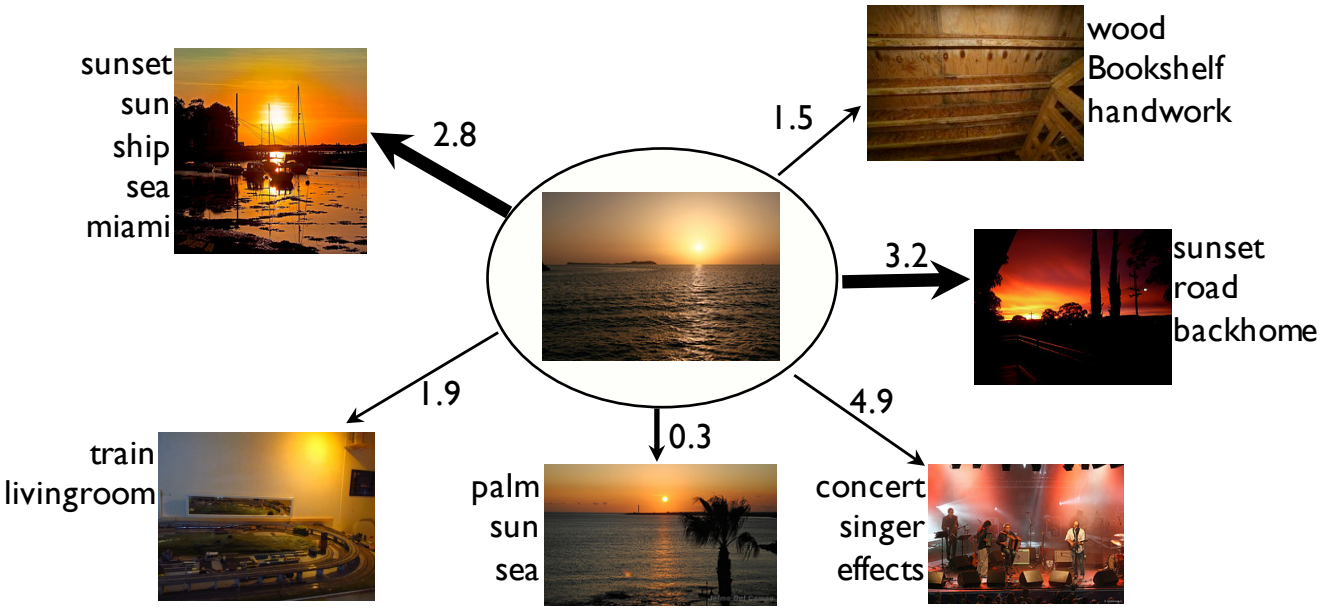
- k_t is the number of images with t in the visual neighborhood of x.
- n_t is the frequency of tag t in S.

- Like KNN, user tags on test image are not used. Not applicable to Tag Refinement.

- Complexity O(d · |S| + k · log|S|) – same complexity as KNN
- Memory O(d · |S|)

TAGPROP

[Guillaumin et al. 2009] Model-Based Tag + Image



- Differently from KNN, it gives different weights to images of the neighborhood.
- Probabilistic metric learning on image ranks or distance.

Probability of tag w on image I

$$p(y_{iw} = +1) = \sum_j \pi_{ij} p(y_{iw} = +1|j),$$

Probability of tag w on neighbor J

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

TAGPROP

[Guillaumin et al. 2009] Model-Based Tag + Image

$$f_{TagProp}(x,t) := \sum_{j}^{k} \pi_j \cdot \mathbf{I}(x_j,t),$$

- $I(x_j,t)$ returns 1 if x_j is labeled with t, 0 otherwise.

Rank weights

$$\pi_{ij} = \gamma_k$$

Distance weights

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

TAGPROP

[Guillaumin et al. 2009] Model-Based Tag + Image

- A logistic regressor per tag upon f_{TagProp}, is added to promote rare tags and penalize frequent ones.

$$f_{\text{TagProp}}(x,t) := \sigma\left(a_t \cdot \left(\sum_j^k \pi_j \cdot \mathbf{I}(x_j,t)\right) + b_t\right) \qquad \sigma(z) = \frac{1}{1+e^{-z}}$$

- User tags on test image are not used. Not applicable to Tag Refinement.

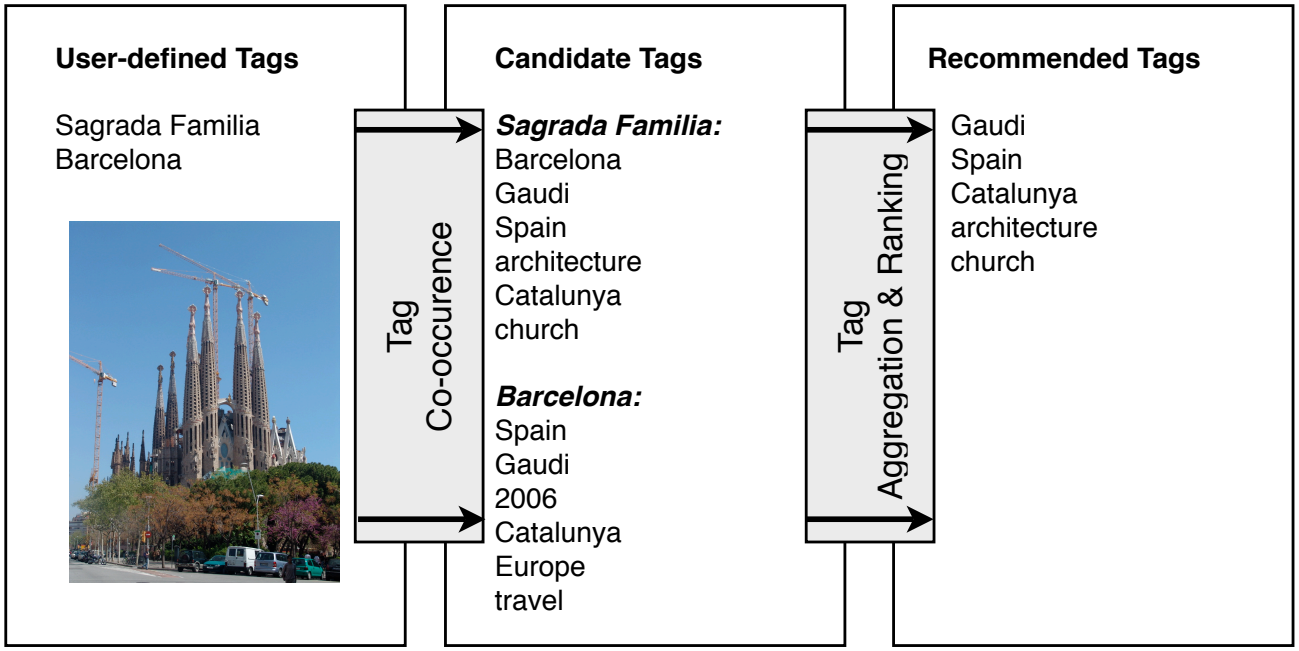
- Complexity O(l · m · k) – l steps of gradient descent
- Memory O(d · |S|) – same as KNN, extra 2m for logistic regression

KEY METHODS

Media \ Learning	Instance Based	Model Based	Transductive Based
Tag	**SemanticField** [Zhu et al. 2012] **TagCooccur** [Sigurbjörnsson and van Zwol 2008]		
Tag + Image	**TagRanking** [Liu et al. 2009] **KNN** [Makadia et al. 2010]	**TagProp** [Guillaumin et al. 2009] **TagFeature** [Chen et al. 2012] **RelExample** [Li and Snoek 2013]	**RobustPCA** [Zhu et al. 2010]
Tag + Image + User	**TagVote** **TagCooccur+** [Li et al. 2009b]		**TensorAnalysis** [Sang et al. 2012a]

TAGCOOCCUR

[Sigurbjörnsson and van Zwol 2008] Instance-Based Tag



- Refines user tags by looking for co-occurrences in training set.
- Tags are given a score based on an heuristic that takes into account ranks, stability and frequency of tags.

TAGCOOCCUR

[Sigurbjörnsson and van Zwol 2008] Instance-Based Tag

$$f_{tagcooccur}(x,t) = descriptive(t) \sum_{i=1}^{l_x} vote(t_i,t) \cdot rank\text{-}promotion(t_i,t) \cdot stability(t_i),$$

- *Descriptive* lowers the contribution of very high frequency tags.
- *Rank-promotion* measures tags contribution w.r.t tag ranks.
- *Stability* promotes tags for which statistics are more stable.
- *Vote* is 1 if t is among the 25 top ranked tags of t_i, 0 otherwise.

- Depends on user tags of the test image, not applicable to Tag Assignment.

- Complexity $O(m \cdot l_x)$ – same as SemanticField
- Memory $O(m^2)$

KEY METHODS

Media \ Learning	Instance Based	Model Based	Transductive Based
Tag	**SemanticField** [Zhu et al. 2012] **TagCooccur** [Sigurbjörnsson and van Zwol 2008]		
Tag + Image	**TagRanking** [Liu et al. 2009] **KNN** [Makadia et al. 2010]	**TagProp** [Guillaumin et al. 2009] **TagFeature** [Chen et al. 2012] **RelExample** [Li and Snoek 2013]	**RobustPCA** [Zhu et al. 2010]
Tag + Image + User	**TagVote** **TagCooccur+** [Li et al. 2009b]		**TensorAnalysis** [Sang et al. 2012a]

TagCooccur+

[Li et al. 2009b] | Instance-Based | Tag + Image

- A variant of TagCooccur that is improved by considering the image content in addition to solely user tags.
- The heuristic is updated by multipling TagCooccur score with a corrective factor based on Tag Vote scores.

$$f_{tagcooccur+}(x,t) = f_{tagcooccur}(x,t) \cdot \frac{k_c}{k_c + r_c(t) - 1},$$

- r_c is the rank of t when sorting $f_{tagvote}(x,t)$ in descending order. k_c is a positive weighting parameter.
- Complexity $O(d \cdot |S| + k \cdot \log|S|)$ – same complexity as TagVote
- Memory $O(d \cdot |S|)$

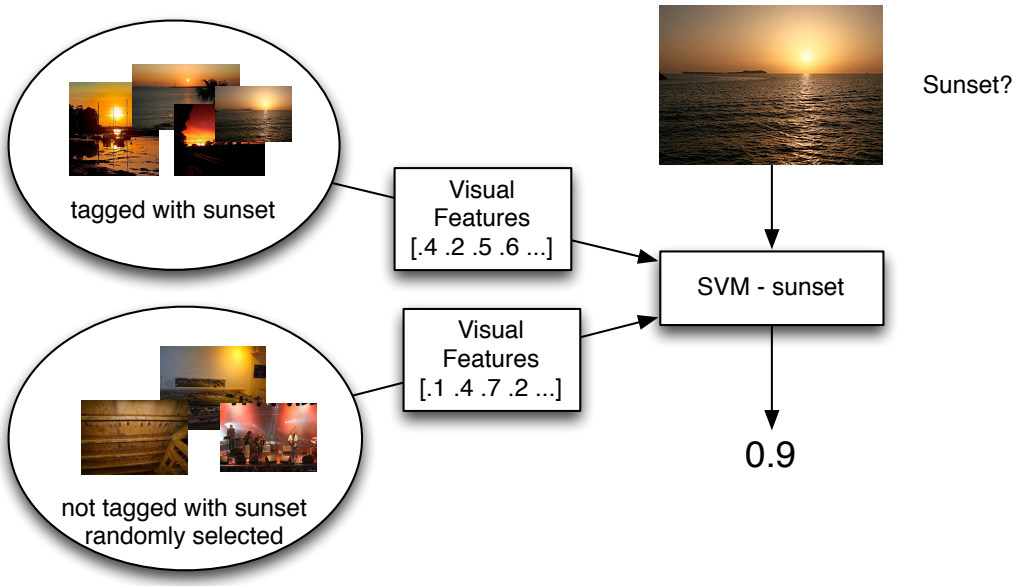
KEY METHODS

Media \ Learning	Instance Based	Model Based	Transductive Based
Tag	**SemanticField** [Zhu et al. 2012] **TagCooccur** [Sigurbjörnsson and van Zwol 2008]		
Tag + Image	**TagRanking** [Liu et al. 2009] **KNN** [Makadia et al. 2010]	**TagProp** [Guillaumin et al. 2009] **TagFeature** [Chen et al. 2012] **RelExample** [Li and Snoek 2013]	**RobustPCA** [Zhu et al. 2010]
Tag + Image + User	**TagVote** **TagCooccur+** [Li et al. 2009b]		**TensorAnalysis** [Sang et al. 2012a]

TAGFEATURE

[Chen et al. 2012] Model-Based Tag + Image

- Train per-tag classifier with tagged images as positive examples and random untagged images as negative examples.

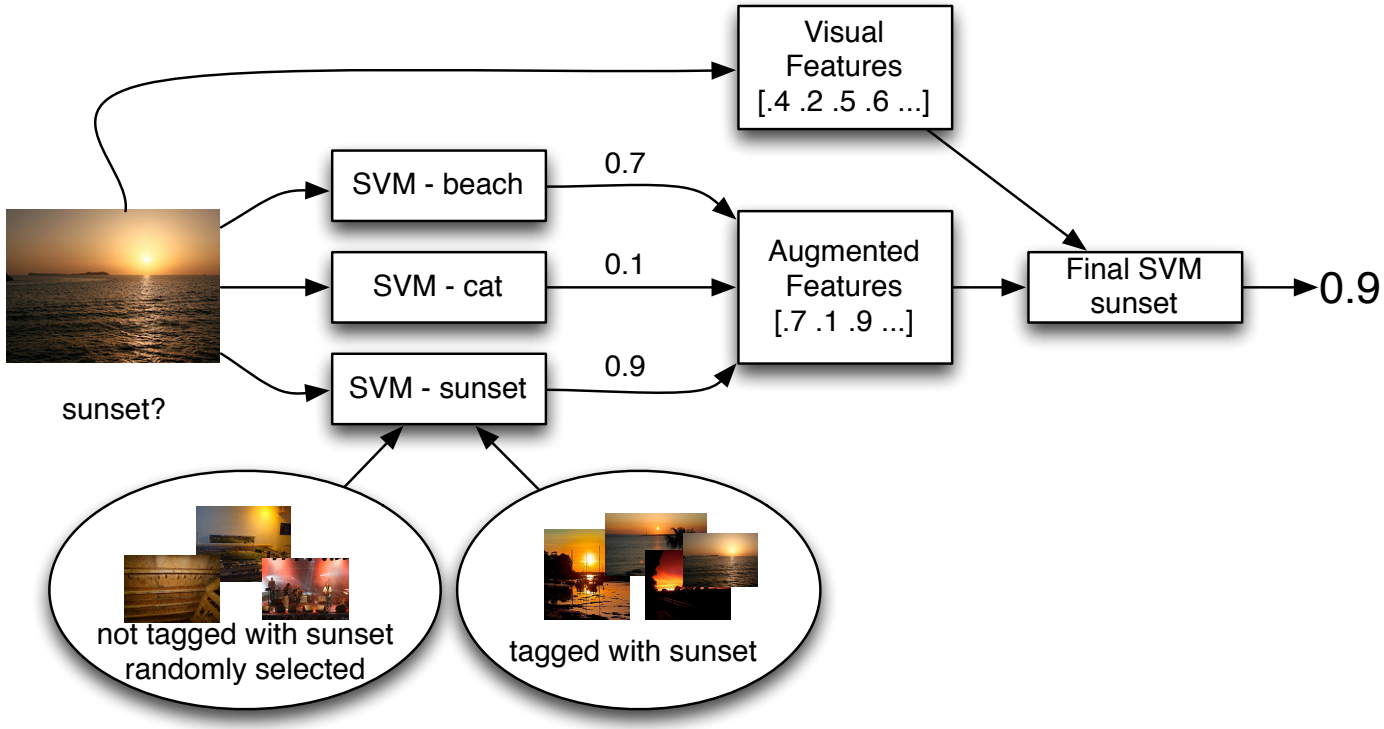


- Since rare tags are only associated with a limited number of positive training images, they may degrade SVM classifiers performance.

TAGFEATURE

[Chen et al. 2012] Model-Based Tag + Image

- TagFeature idea is to enrich visual features with tag augmented features, derived from prelearned SVM classifiers of popular concepts.



TAGFEATURE

[Chen et al. 2012] | Model-Based | Tag + Image

$$f_{TagFeature}(x,t) := b + < x_t, x >,$$

- Linear classifiers are used to reduce computational cost.
- It allows to sum up all the support vectors into a single vector x_t.
- d visual features and d' tag features i.e. svm classifiers.

- User tags on test image are not used. Not applicable to Tag Refinement.

- Complexity O((d + d') nm) – n images, m tags.
- Memory O(m (d + d')).

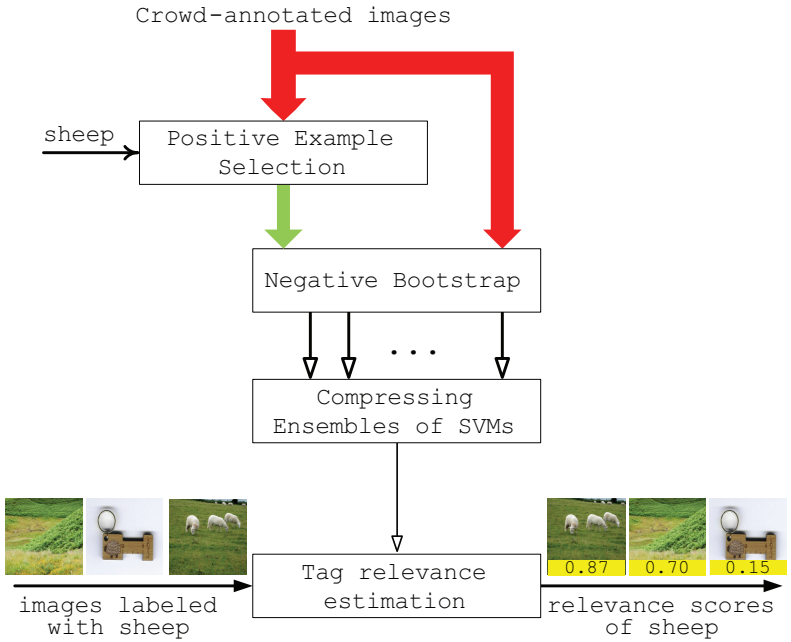
KEY METHODS

Media \ Learning	Instance Based	Model Based	Transductive Based
Tag	**SemanticField** [Zhu et al. 2012] **TagCooccur** [Sigurbjörnsson and van Zwol 2008]		
Tag + Image	**TagRanking** [Liu et al. 2009] **KNN** [Makadia et al. 2010]	**TagProp** [Guillaumin et al. 2009] **TagFeature** [Chen et al. 2012] **RelExample** [Li and Snoek 2013]	**RobustPCA** [Zhu et al. 2010]
Tag + Image + User	**TagVote** **TagCooccur+** [Li et al. 2009b]		**TensorAnalysis** [Sang et al. 2012a]

RELEXAMPLE

[Li and Snoek 2013] Model-Based Tag + Image

- A classifier tends to misclassify negative examples which are visually similar to positive examples.
- RelExample exploits positive and negative training examples which are deemed to be more relevant with respect to the test tag t.

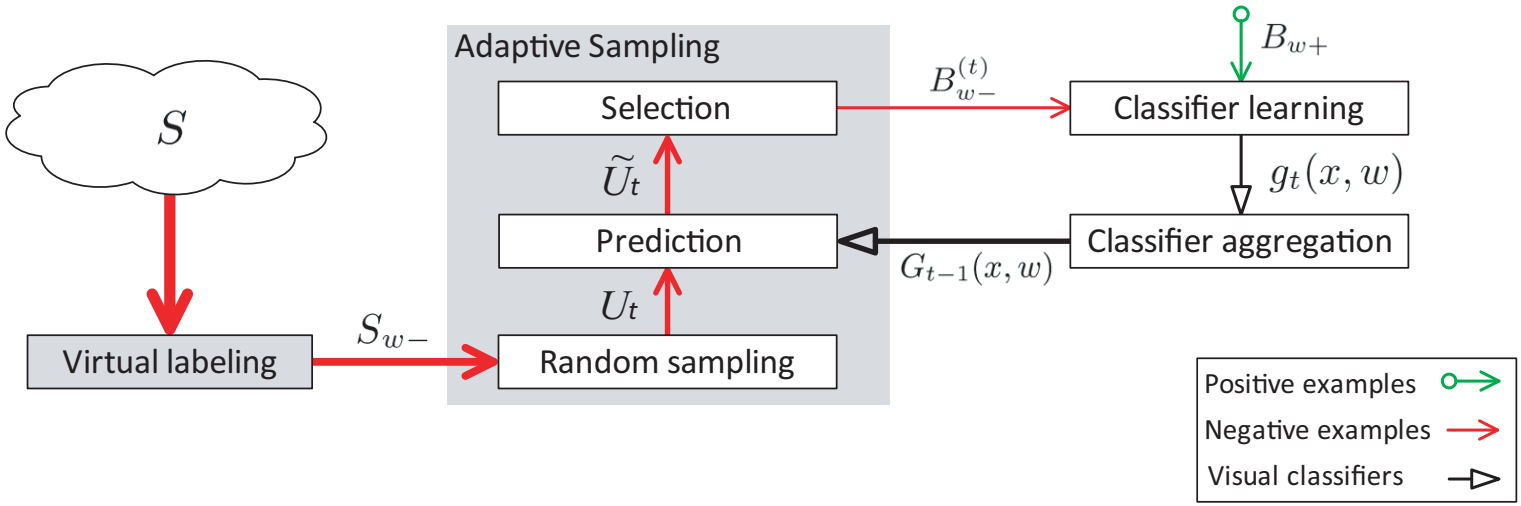


- Positive examples are selected by taking the top-ranked images by TagVote and SemanticField.
- Negative examples are selected by Negative Bootstrap [Li et al. 2013].

RELEXAMPLE

[Li and Snoek 2013] Model-Based Tag + Image

- Negative Bootstrap [Li et al. 2013] trains a series of classifiers g_t that explicitly address mis-classified examples at previous step.



$$G_t(x,w) = \frac{t-1}{t} G_{t-1}(x,w) + \frac{1}{t} g_t(x,w).$$

RELEXAMPLE

[Li and Snoek 2013] | Model-Based | Tag + Image

$$f_{RelExample}(x,t) := \frac{1}{T}\sum_{l=1}^{T}(b_l + \sum_{j=1}^{n_l}\alpha_{l,j}\cdot y_{l,j}\cdot \mathcal{K}(x,x_{l,j})),$$

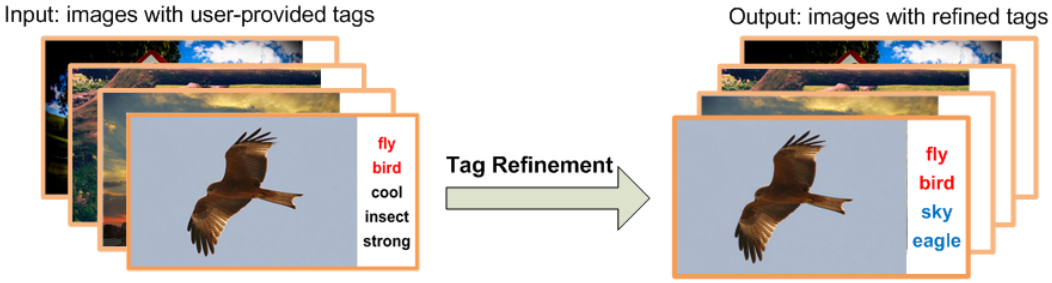
- T iterations for a corresponding number of trained classifiers.
- User tags on test image are not used. Not applicable to Tag Refinement.
- Complexity $O(Tdp^2)$ – training T svm classifiers
- Memory $O(dp + dq)$ – d visual features, p pos and q neg examples.

KEY METHODS

Media \ Learning	Instance Based	Model Based	Transductive Based
Tag	**SemanticField** [Zhu et al. 2012] **TagCooccur** [Sigurbjörnsson and van Zwol 2008]		
Tag + Image	**TagRanking** [Liu et al. 2009] **KNN** [Makadia et al. 2010]	**TagProp** [Guillaumin et al. 2009] **TagFeature** [Chen et al. 2012] **RelExample** [Li and Snoek 2013]	**RobustPCA** [Zhu et al. 2010]
Tag + Image + User	**TagVote** **TagCooccur+** [Li et al. 2009b]		**TensorAnalysis** [Sang et al. 2012a]

RobustPCA

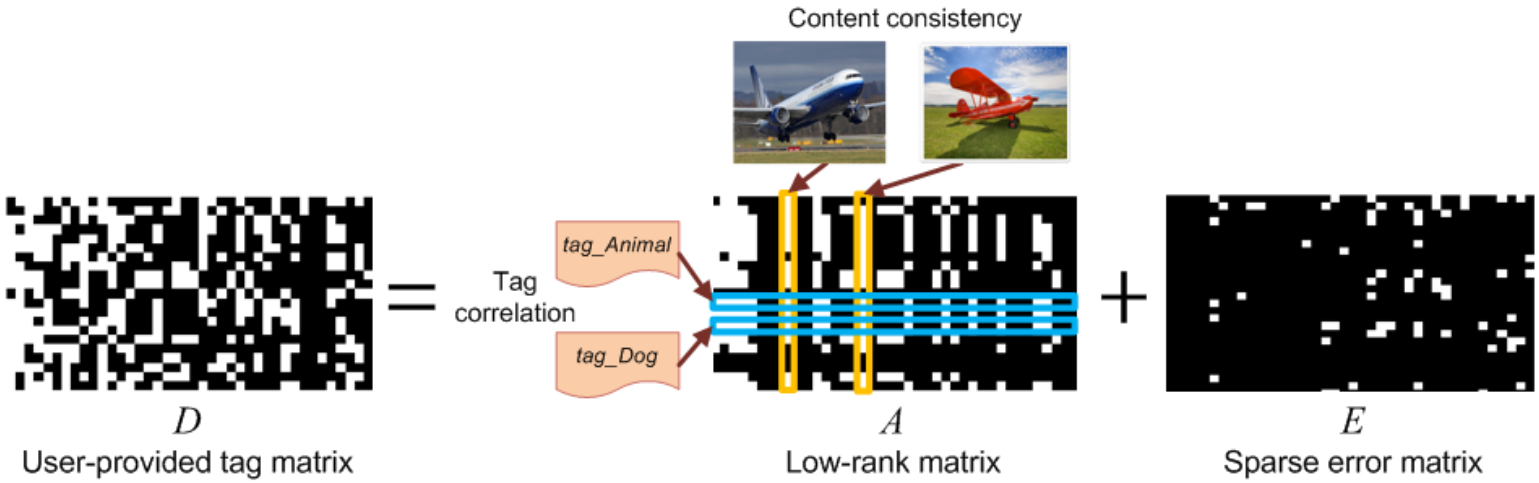
[Zhu et al. 2010] Transduction-Based Tag + Image



- Based on a few assumptions on tag characteristics:
  - *low-rank property*: the semantic space spanned by tags can be approximated by a smaller subset of salient words derived from the original space.
  - *tag correlation*: semantic tags are correlated.
  - *visual consistency*: visually similar images have similar tags.
  - *error sparsity for the image-tag matrix*: user's tagging is reasonably accurate and one image usually is labelled with few tags.

RobustPCA

[Zhu et al. 2010] | Transduction-Based | Tag + Image



- RobustPCA factorize the tag matrix D into a low-rank matrix A and a sparse error matrix E.
- Explicitly enforces content consistency and tag correlation with Laplacian graph-based regularizers.

RobustPCA

[Zhu et al. 2010] | Transduction-Based | Tag + Image

$$\min_{A,E} \quad ||A||_* + \lambda_1||E||_1 + \lambda_2[T_c(A) + T_t(A)]$$
$$\text{subject to} \quad D = A + E$$

- The problem reduces to recover the noise-free matrix A, so each column vector can be used to represent the corresponding images.
- T_c and T_t are regularizer based respectively on the similarity of images and tags.
- Complexity $O(cm^2n+c'n^3)$ – SVD computation
- Memory $O(cn \cdot m + c' \cdot (n^2 + m^2))$ – Full matrix D, tag and image similarity matrices.

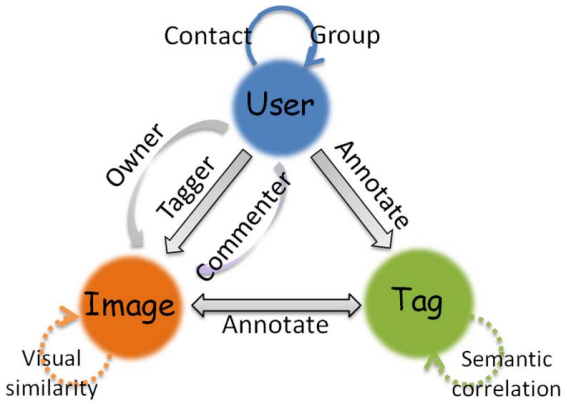
KEY METHODS

Media \ Learning	Instance Based	Model Based	Transductive Based
Tag	**SemanticField** [Zhu et al. 2012] **TagCooccur** [Sigurbjörnsson and van Zwol 2008]		
Tag + Image	**TagRanking** [Liu et al. 2009] **KNN** [Makadia et al. 2010]	**TagProp** [Guillaumin et al. 2009] **TagFeature** [Chen et al. 2012] **RelExample** [Li and Snoek 2013]	**RobustPCA** [Zhu et al. 2010]
Tag + Image + User	**TagVote** **TagCooccur+** [Li et al. 2009b]		**TensorAnalysis** [Sang et al. 2012a]

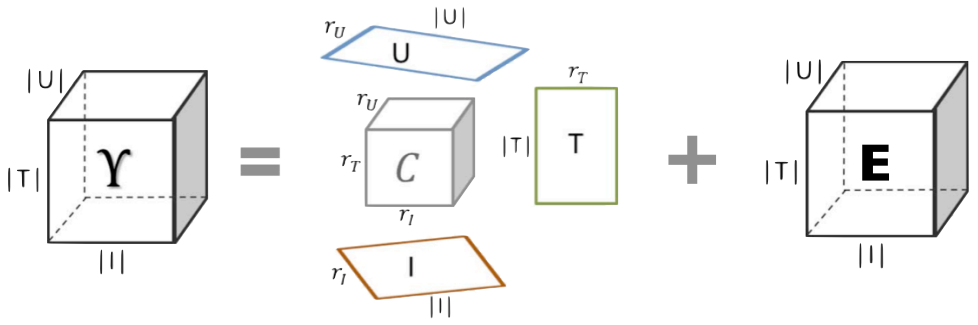
TENSORANALYSIS

[Sang et al. 2012a] Transduction-Based Tag + Image + User

- The method considers that, on top of visual appearance, images tagged by similar users can capture more semantic correlations.
- Jointly models the ternary relations between users, tags and images.
- It uses a tensor-based representation and Tucker decomposition to inference latent subspaces for the latent factors.



$$\text{tag}(u, i, t) \subseteq U \times I \times V_T$$



$$\underline{\mathbf{Y}} = \underline{\hat{\mathbf{C}}} \times_u \hat{\mathbf{U}} \times_i \hat{\mathbf{I}} \times_t \hat{\mathbf{T}} + \underline{\mathbf{E}}$$

$$y_{u,i,t} = \sum_{\tilde{u}} \sum_{\tilde{i}} \sum_{\tilde{t}} c_{\tilde{u},\tilde{i},\tilde{t}} \cdot u_{u,\tilde{u}} \cdot i_{i,\tilde{i}} \cdot t_{t,\tilde{t}}$$

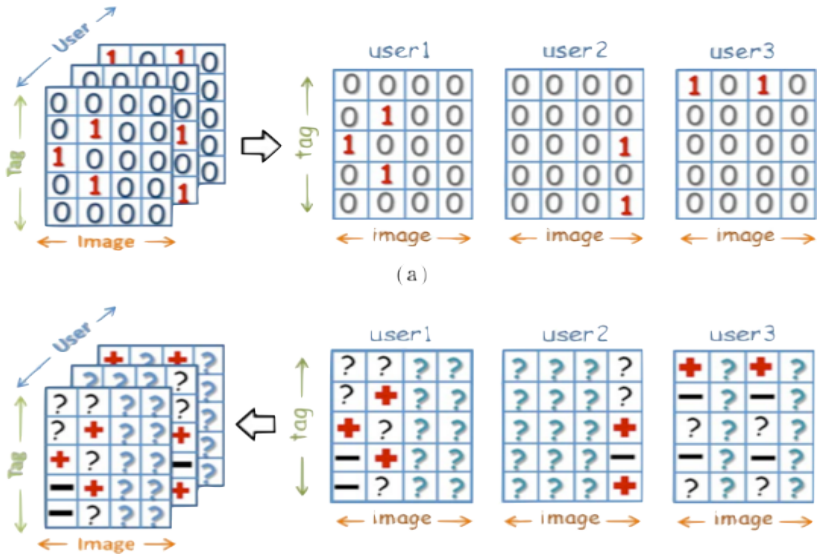
TENSORANALYSIS

[Sang et al. 2012a] | Transduction-Based | Tag + Image + User

- Only qualitative differences are important. The task is cast into a ranking problem to determine which tag is more relevant for a user to describe an image.

- Thus the method adopt a three state logic:
  - *positive tags*: tags assigned by the users,
  - *negative tags*: dissimilar tags that do not occur together with positive tags.
  - *neutral tags:* the other tags, removed from the learning process

Binary vs ternary logic



TensorAnalysis

[Sang et al. 2012a] | Transduction-Based | Tag + Image + User

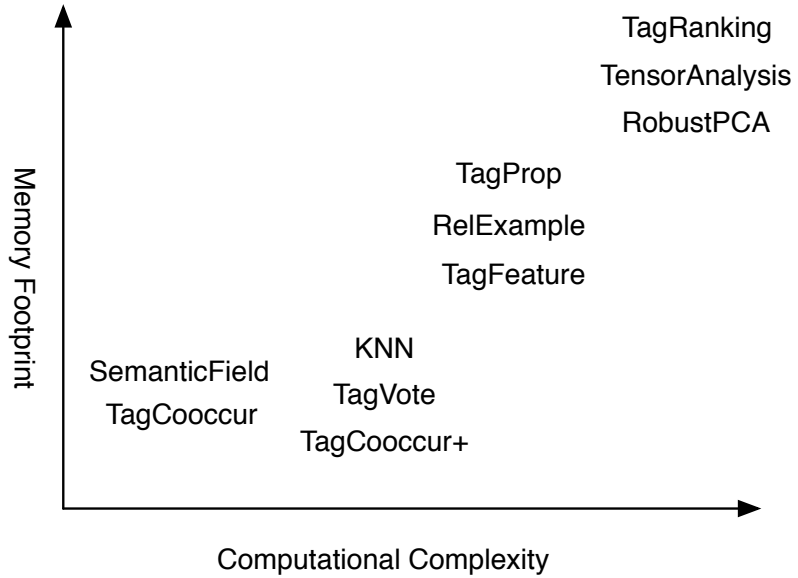
$$\underset{\theta}{\operatorname{argmin}} \sum_{t^+ \in T^+} \sum_{t^- \in T^-} H(\hat{y}_{t^-} - \hat{y}_{t^+}) \quad + \lambda_1(||\theta||^2) + \lambda_2(T_U(\theta) + T_I(\theta) + T_T(\theta))$$

$$\theta = \{U, I, T\}$$

- H is the heaviside function, $T_{\{U,I,T\}}$ are laplacian graph-based regularizers.
- Optimization is performed iteratively using stochastic gradient descent, one latent matrix at a time.

- Complexity $O(|P_1| \cdot (r_T \cdot m^2 + r_U \cdot r_I \cdot r_T))$ – P_1 is the ones in D, $r_{\{U,I,T\}}$ are latent matrices dimensionalities.
- Memory $O(n^2 + m^2 + u^2)$ – the three regularizers matrices.

COMPLEXITY CONSIDERATIONS



- SemanticField and TagCooccur have the best scalability with respect to both computation and memory.
- The model-based methods require less memory and run faster in the test stage, but at the expense of SVM model learning in the training stage.
- The two transduction-based methods have limited scalability, and can operate only on small sized S.

EVALUATION

- We tested the eleven methods on the proposed testbed.

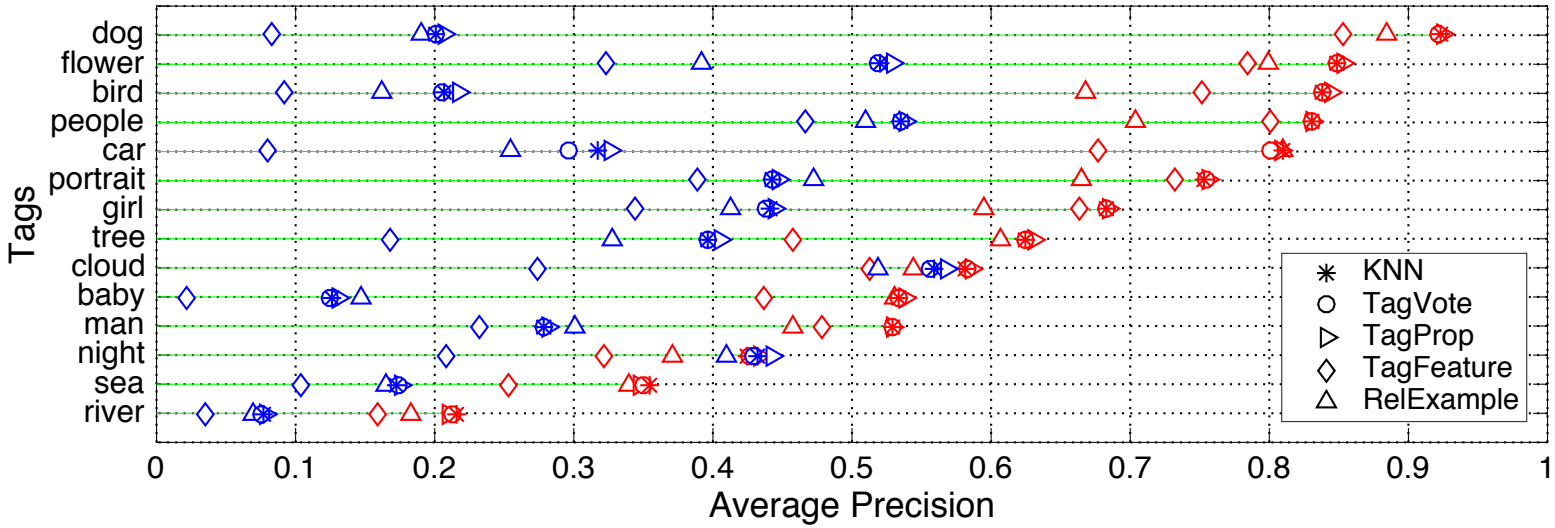
	Assignment	Refinement	Retrieval
KNN	X		X
TagVote	X		X
TagProp	X		X
TagFeature	X		X
RelExample	X		X
TagCooccur		X	X
TagCooccur+		X	X
RobustPCA		X	X
TensorAnalysis		X	X
SemanticField			X
TagFeature			X

- Here we discuss few main results. Refer to our survey paper for the complete picture.

TAG ASSIGNMENT

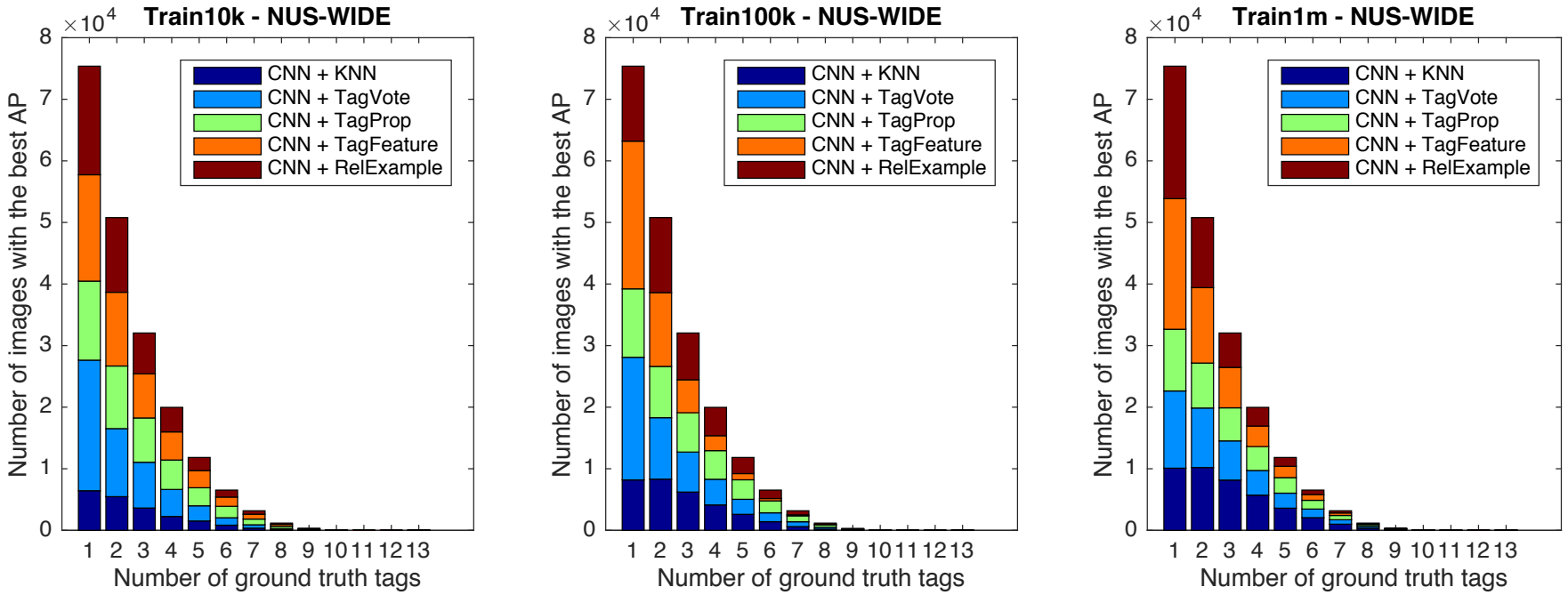
MIRFlickr test set, trained on Train1m.

CNN Features
BovW Features



- All methods benefit from using CNN Features.
- RelExample has better performance than TagFeature due to its filtering component.
- TagProp has the best MAP. Its performance is similar to KNN, TagVote since they all use the same basic nearest-neighbor label propagation.

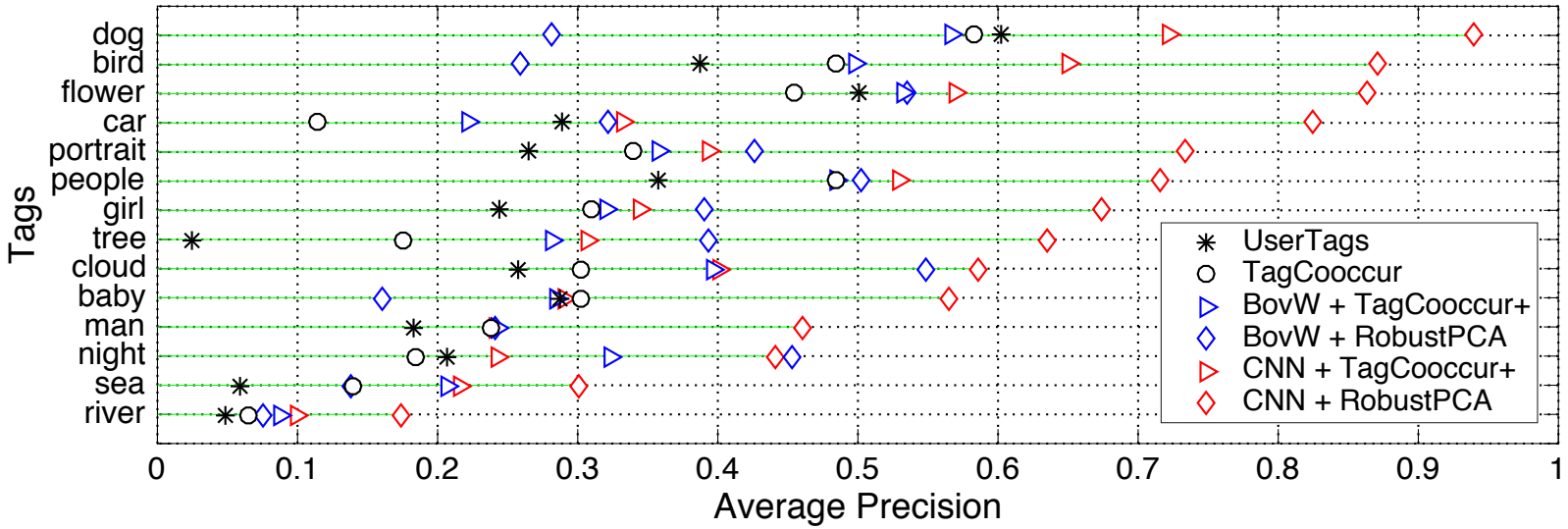
TAG ASSIGNMENT



- Test images are grouped in terms of their number of ground truth tags. The area of a colored bar is proportional to the number of images that the corresponding method scores best.

- When increasing the training set size, the most visible change is that of TagFeature and RelExample on images with one ground truth tag.

TAG REFINEMENT

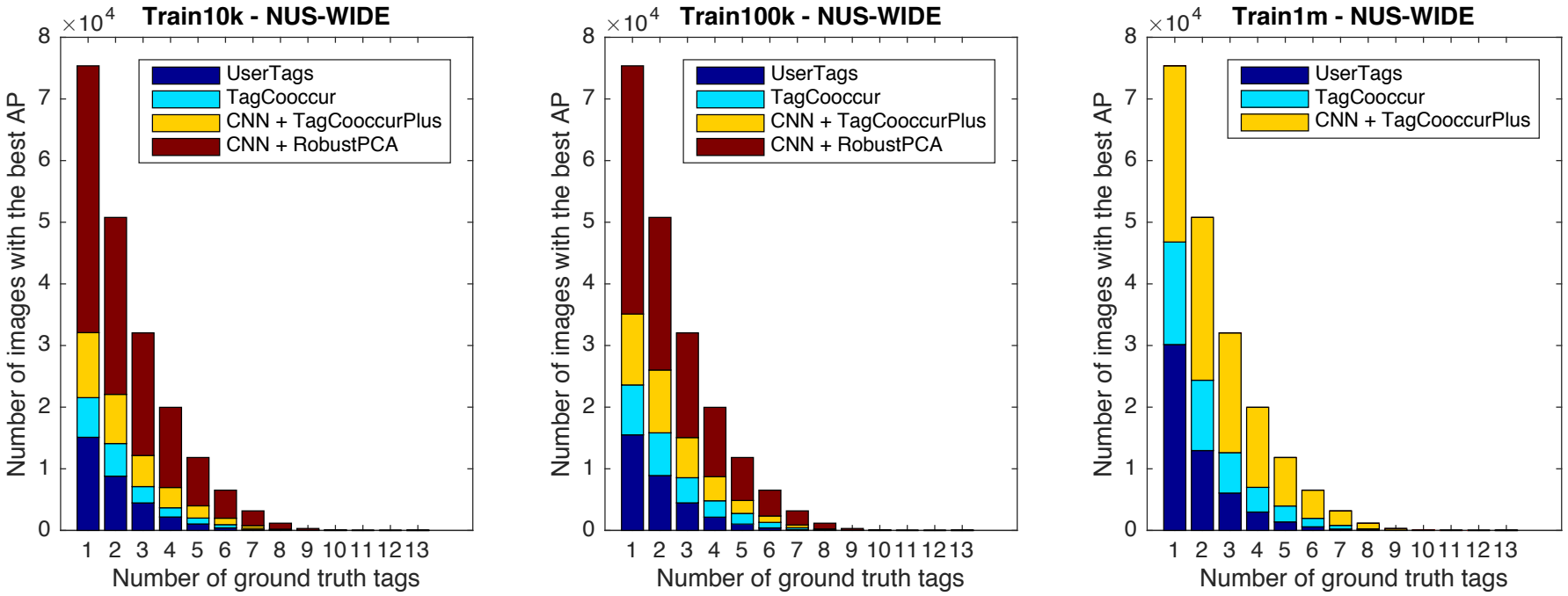


MIRFlickr test set, trained on Train100k.

CNN Features
BovW Features

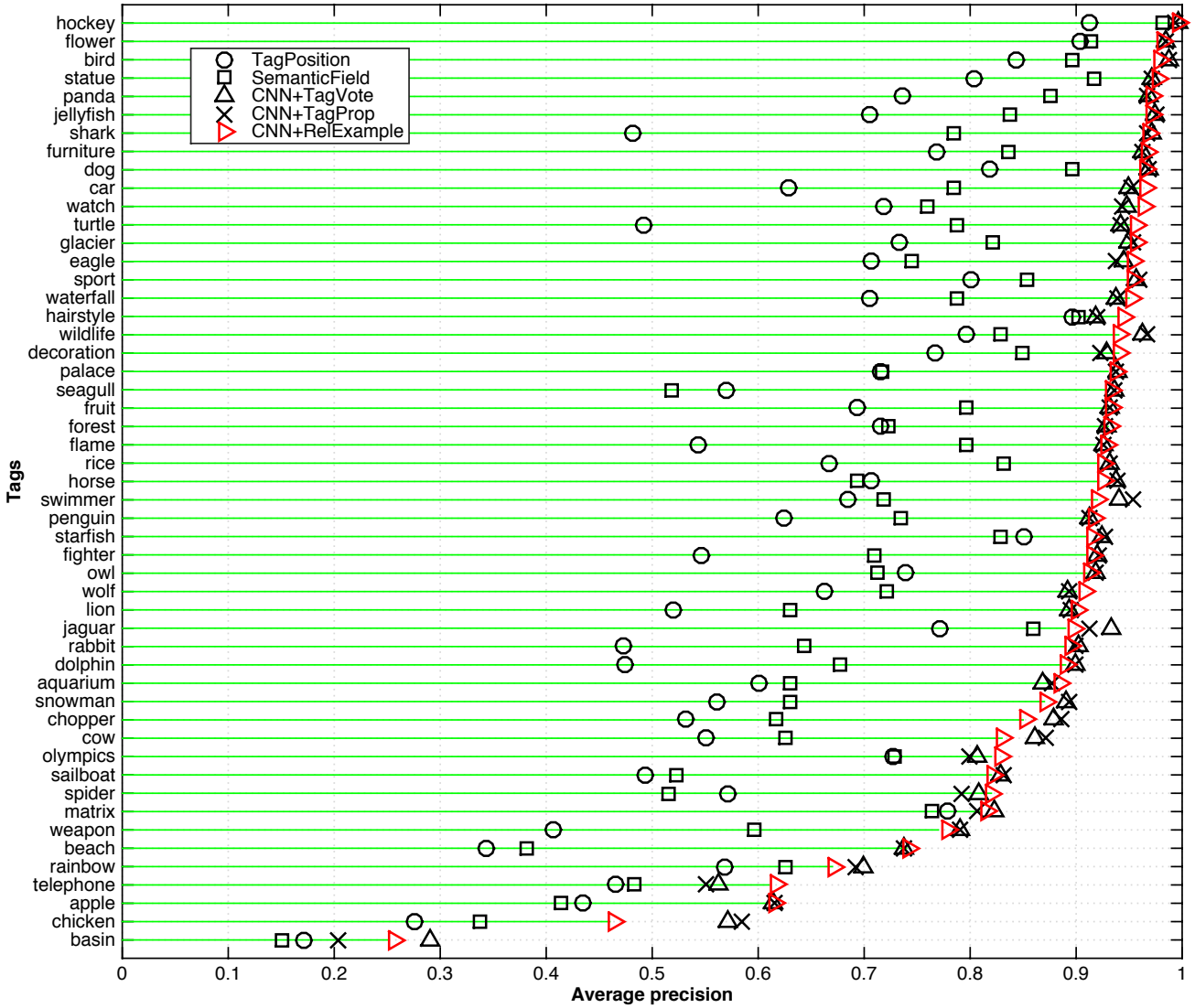
- All methods have performance superior to user tagging.
- The tag + image based methods outperform the tag based TagCooccur.
- RobustPCA provides the best performance.

TAG REFINEMENT



- CNN+RobustPCA has the best performance in every group of images.
- Almost the totality of images with more than 4 ground truth tags are better refined by RobustPCA than the other methods.
- TagCooccur+ refines tags better than TagCoccur.

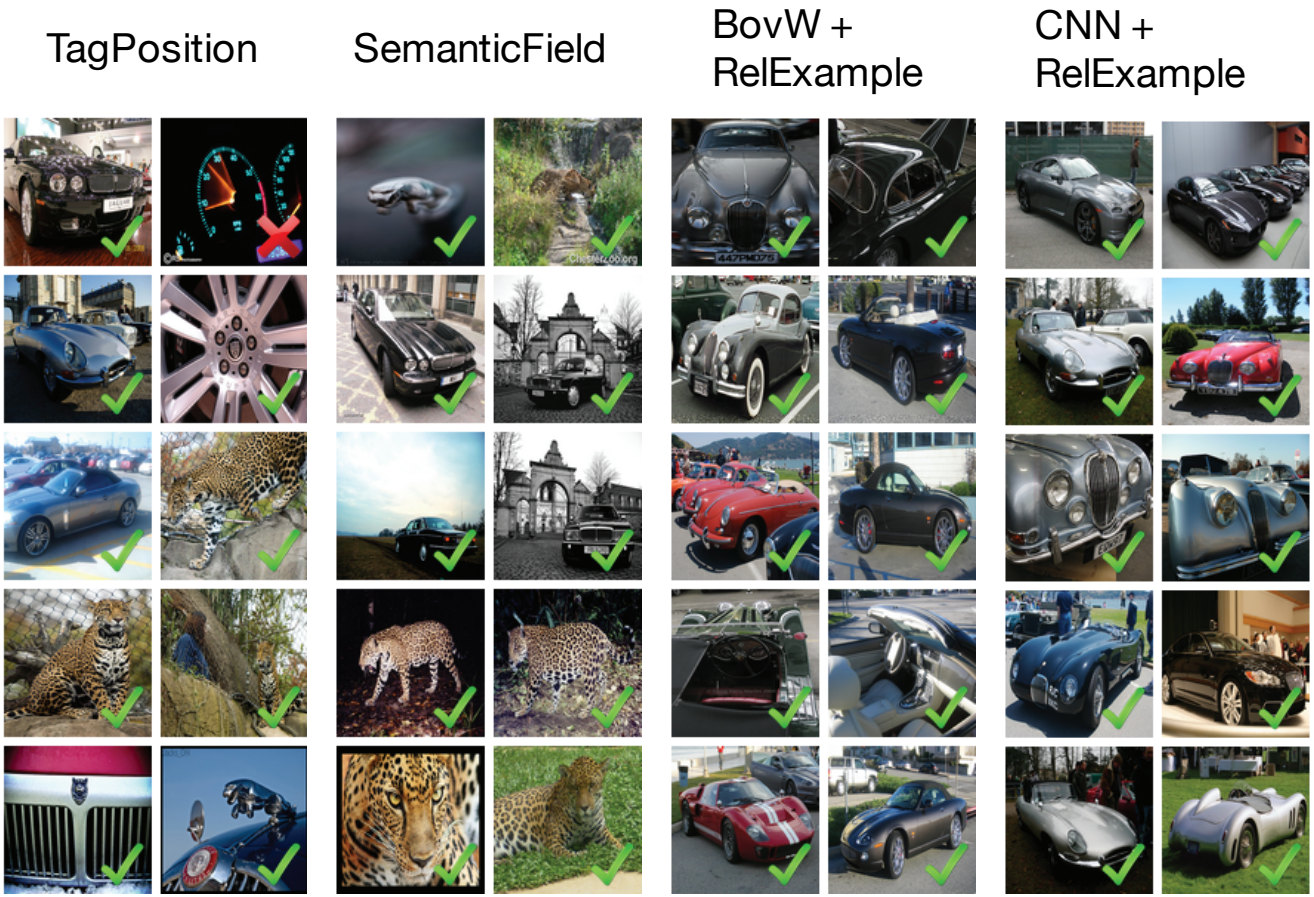
TAG RETRIEVAL



- Like Assignment, TagVote and TagProp provide the best performance.
- For 33 out of the 51 test tags, RelExample exhibits average precision higher than 0.9.

TAG RETRIEVAL

The top 10 ranked images for ‘jaguar’



COMMON PATTERNS

- Some common patterns have emerged, indipendently from the task:
  - All methods benefit from using CNN Features.
  - The more social data for training, the better performance is obtained.
  - With small-scale training sets, tag + image based methods that conducts model-based learning with denoised training examples turn out to be the most effective solution.

IMAGENET AS TRAINING SET

ImageNet already provides labeled examples for over 20k categories. Is it necessary to learn from socially tagged data?

- Some methods can't be run or require modifications:
  - No user information in ImageNet. Tag+Image+User must be able to remove their dependency on user.
  - Tag co-occurrences are limited in ImageNet because images are labelled with a single WordNet synset.

- We ran an empirical evaluation between Train100k, Train1m and ImageNet.

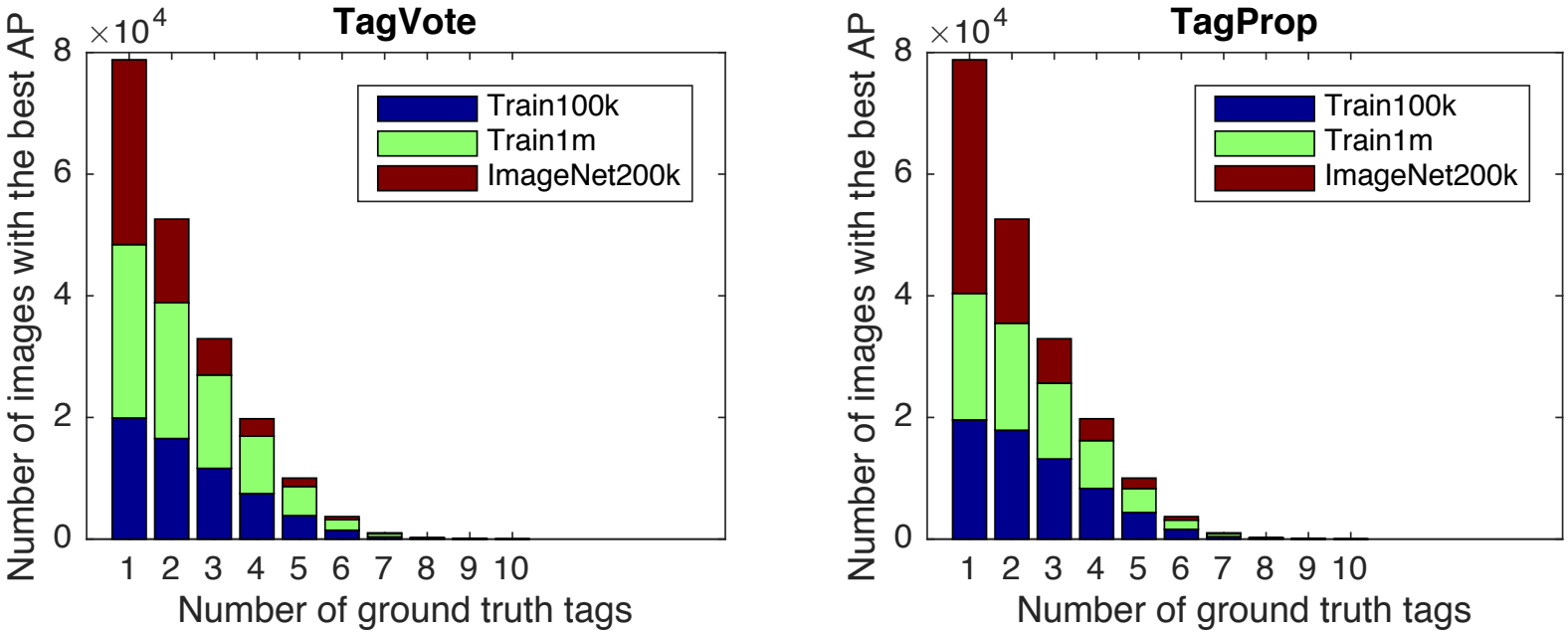
- We tested TagVote (without unique-user constraint) and TagProp, the two methods that reported the best overall performance.

IMAGENET RESULTS

Tag Assignment				
	MIRFlickr		NUS-WIDE	
Training Set	TagVote	TagProp	TagVote	TagProp
MiAP scores:				
Train100k	0.377	0.383	0.392	0.389
Train1M	0.389	**0.392**	**0.414**	0.393
ImageNet200k	0.345	0.304	0.325	0.368
MAP scores:				
Train100k	0.641	0.647	0.386	0.405
Train1M	0.664	**0.668**	**0.429**	0.420
ImageNet200k	0.532	0.532	0.363	0.362

- Methods trained on socially tagged datasets show better performance for tag assignment.

IMAGENET RESULTS



- TagVote and TagProp trained on ImageNet200k have better performance on images with a single relevant tag.
- On the other groups, Train100k and Train1M are a better choice.
- For its single-label nature, ImageNet is less effective for assigning multiple labels to an image.

IMAGENET RESULTS

	Tag Retrieval			
	Flickr51		NUS-WIDE	
Training Set	TagVote	TagProp	TagVote	TagProp
MAP scores:				
Train100k	0.854	0.860	0.742	0.745
Train1M	**0.874**	0.871	0.753	0.745
ImageNet200k	0.873	0.873	**0.762**	**0.762**
$NDCG_{20}$ scores:				
Train100k	0.838	0.863	0.849	0.856
Train1M	0.894	0.851	**0.891**	0.853
ImageNet200k	**0.920**	0.898	0.843	0.847

- For retrieval, in general the two socially tagged yield better performance than ImageNet200k. However, in some cases is not!
- Train100k and Train1m yields better performance on tags where ImageNet examples lack diversity (for instance 'running').
- ImageNet200k performance gain is largely due to a few tags where social tagging is very noisy.

IMAGENET RESULTS

ImageNet already provides labeled examples for over 20k categories. Is it necessary to learn from socially tagged data?

- Yes!
- For tag assignment social media examples are a preferred resource of training data.
- For tag retrieval ImageNet may provide better performance, yet the performance gain is largely due to a few tags where social tagging is very noisy.

CONCLUSIONS

- We went through eleven key methods of various media and learning.
- Take home messages:
  - The more social data for training, the better performance is obtained
  - Substituting BovW for CNN features boosts all methods performance.
  - TagVote and TagProp provide the best overall performance for Assignment and Retrieval.
  - RobustPCA is the choice for Refinement.
  - Given a small sized training set, the model-based RelExample may be a better performance.

REFERENCES

- [Jiang et al. 2009] Jiang, Yu-Gang, Chong-Wah Ngo, and Shih-Fu Chang. "Semantic context transfer across heterogeneous sources for domain adaptive video search." *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 2009.
- [Liu et al. 2011] Liu, Yiming, et al. "Textual query of personal photos facilitated by large-scale web data." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33.5 (2011): 1022-1036.
- [Zhu et al. 2012] S. Zhu, C.-W. Ngo, and Y.-G. Jiang. 2012. "Sampling and Ontologically Pooling Web Images for Visual Concept Learning". *IEEE Transactions on Multimedia* 14, 4 (2012), 1068–1078.
- [Liu et al. 2009] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. 2009. "Tag Ranking". *In Proc. of WWW*. 351–360.
- [Makadia et al. 2010] A. Makadia, V. Pavlovic, and S. Kumar. 2010. "Baselines for Image Annotation". *International Journal of Computer Vision* 90, 1 (2010), 88–105.
- [Li et al. 2009b] X. Li, C. Snoek, and M. Worring. "Learning Social Tag Relevance by Neighbor Voting". *IEEE Transactions on Multimedia* 11, 7 (2009), 1310–1322.
- [Guillaumin et al. 2009] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. 2009. "TagProp: Discriminative Metric Learning in Nearest Neighbor Models for Image Auto-Annotation". *In Proc. of ICCV.* 309–316.

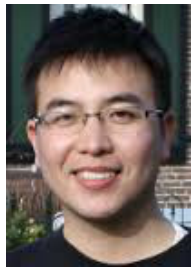
REFERENCES

- [Sigurbjörnsson and van Zwol 2008] B. Sigurbjörnsson and R. van Zwol. 2008. “Flickr tag recommendation based on collective knowledge”. *In Proc. of WWW.* 327–336.
- [Chen et al. 2012] L. Chen, D. Xu, I. Tsang, and J. Luo. 2012. “Tag-Based Image Retrieval Improved by Augmented Features and Group-Based Refinement”. *IEEE Transactions on Multimedia* 14, 4 (2012), 1057–1067.
- [Li and Snoek 2013] X. Li and C. Snoek. 2013. “Classifying tag relevance with relevant positive and negative examples”. *In Proc. of ACM MM*. 485–488.
- [Zhu et al. 2010] G. Zhu, S. Yan, and Y. Ma. 2010. “Image Tag Refinement Towards Low-Rank, Content-Tag Prior and Error Sparsity”. *In Proc. of ACM MM.* 461–470.
- [Sang et al. 2012] J. Sang, C. Xu, and J. Liu. 2012a. “User-Aware Image Tag Refinement via Ternary Semantic Analysis”. *IEEE Transactions on Multimedia* 14, 3 (2012), 883–895.

IMAGE TAG ASSIGNMENT, REFINEMENT AND RETRIEVAL

ACM Multimedia 2015 Tutorial

October 26, 2015



Xirong Li
Renmin University of China

Tiberio Uricchio
University of Florence

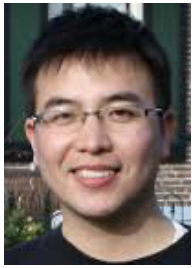
Lamberto Ballan
University of Florence &
Stanford University

Marco Bertini
University of Florence

Cees Snoek
University of Amsterdam &
Qualcomm Research
Netherlands

Alberto Del Bimbo
University of Florence

PART 5
PRACTICES



Xirong Li
Renmin University of China

xirong@ruc.edu.cn

Tiberio Uricchio
University of Florence

tiberio.uricchio@unifi.it

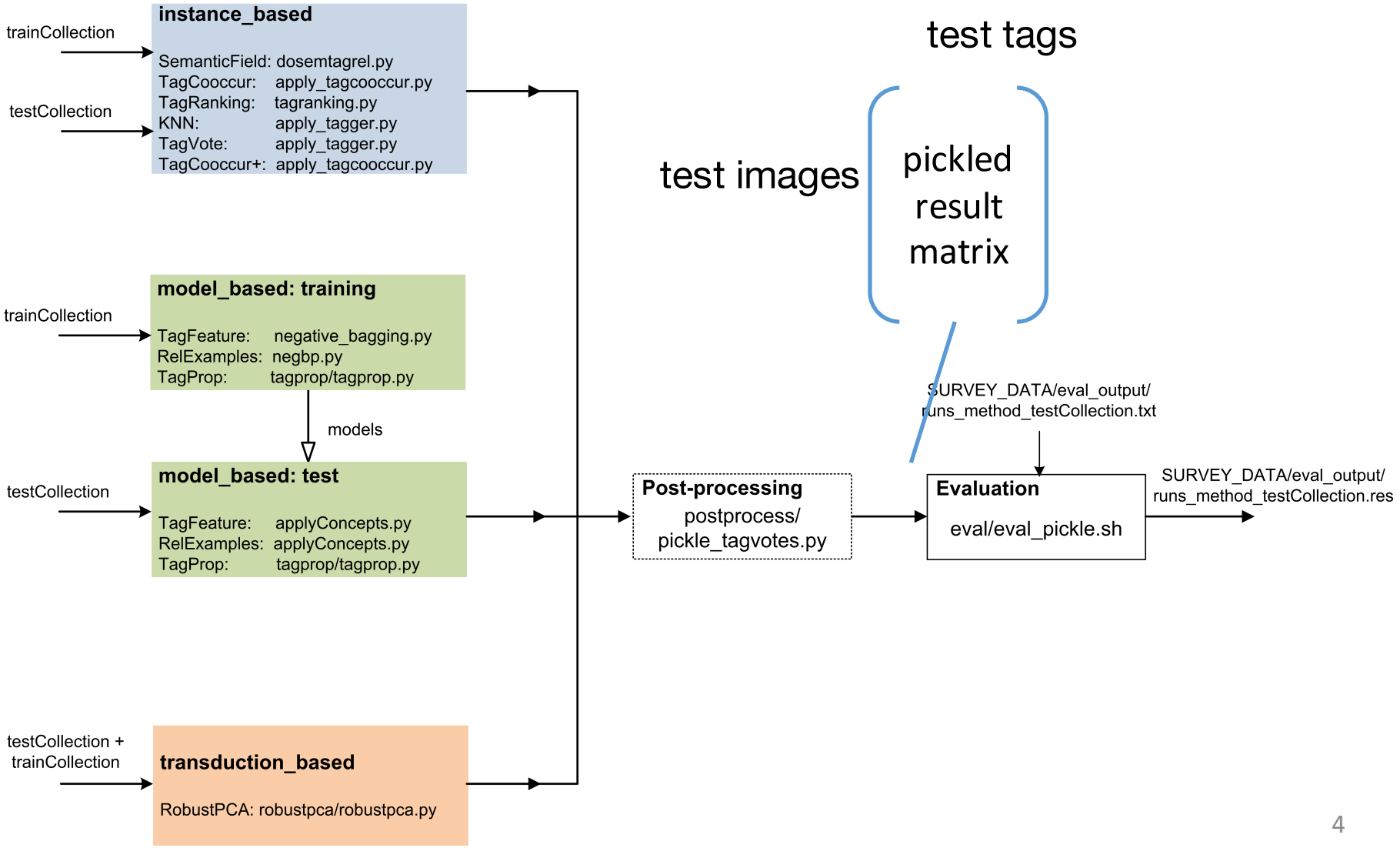
- Introduction to **Jingwei**
  - design
  - API
- Hands on
  - Run TagVote on Train10k + MIRFlickr
  - Learning new tag models on the fly

PRINCIPLES OF DESIGN

- Usability
  - Python APIs
  - cross-platform: linux, window, mac
- Readability
  - Majority of the code is written in Python
- Flexibility
  - Extend easily to new datasets and new visual features

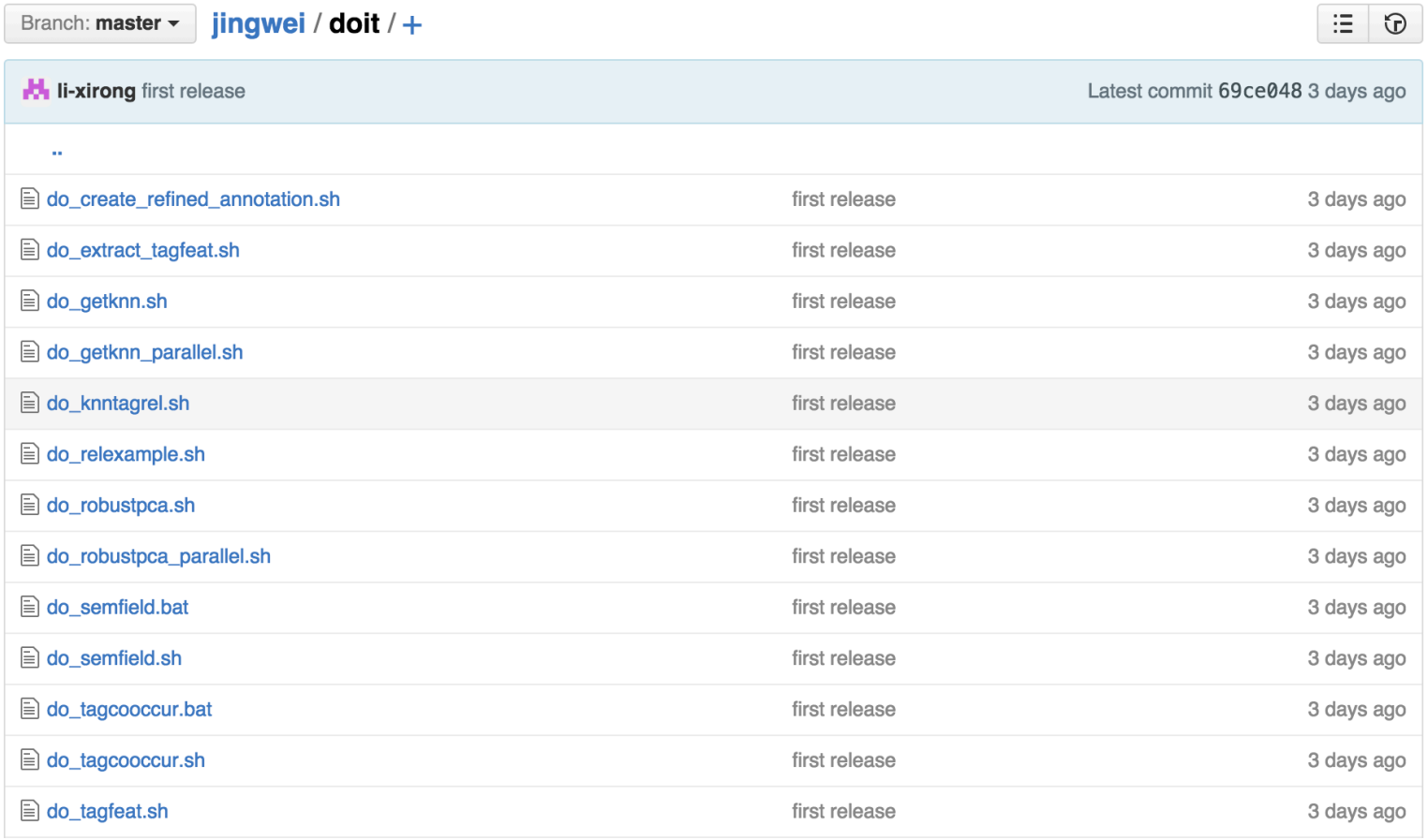
CODE ARCHITECTURE OF JINGWEI

https://github.com/li-xirong/jingwei



RUN A SPECIFIC METHOD

- **doit** series



Branch: **master** ▾ **jingwei** / **doit** / +

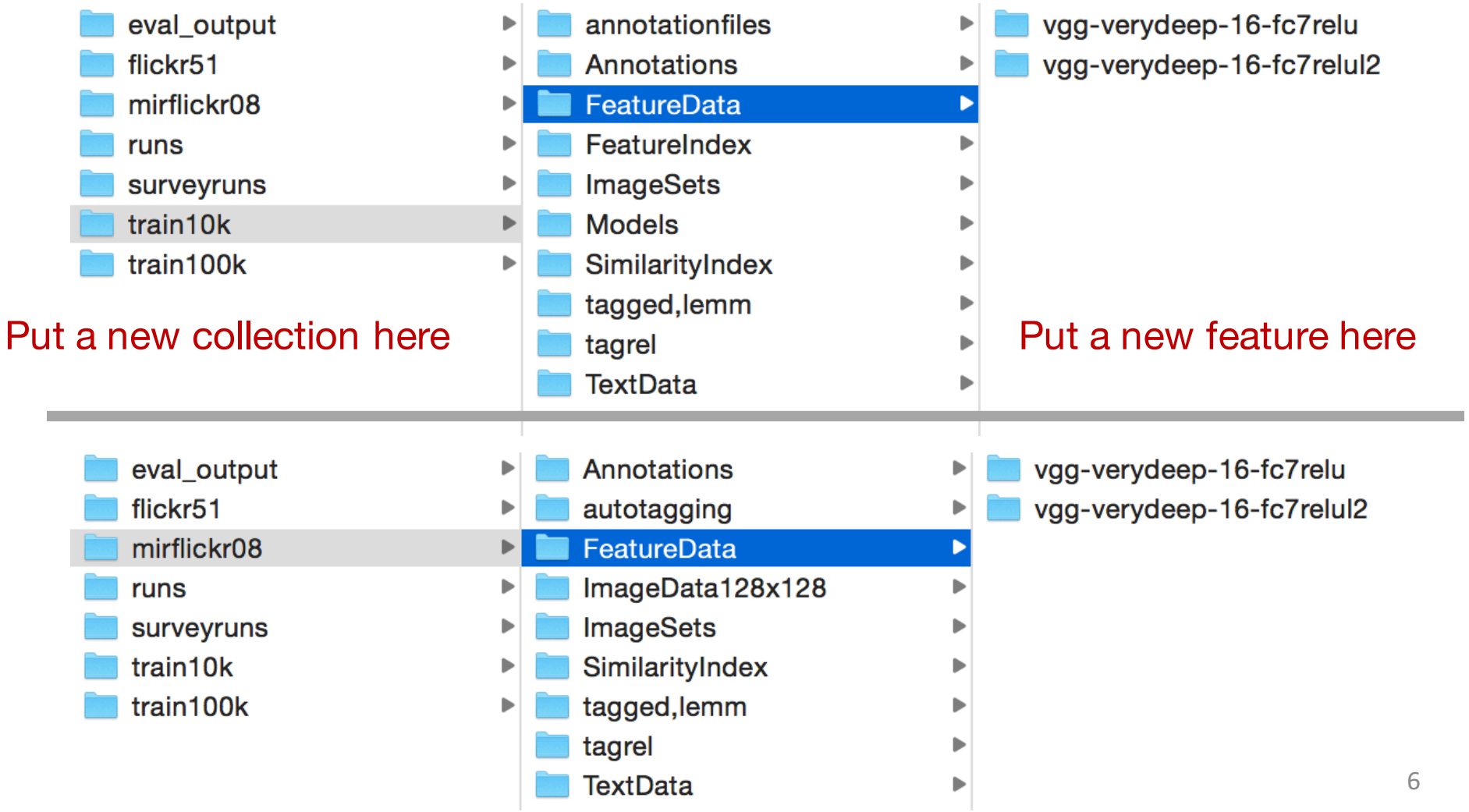
li-xirong first release — Latest commit `69ce048` 3 days ago

..		
do_create_refined_annotation.sh	first release	3 days ago
do_extract_tagfeat.sh	first release	3 days ago
do_getknn.sh	first release	3 days ago
do_getknn_parallel.sh	first release	3 days ago
do_knntagrel.sh	first release	3 days ago
do_relexample.sh	first release	3 days ago
do_robustpca.sh	first release	3 days ago
do_robustpca_parallel.sh	first release	3 days ago
do_semfield.bat	first release	3 days ago
do_semfield.sh	first release	3 days ago
do_tagcooccur.bat	first release	3 days ago
do_tagcooccur.sh	first release	3 days ago
do_tagfeat.sh	first release	3 days ago

https://github.com/li-xirong/jingwei/tree/master/doit

DATA ORGANIZATION

- Training and test collections follow the same data organization



CASE STUDY: TAGVOTE

- In this part, we show how to use the TagVote method

- Datasets
  - Training set: train10k
  - Test set: mirflickr08

- Tasks
  - Tag assignment
  - Tag retrieval

RUN TAGVOTE

- Modify two variables in start.sh according to your machine

  export **SURVEY_CODE**=/Users/xirong/jingwei

  export **SURVEY_DATA**=/Users/xirong/mm15tut

- Go to doit

  ./do_tagvote.sh train10k mirflickr08 color64+dsift

$SURVEY_DATA/surveyruns/train10k_mirflickr08_color64+dsift,tagvote.pkl

IMPLEMENTATION OF TAGVOTE

- Source code: instance_based/tagvote.py

```
class TagVoteTagger:
    def __init__(self, collection, annotationName, feature, distance, tpp=DEFAULT_TPP, rootpath=ROOT_PATH):
        self.concepts = readConcepts(collection, annotationName, rootpath)
        self.nr_of_concepts = len(self.concepts)
        self.concept2index = dict(zip(self.concepts, range(self.nr_of_concepts)))

        feat_dir = os.path.join(rootpath, collection, "FeatureData", feature)
        id_file = os.path.join(feat_dir, 'id.txt')
        shape_file = os.path.join(feat_dir, 'shape.txt')
        self.nr_of_images, feat_dim = map(int, open(shape_file).readline().split())

        self.searcher = simpleknn.load_model(os.path.join(feat_dir, 'feature.bin'), feat_dim, self.nr_of_images, id_file)
        self.searcher.set_distance(distance)
        self.k = DEFAULT_K

        self._load_tag_data(collection, tpp, rootpath)
```

specify concepts

for visual neighbor search

IMPLEMENTATION OF TAGVOTE

- Key functions

Context is optional

```
def predict(self, content, context=None):
    scores = self._compute(content, context)
    return sorted(zip(self.concepts, scores), key=lambda v:v[1], reverse=True)
```

Re-implement this function for your own method

```
def _compute(self, content, context=None):
    users_voted = set()
    vote = [0-self.tagprior(c) for c in self.concepts] # vote only on the given concept list
    voted = 0
    skip = 0

    neighbors = self._get_neighbors(content, context)

    for (name, dist) in neighbors:
        (userid,tags) = self.textstore.get(name, (None, None))
        if tags is None or userid in users_voted:
            skip += 1
            continue
        users_voted.add(userid)
        tagset = set(tags.split())
        for tag in tagset:
            c_idx = self.concept2index.get(tag, -1)
            if c_idx >= 0:
                vote[c_idx] += 1
        voted += 1
        if voted >= self.k:
            break
    #assert(voted >= self.k), 'too many skips (%d) in %d neighbors' % (skip, len(neighbors))
    return vote
```

EVALUATE TAGVOTE

- specify runs (pickle files) to be evaluated in the following file

  $SURVEY_DATA/eval_output/runs_tagvote_mirflickr08.txt

- script
  - eval/eval_pickle.sh mirflickr08 tagvote

$SURVEY_DATA/eval_output/runs_tagvote_mirflickr08.res

TO IMPROVE TAGVOTE

- Now try deep learning features
  - vgg-verydeep-16-fc7relu

./do_tagvote.sh train10k mirflickr08 vgg-verydeep-16-fc7relu

$SURVEY_DATA/surveyruns/train10k_mirflickr08_vgg-verydeep-16-fc7relu,tagvote.pkl

LEARNING NEW TAG MODELS ON THE FLY

- In this part, we show step-by-step how to learn new tag models on the fly using the Jingwei API
- Scenario: To retrieve images from an (unlabeled) collection for a given set of tags, e.g., *child*, *face*, and *insect*.

STEP 1. SPECIFY CONCEPTS

- Generate a new concept file at **train100k/Annotations/conceptsmm15tut.txt**, which has three lines:

child
face
insect

- Obtain labeled examples for these three tags

python util/imagesearch/obtain_labeled_examples.py train100k ~/mm15tut/train100k/Annotations/conceptsmm15tut.txt

STEP 2. CREATE ANNOTATIONS

```
>>> from model_based.dataengine.positiveengine import PositiveEngine
>>> from model_based.dataengine.negativeengine import NegativeEngine
>>> pe = PositiveEngine('train100k')
>>> ne = NegativeEngine('train100k')
>>> pos_set = pe.sample('child', 100)
>>> neg_set = ne.sample('child', 100)
>>> names = pos_set + neg_set
>>> labels = [1] * len(pos_set) + [-1] * len(neg_set)
>>> name2label = dict(zip(names,labels))
```

STEP 3. LOAD FEATURE VECTORS

```
>>> from basic.constant import ROOT_PATH
>>> from util.simpleknn.bigfile import BigFile
>>> feature = "vgg-verydeep-16-fc7relul2"
>>> feat_file = BigFile('%s/train100k/FeatureData/%s' % (ROOT_PATH, feature))
>>> (renamed, vectors) = feat_file.read(names)
>>> y = [name2label[x] for x in renamed]
```

STEP 4. TRAIN A LINEAR SVM MODEL

```
>>> from model_based.svms.fastlinear.liblinear193.python.liblinearutil import train
>>> from model_based.svms.fastlinear.fastlinear import liblinear_to_fastlinear
>>> svm_params = '-s 2 -B -1 -q'
>>> model = train(y, vectors, svm_params)
>>> fastmodel = liblinear_to_fastlinear([model], [1.0], feat_file.ndims)

# optionally save the learned model to disk
>>> from model_based.svms.fastlinear.fastlinear import fastlinear_save_model
>>> import os
>>> model_filename = os.path.join(ROOT_PATH, 'train100k', 'Models',
'conceptsmm15tut.txt', feature, 'fastlinear', 'child.model')
>>> from basic.common import makedirsforfile
>>> makedirsforfile(model_filename)
>>> fastlinear_save_model(model_filename, fastmodel)
```

STEP 5. APPLY THE TRAINED MODEL

```
>>> from basic.util import readImageSet
>>> testCollection = 'mirflickr08’
>>> imset = readImageSet(testCollection)
>>> test_feat_dir = os.path.join(ROOT_PATH, testCollection, 'featureData', feature)
>>> test_feat_file = BigFile(test_feat_dir)
>>> renamed, vectors = test_feat_file.read(imset)
>>> scores = [fastmodel.predict(x) for x in vectors]
>>> ranklist = sorted(zip(renamed, scores), key=lambda v:(v[1],v[0]), reverse=True)
>>> from basic.common import writeRankingResults
>>> resultfile = os.path.join(ROOT_PATH, testCollection, 'SimilarityIndex',
testCollection, 'train100k', 'conceptsmm15tut.txt', '%s,fastlinear'%feature,
'child.txt')
>>> writeRankingResults(ranklist, resultfile)
```

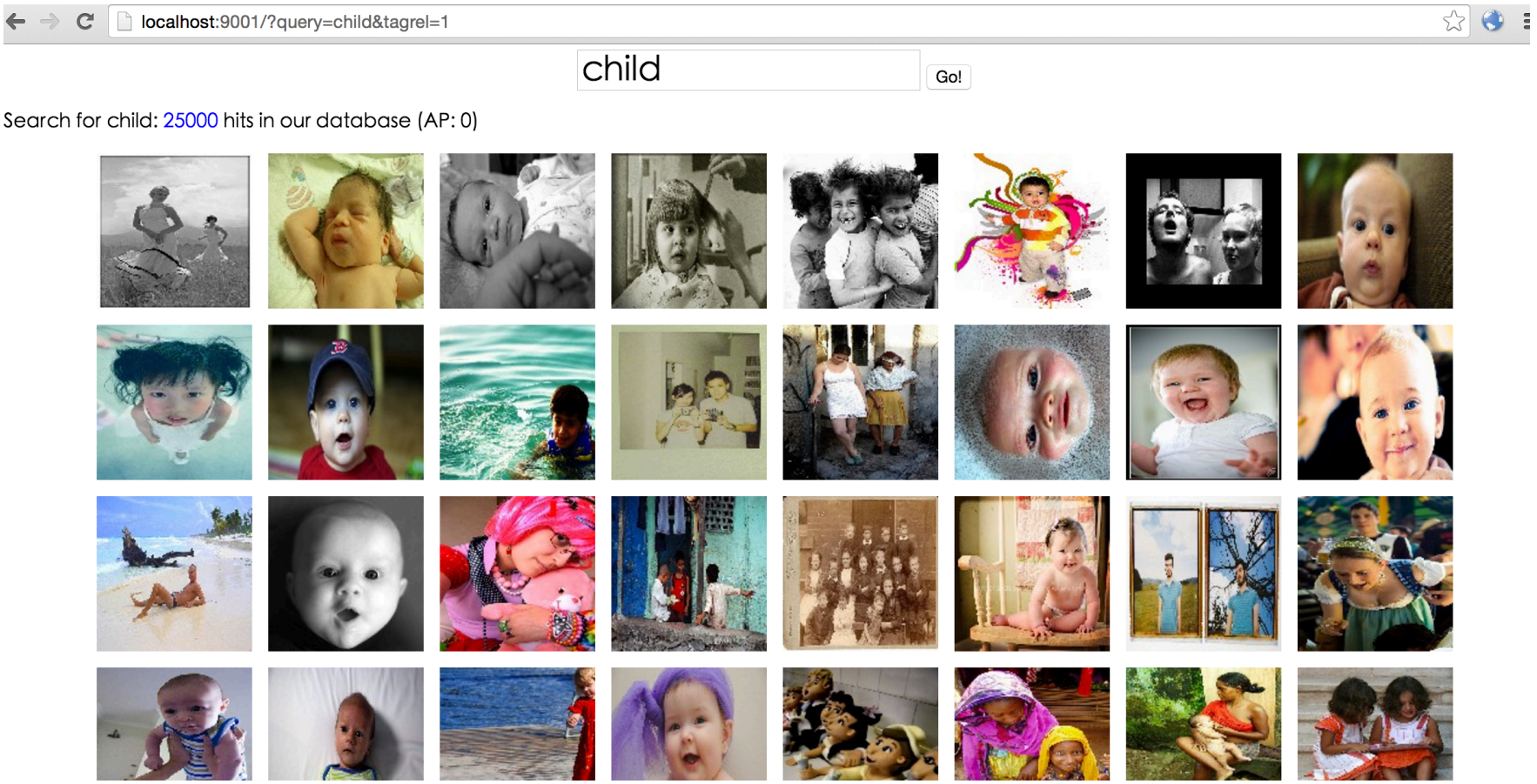
STEP 6. VISUALIZATION

Go to visualize/webdemo, and set config.json

```
{
 "imagedata_path": "/Users/xirong/mm15tut",
 "rootpath": "/Users/xirong/mm15tut",
 "max_hits": 50,
 "collection": "mirflickr08",
 "annotationName": "conceptsmir14.txt",
 "rankMethod": "train100k/conceptsmm15tut.txt/vgg-verydeep-16-fc7relul2,fastlinear",
 "metric": "AP"
}
```

```
python main.py 9001
```

STEP 6. VISUALIZATION



GO BACK TO STEP 2

- Generate better annotations by leveraging tag relevance learning results

python util/imagesearch/sortImages.py train100k conceptsmm15tut.txt tagrel train100k/vgg-verydeep-16-fc7relu,cosineknn,1000,lemm

train100k/SimilarityIndex/train100k/**tagged,lemm/train100k/vgg-verydeep-16-fc7relu,cosineknn,1000,lemm**/child.txt

```
>>> from model_based.dataengine.positiveengine  import SelectivePositiveEngine
>>> spe = SelectivePositiveEngine('train100k', 'tagged,lemm/train100k/vgg-verydeep-16-fc7relu,cosineknn,1000,lemm')
>>> pos_set = spe.sample('child', 100)
```

NOW THE 'CHILD' MODEL IS IMPRVOED

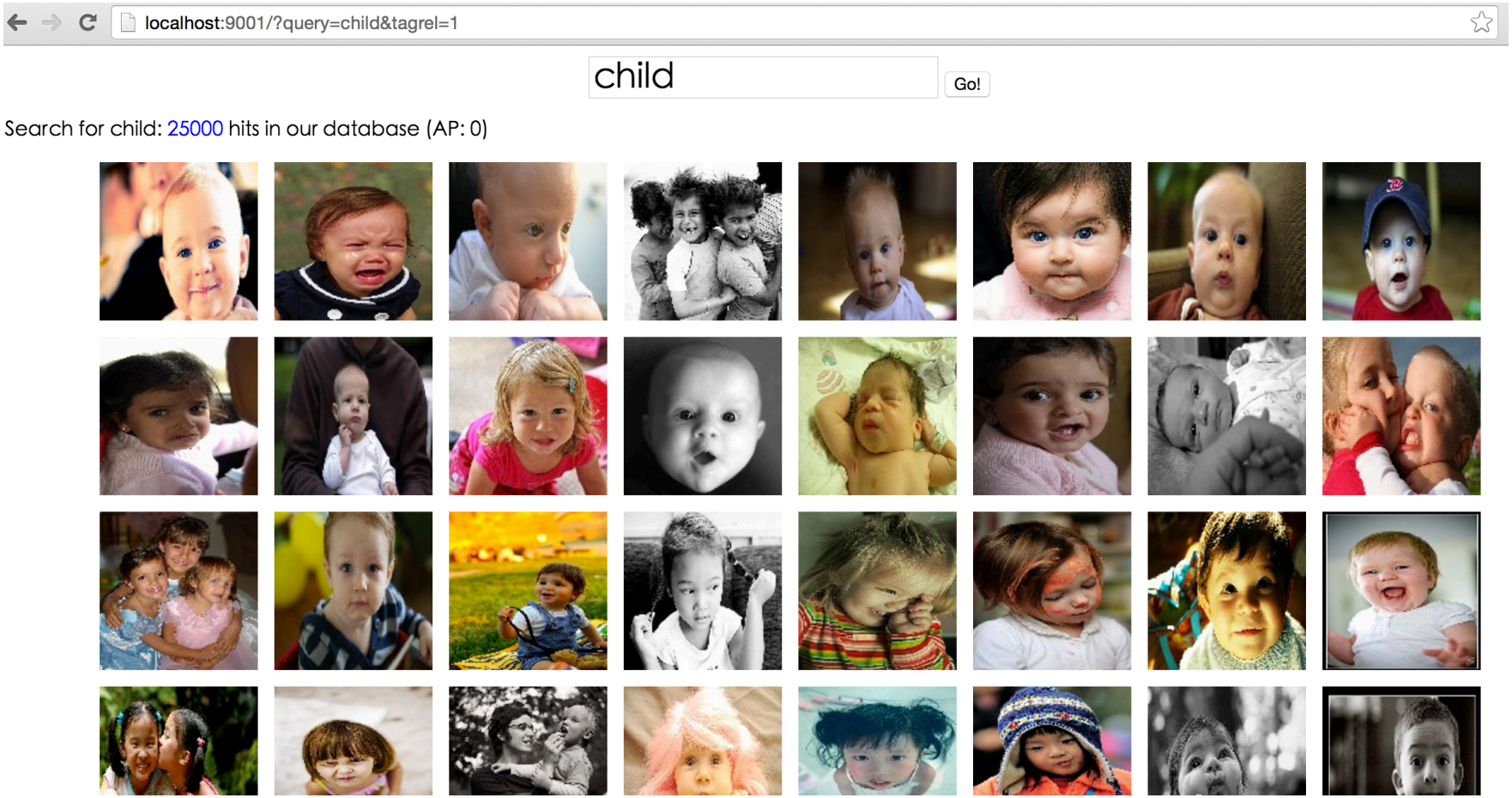
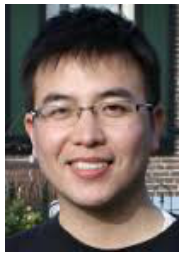


IMAGE TAG ASSIGNMENT, REFINEMENT AND RETRIEVAL

ACM Multimedia 2015 Tutorial

October 26, 2015



Xirong Li
Renmin University of China

Tiberio Uricchio
University of Florence

Lamberto Ballan
University of Florence &
Stanford University

Marco Bertini
University of Florence

Cees Snoek
University of Amsterdam &
Qualcomm Research
Netherlands

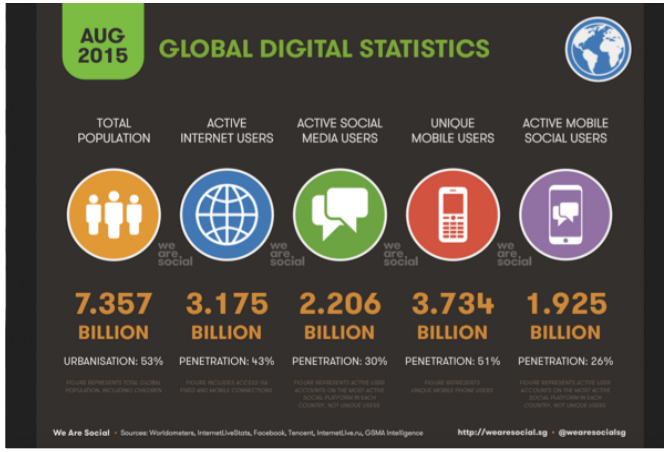
Alberto Del Bimbo
University of Florence

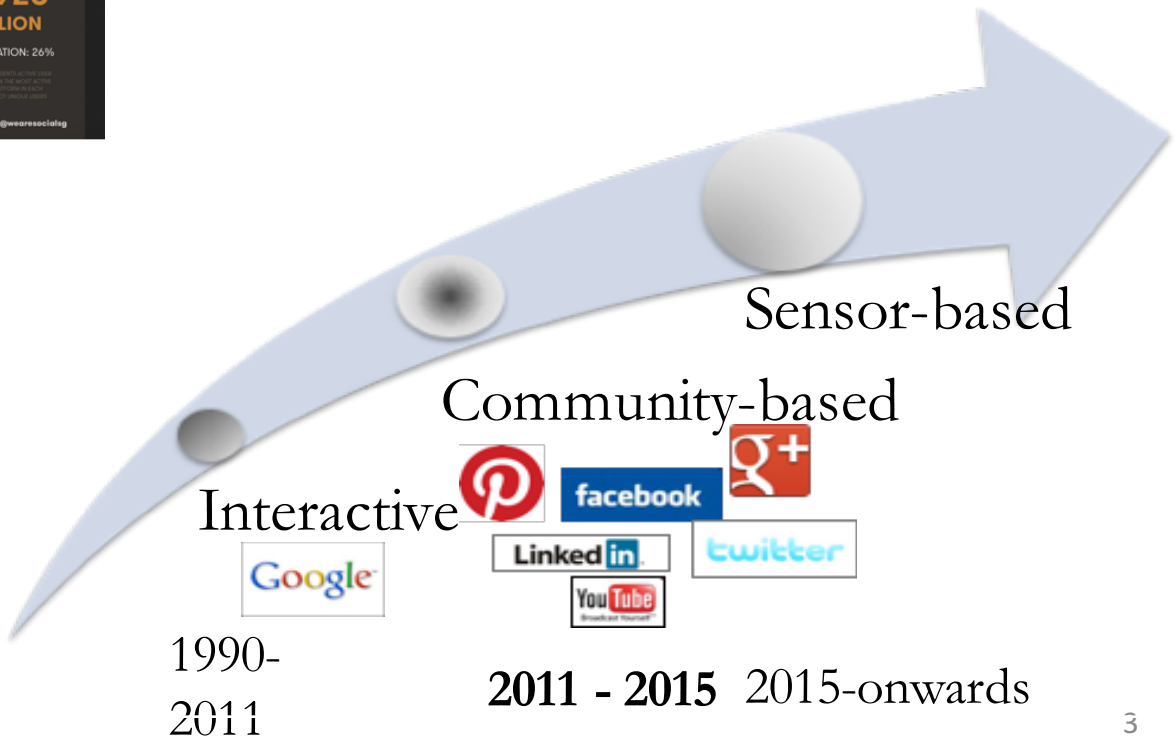
Alberto Del Bimbo
University of Florence

Cees Snoek
University of Amsterdam &
Qualcomm Research
Netherlands

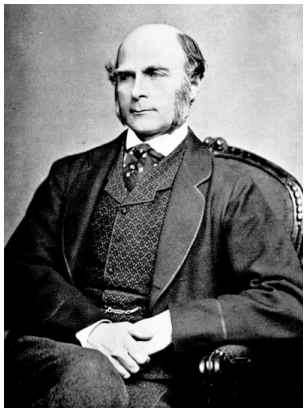
- Future directions

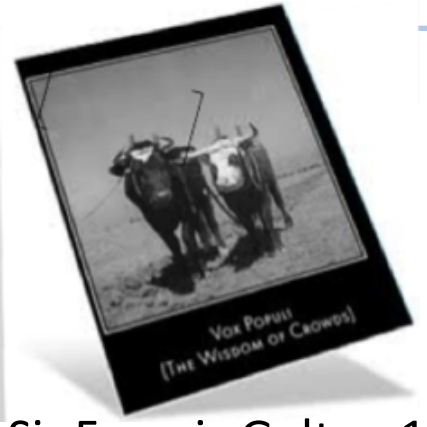
The web evolutionary trend





The wisdom of crowds





Sir Francis Galton 1906

Content-based annotation, refinement, retrieval

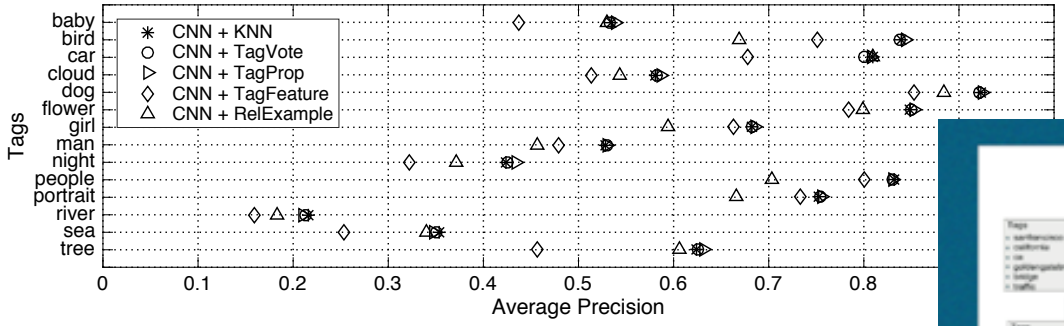
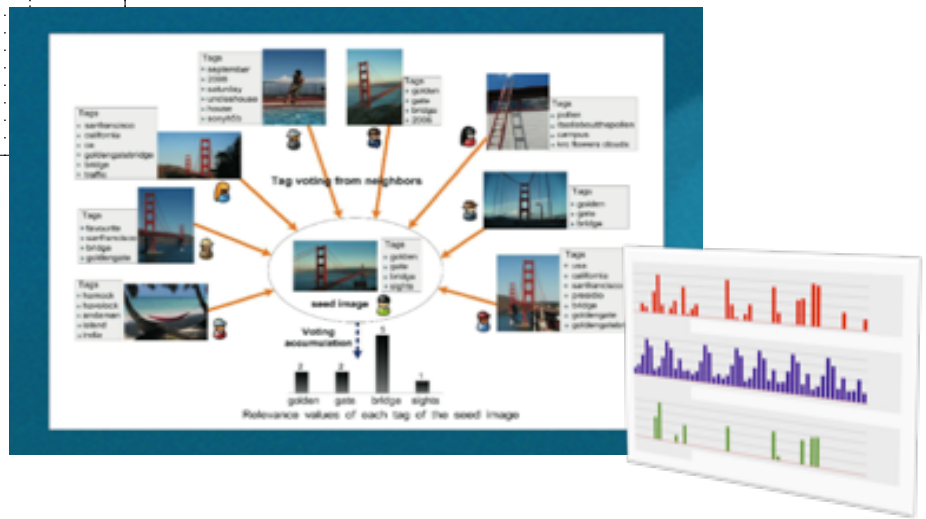
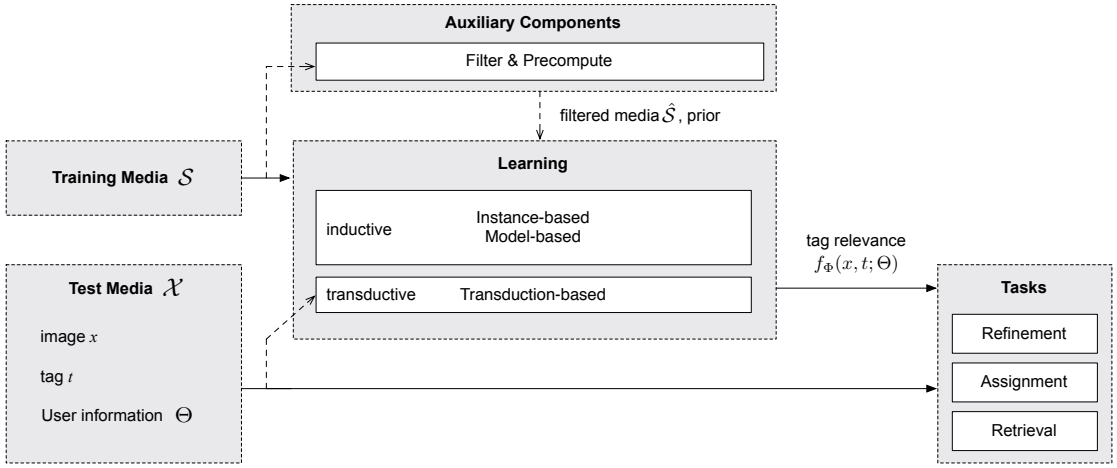
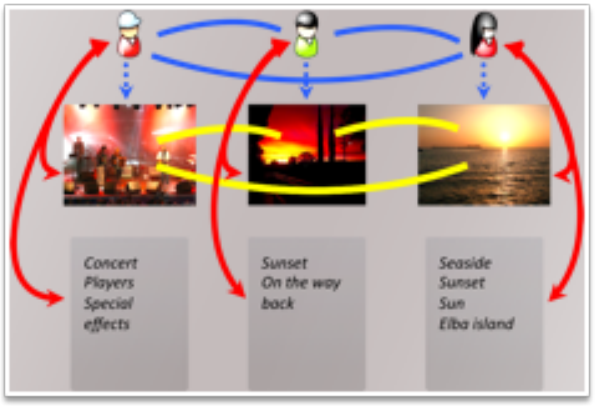


Image collection



The relevance function





Definition of a function f which measures the *relevance* between a given image and a specific tag, stands at the heart of annotation, refinement and retrieval task

Tag relevance learning is based on the visual content (and eventually a set of user information associated with the image)

Media	Learning		
	Instance-based	*Model-based*	*Transduction-based*
tag	**[Sigurbjörnsson and van Zwol 2008]** [Sun et al. 2011] **[Zhu et al. 2012]**	[Xu et al. 2009]	–
tag + image	**[Liu et al. 2009]** **[Makadia et al. 2010]** [Tang et al. 2011] [Wu et al. 2011] [Yang et al. 2011] [Truong et al. 2012] [Qi et al. 2012] [Lin et al. 2013] [Lee et al. 2013] [Uricchio et al. 2013] [Zhu et al. 2014] [Ballan et al. 2014] [Pereira et al. 2014]	[Wu et al. 2009] **[Guillaumin et al. 2009]** [Verbeek et al. 2010] [Liu et al. 2010] [Ma et al. 2010] [Liu et al. 2011b] [Duan et al. 2011] [Feng et al. 2012] [Srivastava and Salakhutdinov 2012] **[Chen et al. 2012]** [Lan and Mori 2013] **[Li and Snoek 2013]** [Li et al. 2013] [Wang et al. 2014] [Niu et al. 2014]	**[Zhu et al. 2010]** [Wang et al. 2010] [Li et al. 2010] [Zhuang and Hoi 2011] [Richter et al. 2012] [Kuo et al. 2012] [Liu et al. 2013] [Gao et al. 2013] [Wu et al. 2013] [Yang et al. 2014] [Feng et al. 2014] [Xu et al. 2014]
tag + image + user	**[Li et al. 2009b]** [Kennedy et al. 2009] [Li et al. 2010] [Znaidia et al. 2013] [Liu et al. 2014]	[Sawant et al. 2010] [Li et al. 2011b] [McAuley and Leskovec 2012] [Kim and Xing 2013] [McParlane et al. 2013b] [Ginsca et al. 2014] [Johnson et al. 2015]	**[Sang et al. 2012a]** [Sang et al. 2012b] [Qian et al. 2015]

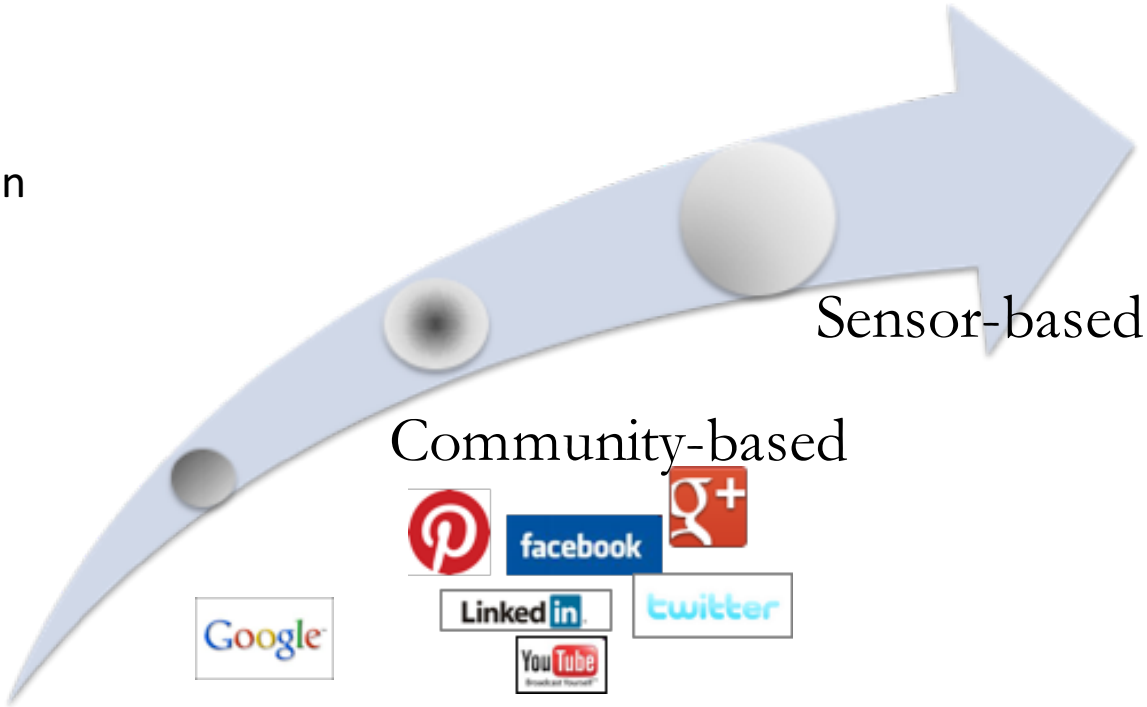
"Socializing the Semantic Gap: A Comparative Survey on Image Tag Assignment, Refinement and Retrieval "

X. LI, T. URICCHIO, L. BALLAN, M. BERTINI, C. SNOEK, A. DEL BIMBO

http://arxiv.org/abs/1503.08248

The wisdom of context

Discover relationships between user, content and concepts, time of use, environment, situation sentiment…….



- Topics are originated by real-world phenomena such as cultural events, real world physical facts….
- User intentions follow facts and change through time…
- Often tags are used to describe a situation rather than a visual content

Research trends and challenges

- Images are not static entities in the cyberspace
  - Mapping cyberspace and real-world beyond multimodal fusion and tag processing…
  - Correlation between visual content and context
  - Video….
- Defining tag importance, beyond tags that merely describe objects visually represented in the image, towards more user-centric and subjective notions such as emotion, sentiment, and preferences….

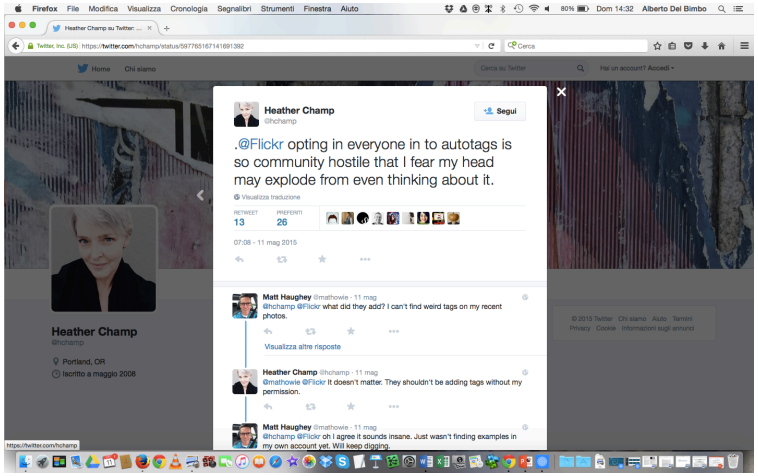
Mapping cyberspace and real-world

- Mapping between cyberspace and real world will be a law about a highly complex system and as such can only be a best approximation of delineated aspects of topic evolution and is beyond multimodal fusion and tag processing

- Content is increasingly personalized and tailored to user tastes, so it is important to understand user tagging behavior and trends
  - User influence is fundamental for prediction, personalized retrieval…
  - Spatial and temporal information are fundamental for prediction, personalized retrieval, topic relevance detection, social trends detection, sub-topic outbreak detection, ….
  - Improving both relevance and diversification is fundamental for personalized retrieval

Personalized image tags

- Personalize generic annotation models by learning from a user's multimedia tagging history
- Personalized tag recommendation by jointly exploring the tagging resources and the geo-location information by learning from user tagging history and geo-location related tagging
- User provided lists treated as having structure: users tend to present their tag lists with an inherent preference order not as a bag-of-words
- Graph learning for enriching the tagging data according to item similarities and tensor factorization for learning coherent ternary relations among users, images and tags



"Personalizing automated image annotation using cross-entropy",
X. LI, et al.
ACM Multimedia 2011

"Personalized Geo-Specific Tag Recommendation for Photos on Social Websites"
J. LIU, Z. LI, J. TANG, YU JIANG, AND H. LU, IEEE TMM 2014

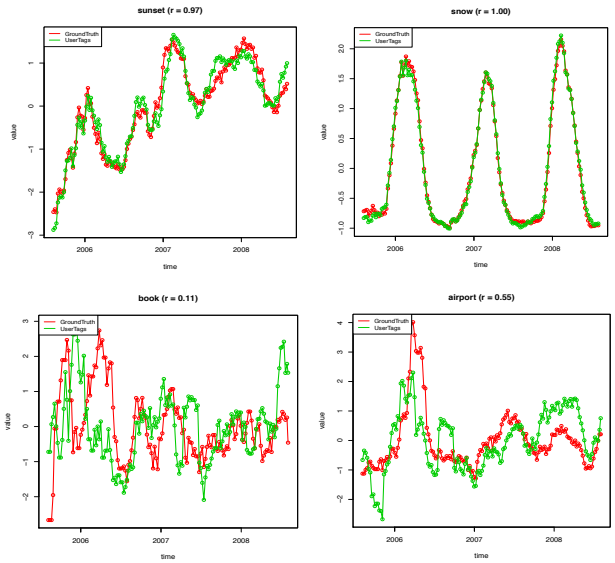
"Towards Understanding User Preferences from User Tagging Behavior",
A. O. NWANA, TSHUAN CHEN,
arXiv:1507.05150, 2015

"Tag Refinement for User-Contributed Images via Graph Learning and Nonnegative Tensor Factorization",
Z QIAN, et al.
Signal Processing Letters, IEEE, 2015

Spatial and temporal information

- Geo-location characterize locations about events, current affairs topic-characteristic patterns

- Topic characteristic temporal patterns:
  - *following a trend, cyclical, periodic, episodic*
  - *….*



"Evaluating Temporal Information for Social Image Annotation and Retrieval"
T. URICCHIO et al. Proc. ICIAP 2013

Event Based Characterization and Comparison of Geosocial Environment
C. KUMAR et al.
Proc. TAIA'14, 2014

"A study on the accuracy of Flickr's geotag data"
C. HAUFF,
Proc. SIGIR'13, 2013

TAIA'14 Workshop on Temporal, Social and Spatially Aware Information Access, 2014

Image diversity

- A set of images is considered to be diverse if it depicts different visual characteristics of the target, i.e., most of the perceived visual information is different from one image to another

- Diversity improved by applying clustering algorithms which rely on textual or/and visual cues

- Diversification based on the social metadata associated with the images or/and on the visual characteristics of the images.

"Visual diversification of image search results",
R. H. VAN LEUKEN et al., in Proc. of WWW 2009

"Retrieving Diverse Social Images"
MEDIAEVAL 2014, Benchmarking Initiative for Multimedia Evaluation.

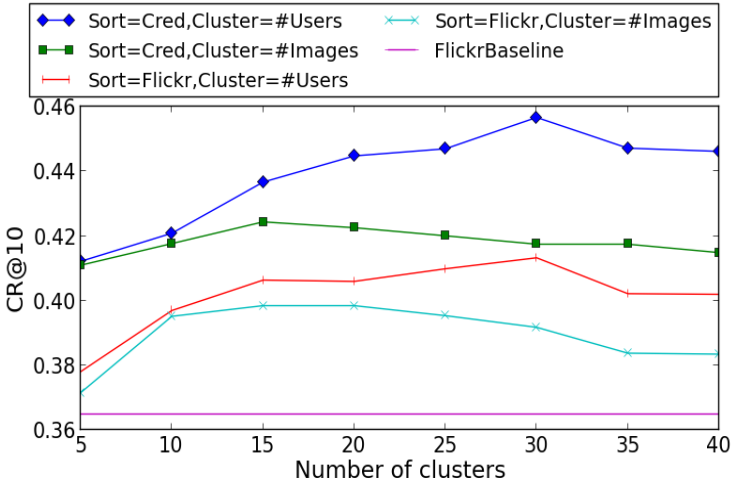
Correlation between visual content and context

- Internet topics are originated by different real-world phenomena:
  - User factors (credibility, groups…)
  - Correlation between visual content and external factors
  - Correlation with social trends
  - Different speed, acceleration, directions… of digital propagation
  - Internet culture or subculture (reposting between different media platforms, remixing, symbolization, repurposing…)
  - ..
- The unit of diffusion keeps getting smaller and smaller, with tweets and images and content fragments

- All of this may result into deviation of image content and tagging and requires modeling the influencing factors. Only a small number of prevalent factors may suffice to explain

User annotation credibility

- The quality of annotations provided by different users can vary strongly
- User credibility determined as an estimation of the quality (correctness) of a particular user's tags
- increase relevance by favoring images uploaded from users with good credibility estimates (user-based reranking)…



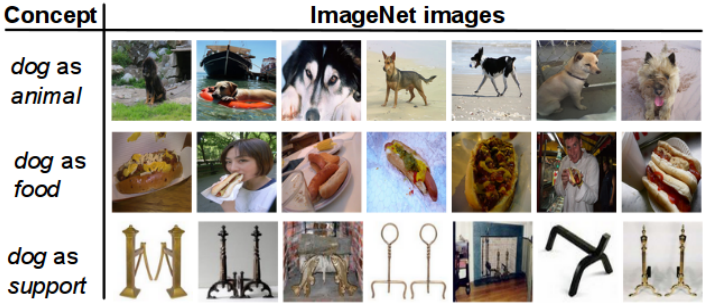
Cluster retrieval @10 performances with DIV400 dataset [A.L: Ginsca]

"Toward estimating user tagging credibility for social image retrieval", A.L. GINSCĂ et al. , ACM Multimedia 2014.

"Learning tag relevance by neighbor voting for social image retrieval", X. LI et al., Proc ACM MIR 2008.

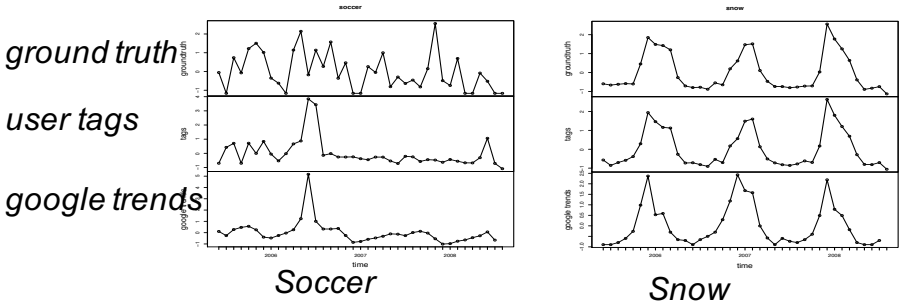
Correlation between visual content and external factors

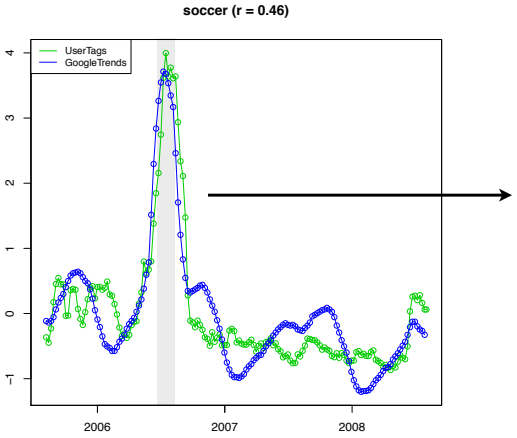
- Social images are related to situations more than simply describing objects: compare the tag-image pair to all models available for the tag and retain only the maximum classification score is too simplistic



[A.L: Ginsca]

- Tags occur frequently in correspondence of special events or have correspondence with event patterns





2006 FIFA World Cup
(9 June - 9 July)

"Evaluating Temporal Information for Social Image Annotation and Retrieval"
T. URICCHIO, et al. Proc. ICIAP 2013

Modeling the influencing factors

- Contents of images that are associated with the same keyword can be variable according to owners and temporal information. Social images reflect different users' experiences and preferences at different times.
  - Occurrence of a media document correlated to multiple Influencing factors
  - A generalized model may describe the time series

- Learn a model of the image occurrences with related factors and then sample the images based on the learned model

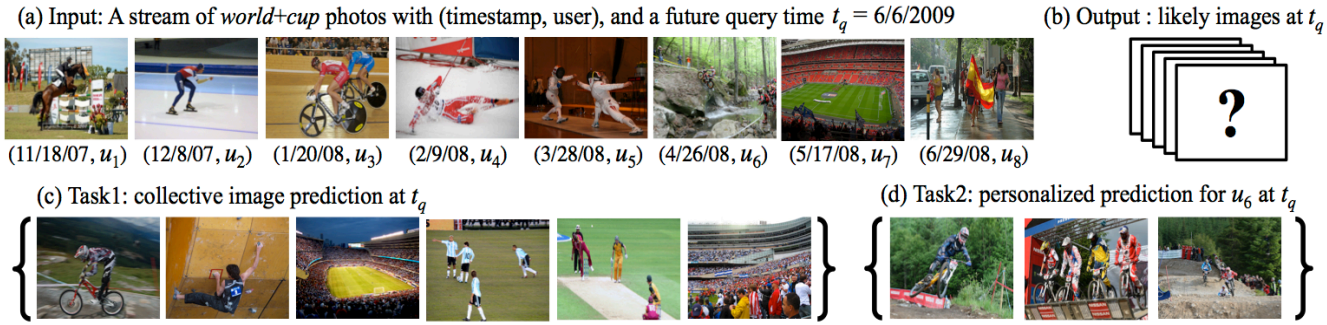


Figure 1: (a) Given an image sequence of *world+cup* up to 12/31/2008, can we guess what images are likely to appear at a future time point t_q=6/6/2009? (c) Collective image prediction. The *world+cup* usually refers to the soccer event, so a soccer scene can be a reasonable guess. However, the actual Web images are diverse because they reflect different users' experiences and preferences. (d) Personalized image prediction for user u_6. A user's unique angle of seeing the topic can make the prediction more focused.

"Web Image Prediction Using Multivariate Point Processes", G. KIM et al. , Proc KDD'12, 2012

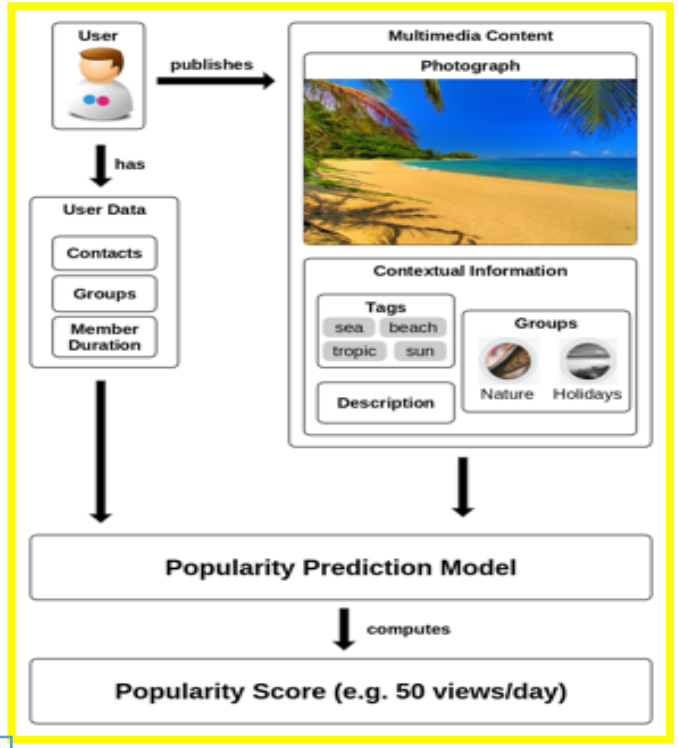
Trendy applications

- Personalized prediction, recommendation, retrieval of multimedia information….
- Popularity prediction based based on user sentiments
- Extending to video, tag localization…….
- Tracking web information sources and their correlation
- Connecting social and mobile contexts to media sensemaking...
- ….

Popularity prediction

Prediction by exploiting user data and image contextual information expressed by associated sentiments

The impact of visual attributes on online image diffusion: visual properties have low predictive power compared that of social cues. However, after factoring-out social influence, visual features show considerable predictive power….
L. Totti et al. 2014



Visual Features
- Image content: 1000d representation from CNN
- Sentibank features (1200 ANPs from the image)

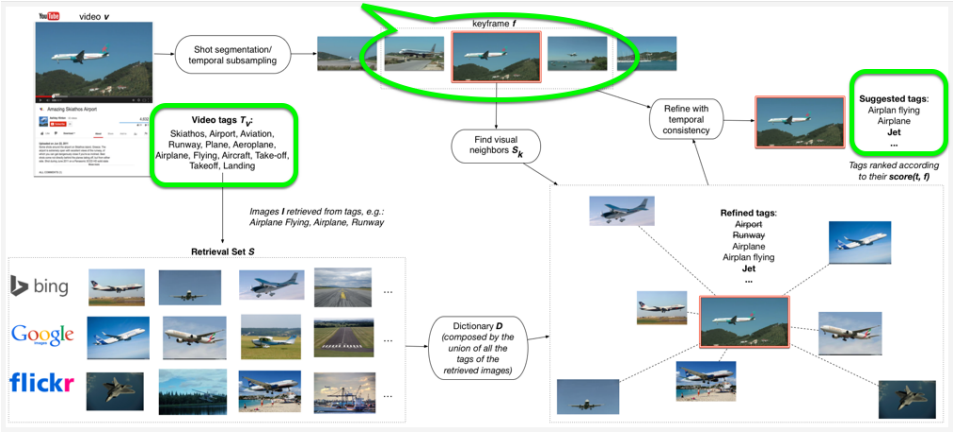
Social Features
- User Based: mean views, # photos, # contacts, # groups…
- Context Based: Freebase (topic notable type), named entity

"What makes an image popular?" A. Khosla et al., 2014.

"*Image Popularity Prediction in Social Media using Sentiment and Context Features*"
F. GELLI ET AL. 2015

Video automatic annotation and tag localization

- Intra-video indexing and search



Events	Objects	Activities	Scenes	Sites	**Avg**
64,8	57,8	66,1	76,1	67,5	**65,3**

Precision @1 (DUT-WEBV - YOUTUBE video)

Method	Precision@5	Precision@10
Random	6.1	4.5
Our	33.4	30.4

Annotation "in the wild", using an open vocabulary.

"A data-driven approach for tag refinement and localization in web videos"
L. BALLAN ET AL. 2014

A few recommended papers

(among the many others…)

Diversification and user credibility

Learning tag relevance by neighbor voting for social image retrieval, X. Li and al., Proc. ACM MIR 2008

Visual diversification of image search results R. H. van Leuken and al, Proc. of WWW 2009.

Socialsensor: Finding diverse images at mediaeval 2013, D. Corney and al, In Proc. of MediaEval Wksp. 2013.

Toward estimating user tagging credibility for social image retrieval, A.L. Gınsca˘, A. Popescu, B. Ionescu, A. Armagan, and I. Kanellos, ACM Multimedia 2014.

Div400: A social image retrieval result diversification dataset, B. Ionescu and al., ACM MMSys 2014.

Correlations with other factors

A study on the accuracy of Flickr's geotag data, C. Hauff, Proc. SIGIR'13, 2013

Personalized Geo-Specific Tag Recommendation for Photos on Social Websites J. Liu, Z. Li, J. Tang, Yu Jiang, and H. Lu, IEEE TMM 2014

Time modeling

Dynamic Topic Modeling for Monitoring Market Competition from Online Text and Image Data H. Zhang, G. Kim, E. P. Xing, Proc. KDD'15, 2015

Web Image Prediction Using Multivariate Point Processes, G. Kim Li Fei-Fei E. P. Xing, *Proc. KDD'12, 2012*

Model-Parallel Inference for Big Topic Models , X. Zheng, J. K. Kim, Q. Ho, E. P. Xing

Modeling and Analysis of Dynamic Behaviors of Web Image Collections , Gunhee Kim1, E. P. Xing1, and A. Torralba

Prediction and popularity

Understanding the Interaction between Interests, Conversations and Friendships in Facebook Q. Ho, R. Yan, R. Raina, E. P. Xing

What makes an image popular? A. Khosla, A. Das Sarma, R. Hamid, Proc. of WWW, 2014

The impact of visual attributes on online image diffusion L. C. Totti, F. A. Costa, S. Avila, E. Valle, W. Meira Jr, and V. Almeida, Proc. of WebSci 2014

DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks, T. Chen, D. Borth, T. Darrell, and S.-F. Chang. . arXiv:1410.8586, 2014.

Image Popularity Prediction in Social Media using Sentiment and Context Features, F. Gelli, T. Uricchio, M. Bertini, A. Del Bimbo, Shih Fu Chang, Proc. ACMMM'15, 2015

Tag localization in video streams

A data-driven approach for tag refinement and localization in web videos L. Ballan, M. Bertini, G. Serra, A. Del Bimbo, ArXiv 1407.0623, 2015

……