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
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 differences, 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.
Part I
Introduction

• Problem statement
• Course organization
PROGRESS IN IMAGE RETRIEVAL

- Query-by-Image content
PROGRESS IN IMAGE RETRIEVAL

- Query-by-sketch
PROGRESS IN IMAGE RETRIEVAL

- By 2000 problem well understood

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

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

% of Users on Each Platform Who Utilize to Find / Shop for Products, USA, 4/16

- Pinterest: 55%
- Facebook: 12%
- Instagram: 12%
- Twitter: 9%
- LinkedIn: 5%
- Snapchat: 3%

‘What Do You Use Pinterest For?’ (% of Respondents), USA, 4/16

- Viewing photos: 60%
- Finding / shopping for products: 55%
- Sharing photos / videos / personal messages: 24%
- Watching videos: 15%
- News: 10%
- Networking / promotion: 10%
AVERAGE DAILY TIME SPENT PER USA USER

<table>
<thead>
<tr>
<th>Platform</th>
<th>11/14</th>
<th>6/15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>42</td>
<td>41</td>
</tr>
<tr>
<td>Instagram</td>
<td>21</td>
<td>25</td>
</tr>
<tr>
<td>OfferUp</td>
<td>13</td>
<td>25</td>
</tr>
<tr>
<td>Snapchat</td>
<td>17</td>
<td>25</td>
</tr>
<tr>
<td>Pinterest</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>Twitter</td>
<td>17</td>
<td>21</td>
</tr>
</tbody>
</table>
EXAMPLES
EXAMPLES
EXAMPLES
PROBLEMS OF TAGS: IRRELEVANCE

- Tags are few, imprecise, ambiguous, and overly personalized
PROBLEMS OF TAGS: DYNAMICS

- In a social network, users continuously add images and create new terms given the freedom of tagging.
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.
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
- A few photos are exceptionally well tagged
- 64% of photos have 1, 2 or 3 tags only.
- 48% of 3.7M tags could not be matched.
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/](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).
**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.
**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**

![Image of a bride and groom](image1.jpg)

**Tags:**
- stealing
- sonnet
- photoshooting
- pentaxk10d
- 31mm
- bride
- Chinese

---

![Image of a bride and groom](image2.jpg)

**Tags:**
- wedding
- father of the bride
- bride
- puglia
- italianwedding
- romance
- romantic
- bridegroom

Photo courtesy of Nicola Bertini (Flickr member: nK10d).
• Foundations
  • tag relevance

• A two-dimensional taxonomy
  • Media for tag relevance
  • Learning for tag relevance
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 $X$. 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 $\Theta$ (e.g. user contacts, geo-localization, etc.)
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 $x$, tag $t$ and user information $\Theta$:

$$f_\phi(x, t; \Theta)$$
EXAMPLE FOR TAG REFINEMENT

Li et al. TMM 2009
UNIFIED FRAMEWORK

Test Media

Tasks
- Assignment
- Refinement
- Retrieval

Filtered media \cdot Prior Learning

\tilde{S} = \text{Transduction-based Training Media}

Image

Auxiliary Components

Instance-based

Model-based

Inductive

Transductive
UNIFIED FRAMEWORK

Test Media → Learning → Tasks

Tag Relevance $f_{\Phi}(x, t; \Theta)$

Tasks
- Assignment
- Refinement
- Retrieval

Filtered media, prior learning

Assignment

Auxiliary Components
- Instance-based
- Model-based
- Inductive
- Transductive

Learning

Assignment

Refinement

Retrieval

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
A unique user constraint prevents ‘spam’ from batch tagging

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**

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

\[
\text{NGD}(x, y) = \frac{\max\{\log h(x), \log h(y)\} - \log h(x, y)}{\log N - \min\{\log h(x), \log h(y)\}},
\]

\[
\text{FCS}(x, y) = e^{-\text{NGD}(x,y)/\sigma}
\]

*[Jiang et al. 2009]*
UNIFIED FRAMEWORK

Auxiliary Components

Filter & Precompute

Tag Relevance

Training Media

Test Media

Learning

Tag Relevance

\( f_\Phi(x, t; \Theta) \)

Tasks

Assignment
Refinement
Retrieval
## Taxonomy

<table>
<thead>
<tr>
<th>Media</th>
<th>Learning</th>
<th>Instance</th>
<th>Model</th>
<th>Transductive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag</td>
<td></td>
<td>2</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Tag + Image</td>
<td></td>
<td>13</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>Tag + Image + User</td>
<td></td>
<td>5</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Media</th>
<th>Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Instance</td>
</tr>
<tr>
<td>Tag</td>
<td>2</td>
</tr>
<tr>
<td>Tag + Image</td>
<td>13</td>
</tr>
<tr>
<td>Tag + Image + User</td>
<td>5</td>
</tr>
</tbody>
</table>
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
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, McParlance et al. 2013a]
- Geo-localization [McParlance et al. 2013b]
- Image group memberships [Johnson et al. 2015]
## Taxonomy

<table>
<thead>
<tr>
<th>Media</th>
<th>Learning</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Instance</td>
<td>Model</td>
<td>Transductive</td>
</tr>
<tr>
<td>Tag</td>
<td>2</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Tag + Image</td>
<td>13</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>Tag + Image + User</td>
<td>5</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>

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.
I NSTANCE 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 without 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**

**Auxiliary Components**
- Filter & Precompute

**Learning**
- Inductive
- Instance-based
- Model-based
- Transductive
- Transduction-based

**Tasks**
- Assignment
- Refinement
- Retrieval

**Training Media**
- Image $x$
- Tag $t$
- User Information $\Theta$

**Test Media**
- Image $x$
- Tag $t$
- User Information $\Theta$

**Tag Relevance**
$$f_\Phi(x, t; \Theta)$$
### Taxonomy

<table>
<thead>
<tr>
<th>Media</th>
<th>Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Instance</td>
</tr>
<tr>
<td>Tag</td>
<td>2</td>
</tr>
<tr>
<td>Tag + Image</td>
<td>13</td>
</tr>
<tr>
<td>Tag + Image + User</td>
<td>5</td>
</tr>
</tbody>
</table>

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

## Learning

<table>
<thead>
<tr>
<th>Media</th>
<th>Instance-based</th>
<th>Model-based</th>
<th>Transduction-based</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>tag</strong></td>
<td></td>
<td><strong>TagCooccur</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Sigurbjörnsson and van Zwol 2008]</td>
<td>[Xu et al. 2009]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Sun et al. 2011]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Zhu et al. 2012]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>TagRanking</strong></td>
<td><strong>SemanticField</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Wu et al. 2009]</td>
<td>[Guillaumin et al. 2009]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Liu et al. 2010]</td>
<td>[Verbeek et al. 2010]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Tang et al. 2011]</td>
<td>[Liu et al. 2010]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Wu et al. 2011]</td>
<td>[Ma et al. 2010]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Yang et al. 2011]</td>
<td>[Liu et al. 2011b]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Truong et al. 2012]</td>
<td>[Duan et al. 2011]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Qin et al. 2012]</td>
<td>[Feng et al. 2012]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Lin et al. 2013]</td>
<td>[Srivastava and Salakhutdinov 2012]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Lee et al. 2013]</td>
<td>[Chen et al. 2012]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Uricchio et al. 2013]</td>
<td>[Lan and Mori 2013]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Zhu et al. 2014]</td>
<td>[Li and Snoek 2013]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Ballan et al. 2014]</td>
<td>[Li et al. 2013]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Pereira et al. 2014]</td>
<td>[Wang et al. 2014]</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>TagProp</strong></td>
<td><strong>TagFeature</strong></td>
<td><strong>RelExample</strong></td>
</tr>
<tr>
<td></td>
<td>[Zhu et al. 2010]</td>
<td>[Wang et al. 2010]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Li et al. 2010]</td>
<td>[Li et al. 2010]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Zhuang and Hoi 2011]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Richter et al. 2012]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Kuo et al. 2012]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Liu et al. 2013]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Gao et al. 2013]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Wu et al. 2013]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Zhuang et al. 2014]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Xu et al. 2014]</td>
<td></td>
</tr>
<tr>
<td><strong>tag + image</strong></td>
<td></td>
<td><strong>TagVote</strong></td>
<td><strong>TensorAnalysis</strong></td>
</tr>
<tr>
<td></td>
<td>[Li et al. 2009b]</td>
<td>[Sawant et al. 2010]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Kennedy et al. 2010]</td>
<td>[Li et al. 2011b]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Li et al. 2010]</td>
<td>[McAuley and Leskovec 2012]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Znaidia et al. 2013]</td>
<td>[Kim and Xing 2013]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Liu et al. 2014]</td>
<td>[McParlane et al. 2013b]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Ginsca et al. 2014]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Johnson et al. 2015]</td>
<td></td>
</tr>
<tr>
<td><strong>tag + image + user</strong></td>
<td></td>
<td><strong>TagCooccur</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Sang et al. 2012a]</td>
<td>[Sang et al. 2012b]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Qian et al. 2015]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
• 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
SOCIALY-TAGGED TRAINING DATA

- Data gathering procedure [Li et al. 2012]
  - using WordNet nouns as query 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 IMAGE NET 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

Credit: figure from [Vreeswijk et al. 2012]
TEST DATA

- Three test datasets
  - contributed by distinct research groups

<table>
<thead>
<tr>
<th>Test dataset</th>
<th>Contributors</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIRFlickr [Huiskes 2010]</td>
<td>LIACS Medialab, Leiden University</td>
</tr>
<tr>
<td>NUS-WIDE [Chua 2009]</td>
<td>LMS, National University of Singapore</td>
</tr>
<tr>
<td>Flickr51 [Wang 2010]</td>
<td>Microsoft Research Asia</td>
</tr>
</tbody>
</table>
MIRFLICKR

- Image collection
  - 25,000 high-quality photographic images from Flickr

- Labeling criteria
  - Potential labels: visibly 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

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

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

**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]
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
    \[ NDCG_h(t) := \frac{DCG_h(t)}{IDCG_h(t)}, \quad DCG_h(t) = \sum_{i=1}^{h} \frac{2^{rel_i} - 1}{\log_2(i+1)} \]
SUMMARY

<table>
<thead>
<tr>
<th>Media</th>
<th># images</th>
<th># tags</th>
<th># users</th>
<th># test tags</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training media ( S ):</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train10k</td>
<td>10,000</td>
<td>41,253</td>
<td>9,249</td>
<td>–</td>
<td>✓</td>
</tr>
<tr>
<td>Train100k</td>
<td>100,000</td>
<td>214,666</td>
<td>68,215</td>
<td>–</td>
<td>✓</td>
</tr>
<tr>
<td>Train1m [Li et al. 2012]</td>
<td>1,198,818</td>
<td>1,127,139</td>
<td>347,369</td>
<td>–</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Test media ( \mathcal{X} ):</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIRFlickr [Huiskes et al. 2010]</td>
<td>25,000</td>
<td>67,389</td>
<td>9,862</td>
<td>14</td>
<td>✓</td>
</tr>
<tr>
<td>Flickr51 [Wang et al. 2010]</td>
<td>81,541</td>
<td>66,900</td>
<td>20,886</td>
<td>51</td>
<td>–</td>
</tr>
</tbody>
</table>

Data servers

LIMITATIONS IN OUR PROTOCOL

- Tag informativeness in tag assignment

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

<table>
<thead>
<tr>
<th>SemanticField</th>
<th>RelExamples</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="SemanticField image" /></td>
<td><img src="image2" alt="RelExamples image" /></td>
</tr>
<tr>
<td><img src="image1" alt="SemanticField image" /></td>
<td><img src="image2" alt="RelExamples image" /></td>
</tr>
<tr>
<td><img src="image1" alt="SemanticField image" /></td>
<td><img src="image2" alt="RelExamples image" /></td>
</tr>
<tr>
<td><img src="image1" alt="SemanticField image" /></td>
<td><img src="image2" alt="RelExamples image" /></td>
</tr>
<tr>
<td><img src="image1" alt="SemanticField image" /></td>
<td><img src="image2" alt="RelExamples image" /></td>
</tr>
<tr>
<td><img src="image1" alt="SemanticField image" /></td>
<td><img src="image2" alt="RelExamples image" /></td>
</tr>
<tr>
<td><img src="image1" alt="SemanticField image" /></td>
<td><img src="image2" alt="RelExamples image" /></td>
</tr>
</tbody>
</table>

Need fine-grained annotation

---

REFERENCES


• [Li 2007] M. Li, Texture Moment for Content-Based Image Retrieval, ICME 2007


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

\[
\begin{align*}
  f(x_1, t_1) & \quad f(x_1, t_2) & \ldots & \quad f(x_1, t_m) \\
  f(x_2, t_1) & \quad f(x_2, t_2) & \ldots & \quad f(x_2, t_m) \\
  \vdots & \quad \vdots & \ddots & \quad \vdots \\
  f(x_n, t_1) & \quad f(x_n, t_2) & \ldots & \quad f(x_n, t_m)
\end{align*}
\]

n images

m tags
## Key Methods

<table>
<thead>
<tr>
<th>Media \ Learning</th>
<th>Instance Based</th>
<th>Model Based</th>
<th>Transductive Based</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tag</strong></td>
<td><strong>SemanticField</strong> [Zhu et al. 2012]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>TagCooccur</strong> [Sigurbjörnsson and van Zwol 2008]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>KNN</strong> [Makadia et al. 2010]</td>
<td><strong>TagFeature</strong> [Chen et al. 2012]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>RelExample</strong> [Li and Snoek 2013]</td>
<td></td>
</tr>
<tr>
<td><strong>Tag + Image + User</strong></td>
<td><strong>TagVote</strong> [Li et al. 2009b]</td>
<td></td>
<td><strong>TensorAnalysis</strong> [Sang et al. 2012a]</td>
</tr>
</tbody>
</table>
# Key Methods

<table>
<thead>
<tr>
<th>Media \ Learning</th>
<th>Instance Based</th>
<th>Model Based</th>
<th>Transductive Based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag</td>
<td><strong>SemanticField</strong> [Zhu et al. 2012]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>TagCooccur</strong> [Sigurbjörnsson and van Zwol 2008]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>KNN</strong> [Makadia et al. 2010]</td>
<td><strong>TagFeature</strong> [Chen et al. 2012]</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>RelExample</strong> [Li and Snoek 2013]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tag + Image + User</td>
<td><strong>TagVote</strong> [Li et al. 2009b]</td>
<td></td>
<td><strong>TensorAnalysis</strong> [Sang et al. 2012a]</td>
</tr>
</tbody>
</table>
• 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

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

\[
\text{NGD}(x, y) = \frac{\max\{\log h(x), \log h(y)\} - \log h(x, y)}{\log N - \min\{\log h(x), \log h(y)\}},
\]

\[
\text{FCS}(x, y) = e^{-\text{NGD}(x, y)/\sigma}
\]

**WU-PALMER SIMILARITY**

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

- 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.

---

## SEMANTICFIELD

<table>
<thead>
<tr>
<th>[Zhu et al. 2012]</th>
<th>Instance-Based</th>
<th>Tag</th>
</tr>
</thead>
</table>

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

- \textit{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

<table>
<thead>
<tr>
<th>Media \ Learning</th>
<th>Instance Based</th>
<th>Model Based</th>
<th>Transductive Based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag</td>
<td><strong>SemanticField</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Zhu et al. 2012]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>TagCooccurs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Sigurbjörnsson and van Zwol 2008]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tag + Image</td>
<td><strong>TagRanking</strong></td>
<td><strong>TagProp</strong></td>
<td><strong>RobustPCA</strong></td>
</tr>
<tr>
<td></td>
<td><strong>KNN</strong></td>
<td><strong>TagFeature</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Makadia et al. 2010]</td>
<td>[Chen et al. 2012]</td>
<td></td>
</tr>
<tr>
<td>Tag + Image + User</td>
<td><strong>TagVote</strong></td>
<td><strong>RelExample</strong></td>
<td><strong>TensorAnalysis</strong></td>
</tr>
<tr>
<td></td>
<td><strong>TagCooccurs+</strong></td>
<td>[Li and Snoek 2013]</td>
<td>[Sang et al. 2012a]</td>
</tr>
<tr>
<td></td>
<td>[Li et al. 2009b]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**TagRanking**

[Liu et al. 2009]  

<table>
<thead>
<tr>
<th>Instance-Based</th>
<th>Tag + Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>![image](flower tree bird sky)</td>
<td>![image](Gaussian Kernel Density Estimation)</td>
</tr>
<tr>
<td><img src="S(bird)" alt="image" /></td>
<td><img src="S(tree)" alt="image" /></td>
</tr>
<tr>
<td><img src="bird" alt="image" /></td>
<td><img src="tree" alt="image" /></td>
</tr>
<tr>
<td><img src="flower" alt="image" /></td>
<td><img src="sky" alt="image" /></td>
</tr>
<tr>
<td><img src="S(flower)" alt="image" /></td>
<td><img src="S(sky)" alt="image" /></td>
</tr>
</tbody>
</table>

- 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**

<table>
<thead>
<tr>
<th>[Liu et al. 2009]</th>
<th>Instance-Based</th>
<th>Tag + Image</th>
</tr>
</thead>
</table>

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

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

- \( l_x \) 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

<table>
<thead>
<tr>
<th>Media \ Learning</th>
<th>Instance Based</th>
<th>Model Based</th>
<th>Transductive Based</th>
</tr>
</thead>
</table>
| Tag             | **SemanticField**  
[Zhu et al. 2012] |  |  |
|                 | **TagCooccur**  
[Sigurbjörnsson and van Zwol 2008] |  |  |
| Tag + Image     | **TagRanking**  
[Liu et al. 2009] | **TagProp**  
[Guillaumin et al. 2009] | **RobustPCA**  
[Zhu et al. 2010] |
|                 | **KNN**  
[Makadia et al. 2010] |  |  |
| Tag + Image + User | **TagVote**  
[Li et al. 2009b] | **TagFeature**  
[Chen et al. 2012] | **RelExample**  
[Li and Snoek 2013] |
|                 | **TagCooccur**  
[Li et al. 2009b] |  |  |
|                 | **RobustPCA**  
[Zhu et al. 2010] |  |  |
|                 | **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

\[ f_{\text{KNN}}(x, t) := k_t, \]

- \( k_t \) 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
Fig. 2. Learning tag relevance by neighbor voting. The tag relevance value of each tag is estimated by accumulating the neighbor votes it receives from visually similar images of the seed image. In this example, since four neighbor images are labeled with "bridge", the tag relevance value of "bridge" with respect to the seed image is 4. Hence, we update the tag frequency of "bridge" from 1 to 4.

Query expansion methods augment the original query by automatically adding relevant terms [30]–[32]. In [31], for instance, the authors use synonyms from a dictionary, whereas in [30], the authors select strongly related terms from text snippets returned by web search engines. Another example is [32], where the authors use clustering methods to find correlated tags. Though adding more query terms may retrieve more relevant results, how to choose appropriate expansion terms requires further research [37].

In summary, the reranking and query expansion methods try to rank images relevant with respect to a query ahead of irrelevant images. However, the methods leave the fundamental problem of subjective user tagging unaddressed. Though we have witnessed great efforts devoted into improving both image tagging and image retrieval, the efforts are almost disconnected. Recent research, e.g., [38]–[41], investigates the potential of leveraging automatic tagging results for image and video retrieval. To the best of our knowledge, however, up until now, the solutions to the two problems are still separated, including our previous works [11], [22] which deal with social image retrieval and social image tagging, respectively. This work is an attempt to solve image ranking and tag ranking in a unified tag relevance learning framework. In contrast to approaches for image ranking which are query-dependent, e.g., [25] and [28], our algorithm is query-independent. This advantage allows us to run the algorithm offline without imposing extra waiting time on users. Further, by updating tag frequency with the learned tag frequency, we seamlessly embed visual information into current tag-based social image retrieval paradigms. For automatic image tagging, our algorithm shares similarities with the model-free approaches, e.g., [7], [8], and [21], since they can be regarded as propagating tags between neighbor images. Note, however, that our algorithm is more general as it is applicable to both image retrieval and tagging. Moreover, we provide a formal analysis which is missing in previous studies.

III. LEARNING TAG RELEVANCE BY NEIGHBOR VOTING

In order to fulfill image retrieval, we seek a tag relevance measurement such that images relevant with respect to a tag are ranked ahead of images irrelevant with respect to the tag. Meanwhile, to fulfill image tagging, the measurement should rank tags relevant with respect to an image ahead of tags irrelevant with respect to the image. Recall the intuition that if different persons label visually similar images using the same tags, these tags are likely to reflect objective aspects of the visual content. This intuition suggests that the relevance of a tag given an image might be inferred from how visual neighbors of that image are tagged: the more frequent the tag occurs in the neighbor set, the more relevant it might be, as illustrated in Fig. 2. However, some frequently occurring tags, such as "2007" and "2008", are unlikely to be relevant to the majority of images. Hence, a good tag relevance measurement should take into account the distribution of a tag in the neighbor set and in the entire collection, simultaneously. Motivated by the informal analysis above, we propose a neighbor voting algorithm for learning tag relevance, as depicted in Fig. 2. Though the proposed algorithm is simple, we deem it important to gain insight into the rationale for the algorithm. The following two subsections serve for this purpose. Concretely, we first define in Section III-A two criteria to describe the general objective of tag relevance learning. Then, in Section III-B, we provide a formal analysis of user tagging and content-based nearest neighbor search. We see how our algorithm is naturally derived from the analysis. Finally, we describe in detail the algorithm in Section III-C.

• Adds two improvements to KNN-voting:
  - Unique-user constraint
  - Tag prior frequency
Tag Vote

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

\[ f_{Tag\, Vote}(x, t) := k_t - k \frac{n_t}{|S|}, \]

- \( k_t \) 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 \( O(d \cdot |S| + k \cdot \log|S|) \) – same complexity as KNN
- Memory \( O(d \cdot |S|) \)
TagProp

[Guillaumin et al. 2009]

Model-Based

Tag + Image

• Key improvement: give different weights to image neighborhoods

• 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

<table>
<thead>
<tr>
<th>Model-Based</th>
<th>Tag + Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Guillaumin et al. 2009]</td>
<td></td>
</tr>
</tbody>
</table>

\[
\hat{f}_{\text{TagProp}}(x, t) := \sum_{j}^{k} \pi_j \cdot I(x_j, t),
\]

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

**Rank weights**

\[
\pi_{ij} = \frac{\gamma}{k}
\]

**Distance weights**

\[
\pi_{ij} = \frac{\exp(-d_\theta(i, j))}{\sum_{j'} \exp(-d_\theta(i, j'))},
\]
A logistic regressor per tag upon $f_{\text{TagProp}}$, is added to promote rare tags and penalize frequent ones.

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

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

- Complexity $O(l \cdot m \cdot k)$: $l$ steps of gradient descent
- Memory $O(d \cdot |S|)$: same as KNN, extra $2m$ for logistic regression
### Key Methods

<table>
<thead>
<tr>
<th>Media \ Learning</th>
<th>Instance Based</th>
<th>Model Based</th>
<th>Transductive Based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag</td>
<td><strong>SemanticField</strong> [Zhu et al. 2012]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>TagCooccur</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Sigurbjörnsson and van Zwol 2008]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tag + Image</td>
<td><strong>TagRanking</strong></td>
<td><strong>TagProp</strong> [Guillaumin et al. 2009]</td>
<td><strong>RobustPCA</strong> [Zhu et al. 2010]</td>
</tr>
<tr>
<td></td>
<td>[Liu et al. 2009]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>KNN</strong> [Makadia et al. 2010]</td>
<td><strong>TagFeature</strong> [Chen et al. 2012]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>RelExample</strong> [Li and Snoek 2013]</td>
<td></td>
</tr>
<tr>
<td>Tag + Image + User</td>
<td><strong>TagVote</strong> [Li et al. 2009b]</td>
<td></td>
<td><strong>TensorAnalysis</strong> [Sang et al. 2012a]</td>
</tr>
<tr>
<td></td>
<td><strong>TagCooccur+</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• 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]

<table>
<thead>
<tr>
<th>Instance-Based</th>
<th>Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

- **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

<table>
<thead>
<tr>
<th>Media \ Learning</th>
<th>Instance Based</th>
<th>Model Based</th>
<th>Transductive Based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag</td>
<td><strong>SemanticField</strong>&lt;br&gt;[Zhu et al. 2012]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>TagCooccur</strong>&lt;br&gt;[Sigurbjörnsson and van Zwol 2008]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>KNN</strong>&lt;br&gt;[Makadia et al. 2010]</td>
<td><strong>TagFeature</strong>&lt;br&gt;[Chen et al. 2012]</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>RelExample</strong>&lt;br&gt;[Li and Snoek 2013]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tag + Image + User</td>
<td><strong>TagVote</strong>&lt;br&gt;[Li et al. 2009b]</td>
<td></td>
<td><strong>TensorAnalysis</strong>&lt;br&gt;[Sang et al. 2012a]</td>
</tr>
</tbody>
</table>
**TagCooccurr+**

<table>
<thead>
<tr>
<th>[Li et al. 2009b]</th>
<th>Instance-Based</th>
<th>Tag + Image</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A variant of TagCooccurr that is improved by considering the image content in addition to solely user tags</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The heuristic is updated by multiplying TagCooccurr score with a corrective factor based on Tag Vote scores</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[ f_{tagcooccurr+}(x, t) = f_{tagcooccurr}(x, t) \cdot \frac{k_c}{k_c + r_c(t) - 1} ]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( r_c(t) ) is the rank of ( t ) when sorting ( f_{tagvote}(x, t) ) in descending order. ( k_c ) is a positive weighting parameter</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Complexity ( O(d \cdot</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>Memory ( O(d \cdot</td>
<td>S</td>
</tr>
</tbody>
</table>

# Key Methods

<table>
<thead>
<tr>
<th>Media \ Learning</th>
<th>Instance Based</th>
<th>Model Based</th>
<th>Transductive Based</th>
</tr>
</thead>
</table>
| 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.

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

<table>
<thead>
<tr>
<th>Model-Based</th>
<th>Tag + Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen et al. 2012</td>
<td></td>
</tr>
</tbody>
</table>

\[
f_{\text{TagFeature}}(x, t) := b + < x_t, x >,
\]

- 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 \( O((d + d') \text{ nm}) \), \( n \) images, \( m \) tags
- Memory \( O(m (d + d')) \)
# Key Methods

<table>
<thead>
<tr>
<th>Media \ Learning</th>
<th>Instance Based</th>
<th>Model Based</th>
<th>Transductive Based</th>
</tr>
</thead>
</table>
| Tag             | **SemanticField**  
[Chen 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] | **RobustPCA**  
[Zhu et al. 2010] |
| Tag + Image + User | **TagVote**  
**TagCooccur+**  
[Li et al. 2009b] |  | **TensorAnalysis**  
[Sang et al. 2012a] |
|                 |                | **RelExample**  
[Li and Snoek 2013] |                |
[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].
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

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

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

<table>
<thead>
<tr>
<th>Media \ Learning</th>
<th>Instance Based</th>
<th>Model Based</th>
<th>Transductive Based</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tag</strong></td>
<td><em>SemanticField</em> [Zhu et al. 2012]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>TagCooccur</em> [Sigurbjörnsson and van Zwol 2008]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>RelExample</em> [Li and Snoek 2013]</td>
<td></td>
</tr>
<tr>
<td><strong>Tag + Image + User</strong></td>
<td><em>TagVote</em> [Li et al. 2009b]</td>
<td><em>TensorAnalysis</em> [Sang et al. 2012a]</td>
<td></td>
</tr>
</tbody>
</table>
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
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.
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 $O(c m^2 n + c' n^3)$: SVD computation

Memory $O(c n \cdot m + c' \cdot (n^2 + m^2))$: Full matrix $D$, tag and image similarity matrices.
# Key Methods

<table>
<thead>
<tr>
<th>Media \ Learning</th>
<th>Instance Based</th>
<th>Model Based</th>
<th>Transductive Based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag</td>
<td><strong>SemanticField</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Zhu et al. 2012]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>TagCooccur</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Sigurbjörnsson and van Zwol 2008]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tag + Image</td>
<td><strong>TagRanking</strong></td>
<td>TagProp</td>
<td>RobustPCA</td>
</tr>
<tr>
<td></td>
<td><strong>KNN</strong></td>
<td>TagFeature</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Makadia et al. 2010]</td>
<td>[Chen et al. 2012]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RelExample</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Li and Snoek 2013]</td>
<td></td>
</tr>
<tr>
<td>Tag + Image + User</td>
<td><strong>TagVote</strong></td>
<td></td>
<td><strong>TensorAnalysis</strong></td>
</tr>
<tr>
<td></td>
<td>[Li et al. 2009b]</td>
<td></td>
<td>[Sang et al. 2012a]</td>
</tr>
</tbody>
</table>
**Tensor Analysis**

- 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.

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

\[ Y = \hat{C} \times_u \hat{U} \times_i \hat{I} \times_t \hat{T} + E \]

\[ y_{u,i,t} = \sum_{\tilde{u}} \sum_{\tilde{i}} \sum_{\tilde{t}} c_{\tilde{u},\tilde{i},\tilde{t}} \cdot u_{u,\tilde{u}} \cdot i_{i,\tilde{i}} \cdot t_{t,\tilde{t}} \]
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
**TENSOR ANALYSIS**

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

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

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

- $H$ 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.
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Assignment</th>
<th>Refinement</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>TagVote</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TagProp</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TagFeature</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RelExample</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TagCooccur</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>TagCooccur+</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>RobustPCA</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>TensorAnalysis</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>SemanticField</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>TagFeature</td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

- 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
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**

![Graph showing tag refinement on MIRFlickr test set, trained on Train100k.](image)

- 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 **TagCooccur**.
**TAG RETRIEVAL**

![Graph showing average precision for various tags and methods]

- 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.
The top 10 ranked images for ‘jaguar’

<table>
<thead>
<tr>
<th>TagPosition</th>
<th>SemanticField</th>
<th>BovW + RelExample</th>
<th>CNN + RelExample</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="TagPosition" /></td>
<td><img src="image2" alt="SemanticField" /></td>
<td><img src="image3" alt="BovW + RelExample" /></td>
<td><img src="image4" alt="CNN + RelExample" /></td>
</tr>
</tbody>
</table>

While both RelExample and SemanticField outperform the TagPosition baseline, the results of SemanticField show more diversity for this ambiguous tag. The difference between (c) and (d) suggests that the results of RelExample can be diversified by varying the visual feature in use.

5.4. Flickr versus ImageNet

To address the question of whether one shall resort to an existing resource such as ImageNet for tag relevance learning, this section presents an empirical comparison between our Flickr based training data and ImageNet. A number of methods do not work with ImageNet or require modifications. For instance, tag + image + user information based methods must be able to remove their dependency on user information, as such information is unavailable in ImageNet. Tag co-occurrences are also strongly limited, because an ImageNet example is annotated with a single label. Because of these limitations, we evaluate only the two best performing methods, TagVote and TagProp. TagProp can be directly used since it comes from classic image annotation, while TagVote is slightly modified by removing the unique user constraint. The CNN feature is used for its superior performance against the BovW feature.

To construct a customized subset of ImageNet that fits the three test sets, we take ImageNet examples whose labels precisely match with the test tags. Notice that some test tags, e.g., ‘portrait’ and ‘night’, have no match, while some other tags, e.g., ‘car’ and ‘dog’, have more than one match. In particular, MIRFlickr has 2 missing tags, while
COMMON PATTERNS

• Some common patterns have emerged, independently 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 conduct 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
**IMAGE-NET RESULTS**

<table>
<thead>
<tr>
<th>Training Set</th>
<th>MIRFlickr</th>
<th>NUS-WIDE</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TagVote</td>
<td>TagProp</td>
<td>TagVote</td>
<td>TagProp</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MiAP scores:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train100k</td>
<td>0.377</td>
<td>0.383</td>
<td>0.392</td>
<td>0.389</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train1M</td>
<td>0.389</td>
<td><strong>0.392</strong></td>
<td><strong>0.414</strong></td>
<td>0.393</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ImageNet200k</td>
<td>0.345</td>
<td>0.304</td>
<td>0.325</td>
<td>0.368</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP scores:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train100k</td>
<td>0.641</td>
<td>0.647</td>
<td>0.386</td>
<td>0.405</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train1M</td>
<td>0.664</td>
<td><strong>0.668</strong></td>
<td><strong>0.429</strong></td>
<td>0.420</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ImageNet200k</td>
<td>0.532</td>
<td>0.532</td>
<td>0.363</td>
<td>0.362</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Methods trained on socially tagged datasets show better performance for tag assignment.
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.
### IMAGE NET RESULTS

#### Tag Assignment

<table>
<thead>
<tr>
<th>Training Set</th>
<th>TagVote</th>
<th>TagProp</th>
<th>TagVote</th>
<th>TagProp</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MAP scores:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train100k</td>
<td>0.377</td>
<td>0.383</td>
<td>0.392</td>
<td>0.389</td>
</tr>
<tr>
<td>Train1M</td>
<td>0.389</td>
<td>0.392</td>
<td>0.414</td>
<td>0.393</td>
</tr>
<tr>
<td>ImageNet200k</td>
<td>0.345</td>
<td>0.304</td>
<td>0.325</td>
<td>0.368</td>
</tr>
</tbody>
</table>

| **MAP scores:** |         |         |         |         |
| **NDCG20 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   |

#### Tag Retrieval

<table>
<thead>
<tr>
<th>Training Set</th>
<th>Flickr51</th>
<th>NUS-WIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TagVote</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train100k</td>
<td>0.854</td>
<td>0.860</td>
</tr>
<tr>
<td>Train1M</td>
<td><strong>0.874</strong></td>
<td>0.871</td>
</tr>
<tr>
<td>ImageNet200k</td>
<td>0.873</td>
<td>0.873</td>
</tr>
<tr>
<td><strong>TagProp</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train100k</td>
<td>0.742</td>
<td>0.745</td>
</tr>
<tr>
<td>Train1M</td>
<td>0.753</td>
<td>0.745</td>
</tr>
<tr>
<td>ImageNet200k</td>
<td><strong>0.762</strong></td>
<td><strong>0.762</strong></td>
</tr>
</tbody>
</table>

#### 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](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

**trainCollection**
- SemanticField: dosemtagrel.py
- TagCooccur: apply_tagcooccur.py
- TagRanking: tagranking.py
- KNN: apply_tagger.py
- TagVote: apply_tagger.py
- TagCooccur+: apply_tagcooccur.py

**testCollection**

**instance_based**

**model_based: training**
- TagFeature: negative_bagging.py
- RelExamples: negbp.py
- TagProp: tagprop/tagprop.py

**model_based: test**
- TagFeature: applyConcepts.py
- RelExamples: applyConcepts.py
- TagProp: tagprop/tagprop.py

**transduction_based**
- RobustPCA: robustpca/robustpca.py

**test tags**

**test images**

**pickled result matrix**

**Evaluation**
- eval/eval_pickle.sh

**Post-processing**
- postprocess/pickle_tagvotes.py

**SURVEY_DATA/eval_output/runs_method_testCollection.txt**

**SURVEY_DATA/eval_output/runs_method_testCollection.res**


# Organization of the Tutorial

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30 – 9:30</td>
<td>Part 1: Introduction</td>
</tr>
<tr>
<td></td>
<td>Part 2: Taxonomy</td>
</tr>
<tr>
<td>9:30 – 10:00</td>
<td>Part 3: Experimental protocol</td>
</tr>
<tr>
<td></td>
<td>Part 4: Evaluation</td>
</tr>
<tr>
<td>10:00 – 10:45</td>
<td>Coffee break</td>
</tr>
<tr>
<td>10:45 – 12:00</td>
<td>Part 4: Evaluation cont’d</td>
</tr>
<tr>
<td>12:00 – 12:30</td>
<td>Part 5: Conclusion and future directions</td>
</tr>
</tbody>
</table>
PART 5
CONCLUSION AND FUTURE DIRECTIONS

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

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.
SUMMARY: UNIFIED FRAMEWORK

Auxiliary Components
- Filter & Precompute

Learning
- Inductive
- Model-based
- Instance-based
- Transductive
- Transduction-based

Tag Relevance $f_\Phi(x, t; \Theta)$

Tasks
- Assignment
- Refinement
- Retrieval

Training Media
- Image $x$
- Tag $t$
- User Information $\Theta$

Test Media
- Image $x$
- Tag $t$
- User Information $\Theta$
**SUMMARY: TAXONOMY**

<table>
<thead>
<tr>
<th>Media</th>
<th>Learning</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Instance</td>
<td>Model</td>
<td>Transductive</td>
<td></td>
</tr>
<tr>
<td>Tag</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Tag + Image</td>
<td>13</td>
<td>15</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Tag + Image + User</td>
<td>5</td>
<td>7</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Taxonomy structures 60 papers along **Media** and **Learning** dimensions
**Summary: Key Methods**

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


SUMMARY: OPEN-SOURCE TESTBED

<table>
<thead>
<tr>
<th>Media</th>
<th>Media characteristics</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># images</td>
<td># tags</td>
</tr>
<tr>
<td>Training media $S$:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train10k</td>
<td>10,000</td>
<td>41,253</td>
</tr>
<tr>
<td>Train100k</td>
<td>100,000</td>
<td>214,666</td>
</tr>
<tr>
<td>Train1m [Li et al. 2012]</td>
<td>1,198,818</td>
<td>1,127,139</td>
</tr>
<tr>
<td>Test media $\mathcal{X}$:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIRFlickr [Huiskes et al. 2010]</td>
<td>25,000</td>
<td>67,389</td>
</tr>
<tr>
<td>Flickr51 [Wang et al. 2010]</td>
<td>81,541</td>
<td>66,900</td>
</tr>
</tbody>
</table>

Data servers
[1] [http://www.micc.unifi.it/tagsurvey](http://www.micc.unifi.it/tagsurvey)

Jingwei, a framework for evaluating image tag assignment, tag refinement and tag-based image retrieval:
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 INFORMATION FROM THE IMAGE AND ITS NEIGHBORS

Sample from nearest neighbors

Pooling

Class scores

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

\[
L_{p\&n} = \sum_{i} \sum_{j} \left[ \left| \Delta(y_i, y_j) - f_{s+}(x_i) + f_{s+}(x_j) \right|_+ + \left| \Delta(y_i, y_j) - f_{s-}(x_j) + f_{s-}(x_i) \right|_+ \right]
\]

(popular senses)

(unpopular senses)

- Popularity and unpopularity learned independently at train time
- Single popularity score calculated at test time
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
**PROBLEM:** **EVENT DETECTION IN VIDEO**

- **Dog show**
- **Felling a tree**
- **Wedding dance**
**TagBook: Derived from Social-tagged Video**

### Source set: Social-tagged web videos

<table>
<thead>
<tr>
<th>Video data</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Video data images" /></td>
<td>woman, outdoor, metal-crafts-project, welding machine</td>
</tr>
<tr>
<td><img src="image2.png" alt="Video data images" /></td>
<td>man, kitchen, metallic, cleaning, oven, spray, glasses,</td>
</tr>
<tr>
<td><img src="image3.png" alt="Video data images" /></td>
<td>man, snowboard, snow, board-trick,</td>
</tr>
<tr>
<td><img src="image4.png" alt="Video data images" /></td>
<td>man, climb-on, wall, gym, rock-climbing</td>
</tr>
</tbody>
</table>

TagBook = \{woman, outdoor, metal-crafts-project, welding machine, man, kitchen, ..., wall, gym, rock-climbing\}
TAGBOOK: NEW VIDEO REPRESENTATION
BEYOND TAGS: EMOJI

- Visual grammar of interaction
- Language independent
- Age accessible
- Widely supported
- Semantically diverse
- Easy form factor for smart phones and watches
IMAGE2EMOJI

ImageNet Training Data → ConvNet Classifier → Semantic Embedding → Emoji Scoring

Embedding Corpus

Emoji Prediction Scores

Dataset

Text

Emoji Names

(00:08.33) (00:16.67) (00:25.00) (00:33.33) (00:41.67) Entire Video

Cappallo et al. MM 2015
**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