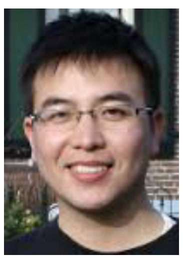
IMAGE TAG ASSIGNMENT, REFINEMENT AND RETRIEVAL

CVPR 2016 Tutorial

June 26, 2016



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

Organization of the tutorial

8:30 – 9:30	Part 1: Introduction
	Part 2: Taxonomy
9:30 – 10:00	Part 3: Experimental protocol
	Part 4: Evaluation
10:00 – 10:45	*Coffee break*
10:45 – 12:00	Part 4: Evaluation cont’d
12:00 – 12:30	Part 5: Conclusion and future directions

READING MATERIAL

Socializing the Semantic Gap: A Comparative Survey on Image Tag Assignment, Refinement and Retrieval,
ACM Computing Surveys, 49(1):14, June 2016.

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

XIRONG LI, Renmin University of China
TIBERIO URICCHIO, University of Florence
LAMBERTO BALLAN, University of Florence, Stanford University
MARCO BERTINI, University of Florence
CEES G. M. SNOEK, University of Amsterdam, Qualcomm Research Netherlands
ALBERTO DEL BIMBO, University of Florence

Where previous reviews on content-based image retrieval emphasize what can be seen in an image to bridge the semantic gap, this survey considers what people tag about an image. A comprehensive treatise of three closely linked problems (i.e., image tag assignment, refinement, and tag-based image retrieval) is presented. While existing works vary in terms of their targeted tasks and methodology, they rely on the key functionality of tag relevance, that is, estimating the relevance of a specific tag with respect to the visual content of a given image and its social context. By analyzing what information a specific method exploits to construct its tag relevance function and how such information is exploited, this article introduces a two-dimensional taxonomy to structure the growing literature, understand the ingredients of the main works, clarify their connections and difference, and recognize their merits and limitations. For a head-to-head comparison with

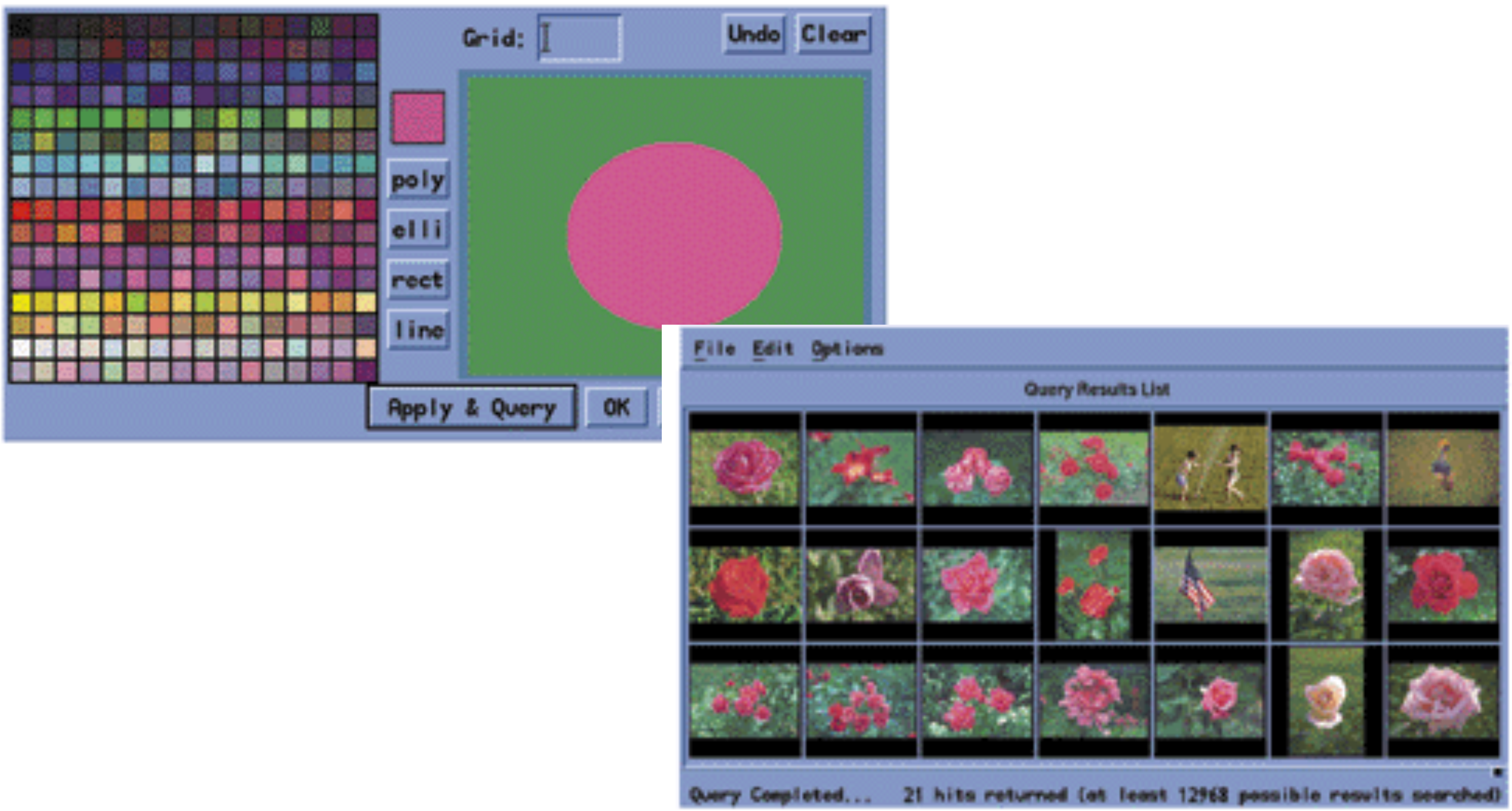
14

PART I
INTRODUCTION

- Problem statement
- Course organization

PROGRESS IN IMAGE RETRIEVAL

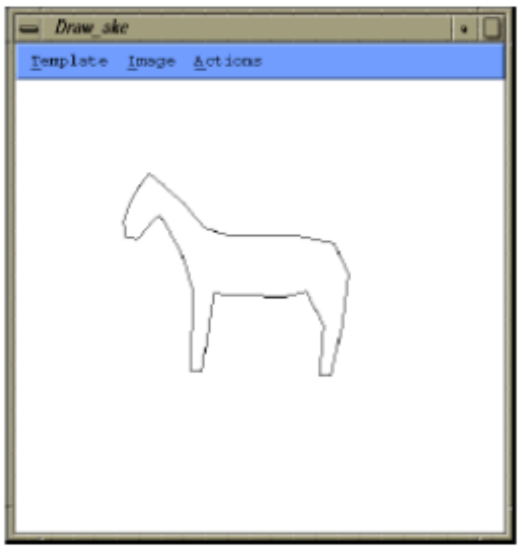
- Query-by-Image content

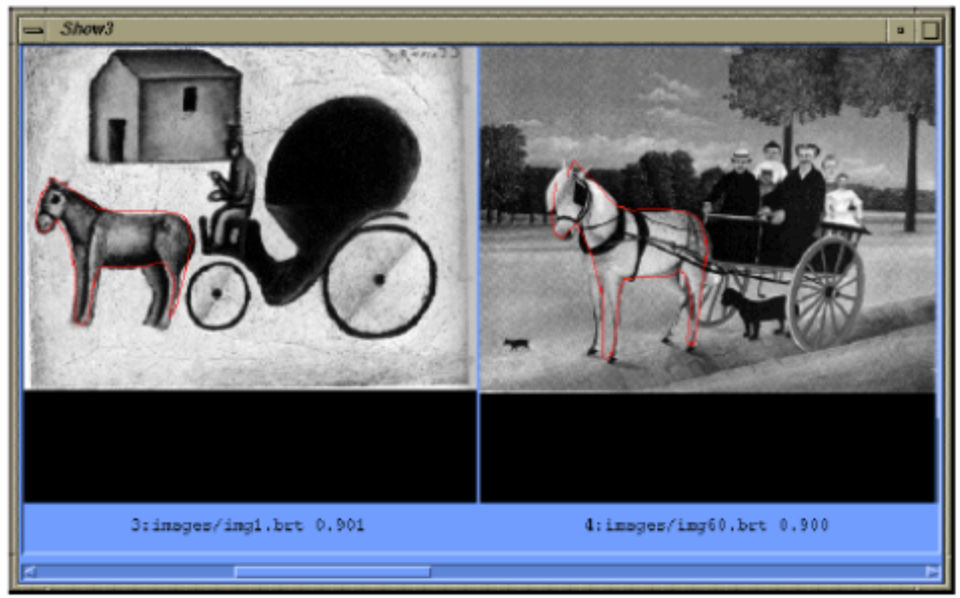


Del Bimbo, PAMI 1997

PROGRESS IN IMAGE RETRIEVAL

- Query-by-sketch





PROGRESS IN IMAGE RETRIEVAL

- By 2000 problem well understood

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 22, NO. 12, DECEMBER 2000 1349

Content-Based Image Retrieval at the End of the Early Years

Arnold W.M. Smeulders, *Senior Member, IEEE*, Marcel Worring, Simone Santini, *Member, IEEE*, Amarnath Gupta, *Member, IEEE*, and Ramesh Jain, *Fellow, IEEE*

Abstract—The paper presents a review of 200 references in content-based image retrieval. The paper starts with discussing the working conditions of content-based retrieval: patterns of use, types of pictures, the role of semantics, and the sensory gap. Subsequent sections discuss computational steps for image retrieval systems. Step one of the review is image processing for retrieval sorted by color, texture, and local geometry. Features for retrieval are discussed next, sorted by: accumulative and global features, salient points, object and shape features, signs, and structural combinations thereof. Similarity of pictures and objects in pictures is reviewed for each of the feature types, in close connection to the types and means of feedback the user of the systems is capable of giving by interaction. We briefly discuss aspects of system engineering: databases, system architecture, and evaluation. In the concluding section, we present our view on: the driving force of the field, the heritage from computer vision, the influence on computer vision, the role of similarity and of interaction, the need for databases, the problem of evaluation, and the role of the semantic gap.

PROGRESS IN IMAGE RETRIEVAL

- By 2008 the field blossomed, but social context mostly ignored

Image Retrieval: Ideas, Influences, and Trends of the New Age

5

RITENDRA DATTA, DHIRAJ JOSHI, JIA LI, and JAMES Z. WANG

The Pennsylvania State University

We have witnessed great interest and a wealth of promise in content-based image retrieval as an emerging technology. While the last decade laid foundation to such promise, it also paved the way for a large number of new techniques and systems, got many new people involved, and triggered stronger association of weakly related fields. In this article, we survey almost 300 key theoretical and empirical contributions in the current decade related to image retrieval and automatic image annotation, and in the process discuss the spawning of related subfields. We also discuss significant challenges involved in the adaptation of existing image retrieval techniques to build systems that can be useful in the real world. In retrospect of what has been achieved so far, we also conjecture what the future may hold for image retrieval research.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*; I.4.9 [**Image Processing and Computer Vision**]: Applications

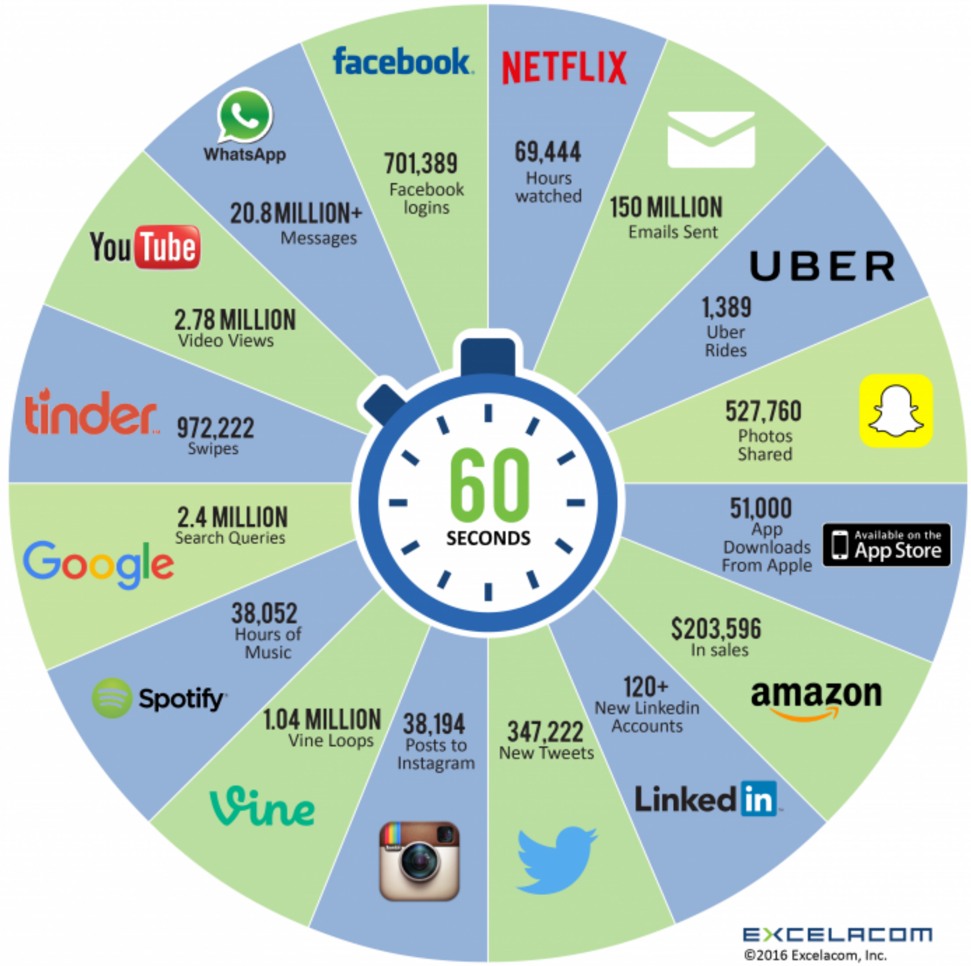
General Terms: Algorithms, Documentation, Performance

Additional Key Words and Phrases: Content-based image retrieval, annotation, tagging, modeling, learning

ACM Reference Format:

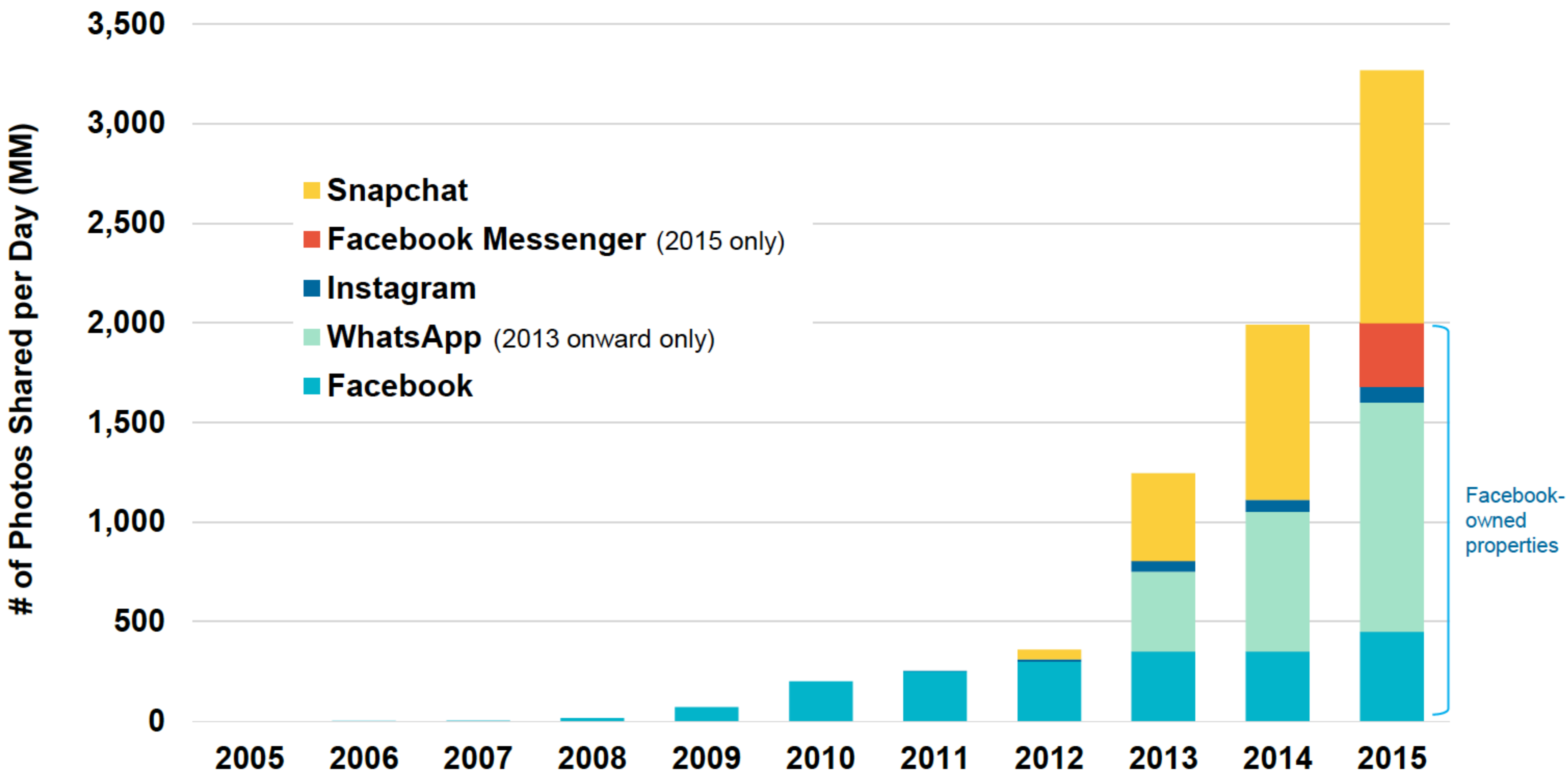
Datta, R., Joshi, D., Li, J., and Wang, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40, 2, Article 5 (April 2008), 60 pages DOI = 10.1145/1348246.1348248 http://doi.acm.org/10.1145/1348246.1348248

Images want to be shared

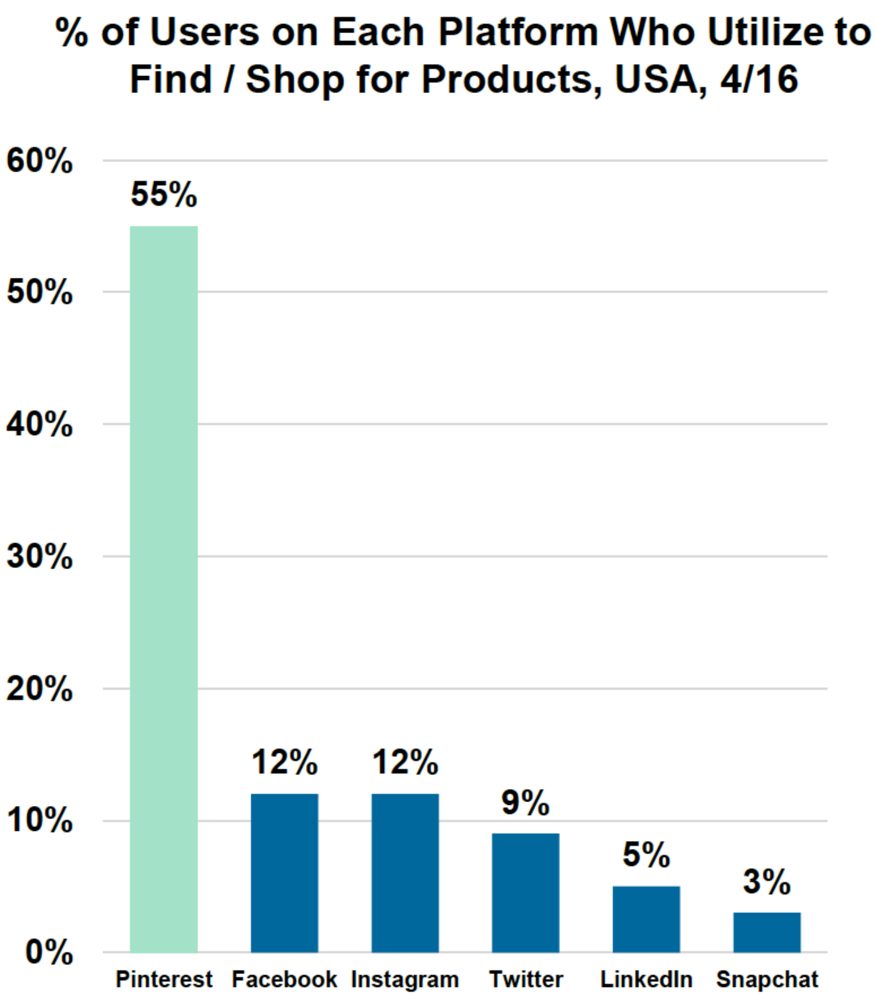


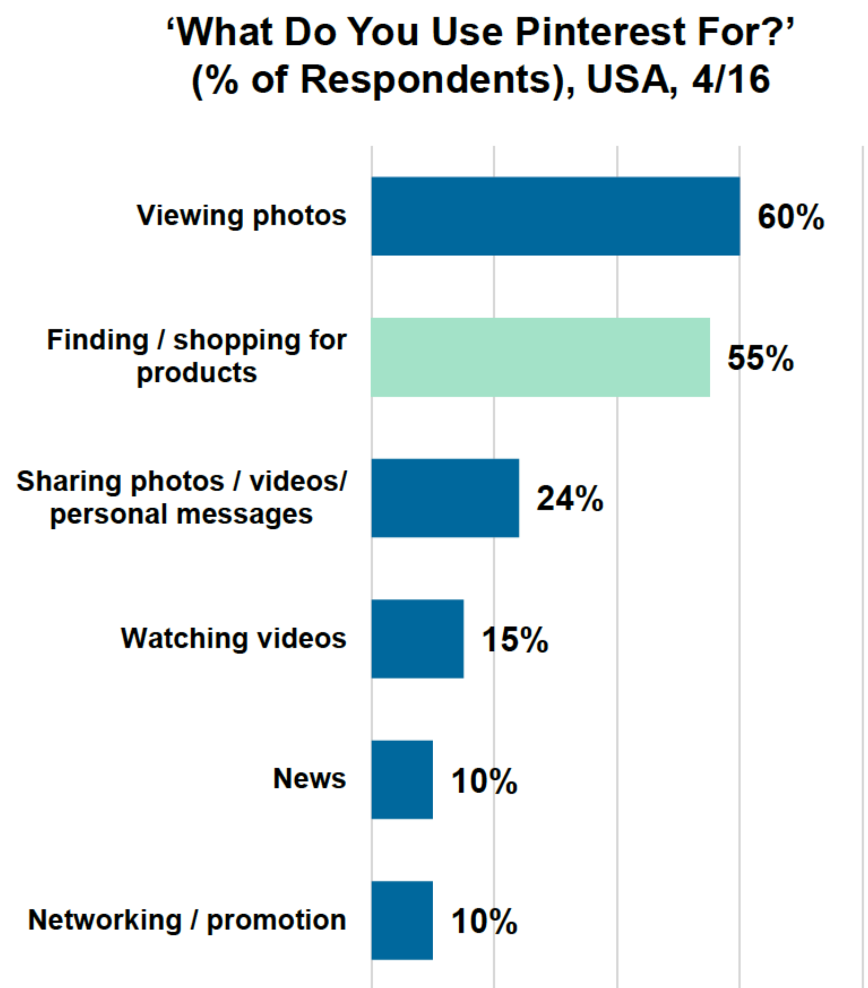
Almost all these services allow users to tag, rate, like, and swipe photos.

Daily Number of Photos Shared on Select Platforms



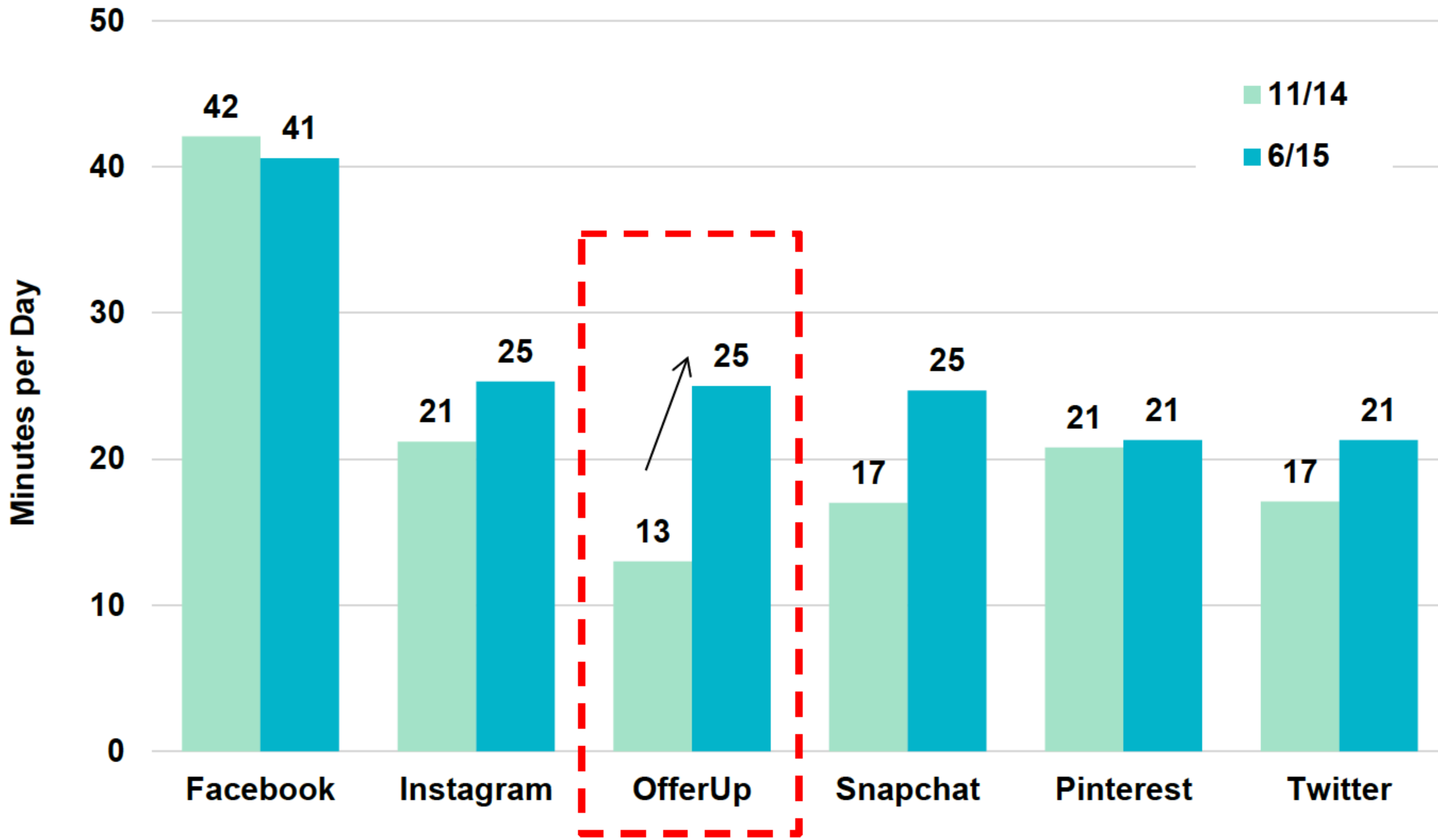
BUSINESS CASE



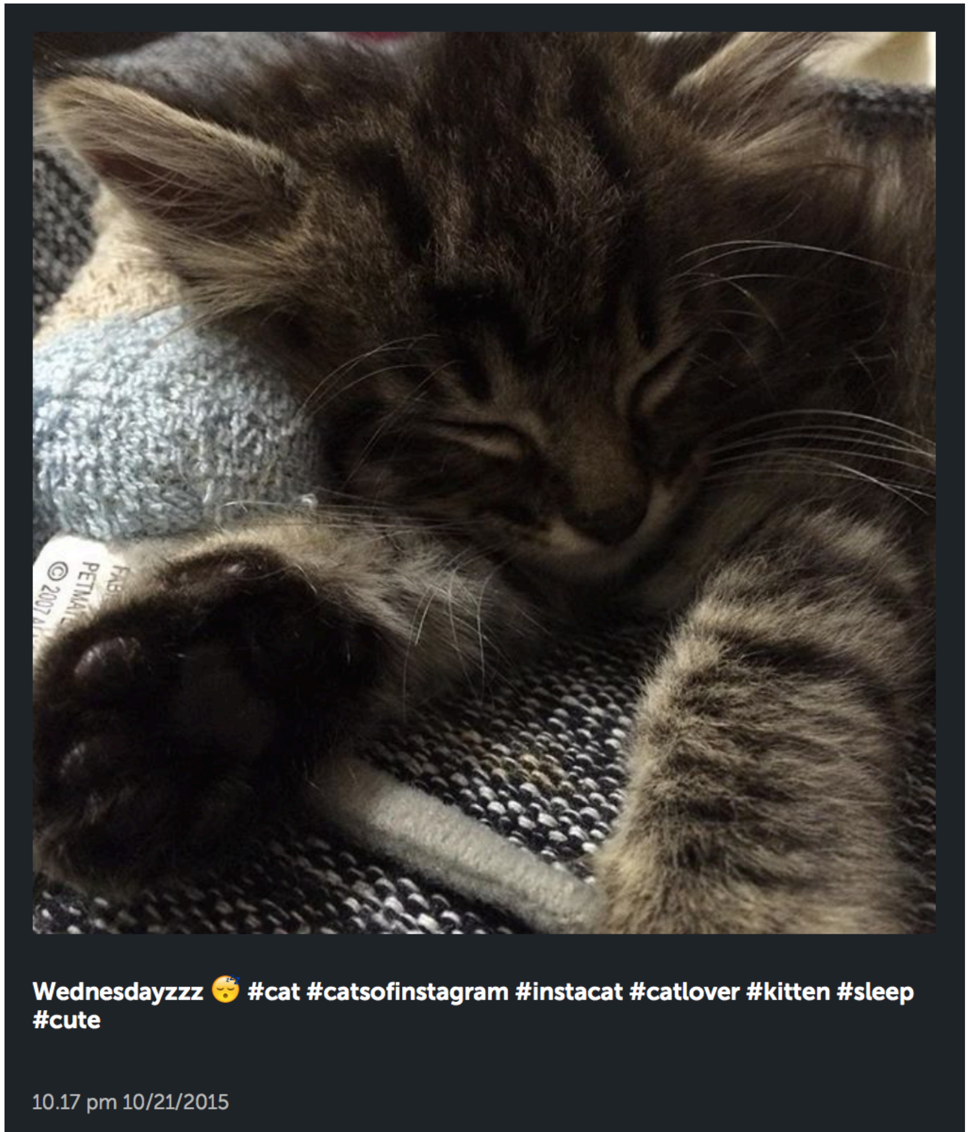


Mary Meeker Internet Trends 2016

AVERAGE DAILY TIME SPENT PER USA USER



EXAMPLES


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

EXAMPLES

Tags BETA ? Add tags

People in photo Add people

pentaxk10d 31mm

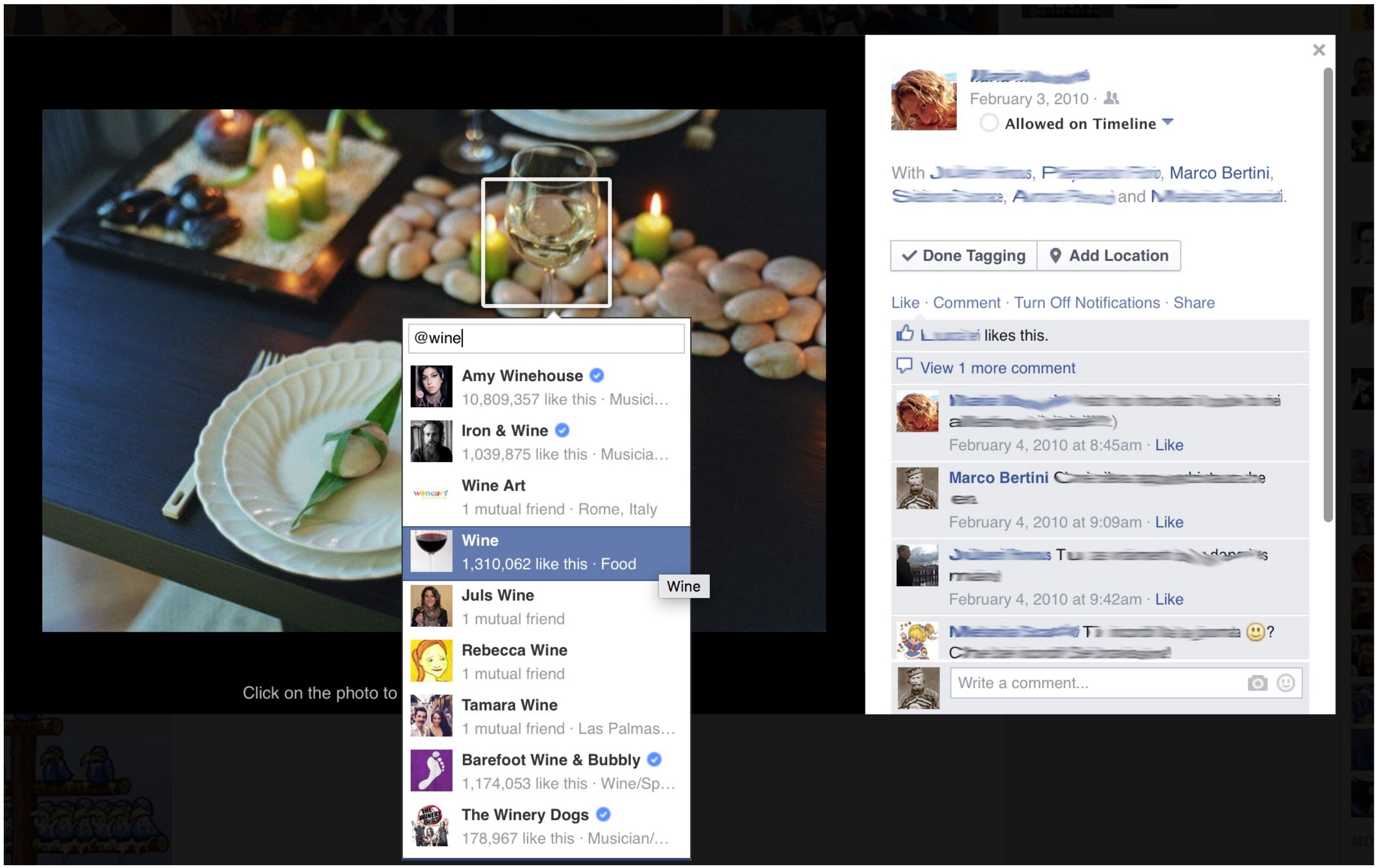
Beijing sonnet 116

stealing photoshooting

chinese bride

Remember That Moment ...

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...

PROBLEMS OF TAGS: **IRRELEVANCE**

- Tags are few, imprecise, ambiguous, and overly personalized

Nikon
Airplane
2016

PROBLEMS OF TAGS: **DYNAMICS**

- In a social network, users continuously add images and create new terms given the freedom of tagging.

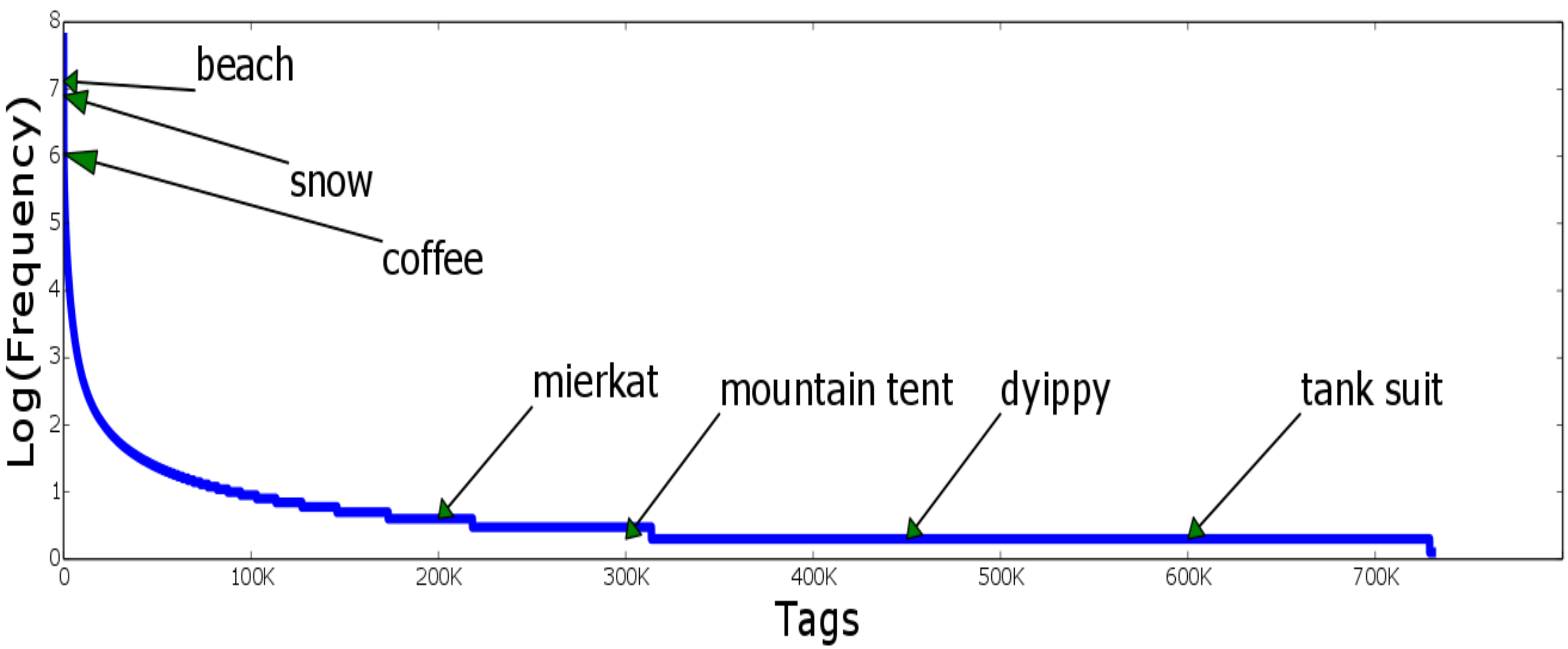
Brexit

PROBLEMS OF TAGS: **SCALE**

- Web-scale quantity of media.

THE LONG TAIL OF IMAGE TAGS

- Some tags are popular and have millions of example images.
- Others are rare, occurring in few images



Kordumova et al. MMM 2016

TAGGING BEHAVIOR

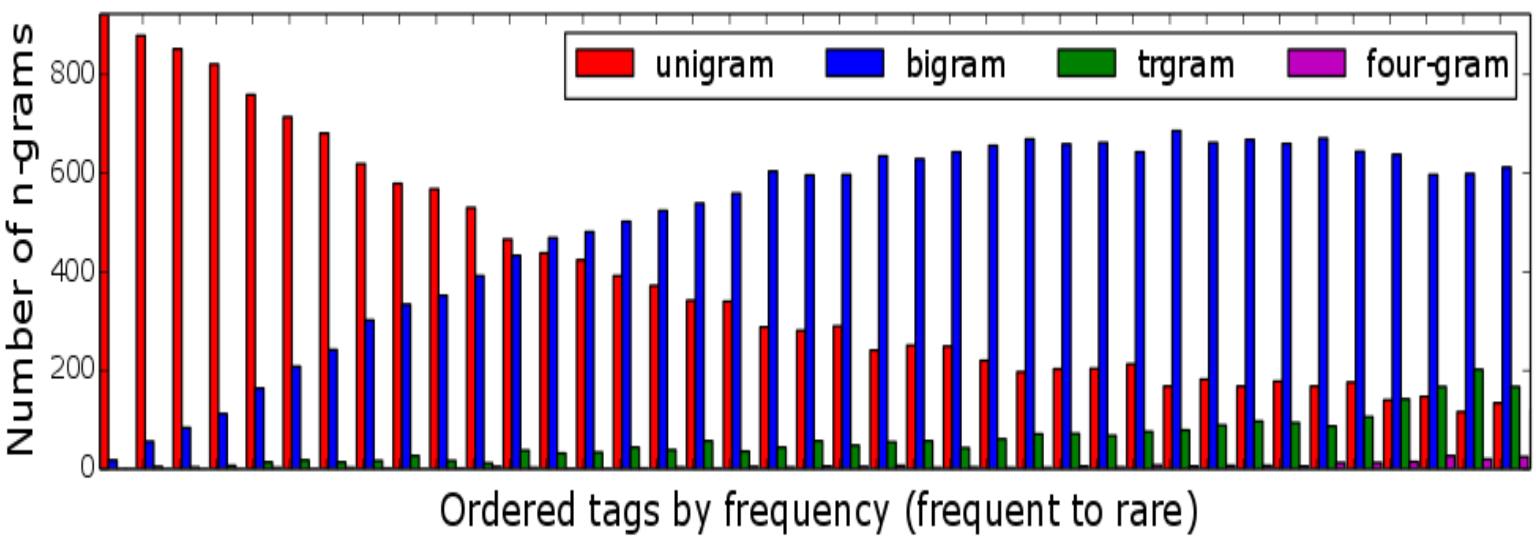
Study by Sigurbjörnsson and van Zwol in WWW 2008 on Flickr

- 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.

An n-gram perspective

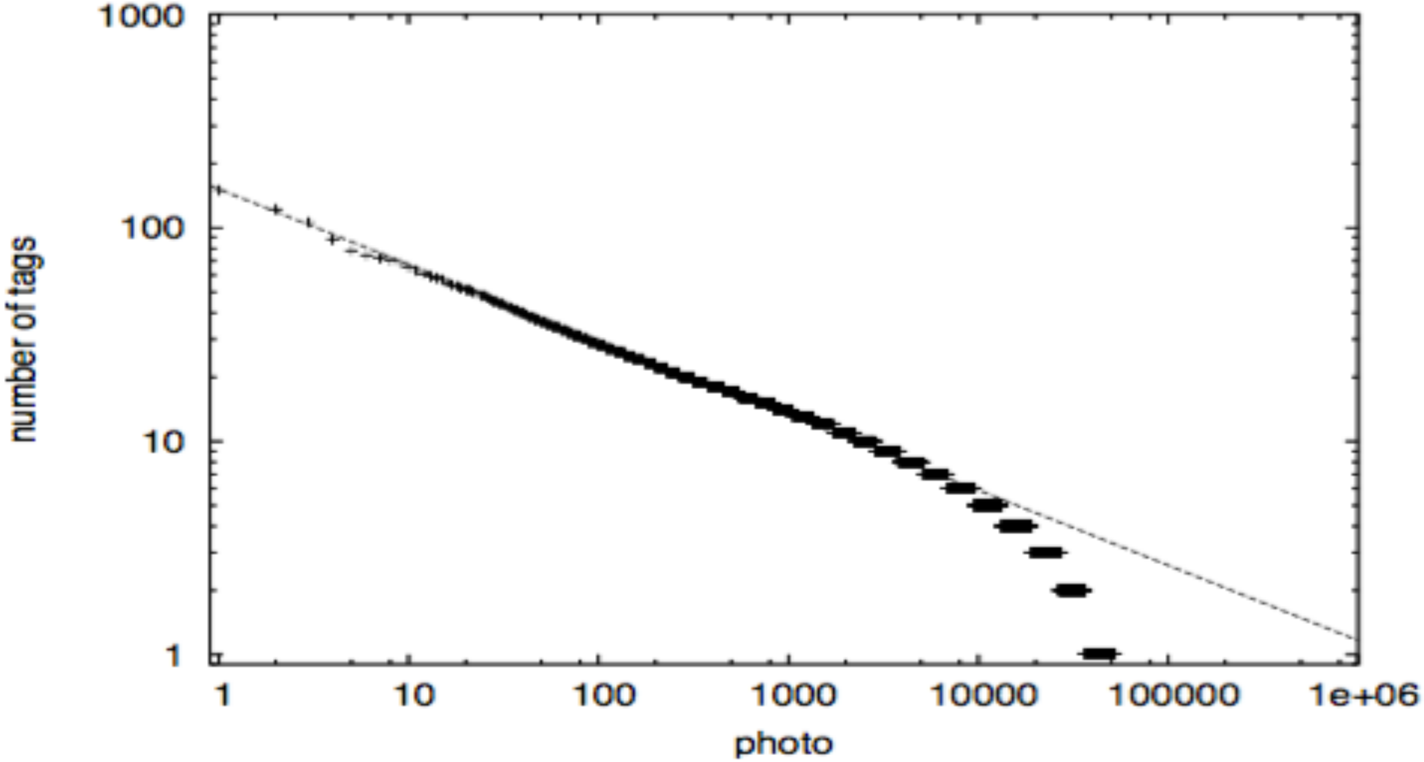
Study by Kordumova et al in MMM 2016 on Flickr

- Most of the frequent tags are unigrams.
- As the frequency goes down more bigrams appear.
- Towards the end trigrams and four-grams occur



christmas tree | kaffir cat | mediterranean water shrew | wine cellar barrel storage

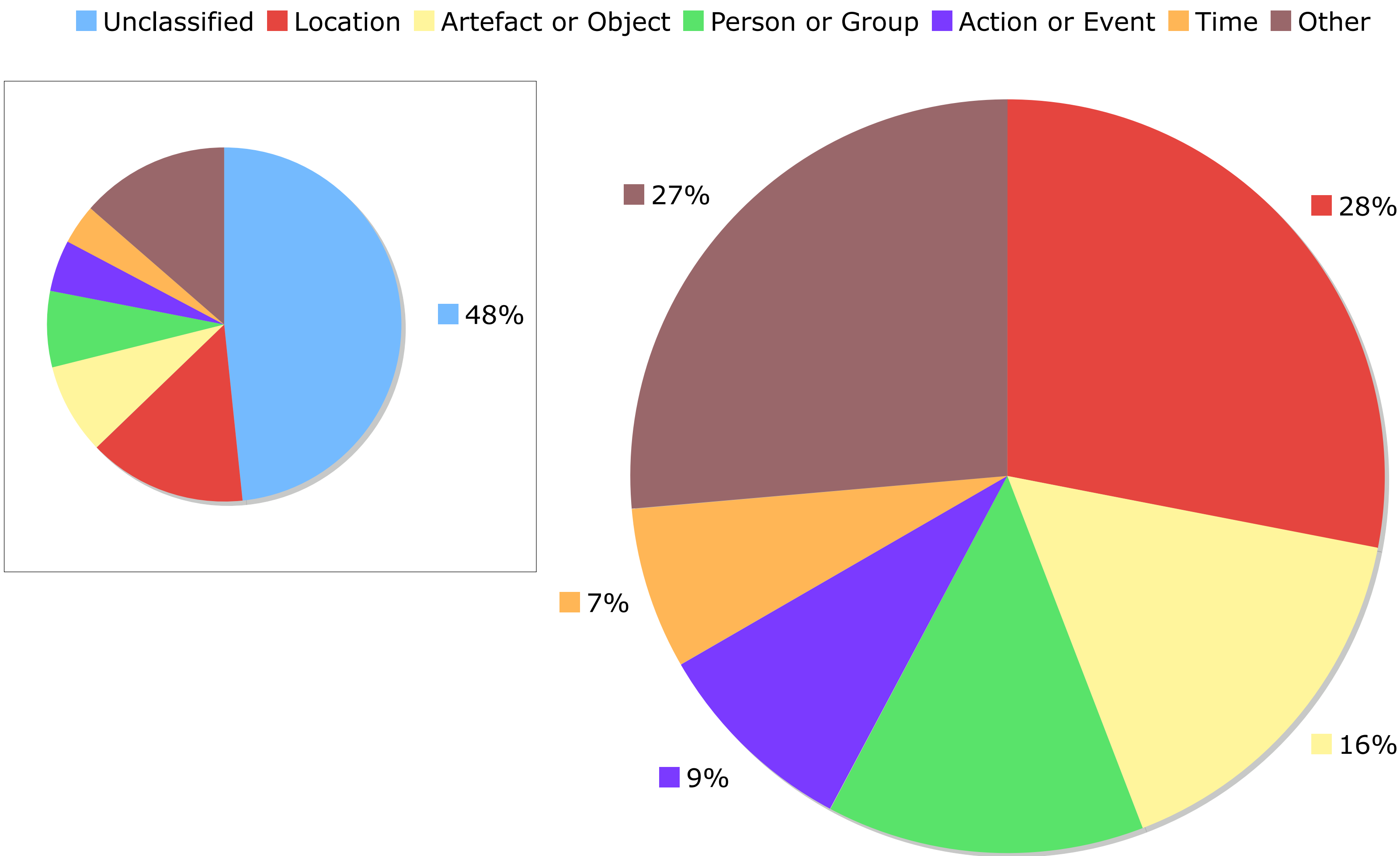
TAGS PER PHOTO (IN 2008)



- A few photos are exceptionally well tagged
- 64% of photos have 1, 2 or 3 tags only.

Sigurbjörnsson and van Zwol, WWW 2008

WordNet Categories of tags



- 48% of 3.7M tags could not be matched.

Sigurbjörnsson and van Zwol, WWW 2008

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 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 11 methods of varied learning complexity.

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

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 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 11 methods of varied learning complexity.

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

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 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 11 methods of varied learning complexity.

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

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 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 11 methods of varied learning complexity.

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

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 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 11 methods of varied learning complexity.

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

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

PART 2
TAXONOMY

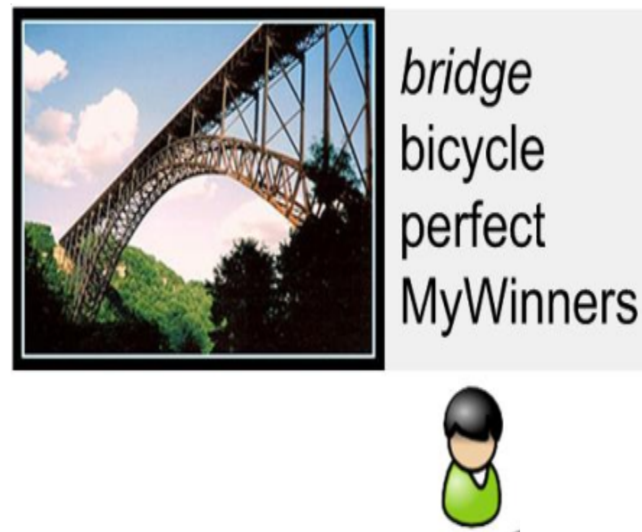
- Foundations
  - tag relevance

- 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***

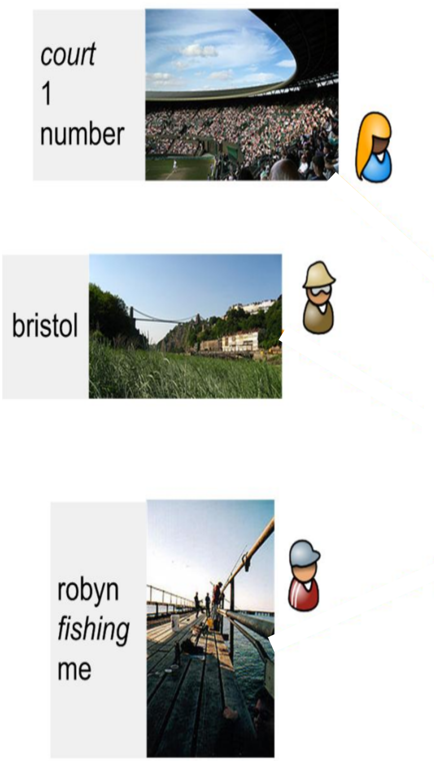


- A user ***u*** can share an image ***x***, assigning tag ***t*** to it

FOUNDATIONS

A set of users ***U*** contributes a set of *n* socially tagged images ***X***.

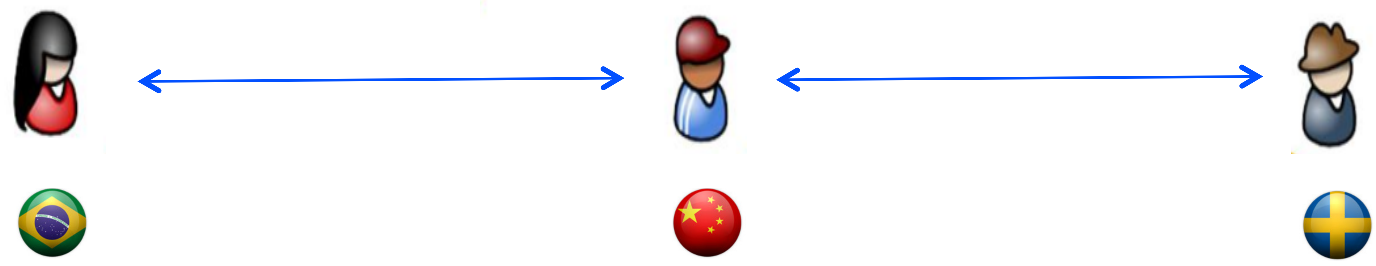
All tags used to describe ***X*** form a vocabulary ***V*** composed of *m* tags.



Vocabulary = {court, 1, number, bristol, roby, fishing, me}

Foundations

- 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.)



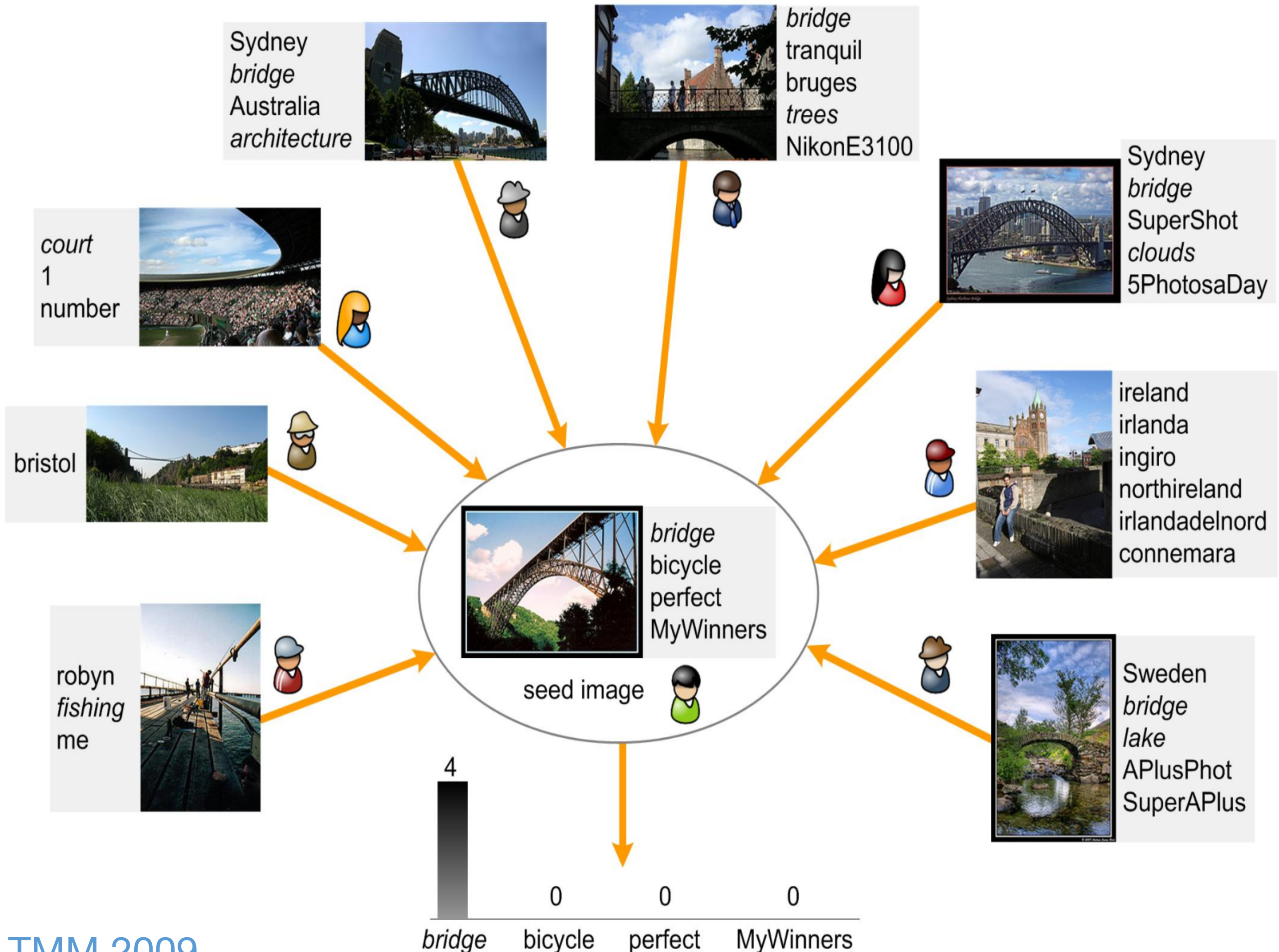
Tag relevance

Tag assignment, refinement and retrieval share an 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 $\boldsymbol{\Theta}$**:**

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

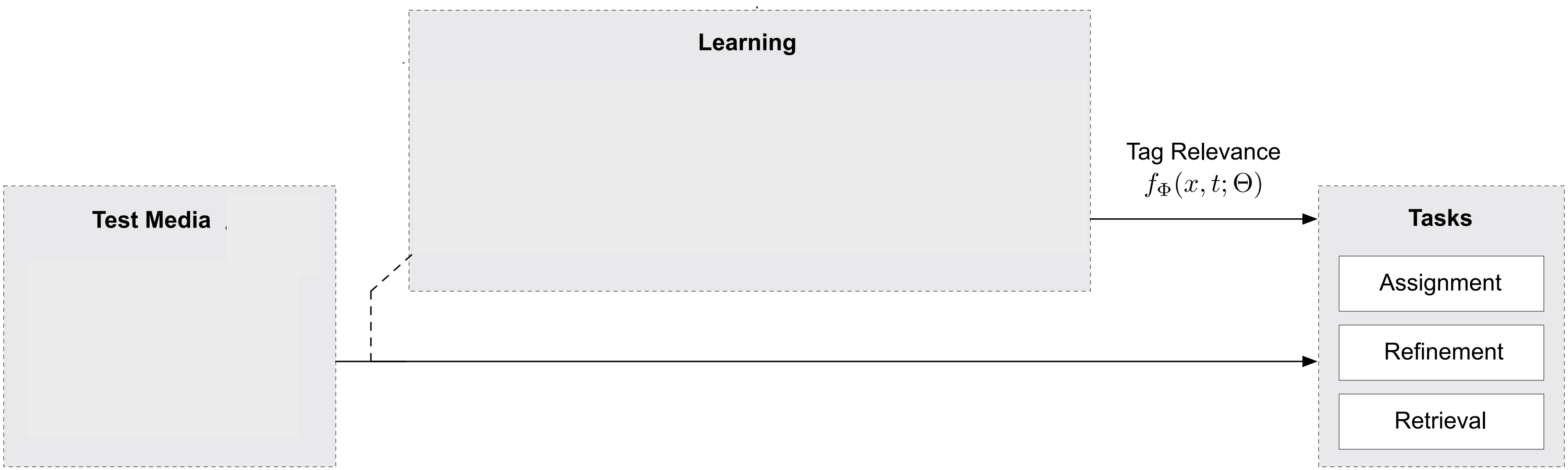
Example for tag refinement



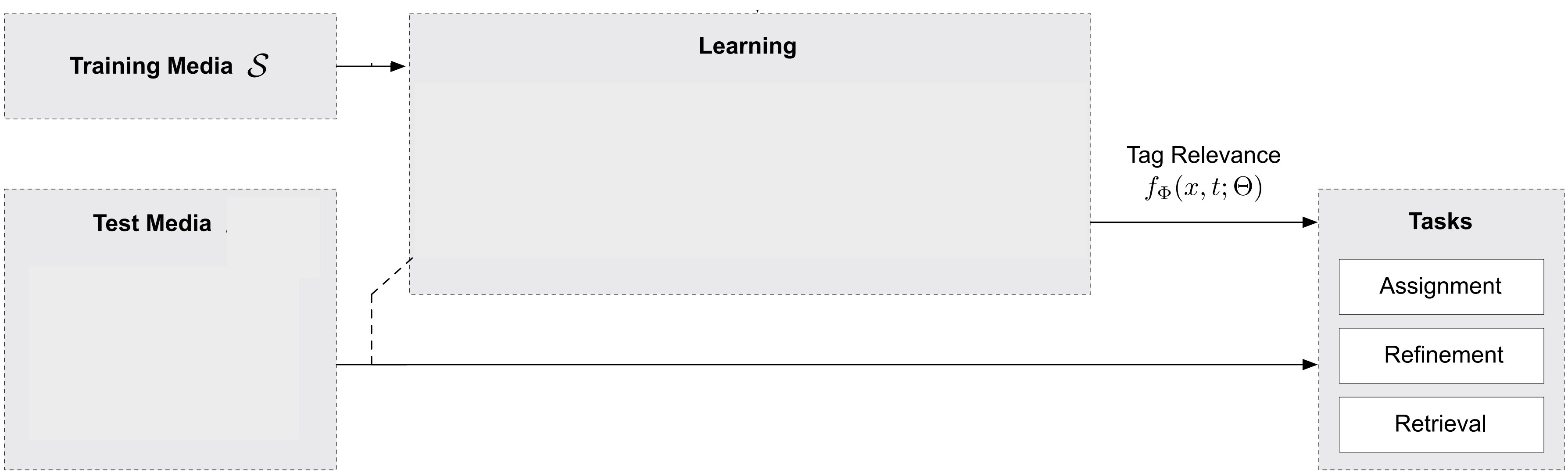
Li et al. TMM 2009

Unified Framework

UNIFIED FRAMEWORK

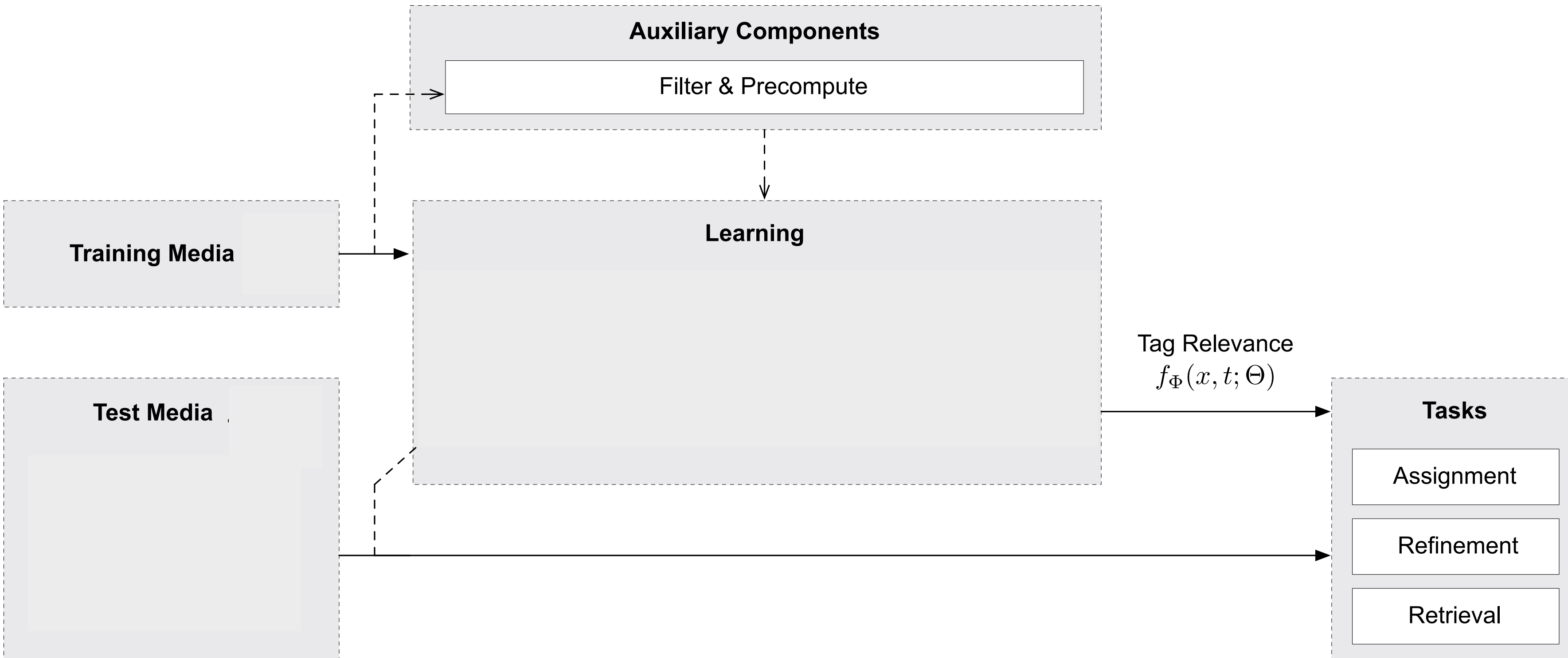


UNIFIED FRAMEWORK



Training media is obtained from social networks, i.e. with unreliable user-generated annotations. It can be filtered to remove unwanted tags or images.

UNIFIED FRAMEWORK



Training media is obtained from social networks, i.e. with unreliable user-generated annotations. It can be filtered to remove unwanted tags or images.

AUXILIARY COMPONENTS: FILTER

- A common practice is to eliminate overly personalized tags like 'hadtopostsomething'
  - e.g. by excluding tags that are not part of WordNet or Wikipedia

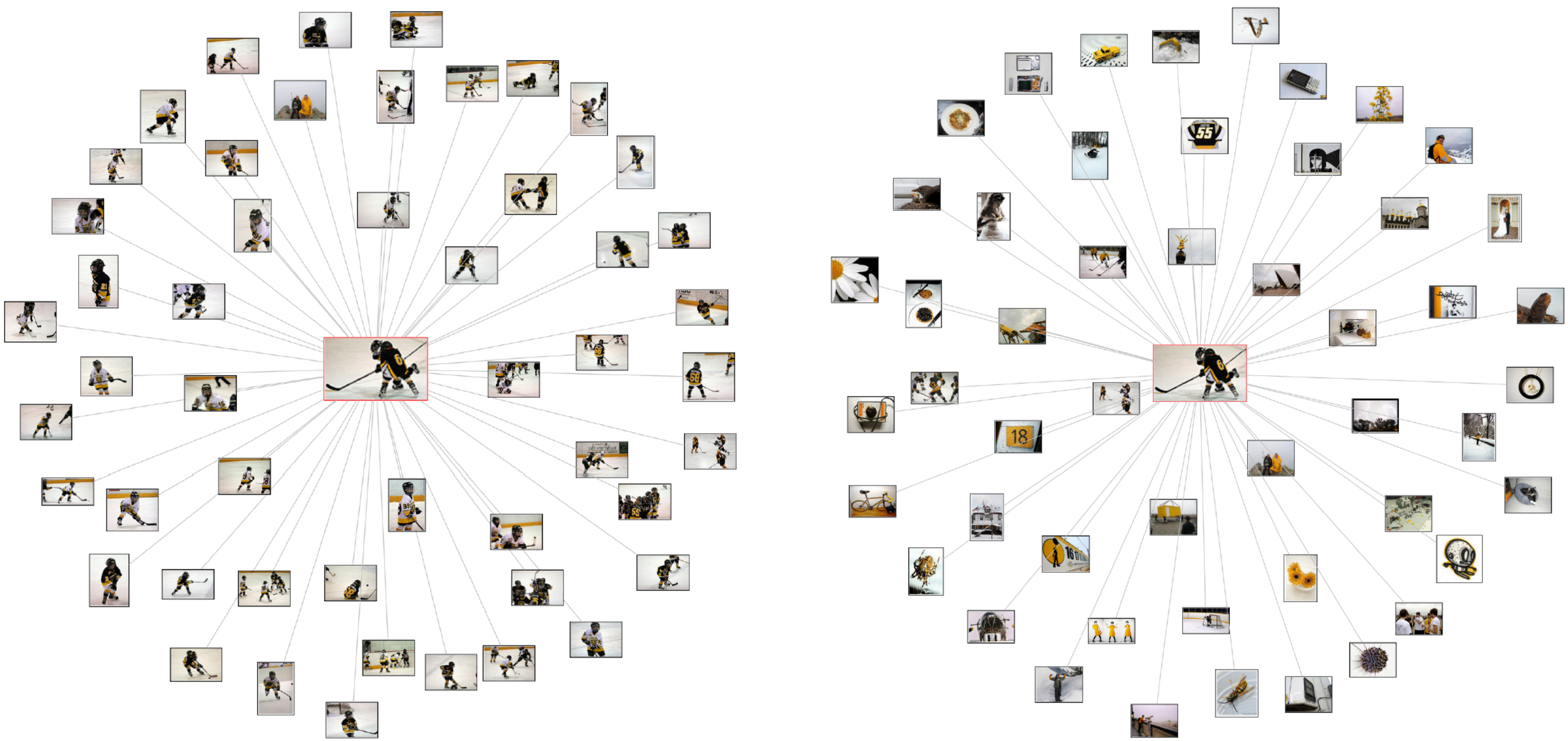
- Often tags that do not appear enough times in the collection are eliminated.

- Reduction of vocabulary size is also important for when using an image-tag association matrix

- Since batch tagging tends to reduce the quality of tags, these types of images can be excluded

BATCH TAGGING

A unique user constraint prevents ‘spam’ from batch tagging



(a) (b)

Li et al. TMM 2009

AUXILIARY COMPONENTS: PRECOMPUTE

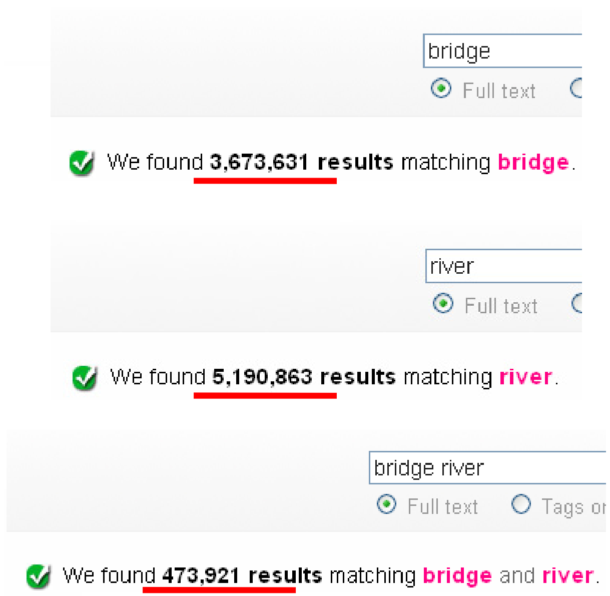
- It is practical to precompute information for the learning.

- A common precomputation is tag occurrence and co-occurrence.

- Occurrence can be used to penalize excessively frequent tags

- Co-occurrence is used to capture semantic similarity of tags directly from users' behavior
  - Semantic similarity typically obtained by Flickr context distance

FLICKR CONTEXT DISTANCE



h(x)

h(y)

h(x,y)

FCS (bridge, river) = 0.65

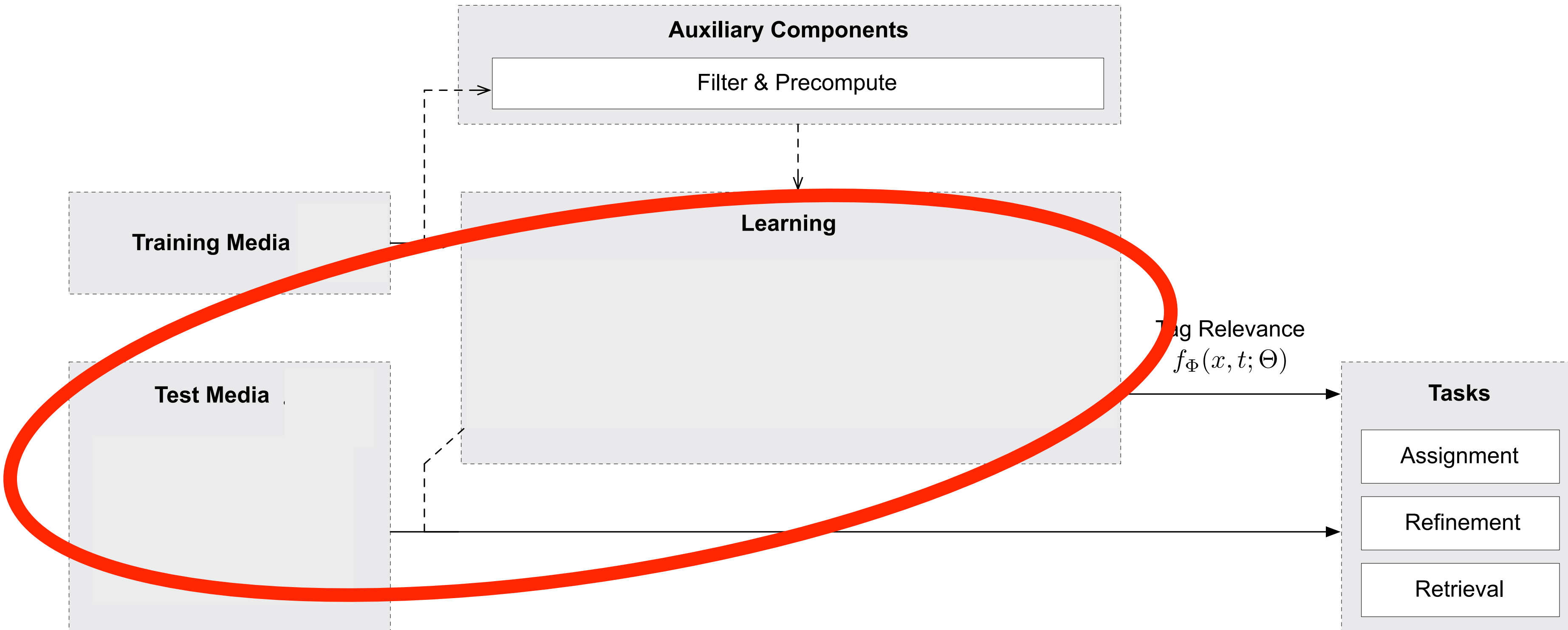
- Based on the Normalized Google Distance.
- Measures the co-occurence of two tags with respect to their single tag occurrencies.
- No semantics is involved, works for any tag.

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

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

[Jiang et al. 2009]

UNIFIED FRAMEWORK



Taxonomy

Media	Learning		
	Instance	Model	Transductive
Tag	2	1	-
Tag + Image	13	15	12
Tag + Image + User	5	7	3

Taxonomy structures 60 papers along **Media** and **Learning** dimensions

TAXONOMY

Media	Learning		
	Instance	Model	Transductive
Tag	2	1	-
Tag + Image	13	15	12
Tag + Image + User	5	7	3

Taxonomy structures 60 papers along **Media** and **Learning** dimensions

MEDIA FOR TAG RELEVANCE

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

- **Tag**
  - e.g. considering ranking of tags as a proxy of user's priorities

- **Tag + image**
  - e.g. considering the set of tags assigned to an image

- **Tag, image + user information**
  - e.g. considering the behaviors of different users tagging similar images

MEDIA: **TAGS**

These methods reduce the problem to text retrieval

Find similarly tagged images by

- user-provided tag ranking [Sun et al. 2011],
- tag co-occurrence [Sigurbjönsson and van Zwol 2008; Zhu et al. 2012] or
- topic modelling [Xu et al. 2009]

These methods assume that test images have already been tagged as well, so unsuited for tag assignment.

MEDIA: **TAGS AND IMAGES**

The main idea of these works is to exploit visual consistency, i.e. the fact that visually similar images should have similar tags.

Three main approaches:

1. Use visual similarity between test image and database
2. Use similarity between images with same tags
3. Learn classifiers from social images + tags

MEDIA: TAGS AND IMAGES

Two tactics to combine the similarity between images and tags

1. **Sequential**: compute visual similarity, then use the tag modality

2. **Simultaneous**: use both modalities at the same time,
   - A unified graph composed by the fusion of a visual similarity graph with an image-tag connection graph [Ma et al. 2010]
   - Tag and image similarities as constraints to reconstruct an image-tag association matrix [Wu et al. 2013; Xu et al. 2014; Zhu et al. 2010]

MEDIA: TAGS, IMAGES AND USER INFO

In addition to tags and images, this group of works exploits user information, motivated from varied perspectives. Such as:

- User identities [Li et al. 2009b],
- Tagging preferences [Sawant et al. 2010],
- User reliability [Ginsca et al. 2014],
- Photo time stamps [Kim and Xing 2013, McParlane et al. 2013a]
- Geo-localization [McParlane et al. 2013b]
- Image group memberships [Johnson et al. 2015]

Taxonomy

Media	Learning		
	Instance	Model	Transductive
Tag	2	1	-
Tag + Image	13	15	12
Tag + Image + User	5	7	3

Taxonomy structures 60 papers along **Media** and **Learning** dimensions

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 compares new test images with training instances.

- There are no parameters and the complexity grows with the number of instances.

- Approaches are typically based on variants of k-NN, with or witout weighted voting

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.

- **Tag-specific:** use linear or fast intersection kernel SVMs trained on features augmented by pre-trained classifiers of popular tags, or relevant positive and negative examples

- **Holistic:** use topic modeling with relevance computed using a topic vector of the image and a topic vector of the tag.

TRANSDUCTION BASED

- This class of methods evaluate tag relevance for a given image-tag pair by minimizing a cost function over a set of images.

- The majority of these methods is based on matrix factorization

PROS AND CONS

Instance-based

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

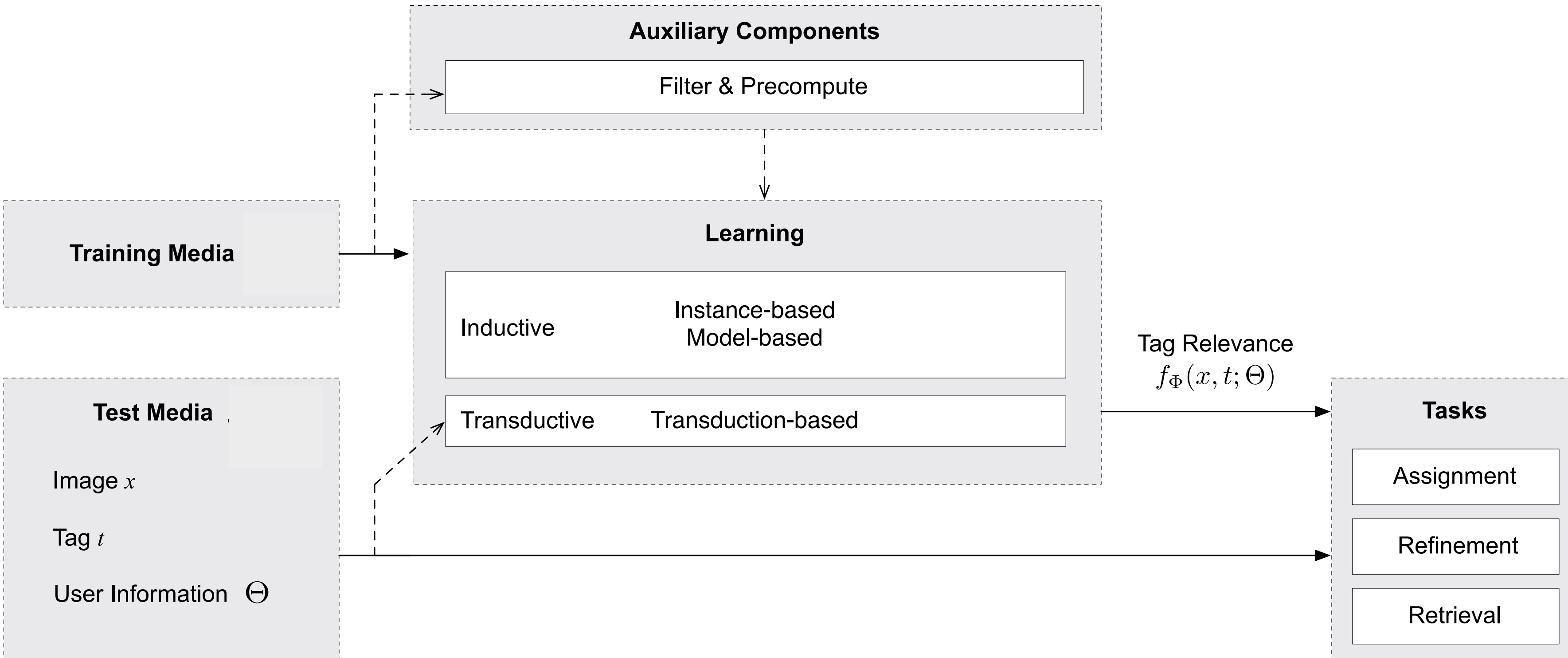
Model-based

- **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.

Transduction-based

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

UNIFIED FRAMEWORK



TAXONOMY

Media	Learning		
	Instance	Model	Transductive
Tag	2	1	-
Tag + Image	13	15	12
Tag + Image + User	5	7	3

Taxonomy structures 60 papers along **Media** and **Learning** dimensions

TAXONOMY

	Learning		
Media	*Instance-based*	*Model-based*	*Transduction-based*
tag	**[Sigurbjörnsson and van Zwol 2008]** TagCooccur [Sun et al. 2011] **[Zhu et al. 2012]** SemanticField	[Xu et al. 2009]	–
tag + image	**[Liu et al. 2009]** TagRanking **[Makadia et al. 2010]** KNN [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]** TagProp [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]** TagFeature [Lan and Mori 2013] **[Li and Snoek 2013]** RelExample [Li et al. 2013] [Wang et al. 2014] [Niu et al. 2014]	**[Zhu et al. 2010]** RobustPCA [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]** TagVote [Kennedy et al. 20 TagCooccur+ [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]	TensorAnalysis **[Sang et al. 2012a]** [Sang et al. 2012b] [Qian et al. 2015]

ORGANIZATION OF THE TUTORIAL

8:30 – 9:30	Part 1: Introduction
	Part 2: Taxonomy
9:30 – 10:00	Part 3: Experimental protocol
	Part 4: Evaluation
10:00 – 10:45	*Coffee break*
10:45 – 12:00	Part 4: Evaluation cont'd
12:00 – 12:30	Part 5: Conclusion and future directions

PART 3
OUR EXPERIMENTAL PROTOCOL

- 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 often not 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)

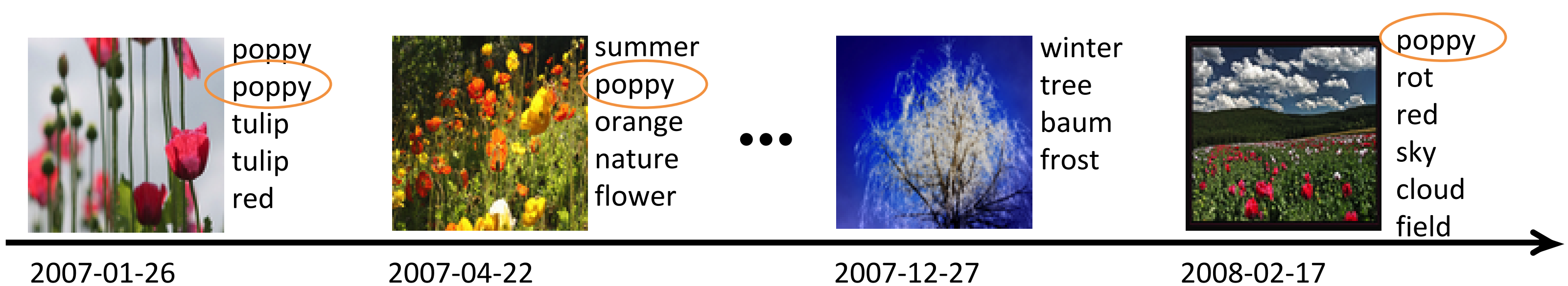
- Training sets of varied size
  - Train1M (a random subset of the collected Flickr images)
  - Train100k (a random subset of Train1m)
  - Train10k (a random subset of Train1m)

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

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

Credit: figure from [Vreeswijk et al. 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

http://press.liacs.nl/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

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

NUS-WIDE

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

- 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

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]

- CNN features
  - 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/REFINEMENT

- 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 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

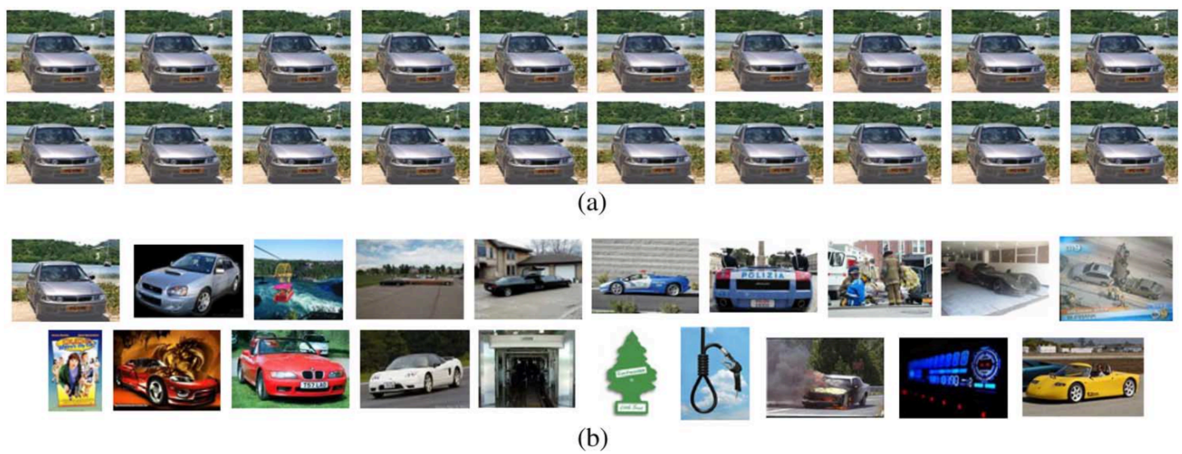


Figure from [Wang et al. 2010]

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

Part 4
Evaluation: Eleven Key Methods

- **Goal:** evaluates key methods based on various Media and Learning paradigm
- Q: What are their key ingredients ?
- Q: What is the computational cost of each of them ?

KEY METHODS

- Covering all published methods is obviously impractical
- We do not consider methods:
  - Which do not show significant improvements or novelties w.r.t. the seminal papers in the field
  - Methods that are difficult to replicate
- We drive our choice by the intention to cover methods that aim for each of the three tasks, exploiting varied modalities and using distinct learning mechanisms
- We select **11 representative methods**

KEY METHODS

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

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

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]

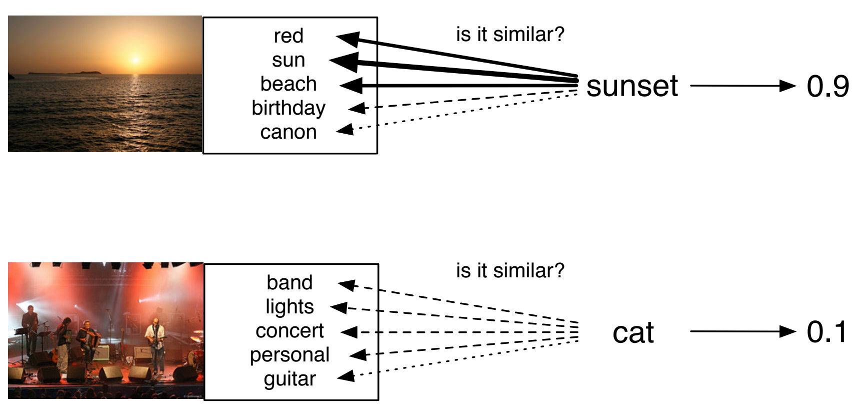
KEY METHODS

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



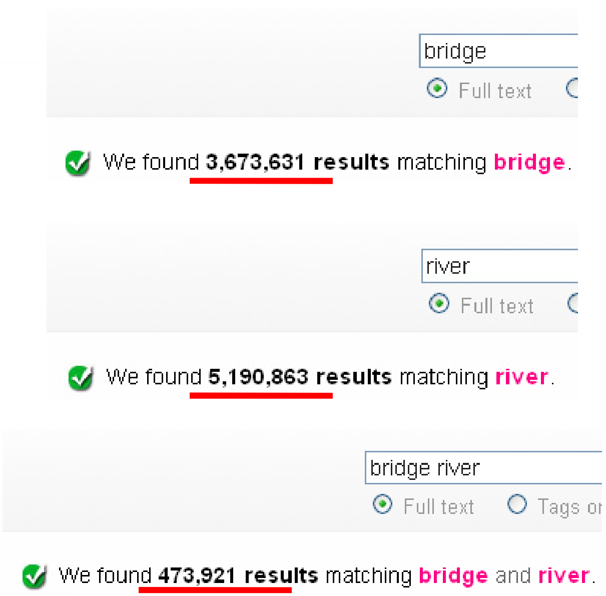
Zhu et al. Sampling and Ontologically Pooling Web Images for Visual Concept Learning. IEEE TMM 2012

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 their single tag occurrencies.
- No semantics is involved, works for any tag.

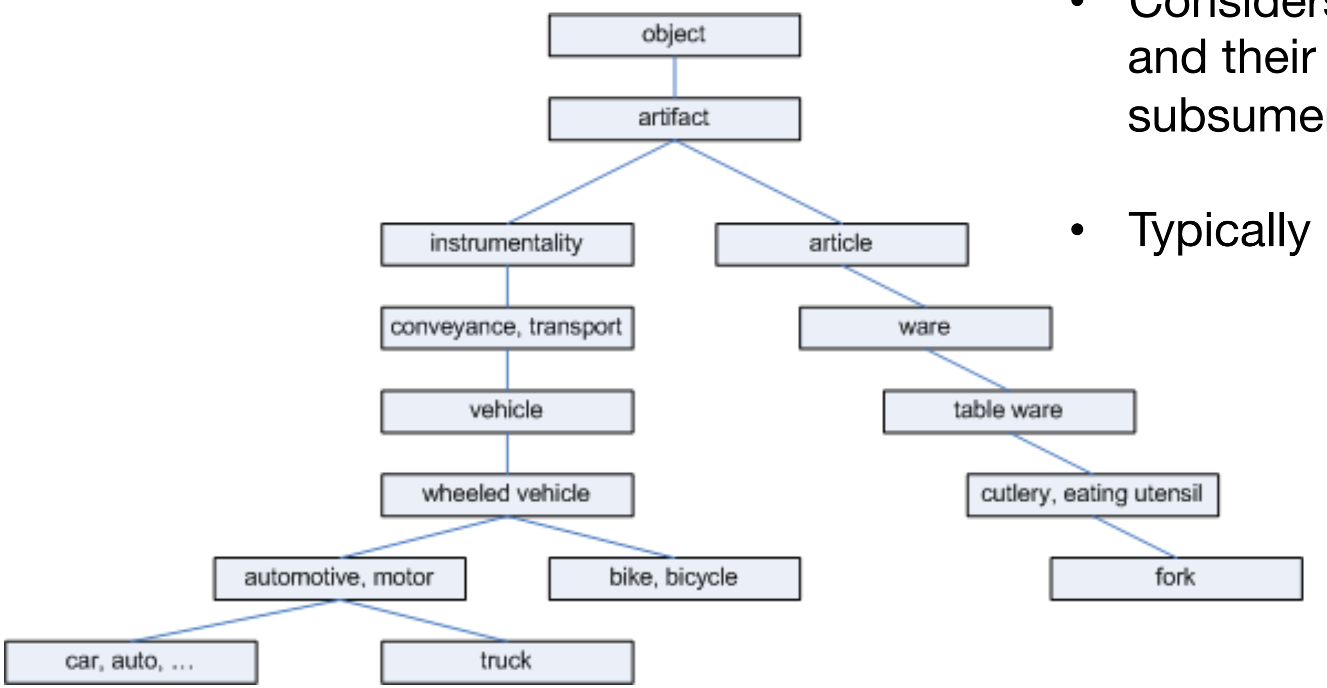
$$\mathrm{NGD}(x,y) = \frac{\max\{\log h(x), \log h(y)\} - \log h(x,y)}{\log N - \min\{\log h(x), \log h(y)\}},$$

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

Y-G. Jiang, C.-W. Ngo, S.-F. Chang. Semantic context transfer across heterogeneous sources for domain adaptive video search. ACM Multimedia 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.

Z. Wu, M. Palmer. Verb semantics and lexical selection. ACL 1994

SemanticField

[Zhu et al. 2012] Instance-Based Tag

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

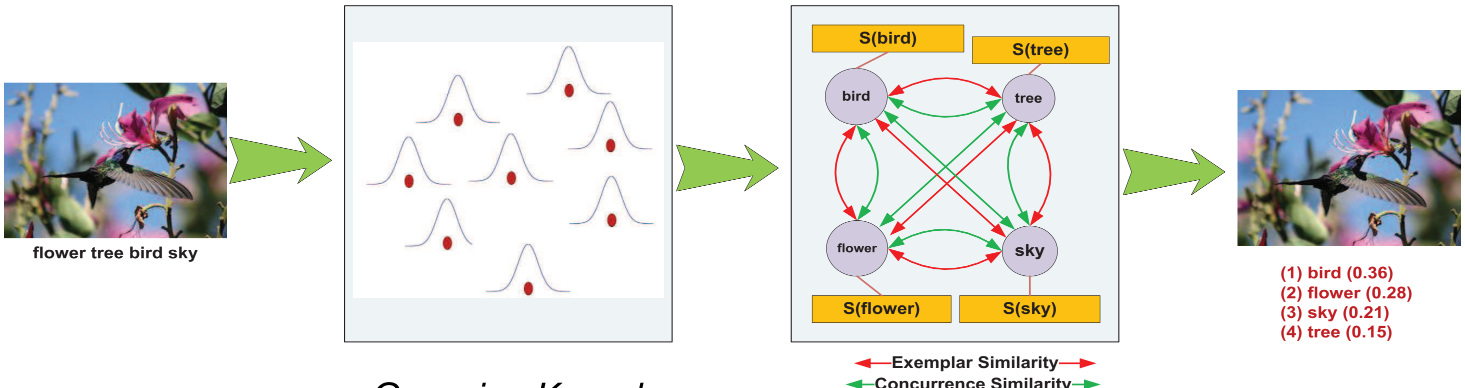
- *Sim* is the similarity between t and the other image tags
- Needs some user tags. Not applicable to Tag Assignment
- Complexity $O(m \cdot l_x)$: the number of image tags l_x times m tags
- Memory $O(m^2)$: quadratic w.r.t. the vocabulary of m tags

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.

D. Liu, X.-S. Hua, L. Yang, M. Wang, H.-J. Zhang. Tag ranking. WWW 2009

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.

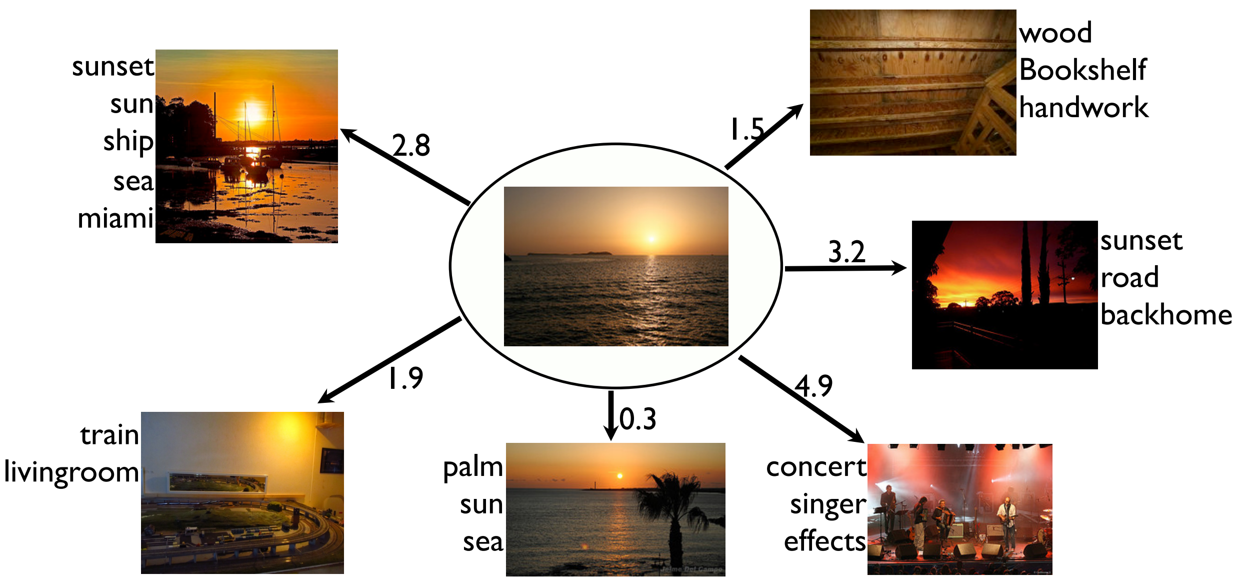
- Complexity $O(m \cdot d \cdot n + L \cdot m^2)$: KDE on n images + L iter random walk
- Memory $O(\max(d \cdot n, m^2))$: max of the two steps

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** **TagCooccur+** [Li et al. 2009b]		**TensorAnalysis** [Sang et al. 2012a]

KNN

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

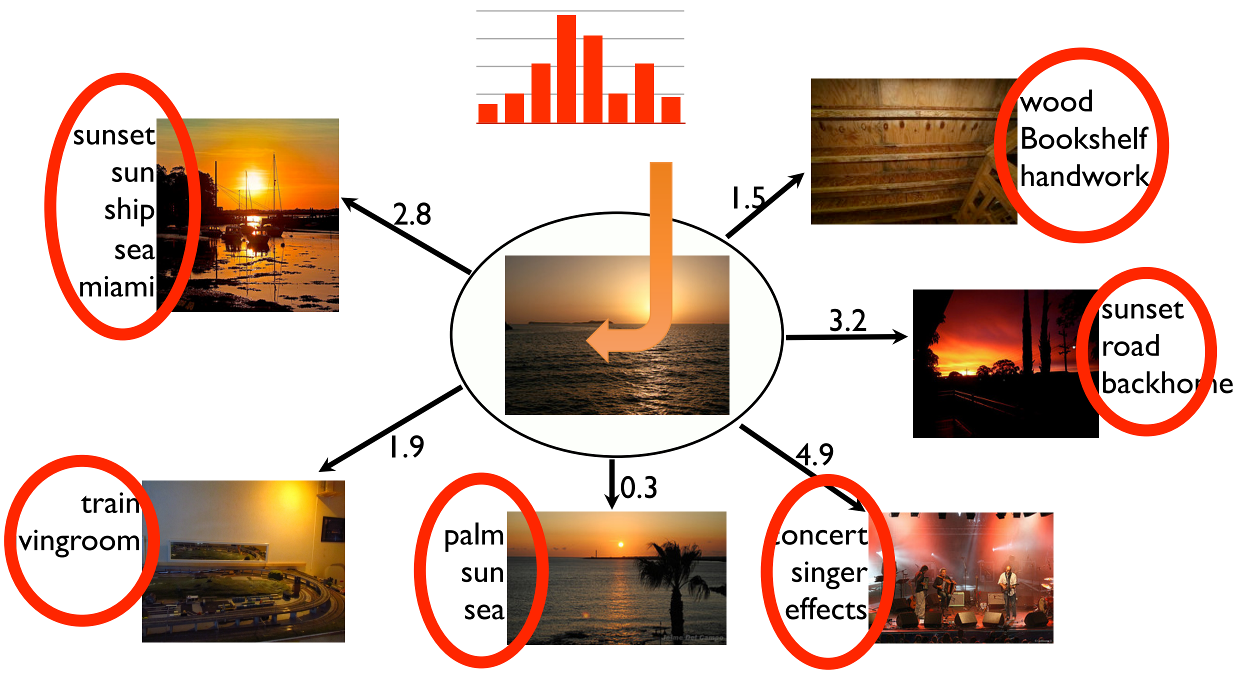


- Similar images share similar tags
- Finds k nearest images with a distance d
- Counts the frequency of tags in the neighborhood
- Assign the top ranked tags to the test image

A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. ECCV 2008

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

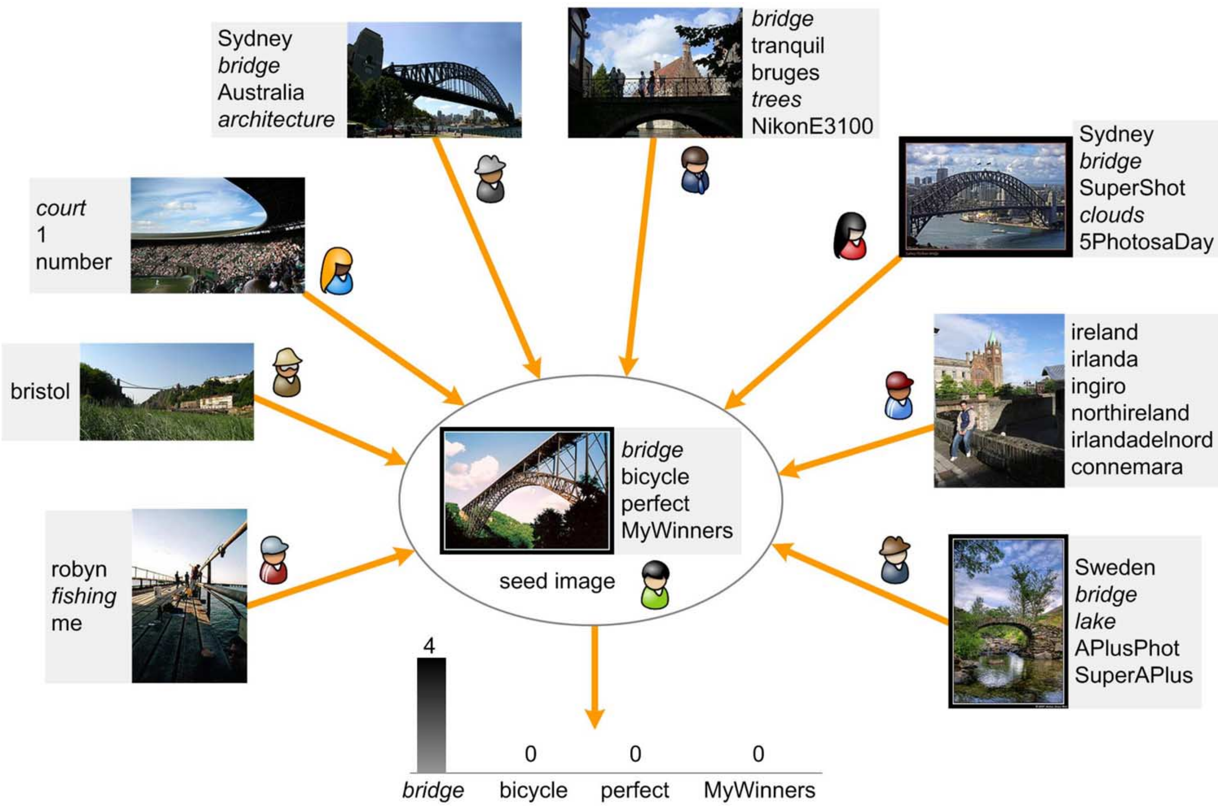
[Makadia et al. 2010] Instance-Based Tag + Image

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

- 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 to KNN-voting:
  - Unique-user constraint
  - Tag prior frequency

X. Li, C. Snoek, M. Worring. Learning Social Tag Relevance by Neighbor Voting. IEEE TMM 2009

TagVote

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

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

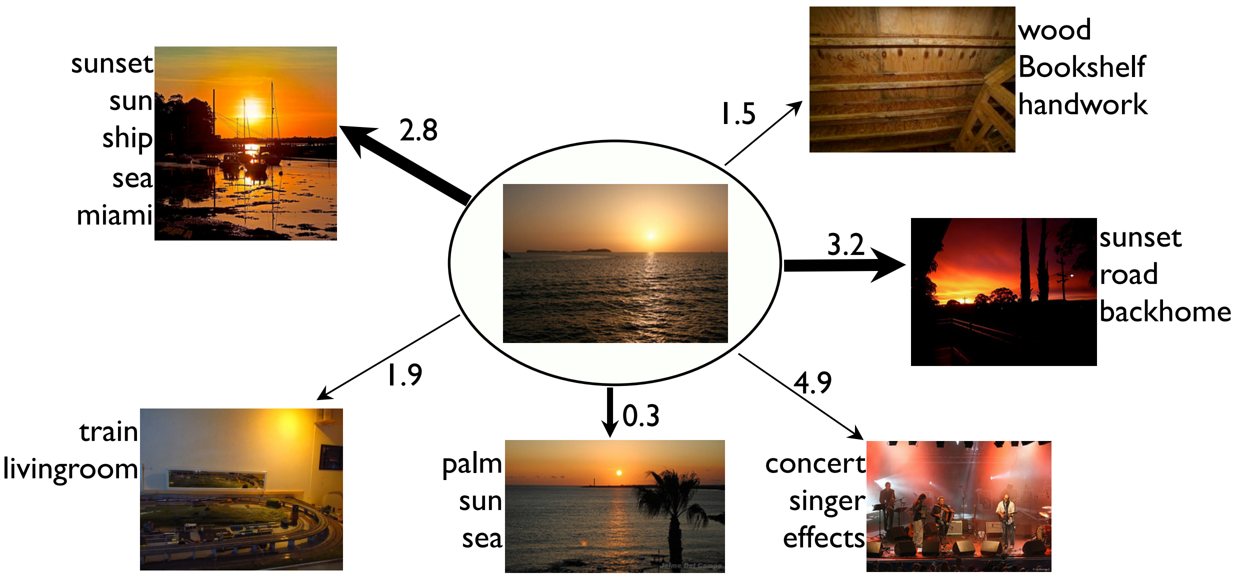
- k_t is the number of images with t in the visual neighborhood of x
- n_t is the frequency of tag t in S

- 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



- Key improvement: give different weights to image neighborhs
- 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

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

M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid. TagProp: Discriminative Metric Learning in Nearest Neighbor Models for Image Auto-Annotation. ICCV 2009

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.

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

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

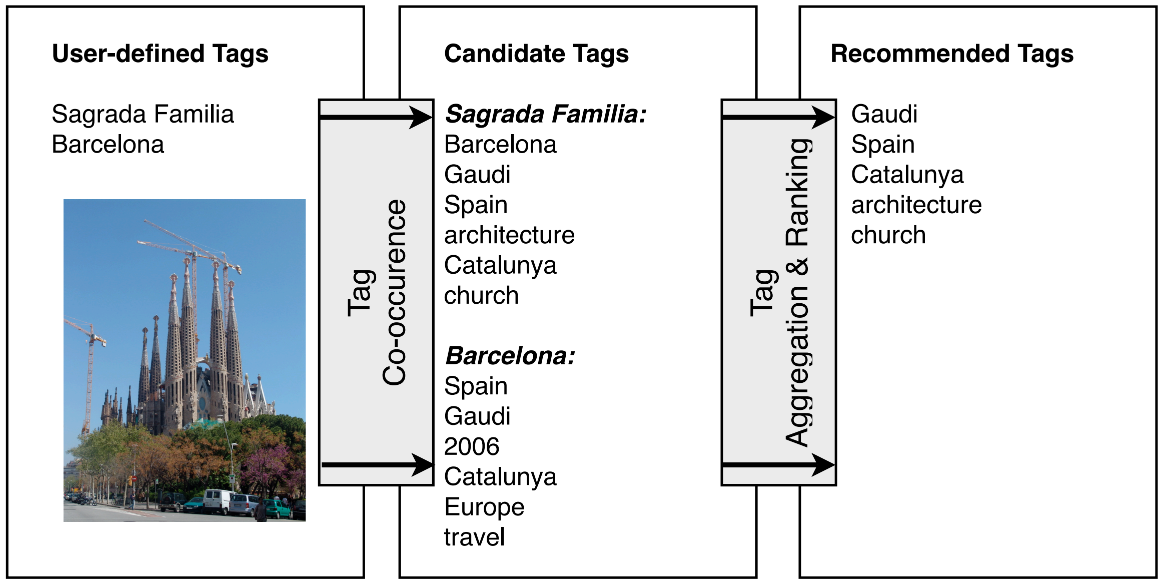
- Complexity O(l · m · k): l steps of gradient descent
- Memory O(d · |S|): same as KNN, extra 2m for logistic regression

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

B. Sigurbjörnsson, R. van Zwol. Flickr tag recommendation based on collective knowledge, WWW 2008

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

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

- $r_c(t)$ is the rank of t when sorting $f_{tagvote}(x,t)$ in descending order. k_c is a positive weighting parameter

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

X. Li, C. Snoek, M. Worring. Learning Social Tag Relevance by Neighbor Voting. IEEE TMM 2009

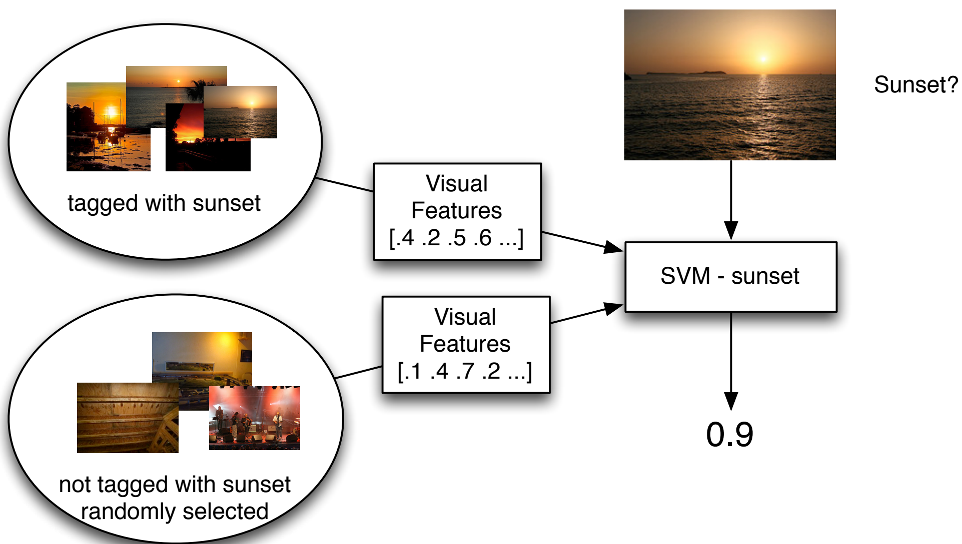
Key Methods

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.



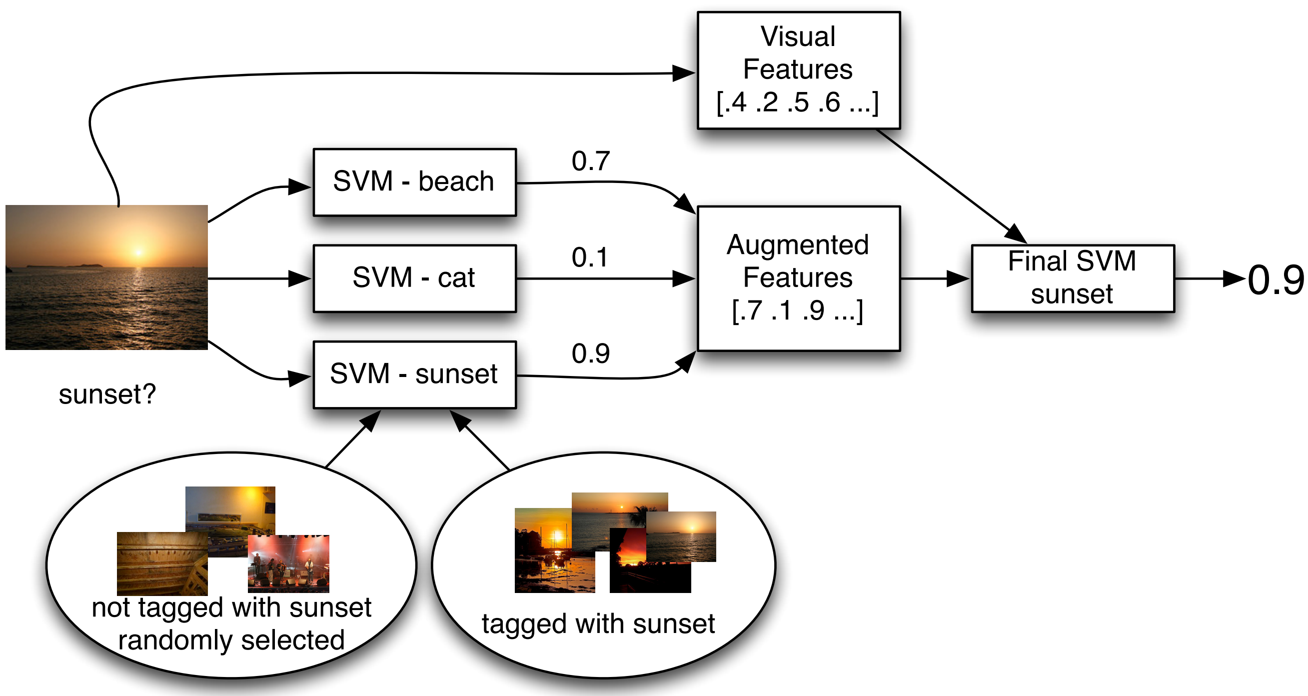
- Since rare tags are only associated with a limited number of positive training images, they may degrade SVMs performance

L. Chen, D. Xu, I. Tsang, J. Luo. Tag-Based Image Retrieval Improved by Augmented Features and Group-Based Refinement. IEEE TMM 2012

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'))

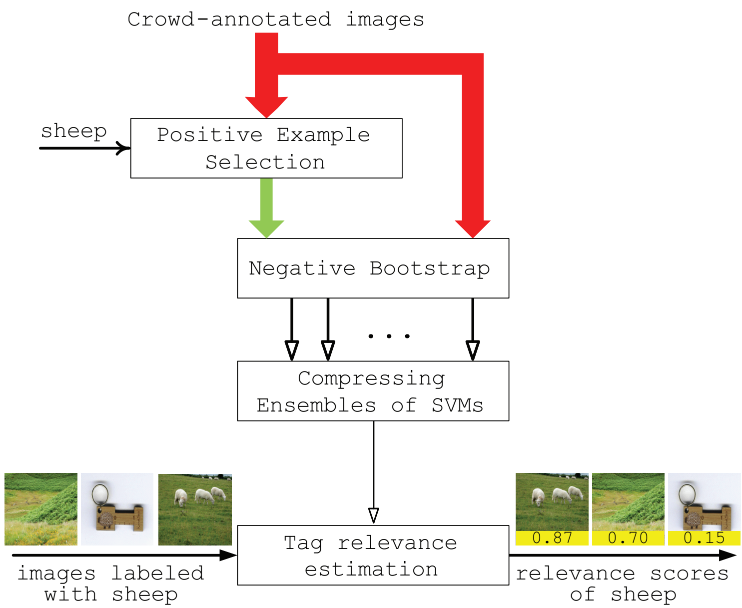
KEY METHODS

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

- Negative examples which are visually similar to positive can be misclassified
- RelExample exploits positive and negative training examples which are deemed to be more relevant with respect to the test tag t



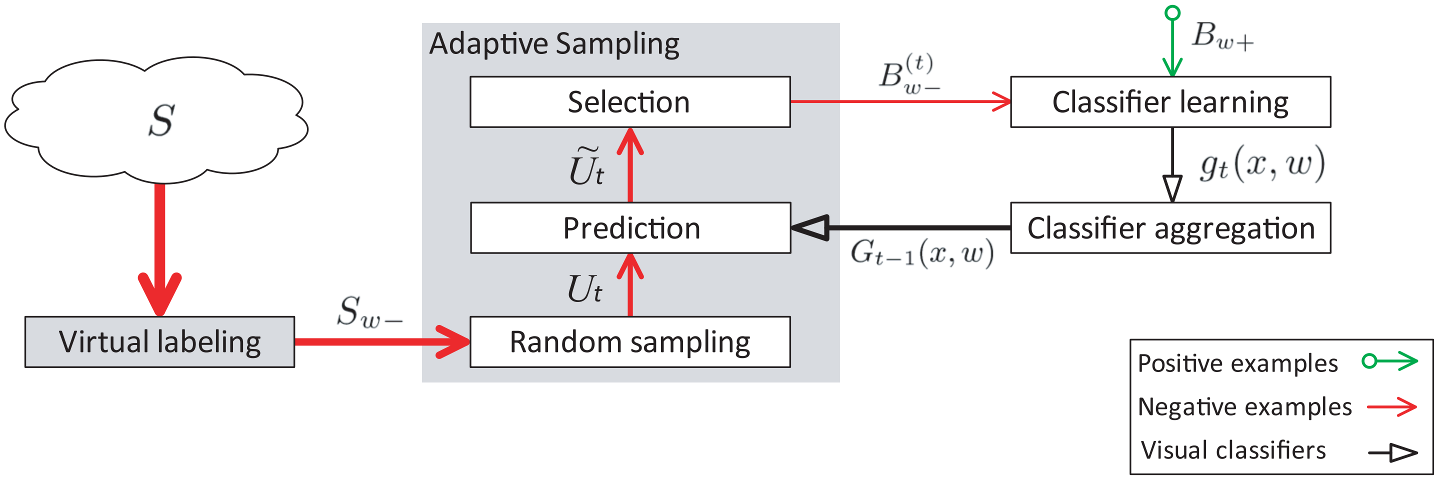
- Positive examples are selected by taking the top-ranked images by TagVote and SemanticField
- Negative examples are selected by Negative Bootstrap [Li et al. 2013]

X. Li, C. Snoek. Classifying tag relevance with relevant positive and negative examples. ACM MM 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

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

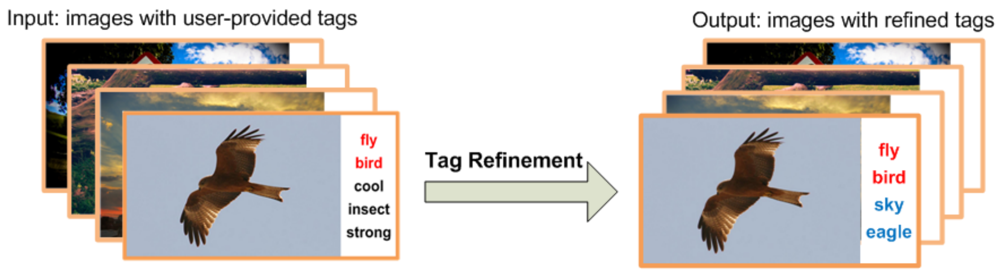
- T iterations for a corresponding number of trained classifiers
- User tags on test image are not used. Not applicable to Tag Refinement.
- Complexity $O(Tdp^2)$: training T SVM classifiers
- Memory $O(dp + dq)$: d visual features, p pos and q neg examples

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

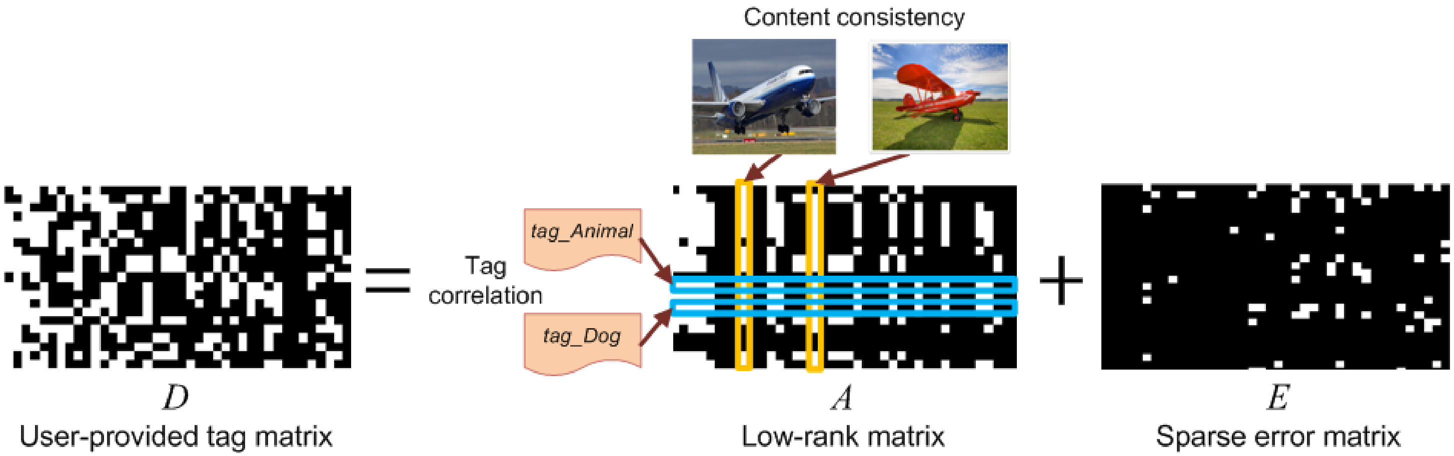


- Based on a few assumptions on tag characteristics:
  - *low-rank property*: the semantic space spanned by tags can be approximated by a smaller subset of salient words derived from the original space
  - *tag correlation*: semantic tags are correlated
  - *visual consistency*: visually similar images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

Zhu et al. Image Tag Refinement Towards Low-Rank, Content-Tag Prior Error Sparsity. ACM MM 2010

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

$$\begin{array}{ll} \min\limits_{A,E} & ||A||_* + \lambda_1 ||E||_1 + \lambda_2 [T_c(A) + T_t(A)] \\ \text{subject to} & D = A + E \end{array}$$

- 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.

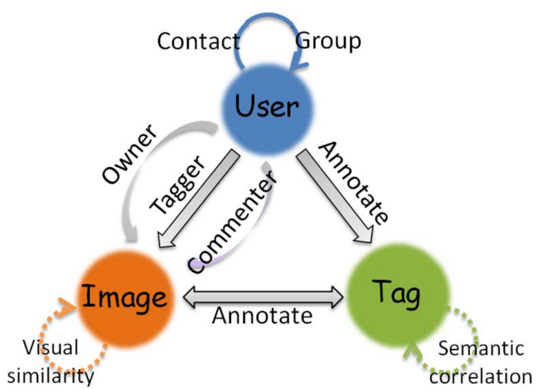
KEY METHODS

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]

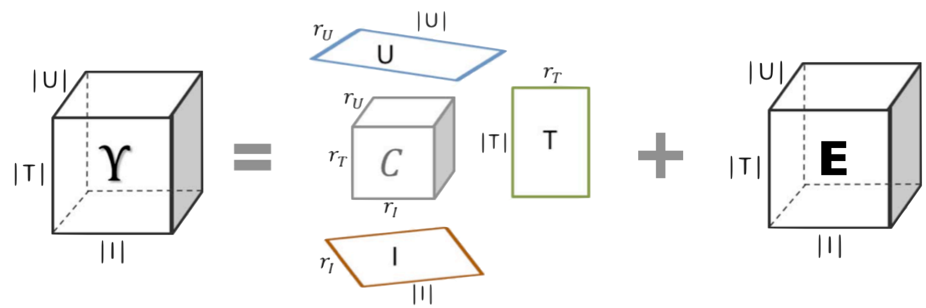
TensorAnalysis

[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



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



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

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

Sang et al. User-Aware Image Tag Refinement via Ternary Semantic Analysis. IEEE TMM 2012

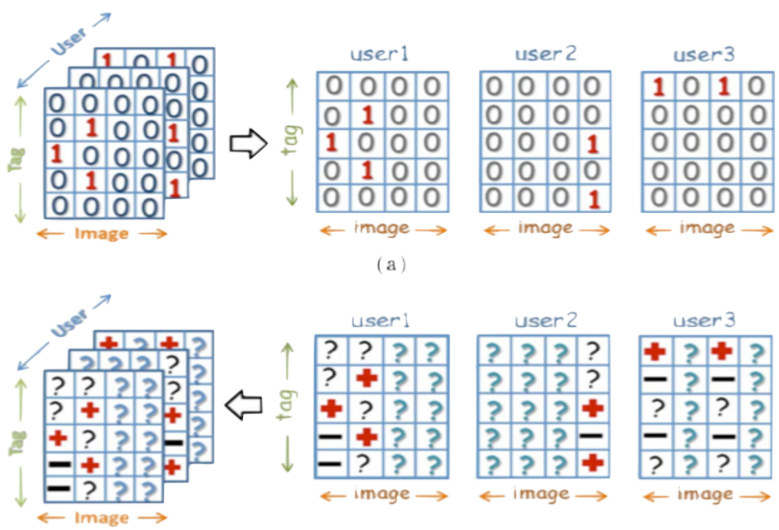
TENSORANALYSIS

[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.

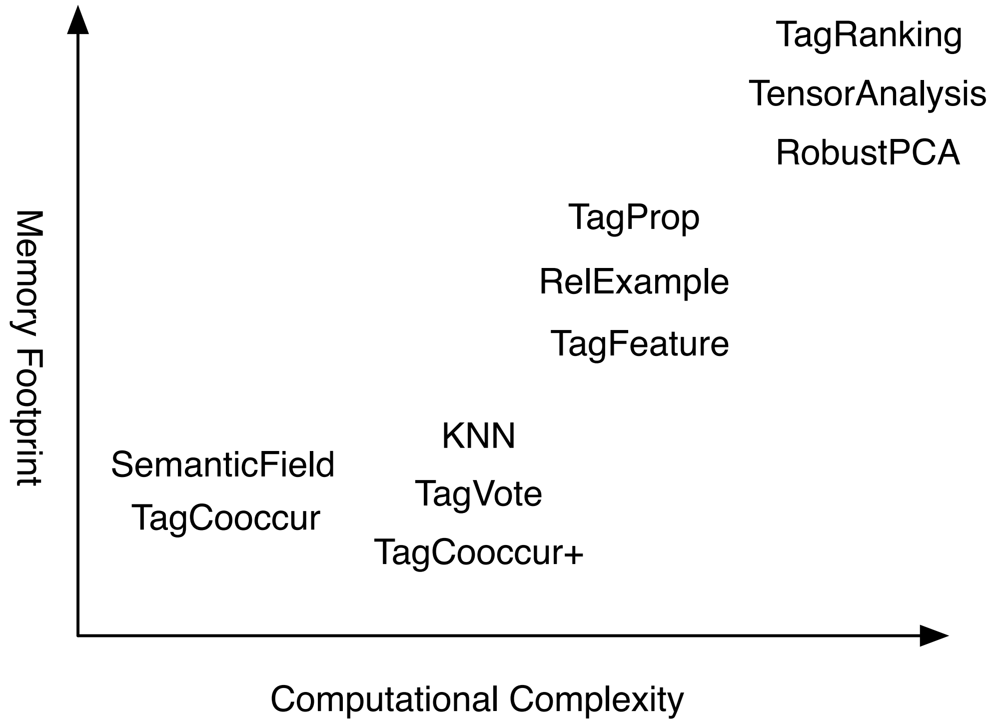
- Complexity $O(|P_1| \cdot (r_T \cdot m^2 + r_U \cdot r_I \cdot r_T))$ – P_1 is the ones in D, $r_{\{U,I,T\}}$ are latent matrices dimensionalities.
- Memory $O(n^2 + m^2 + u^2)$ – the three regularizers matrices.

EVALUATION: EXPERIMENTAL RESULTS

- Q: We evaluate the eleven methods for different tasks and scenarios. What are their performances?

- Q: What is the computational cost of each of them?

ANALYSIS OF COMPLEXITY



- SemanticField and TagCooccur have the best scalability
- 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 report a thorough evaluation of the methods on the proposed testbed

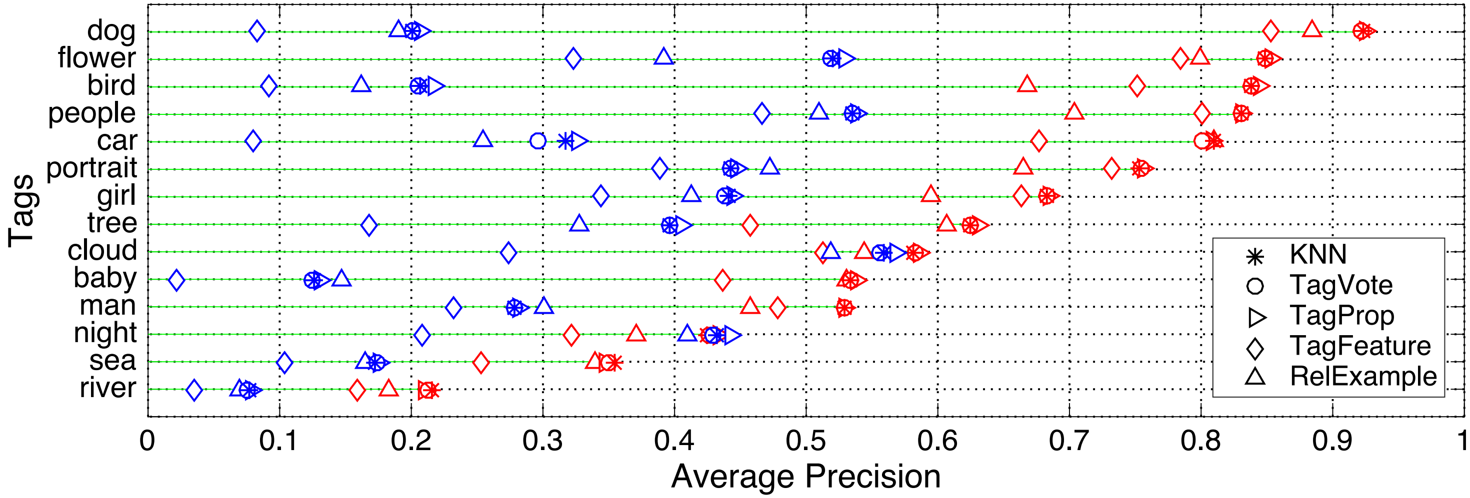
	Assignment	Refinement	Retrieval
KNN	X		X
TagVote	X		X
TagProp	X		X
TagFeature	X		X
RelExample	X		X
TagCooccur		X	X
TagCooccur+		X	X
RobustPCA		X	X
TensorAnalysis		X	X
SemanticField			X
TagFeature			X

- Here we discuss only few main results. Please refer to our survey paper for the full picture.

Tag Assignment

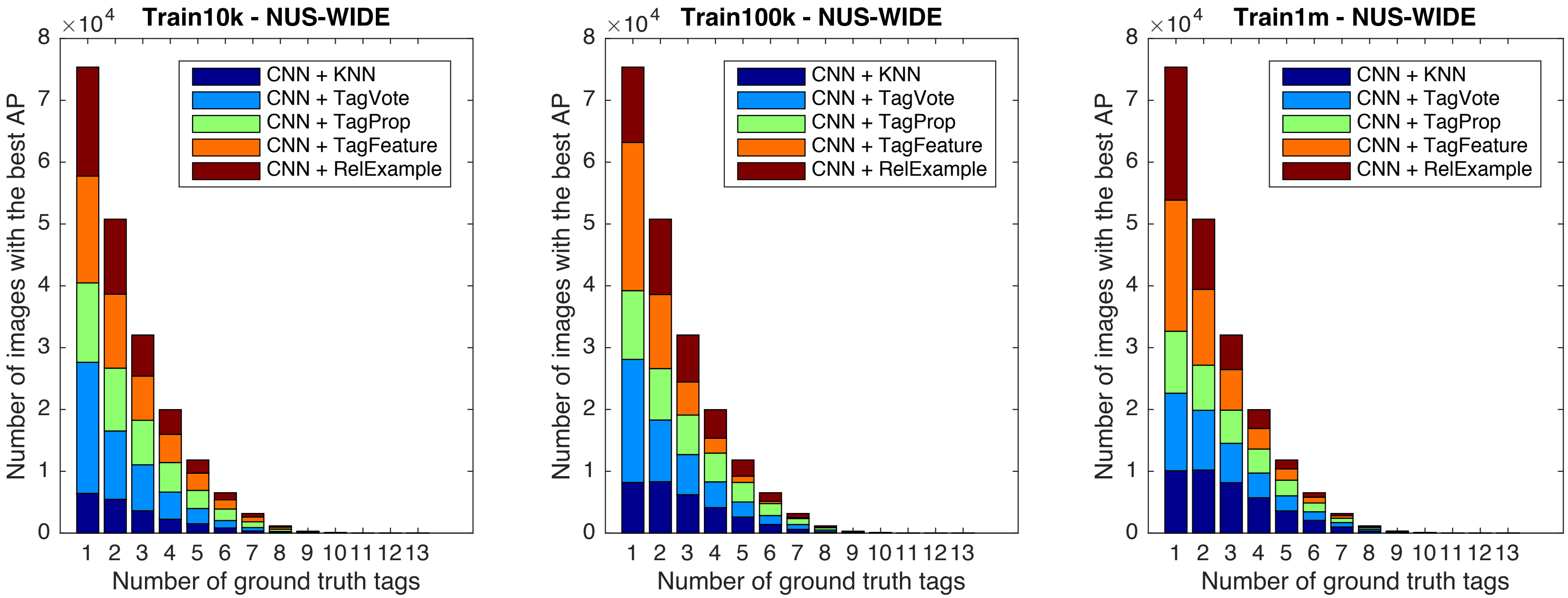
MIRFlickr test set, trained on Train1m.

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

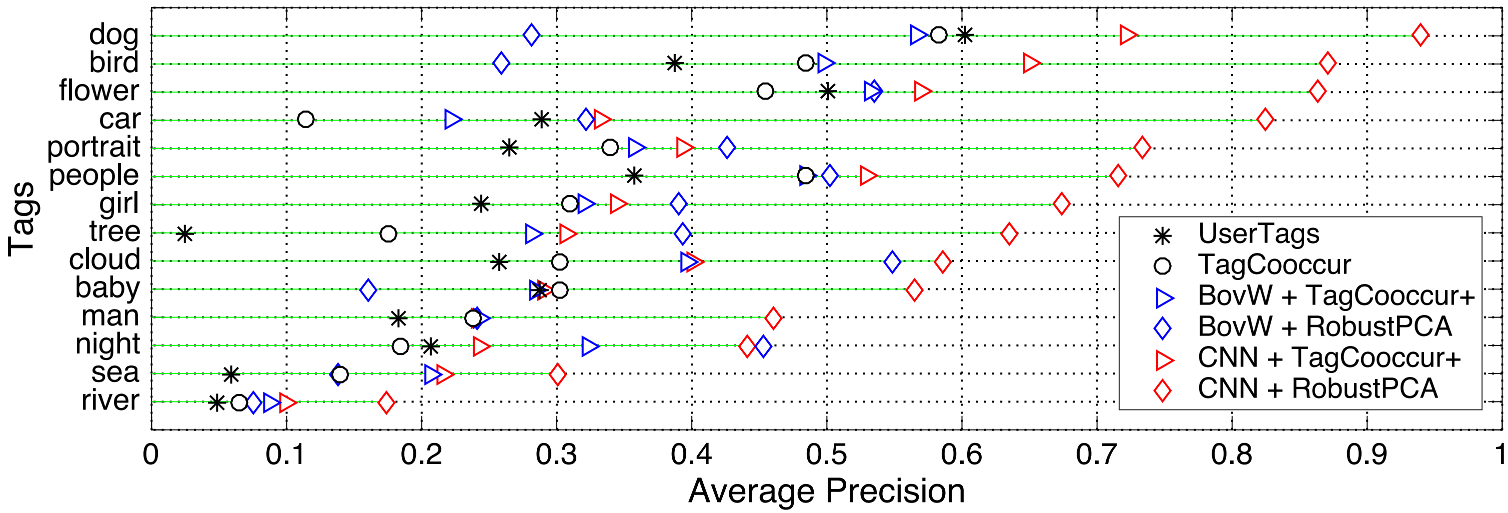
TAG ASSIGNMENT



- 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



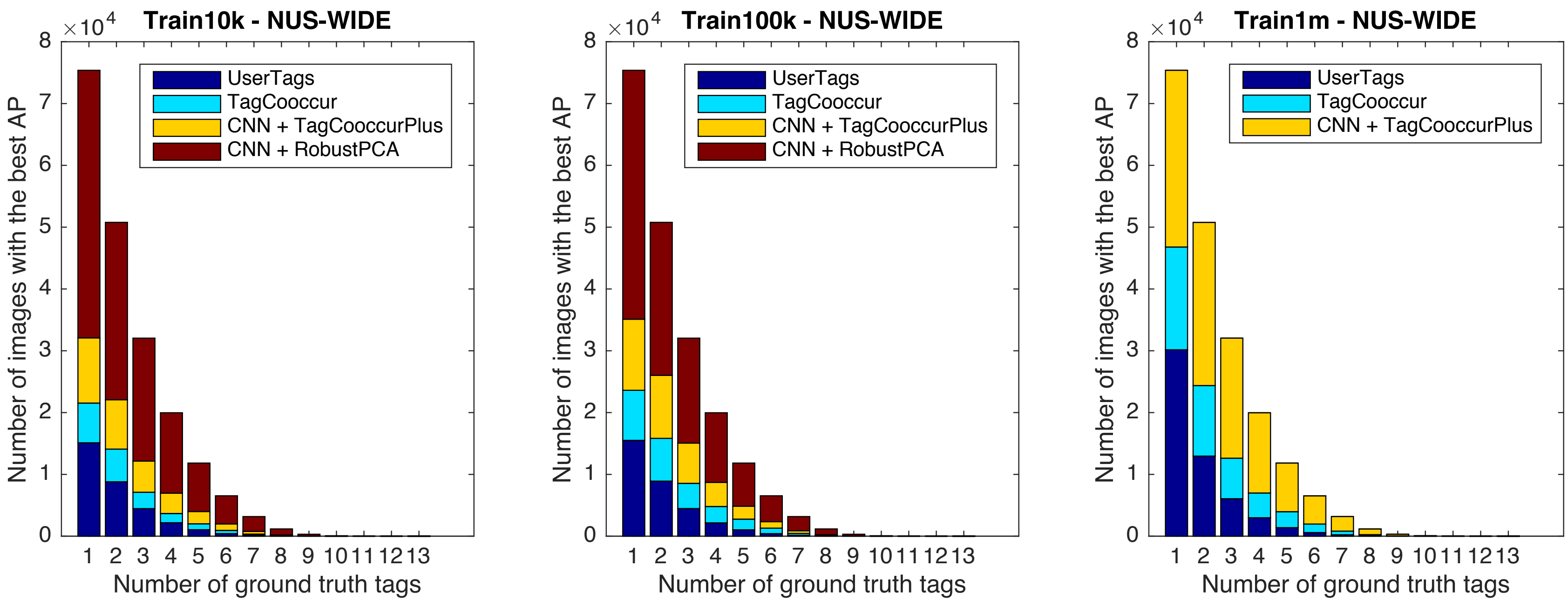
MIRFlickr test set, trained on Train100k.

CNN Features

BovW Features

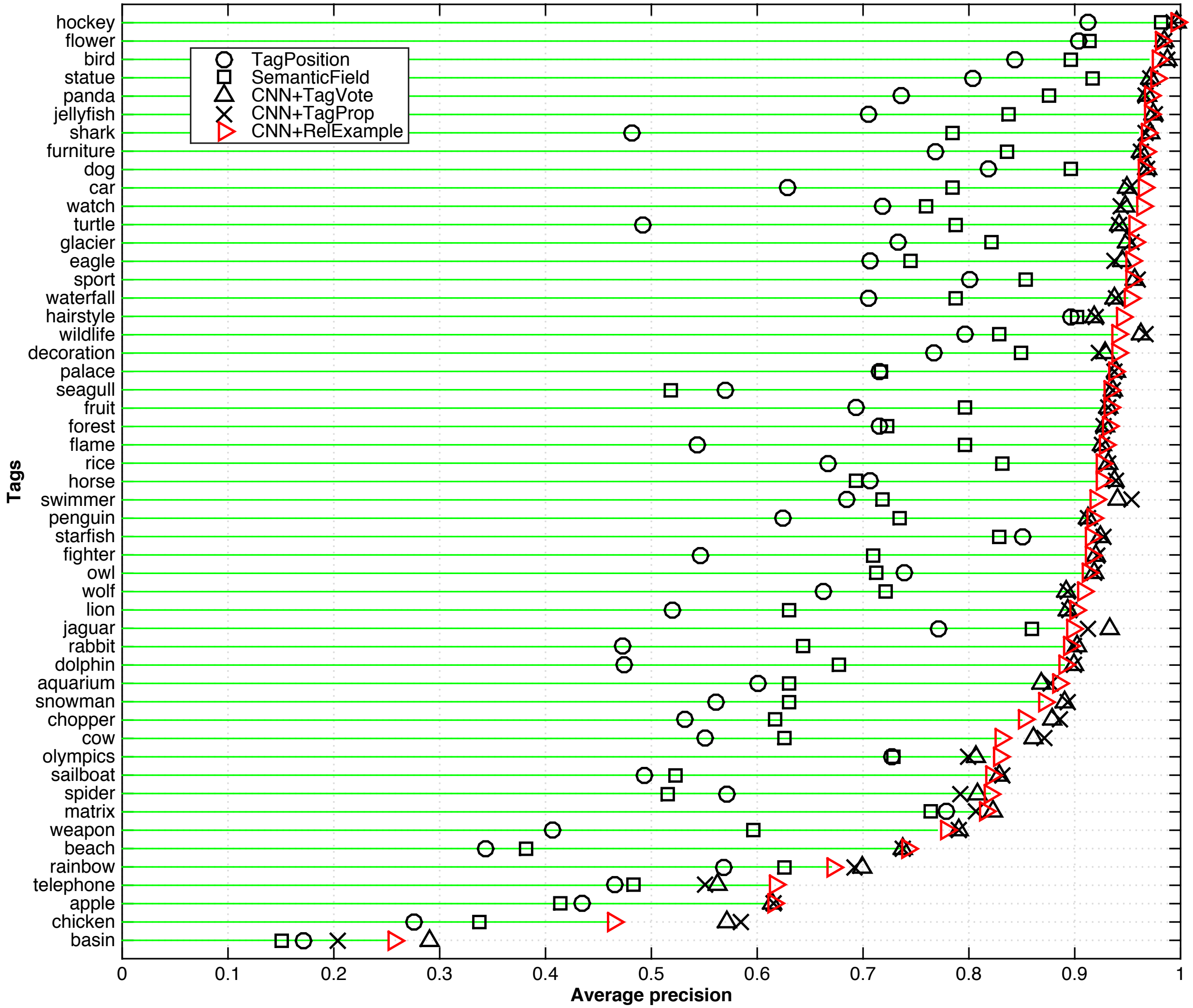
- All methods have performance superior to user tagging
- The tag + image based methods outperform the tag based TagCooccur
- RobustPCA provides the best performance

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

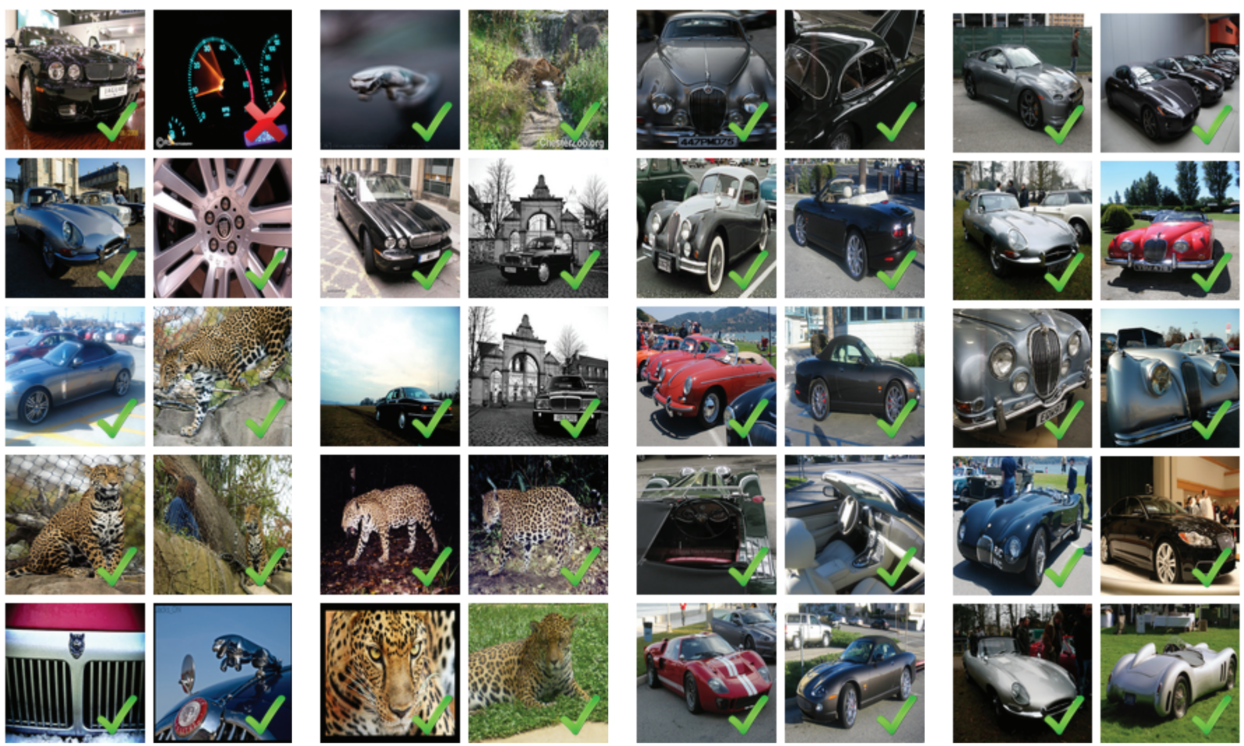


- As for Tag Assignment, TagVote and TagProp provide the best performance
- For 33 out of 51 test tags, RelExample gives average precision higher than 0.9

TAG RETRIEVAL

The top 10 ranked images for ‘jaguar’

TagPosition | *SemanticField* | *BovW + RelExample* | *CNN + RelExample*



Lower diversity

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

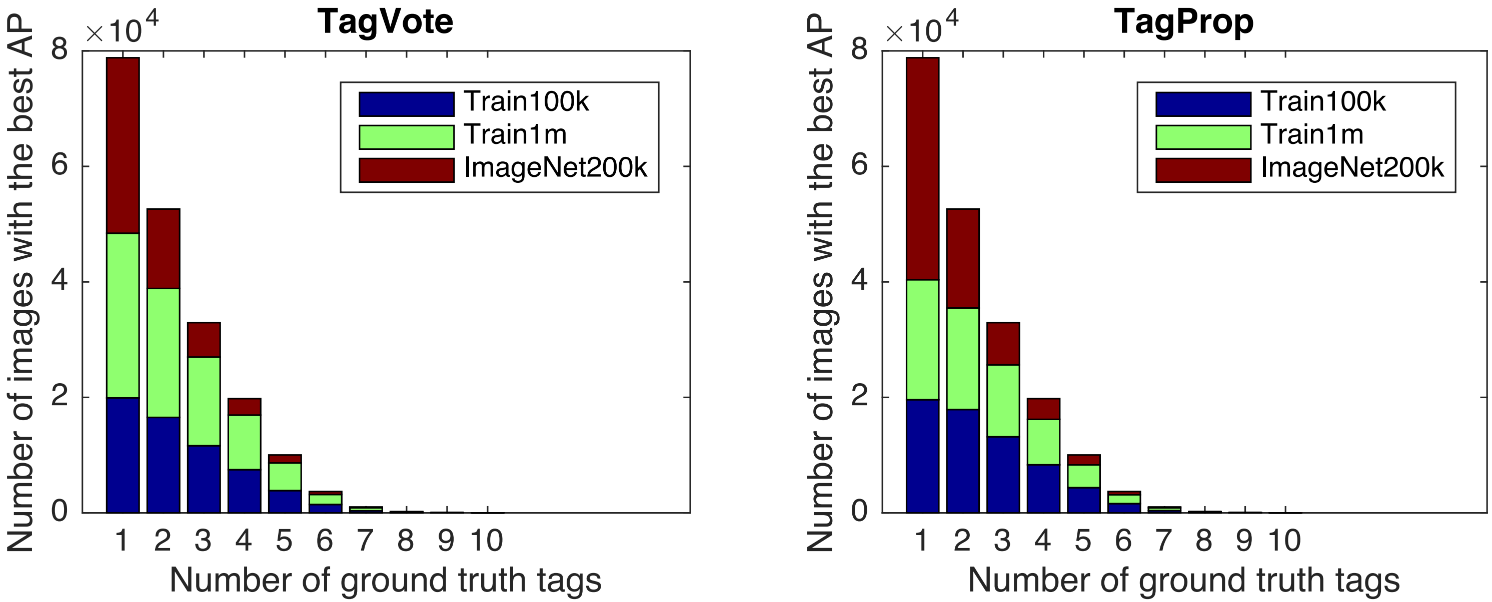
IMAGENET RESULTS

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.

SOFTWARE

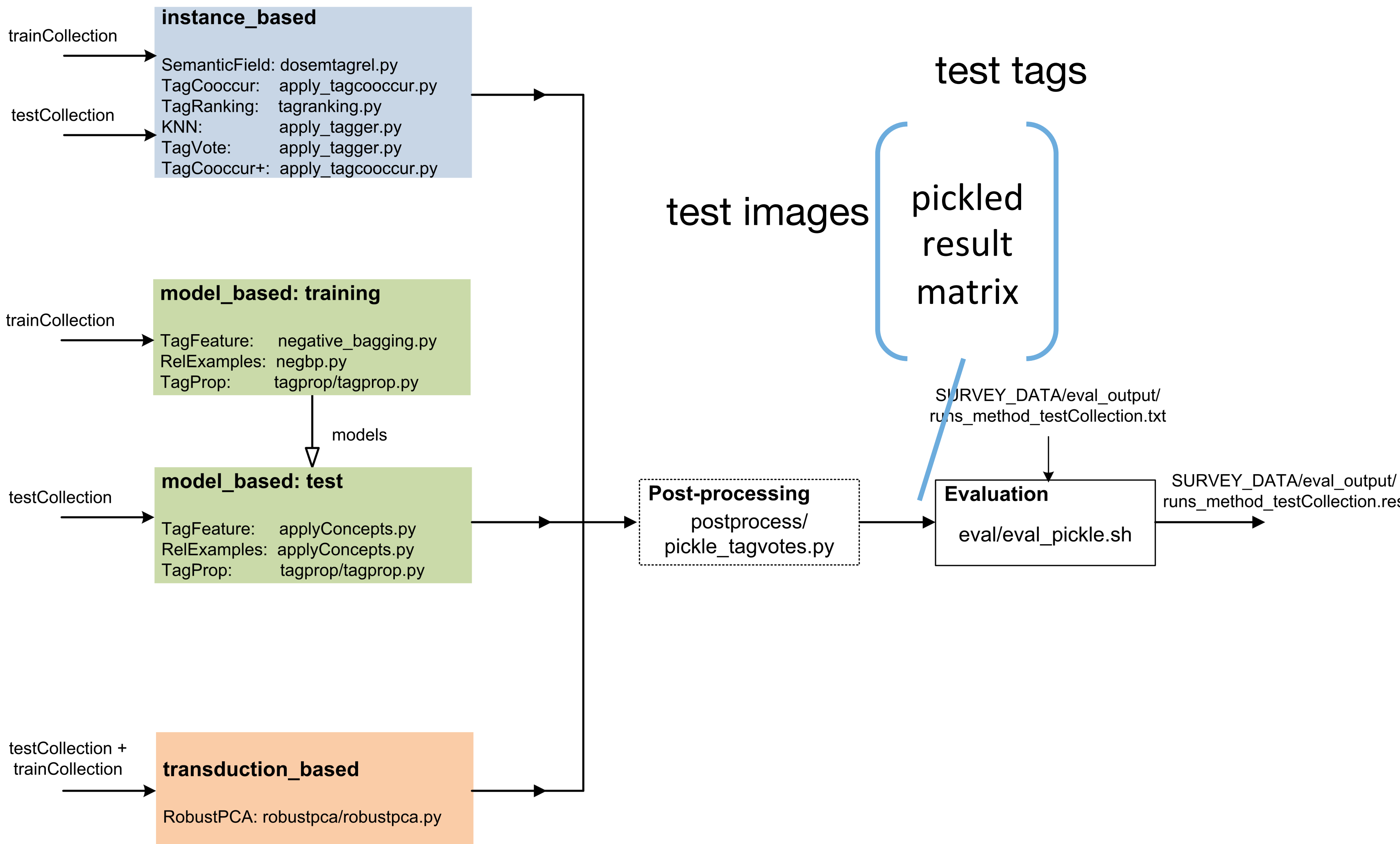
- **Jingwei**, a framework for evaluating image tag assignment, tag refinement and tag-based image retrieval:
  - https://github.com/li-xirong/jingwei

- 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



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.

ORGANIZATION OF THE TUTORIAL

8:30 – 9:30	Part 1: Introduction
	Part 2: Taxonomy
9:30 – 10:00	Part 3: Experimental protocol
	Part 4: Evaluation
10:00 – 10:45	*Coffee break*
10:45 – 12:00	Part 4: Evaluation cont'd
12:00 – 12:30	Part 5: Conclusion and future directions

PART 5
CONCLUSION AND FUTURE DIRECTIONS

- Summary
- Future directions

READING MATERIAL

Socializing the Semantic Gap: A Comparative Survey on Image Tag Assignment, Refinement and Retrieval,
ACM Computing Surveys, 49(1):14, June 2016.

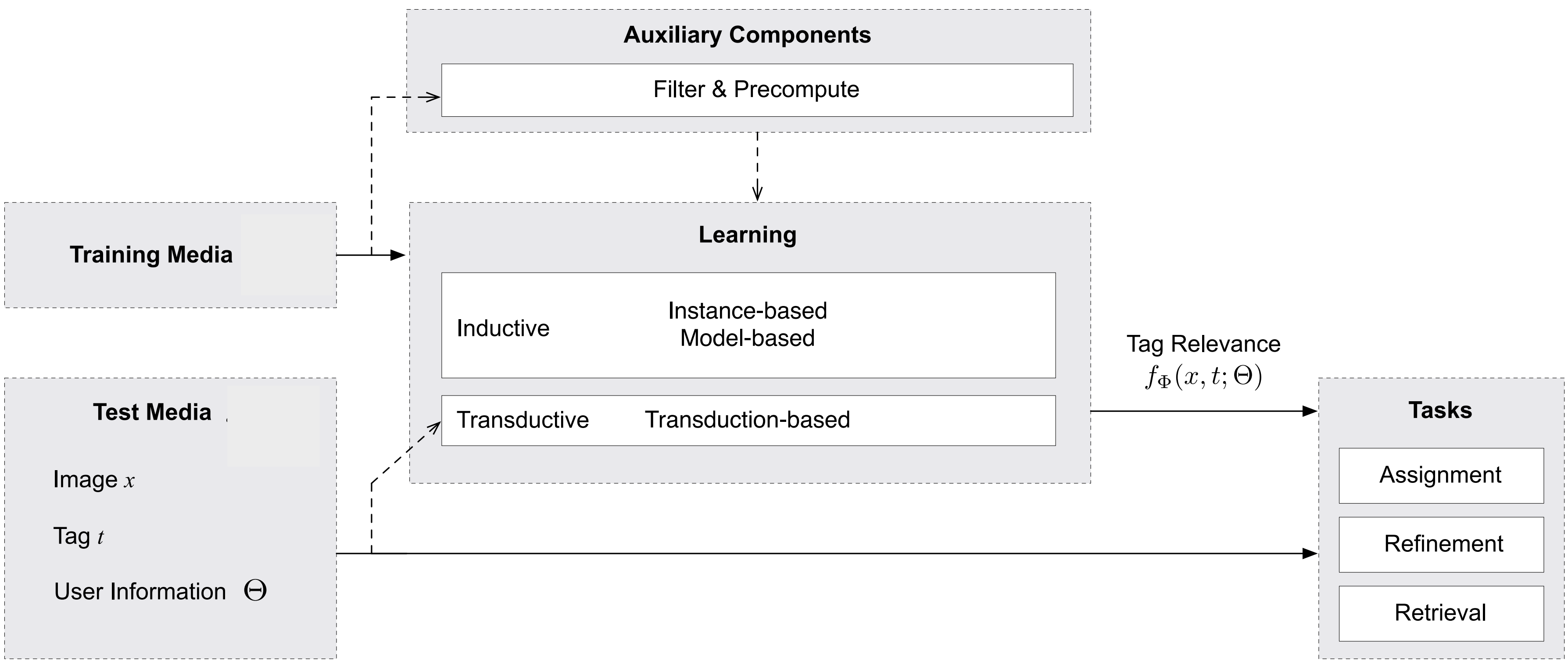
Socializing the Semantic Gap: A Comparative Survey on Image Tag Assignment, Refinement, and Retrieval

XIRONG LI, Renmin University of China
TIBERIO URICCHIO, University of Florence
LAMBERTO BALLAN, University of Florence, Stanford University
MARCO BERTINI, University of Florence
CEES G. M. SNOEK, University of Amsterdam, Qualcomm Research Netherlands
ALBERTO DEL BIMBO, University of Florence

Where previous reviews on content-based image retrieval emphasize what can be seen in an image to bridge the semantic gap, this survey considers what people tag about an image. A comprehensive treatise of three closely linked problems (i.e., image tag assignment, refinement, and tag-based image retrieval) is presented. While existing works vary in terms of their targeted tasks and methodology, they rely on the key functionality of tag relevance, that is, estimating the relevance of a specific tag with respect to the visual content of a given image and its social context. By analyzing what information a specific method exploits to construct its tag relevance function and how such information is exploited, this article introduces a two-dimensional taxonomy to structure the growing literature, understand the ingredients of the main works, clarify their connections and difference, and recognize their merits and limitations. For a head-to-head comparison with the state of the art, a new experimental protocol is presented, with training sets containing 10,000, 100,000

14

SUMMARY: **UNIFIED FRAMEWORK**



SUMMARY: **TAXONOMY**

Media	Learning		
	Instance	Model	Transductive
Tag	2	1	-
Tag + Image	13	15	12
Tag + Image + User	5	7	3

Taxonomy structures 60 papers along **Media** and **Learning** dimensions

Summary: 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]

Summary: Open-source testbed

Media	Media characteristics				Tasks		
	# 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

Jingwei, a framework for evaluating image tag assignment, tag refinement and tag-based image retrieval:

[3] https://github.com/li-xirong/jingwei

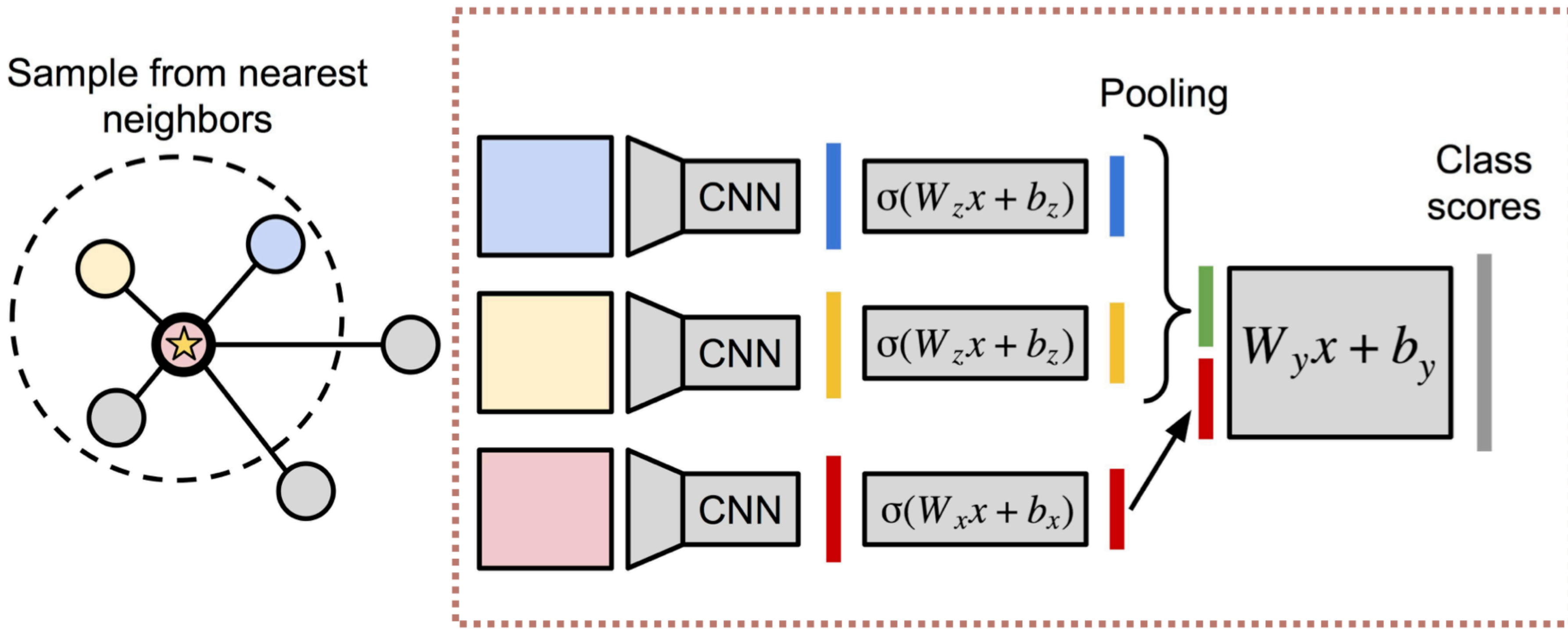
SUMMARY: **TAKE HOME MESSAGES**

- 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.

FUTURE: MUCH REMAINS TO BE DONE

- Novel deep-learning features likely to boost the performance of the tag + image methods further
- Learning strategy capable of jointly exploiting tag, image, and user information in a much more scalable manner than currently feasible.
- The importance of the filter component, which refines socially tagged training examples in advance to learning, is underestimated.
- Image retrieval by multi-tag query is another important yet largely unexplored problem.

CNN THAT BLENDS VISUAL INFORMAITON FROM THE IMAGE AND ITS NEIGHBORS



[J.Johnson*, **L.Ballan***, L.Fei-Fei - ICCV 2015]

QUALITATIVE RESULTS

V-only
animal
water
flowers

Ours
water
swimmers
person

V-only
sky
clouds
person

Ours
police
person
military

Neighborhood

Cappallo et al. ICMR 2015

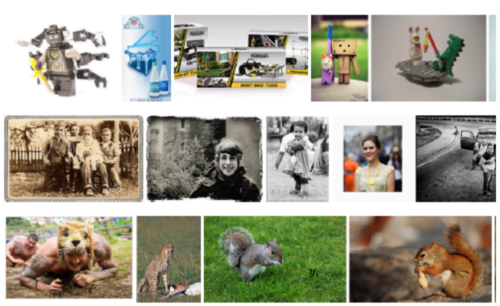
POPULAR *AND* UNPOPULAR LATENT SENSES

- Introduce **latent senses** to capture nuances in popularity
- What makes an image **un**popular is also informative

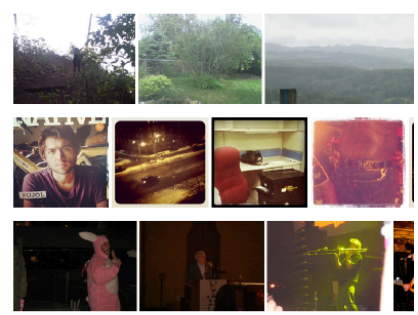
$$L_{p\&n} = \sum_i \sum_j \left[|\Delta(y_i, y_j) - f_{s_+}(x_i) + f_{s_+}(x_j)|_+ + |\Delta(y_i, y_j) - f_{s_-}(x_j) + f_{s_-}(x_i)|_+ \right]$$

(popular senses) (unpopular senses)

- Popularity and unpopularity learned independently at train time
- Single popularity score calculated at test time



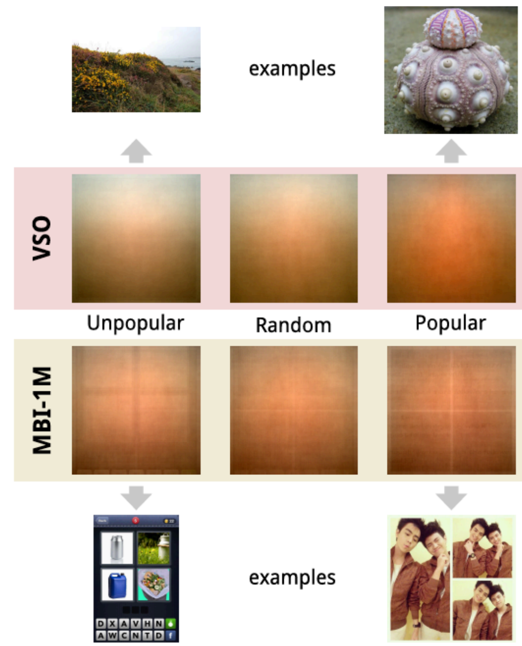
Popular latent senses



Unpopular latent senses

1M Micro-Blog Images

- New, challenging dataset of 1 million images from social media
- Twitter posts containing images from TREC 2013 Microblog track
- Retweet and Favorite counts for popularity prediction research
- Many graphical, non-photographic images



http://staff.fnwi.uva.nl/s.h.cappallo/data.html

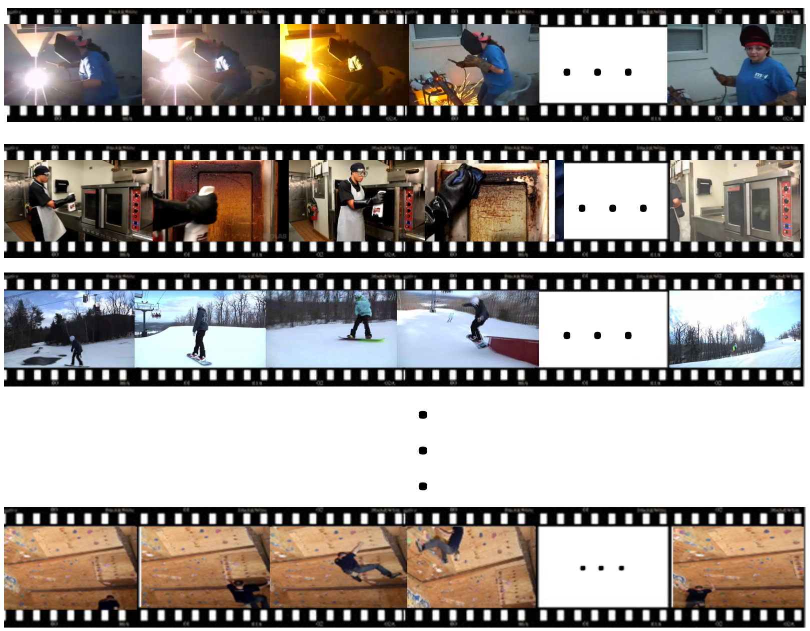
Mazloom et al. TMM 2016

PROBLEM: EVENT DETECTION IN VIDEO

TagBook: derived from social-tagged video

Source set: Social-tagged web videos

Video data



Tags

woman, outdoor, metal-crafts-project, welding machine

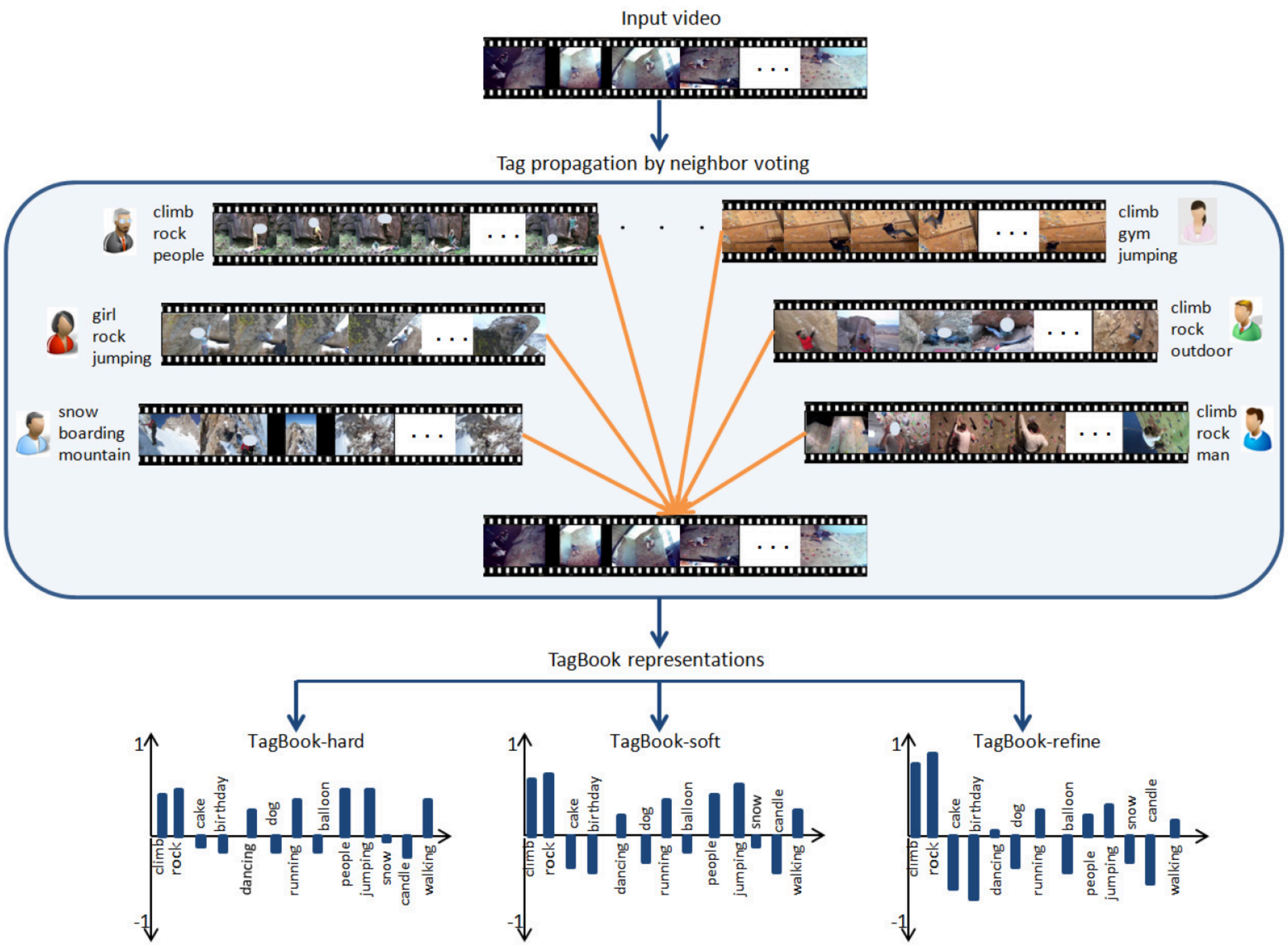
man, kitchen, metallic, cleaning, oven, spray, glasses,

man, snowboard, snow, board-trick,

man, climb-on, wall, gym, rock-climbing

TagBook = {woman, outdoor, metal-crafts-project, welding machine, man, kitchen,…, wall, gym, rock-climbing}

TagBook: New Video Representation



Cappallo et al. MM 2015

BEYOND TAGS: EMOJI

- Visual grammar of interaction
- Language independent
- Age accessible
- Widely supported
- Semantically diverse
- Easy form factor for smart phones and watches

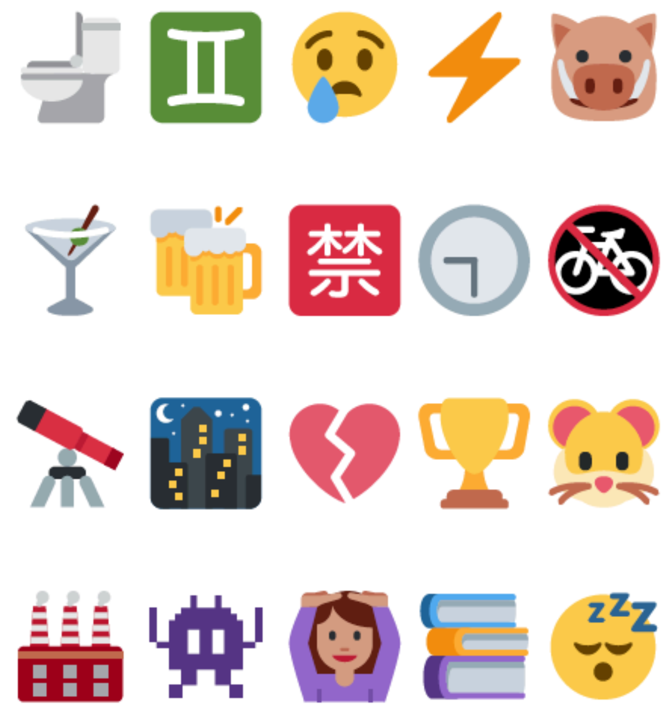
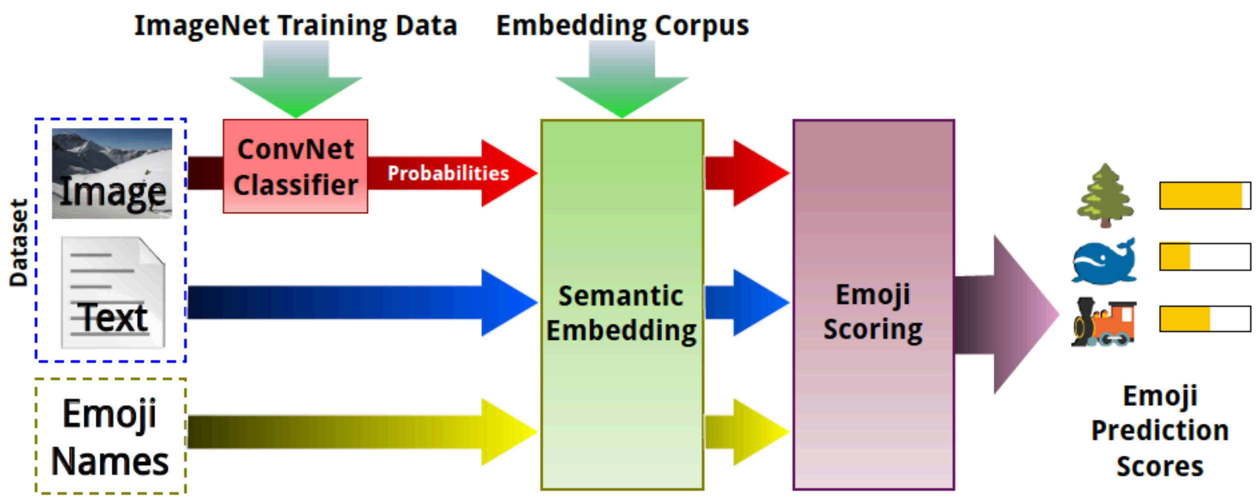


IMAGE2EMOJI



THIS CVPR: Fast Zero-Shot Image Tagging

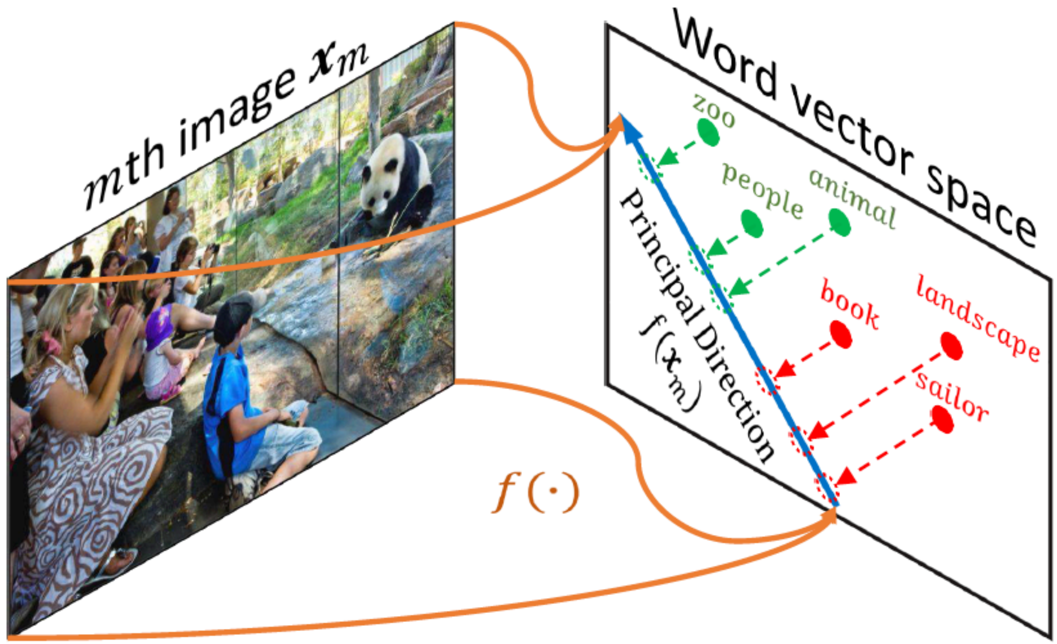
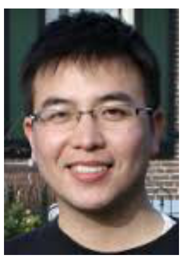


Figure 1: Given an image, its relevant tags' word vectors rank ahead of the irrelevant tags' along some direction in the word vector space. We call that direction the **principal direction** for the image. To solve the problem of image

Wishing you a great conference

CVPR 2016 Tutorial

June 26, 2016



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