MICC - Media Integration
and Communication Center

UNIVERSITÀ
DEGLI STUDI
FIRENZE
DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

Social Media Annotation

Alberto del Bimbo
Università di Firenze
alberto.delbimbo@unifi.it

Credits:
Marco Bertini,
Lamberto Ballan,
Tiberio Uricchio

40 slide presentation

Web 2.0: Social Media

- Facebook
  - 964 million monthly active users on March 2013
  - an average user has 130 friends (Dunbar's number ≅ 150)
  - more than 3.5 billion pieces of content shared per week
- Twitter
  - 200 Millions of monthly active Twitter users
  - 175 Millions of tweets per day sent in 2012 (307 avg user)
- Google+
  - 925,000 new users on Google+ every day
  - favorite among tech industries and engineers.
- Flickr
  - Flickr hosts more than 6.7 billion images
  - ~4 millions new uploads per day
- Youtube
  - More than 4 billion views a day and 60-70 hours of videos uploaded per minute
  - 500 years of YouTube videos are watched on Facebook everyday
- ….

It took to reach 50 million users:
- Radio 38 years
- TV: 13 years,
- Internet: 4 years,
- iPod: 3 years

Source: Social Media Statistics (2012)

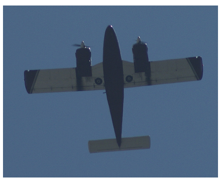
Social Tagging

- People upload, share and annotate huge quantities of multimedia content with tags motivated by contribution and sharing, self presentation, future retrieval:
  - using crowdsourcing options such as *LabelMe* or *Amazon Mechanical Turk*
  - participating to collaborative image labelling games such as *ESP game..*
  - tagging photos on URLs, *Delicious…*
  - tagging blog posts on *Wordpress*, *Livejournal…*
  - tagging on media sharing social networks like *Flickr*, *YouTube*, *Facebook..*
  - ….

Folksonomies and Challenges

- Tags imposed by social networking define soft organizations (*Folksonomies*) on data. Folksonomies pose new opportunities of semantics extraction from visual data, opposed to fixed static taxonomies that are rigid, conservative, and centralized.
- Main challenges:
  - Imprecise and ambiguous tags, order not corresponding to tag relevance and influenced by culture.
  - tags irrelevant to the visual content and overly personalized.
  - spontaneous choice of words with variability among different people, polysemy, synonymy..
  - semantic loss in the textual descriptions: meaningful tags missing.

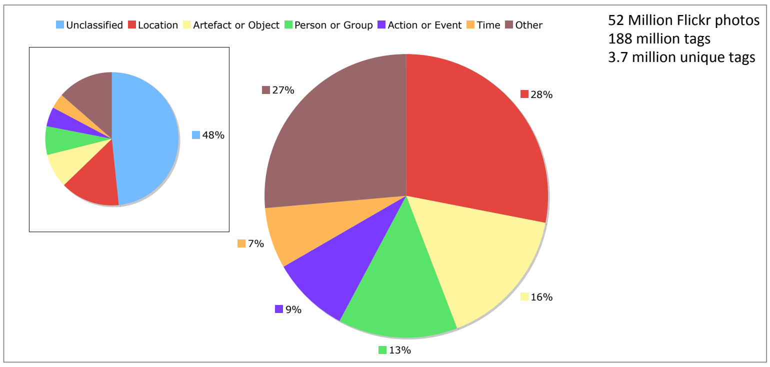
Query tag: airplane



airplane
twin
engine
los angeles

daytime
beach
airplane
ocean

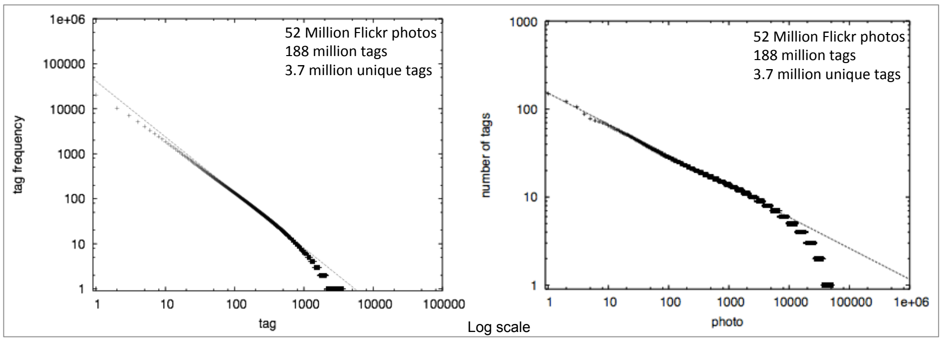
Wordnet Categories of Tags in Flickr



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

From: [Sigurbjörnsson et al.-08]

Flickr Tags Frequency and Number Distribution

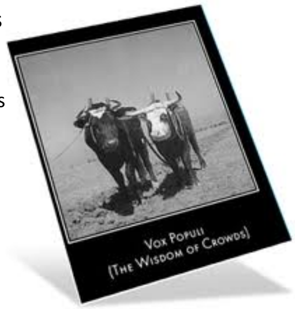


- Both tag frequency and number of tags per photo follow a power distribution.
  - Tag frequency:
    - The head of the distribution contains too generic tags to be useful (*wedding, party,…*).
    - The tail contains the infrequent tags with incidentally occurring terms such as misspellings and complex phrases.
  - Tag number: about 64% of tags have only 1-3 tags.
- These are the cases where tag recommendation can be useful.

From: [Sigurbjörnsson et al.-08]

The Wisdom of Crowds

- *The wisdom of crowds*: the verdict of a group of people is closer to the truth than that of any individual in the group [Galton 1906]
- With social media annotations tag vocabulary reaches a statistical regularity. Mechanisms to convert the opininions into an aggregated verdict:
  - *Tag co-occurrence*: the number of images where several tags are used in the same annotation is the key to tag recommendation
  - *Visual content-tag association*: if different persons label visually similar images with the same tags, these tags are likely to reflect objective aspects of the visual content.
  - *Considering the complex relationships of tags* in a folksonomy



Improving Image Tagging

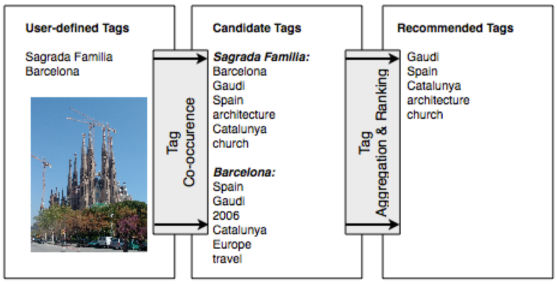
Luigi Torreggiani, CC BY 2.0 license.

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

- The goal is to improve the quality of tagging by removing noisy tags, disambiguating tags and recommending new tags that are relevant to the visual content and the other tags, automatically by exploiting folksonomies
- Includes tag refinement, tag recommendation, and tag re-ranking.

Visual Content-tag Association

- Methods addressing both visual and textual clues are required:
  - Model-based supervised approaches do not scale with social media. The scarcity of training examples and the diversity in visual appearance might make the learned models difficult to generalize.
  - Model-free semi-automatic annotation systems recommend additional tags *co-occurrence statistics, content-based retrieval, nearest neighbor matching* and *clustering*. Require users to supply an initial set of tags for images to be annotated.



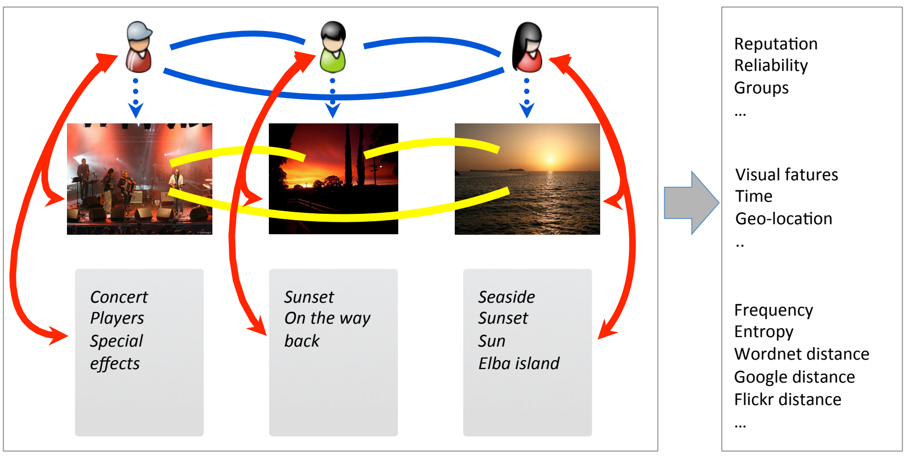
From: [Sawant et al.-11]

Tag Relationships in Folksonomy

- Popular view of a folksonomy: ternary relationship between users, images and tags. Can be modeled as a three-dimensional association matrix

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

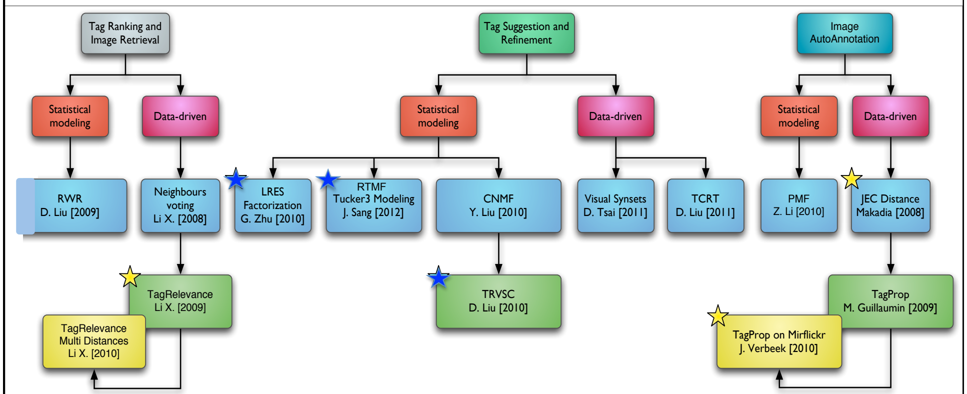
features



Addressing Social Media Annotation

- Since last 5-6 years. Two basic approaches:
  - Statistical modeling matrix factorization: describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables. Can consider users explicitely.
  - Data-driven approaches predict tags from presence/absence of tags among neighbors. Don't consider users explicitely.

Taxonomy of Main Research Contributions



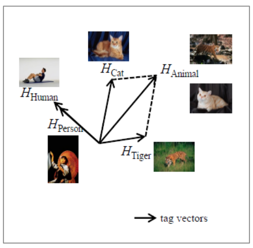
★ Tag Refinement: Statistical Modeling

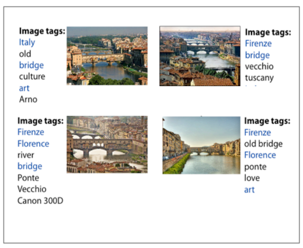
- Techniques based on statistical modeling employ matrix factorization.
- Variations in observed variables might reflect the variations in fewer unobserved variables. The observed variables are modeled as linear combinations of the potential factors, plus error terms variations in response to unobserved latent variables.
  - Low-Rank approximation and Error Sparsity (LR+ES)
  - By visual and semantic consistency (TRVSC)
  - Ranking based Multi-correlation Tensor factorization (RMTF)

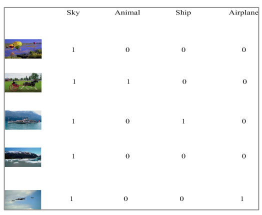
Low-Rank Approximation and Error Sparsity (LR+ES)

G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *Proc. of ACM Multimedia,* 2010.

- 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 (Google distance like).
  - *visual consistency*: visually similar images are annotated with similar tags.
  - *error sparsity* for the image-tag matrix: user's tagging is reasonably accurate and one image usually is labelled with few tags.







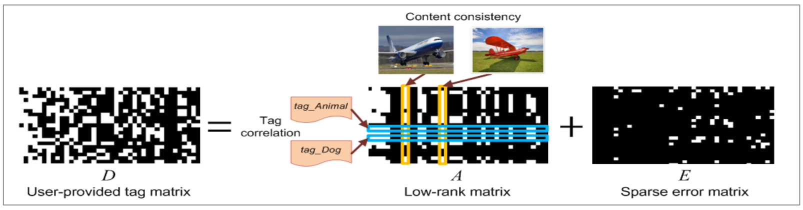
- The problem of tag refinement is cast into a decomposition of the user-provided tag matrix **D** into a low-rank refined matrix **A** and a sparse error matrix **E** subject to a cost term for every observation.

- The problem reduces to recover the noise-free matrix **A**, so each column vector can be used to represent the corresponding images:

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

||A|| cost associated to the rank
||E|| cost to maximize sparseness of E
$T_c(A)$ cost related to image content similarity
$T_t(A)$ cost based on tag correlation (rows of A should be similar if the corresponding tags are semantically near).

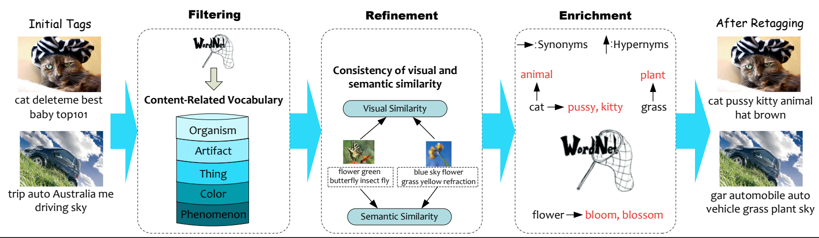
- Accelerated Proximal Gradient Method to iteratively converge to the optimal solution. Image tags are aggregated over all users, so losing important information about individual user's variation in tag usage.



Retagging by Visual and Semantic Consistency (TRVSC)

D Liu, X.-S. Hua, M. Wang, and H.-J. Zhang, "Image retagging," in *Proc. of ACM Multimedia*, 2010.

- Address the gap between tags and image content. Composite solution of tag filtering, refinement and enrichment. Assume consistency between visual and semantic similarity in social images.
  - Tag filtering uses Wordnet to constrain the tag vocabulary within content-related tags.
  - Tag refinement optimizes the filtered tags, by maximizing consistency of visual and semantic similarities between images while minimizing the deviation of the tags from those provided by users.
  - Tag enrichment expands each tag with synonyms and hypernyms in Wordnet.



- Refinement optimization framework:
  - Consistency between visual and semantic similarity: tags of visually close images are expected to be similar
  - User-provided tags are relevant with high probability

$$\min_{\mathbf{Y},\mathbf{D}} ||\mathbf{W} - \mathbf{Y}\mathbf{S}\mathbf{Y}^T||_F^2 + C||\mathbf{Y} - D\hat{\mathbf{Y}}||_F^2$$
$$s.t. \quad Y_{jl}, D_{jj} \geq 0$$

$$W_{ij} = \exp(-\frac{||x_i - x_j||_2^2}{\sigma^2})$$
$$S_{ij} = \frac{2 * IC(lcs(t_i, t_j))}{IC(t_i) + IC(t_j)}$$

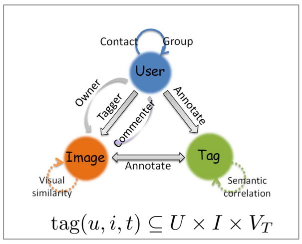
minimizes the difference between image similarity (a Gaussian kernel over visual features) and tag similarity (from WordNet using Lin's Similarity measure).

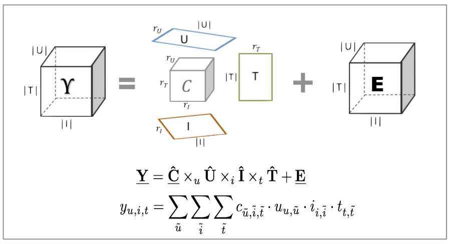
- Efficient iterative optimization algorithm with provable convergent properties. Use of a full visual similarity matrix between all images poses a scalability issue.

Ranking Based Multi-correlation Tensor Factorization (RMTF)

J. Sang, C, Xu, J. Liu
"User-aware image tag refinement via ternary semantic analysis" in IEEE Trans. On Multimedia, 2012.

- 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 and uses a tensor-based representation and Tucker decomposition into latent subspaces for the latent factor inference.



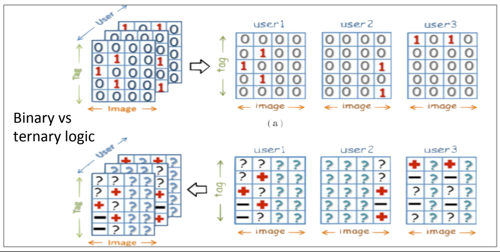


- Only qualitative differences are important. The task is cast into a ranking problem to determine which tag is more relevant for a user to describe an image.
- Ternary logic: tags that are not in the list can be either missing or non relevant:
  - *positive tags* (tags assigned by the users),
  - *negative tags* (dissimilar tags that rarely occur together with positive tags)
  - *neutral tags* (all the other tags) removed from the learning process

- Model parameters are learned by minimizing the penalty for each *positive tag* with rank lower than a *negative tag*:

$$\underset{\theta}{\operatorname{argmin}} \sum_{t^+ \in T^+} \sum_{t^- \in T^-} H(\hat{y}_{t^-} - \hat{y}_{t^+}) \quad \forall \, \text{user}, \text{image}$$

- Optimization is obtained iteratively using stochastic gradient, one latent matrix at a time.



Affine items will be mapped close in the subspace

$$W^U_{i,j} = \frac{n(u_i, u_j)}{n(u_i) + n(u_j)}$$

Based on common Flickr user groups

$$W^T_{i,j} = \lambda \cdot T^c_{i,j} + (1 - \lambda) \cdot T^s_{i,j}$$

Based on Lin similarity in Wordnet and tag cooccurence

$$W^I_{i,j} = e^{-\frac{\mu(I_i, I_j)}{\sigma^2}}$$

Based on visual distance

★ Tag Refinement: Data-driven

- Data driven methods exploit binary image-tag relations. Assume there exist large well labeled dataset where one can find visual duplicates of the image.

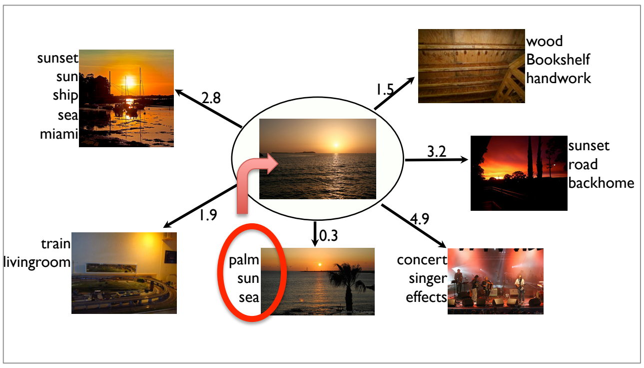
- Ground on the idea of selecting a set of visually similar images and then extract a set of relevant tags associated using a tag transfer procedure. Nearest Neighbor voting is used

  - Simple Label Transfer (SLT)
  - Tag Relevance Learning (TR)
  - Tag Prop weighted NN annotation (TAGProp)

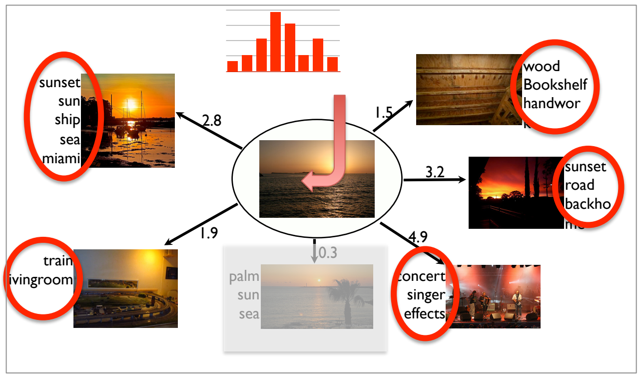
Simple Label Transfer (SLT)

A. Makadia, V. Pavlovic, and S. Kumar,
"A new baseline for image annotation," in *Proc. of ECCV*, 2008.

- Images are ranked according to content similarity distances. Joint equal contribution between distances or Lasso are used.
- The most similar image is selected and its tags are applied



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



Tag Prop weighted NN annotation (TP)

M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. of ICCV*, 2009.

- Learns a weighted nearest neighbor model to find the optimal combination of feature distances. The model is defined over a probabilistic framework.
- The probability that word w is associated to an image i is defined as:

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

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

y_{iw} ∈ {+1,-1} indicates whether tag w is relevant or not for image i

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

- The objective is to maximize the log-likelihood by using an EM-algorithm or a projected gradient descent:

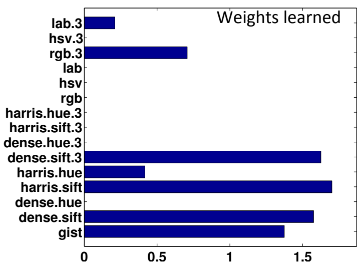
$$\mathcal{L}=\sum_{i,w}\ln p(y_{iw})=\sum_{i,w}\ln\sum_j \pi_{ij}\,p(y_{iw}|j)$$

- Weights can be defined as a function of distance of neighbours images.

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

- Due to the unbalanced frequency of tags, a word-specific logistic discriminant is introduced to boost the probability for rare terms and decrease it for frequent ones.

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

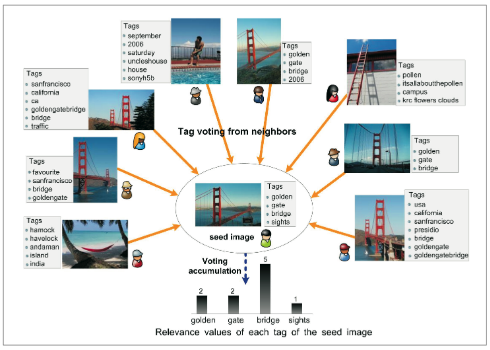


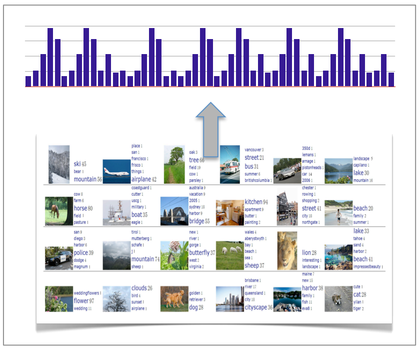
clouds	sky (0.99)
male	clouds (0.99)
people	water (0.69)
sea	structures (0.64)
sky	sea (0.32)
water	tree (0.32)

Tag Relevance Learning (TR)

X. Li, C. G. M. Snoek, and M. Worring,
"Learning social tag relevance by neighbor voting,"
IEEE Transactions on Multi- media, vol. 11, no. 7, pp. 1310–1322, 2009.

- Define a tag relevance measure considering the distribution of the tag in the neighbor set of the image and in the entire collection:
  - the more frequent a tag is in the neighbor set the more relevant it is (if different persons label visually similar images using the same tags, then these tags are more likely to reflect objective aspects of the visual content).
  - frequently occurring tags in the collection are unlikely to be relevant to the majority of images





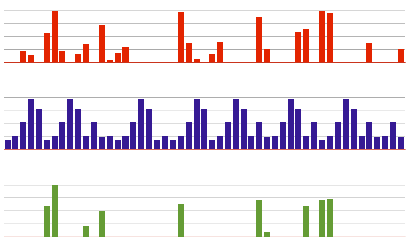
- Images relevant wrt a tag should be ranked ahead of images irrelevant wrt the tag. Simply using a tf-idf like scheme tends to overweight rare tags

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

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

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

Only one image per user is considered



Distribution of each tag in $N_k(I, k)$

minus

Prior of each tag in the collection

Final TagRelevance measure

- Good tag relevance for both image ranking and tag ranking, if we can assume that:
  - probability of correct user tagging is larger than incorrect tagging
  - content based search is better than search at random

A Comparative Analysis

T. Uricchio, L. Ballan, M. Bertini and A. Del Bimbo,
"An evaluation of nearest-neighbor methods for tag refinement,
Proc. of ICME, 2013. (to be presented).

- MIRFlickr dataset:
  - 16 global and local features
  - Distance: combination of L2 and e KL-divergence
  - Performance: macro and micro-average
  - Train 10k and test 15k set

1.3 tag per image avg

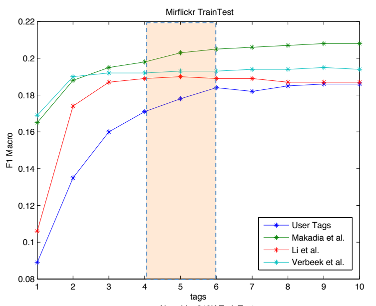
- NUSwide dataset:
  - features 428dim feature vector:
    - Color moments
    - Wavelet coefficients
    - Edge histogram
  - Distance: L2-norm
  - Only tags with direct correspondence with ground truth are retained
  - Performance: macro and micro-average
  - Train 1/3 and test 2/3 set

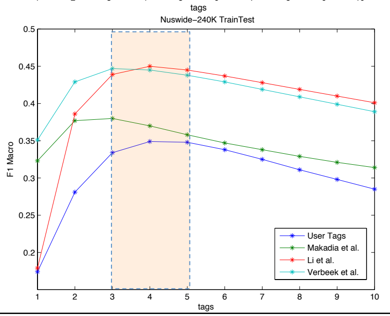
4 tag per image avg

	NUSWIDE-270K	NUSWIDE-240K	MIRFLICKR
Images	269,648	238,251	25.000 Flickr images
Train Set	161,789	158,834	–
Test Set	107,859	79,417	–
Ground truth tags	81	81	27
Users	–	24,625	9.862
Original Tags	5,018	5,018	1386
Filtered Tags Wikipedia	521	–	–
Filtered Tags WordNet	–	684	219

Methods Comparison

- Dependency of precision on number of tags suggested (train dataset)





[4] TRSVC D Liu, X.-S. Hua, M. Wang, and H.-J. Zhang, "Image retagging," *Proc. of ACM Multimedia*, 2010.

[5] LR G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," *Proc. of ACM Multimedia*, 2010.

[11] RWR C. Wang, F. Jing, L. Zhang, and H.-J. Zhang, "Content-based image annotation refinement," *Proc. of CVPR*, 2007.

[7] SLT A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," *Proc. of ECCV*, 2008.

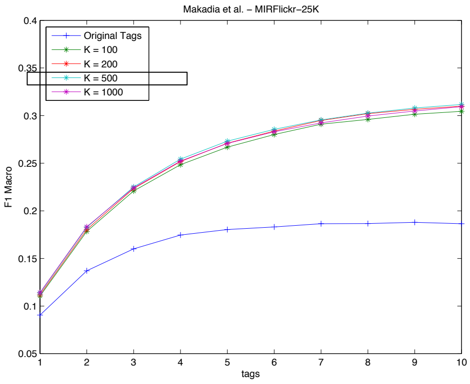
[8] TR X. Li, C. G. M. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting, " *IEEE Transactions on Multi- media*, vol. 11, no. 7, 2009.

[9] TP M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Tag- prop: Discriminative metric learning in nearest neighbor models for image auto-annotation," *Proc. of ICCV*, 2009.

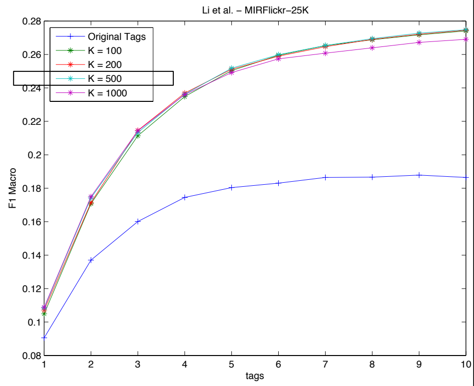
FULL NUSWIDE-270K performance comparison

	UT	RWR [11]	TRVSC [4]	LR [5]
Zhu *et al.* [5]	0.22	0.34	0.41	0.42
Liu *et al.* [16]	0.2	0.31	0.37	-

- Influence of neighbourhood on precision (full dataset)



Simple Label Transfer algorithm Makadia



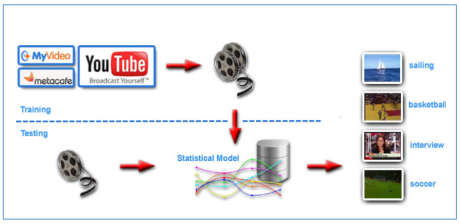
Tag Relevance Learning algorithm Li

Considerations

- Nearest-neighbour methods, when applied to tag refinement, give comparable results to more complex state-of-the-art approaches, despite their simplicity and low computational cost.
- High sparsity and unbalanced annotations are the main difficulties to overcome. Decomposition models are capable of addressing an higher level semantic by exploiting latent relations and fusing several dimensions of multimedia data. This can be an advantage as the dictionary of annotations become bigger and bigger.
- Nearest-neighbor models depends exclusively on the distance over images. High level concepts (with a strong semantic gap) are difficult to predict. A higher level semantic space could be the way to boost performance in NN models.

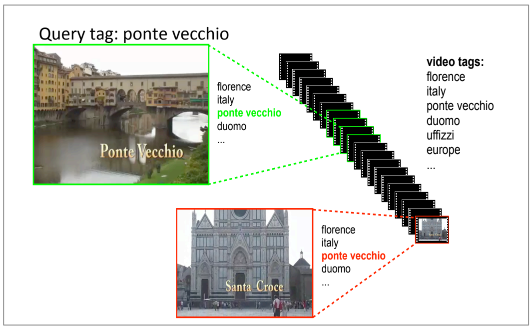
Video Social Annotation

- Tag suggestion in user-generated videos has been less explored:
  - Use YouTube related videos metadata to enrich and re-rank information of a specific video [Wu et al.' 09, Liu et al.' 10]
  - Learn statistical models for the appearance of semantic concepts to tag unseen videos [Ulges et al. '10].
  - Visual near-duplicates for tag-suggestion and video re-ranking [Siersdorfer et al.' 09, Zhao et al.' 10]
- Effective automatic annotation of video must address refinement of the existing tags and temporal localization within video shots.
  - Tag localization based on learning [Chu et al. '11] [Li et al. '11] [Li et al. '13]



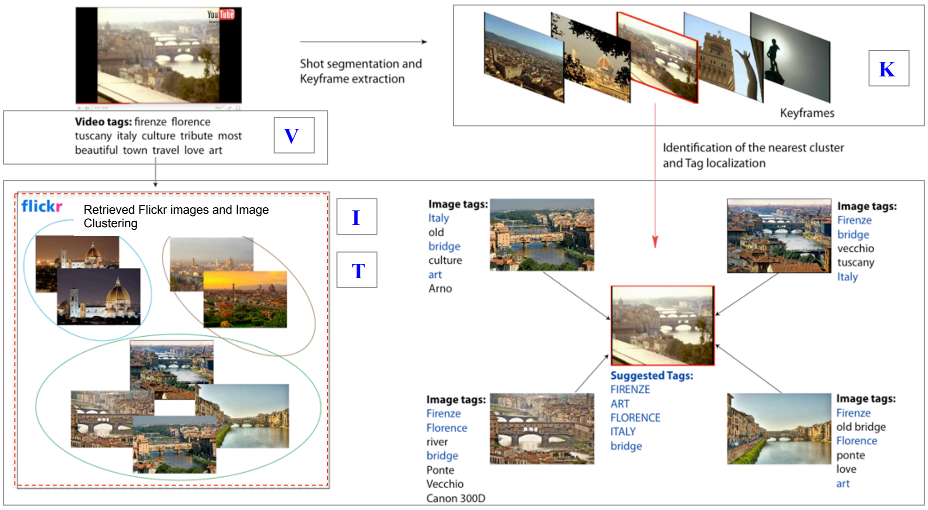
Extending Data-driven Methods to Video

- Key problems with web video social annotation:
  - *Domain complexity*. Online video as a domain shows very high variability of concepts (3000–5000 for news video, higher for general-purpose video [Hauptmann et al. 2009])
  - *Unreliable* and *coarse annotations*. Online video tags are coarse with no shot-accurate annotations. Tags are not localized in the video frames.
  - *Temporal relationships* between tags are highly semantical
- Data-driven methods are appropriate for tag localization in videos



Video Tag Refinement with Tag Localization

L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra
"Enriching and Localizing Semantic Tags in Internet Videos"
Proc. of ACM International Conference on Multimedia, 2011



- Flickr images $\mathbf{I}$ retrieved using video tags are clustered by k-means.
- For each keyframe in $\mathbf{K}$ and image in $\mathbf{I}$ a 72-d visual feature vector is computed
- Cluster centers of images in $\mathbf{I}$ are used as an index to search for similar keyframes in $\mathbf{K}$
- The set of image tags $\mathbf{T}$ of $\mathbf{I}$ is assumed as the dictionary.
- For each keyframe k retrieved, all the images in the clusters are regarded as visual neighbors of k
- Tags of these images are associated to keyframe k : $\mathbf{T_k} = \{\mathbf{v_1, ..., v_n}\}$
- Video tags in $\mathbf{V}$ are assumed as valid only if they are also in $\mathbf{T_k}$ (otherwise are eliminated from the tag list)

- Tag relevance is computed for the tags in $\mathbf{T_k}$ in order to add new tags. Relevance is evaluated by *tag relevance learning*, counting the occurrences of each tag t in the kNN images, minus the prior frequency of t
- WordNet synonyms are added to $\mathbf{T_k}$ and used to download new images from Flickr

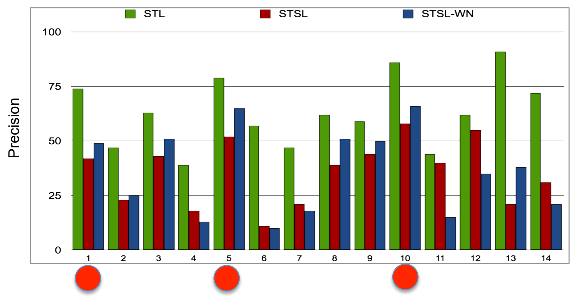
- Tags in $\mathbf{T_k}$ that have high value of co-occurrence with the initial tag set define a tag candidate list $\mathbf{C}$
- For each candidate tag c in $\mathbf{C}$ a *score*(c,$\mathbf{T_k}$) is computed according to the Vote⁺ algorithm. For each candidate tag c and each keyframe k, the *score*(c,k) is computed as:

$$score(c,k) = score(c,T_k) \cdot \frac{\lambda}{\lambda + (rank_c - 1)}$$

- The five most relevant tags according to *score*(c,k) are added at the shot level
- The union of all tags added at shot level are used for annotation at video level

Experimental Results

- YouTube60 dataset http://www.micc.unifi.it 4 YouTube videos for each YouTube category
  - 1135 shots, 3405 keyframes annotated
  - all the videos tagged by YouTube users (min. 3, max 26 tags)
- For each YouTube tag 15 Flickr images downloaded
- 5 additional Flickr images for each WordNet synonym



Shot level Tag Localization (STL)
Accuracy of localization of the user-generated YouTube tags in the correct shots

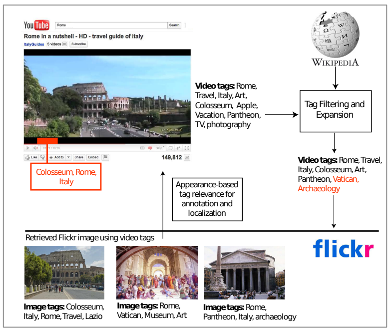
Shot level Tag Localization and Suggestion (STSL)
accuracy of the tag localization at shot level for both user-generated and suggested tags

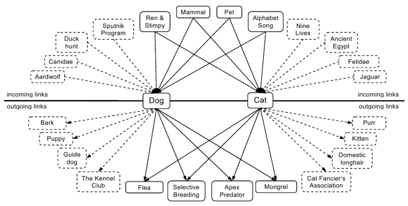
STSL with WordNet expansion (STSL-WN)
accuracy of STSL with WordNet synset expansion of the YouTube tags in the localization process

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

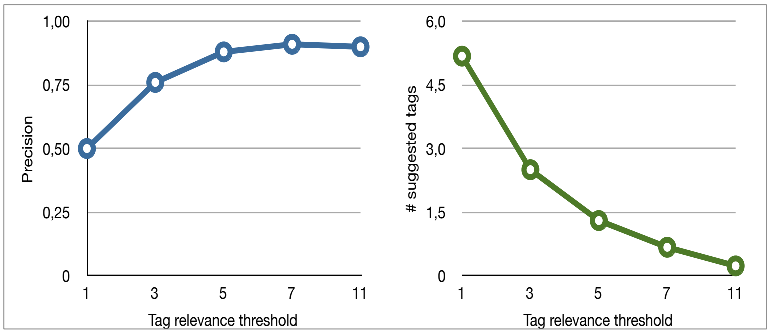
More on Semantic Tag Filtering and Expansion

- Semantic tag expansion can be done done considering:
  - Tags in You Tube related videos with a high number of occurrences.
  - Relevant Wikipedia articles expand semantically the tags: consider the list of anchors as candidate tags.
  - Tag filtering based on tag position in the video title or relevance from You Tube related videos: for each filtered tag, download the first 15 Flickr images ranked according to the relevance criterion of the Flickr API.





Experimental results



- Precision: average tag localization precision as ratio between the number of tags correctly suggested and the total number of suggested tags.
- Tag relevance threshold: how many times a tag should be present in the neighbourhood to be added

- Different performance by video category: "Auto & Vehicle" best; "Film & Animation" difficult to retrieve Flickr images similar to trailer scenes; "Howto & Style" too diverse content, hard to be correctly annotated

Some Examples (a composed clip)

Uppercase: original YouTube tags at video level
Lowercase: suggested new tags at shot level

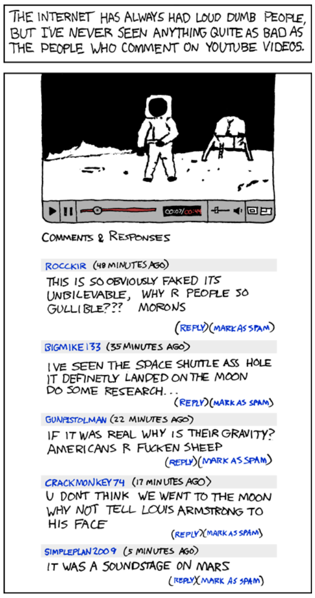
Questions on external knowledge

- Exploit ontologies, ImageNet to search for candidate tags
  - ...but who creates/maintains these ontologies ?
  - ...are they computationally expensive ?
  - ...are they big enough ?
- Exploit knowledge bases like Wikipedia to select candidate tags
  - ...need of algorithms to select tags from Wikipedia pages: use Google-like relevance ? Topic models ? a mix of Google-like + Topic models ?
- Exploit ready-made services like the Youtube related videos
  - ...can we rely on it ? How does it work ?
  - ...isn't it too much application specific ?

Adding community information

- Social graph
  - ...a user may be part of specific groups
    (e.g. Flickr groups, as in [Ulges and Worring 2011])
  - ...a user may have a pattern in the type of material uploaded
  - ...the likes may indicate that the content of a user is interesting for a community

- User comments
  - ...although taking them into account may be much more noisy than tags...



41

41