Tag Completion based on Belief Theory and Neighbor Voting

Amel ZNAIDIA[1,2], Hervé LE BORGNE[1]
Céline HUDELOT[2]

[1] Vision Content Engineering Laboratory
CEA LIST, France

[2] Laboratory of Mathematics Applied to Systems,
Ecole Centrale Paris, France

Motivation

- Users can annotate their photos with their own tags,

flickr

Describe the content

Add tags [?]

Add

Title, description, tags

Title

Image1.jpg

Description

Tags

Motivation

- tags are usually imperfect, only 50% are related to image content.

Tags :
Dog, corgie,50mm, captain, Seattle, SonyA200, Minolta.

No Tag

Motivation

- tags related to shooting conditions,

Tags :
Dog, corgie,50mm, captain,
Seattle, SonyA200, Minolta.

No Tag

Motivation

- tags related to shooting conditions, subjective context,

Tags :
Dog, corgie,50mm, captain, Seattle, SonyA200, Minolta.

No Tag

Motivation

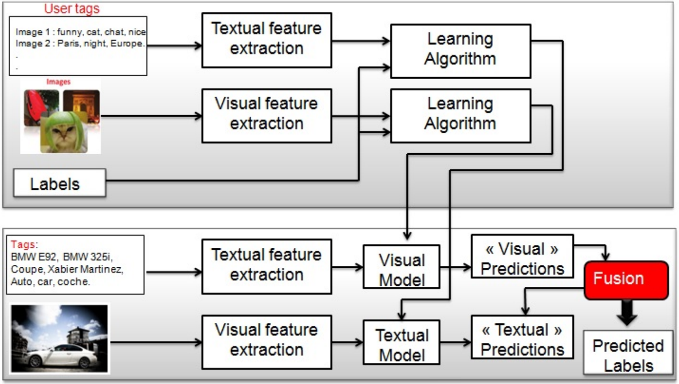
- tags related to shooting conditions, subjective context,
- misspelled and missing tags.

Tags :
Dog, Corgie ,50mm, captain,
Seattle, SonyA200, Minolta.

No Tag

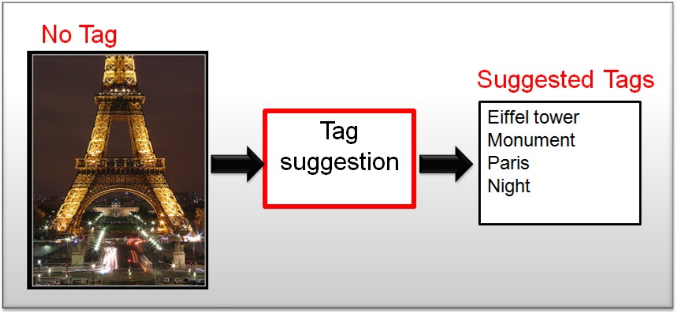
The goal of this work

- Tag completion for :
  1. Multimodal image classification,



The goal of this work

- Tag completion for :
  1. Multimodal image classification,
  2. Tag suggestion.

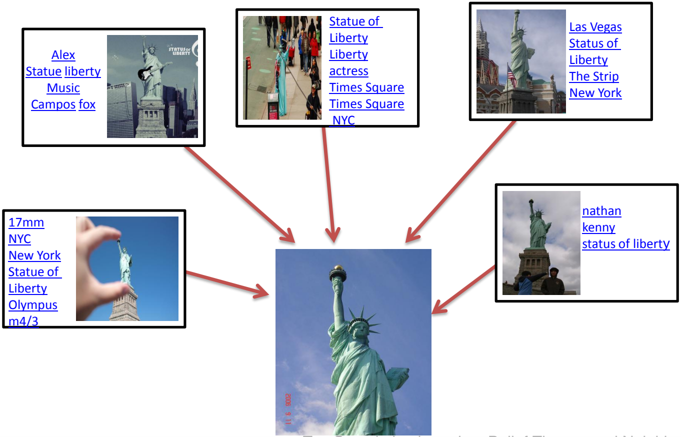


Outline

1. Related works
2. The proposed method
3. Experimental results
4. Conclusions and perspectives

Related works

Intuition: *"if many distincts users use the same tags to label visually similar images then these tags are likely to reflect the visual content of these images"*

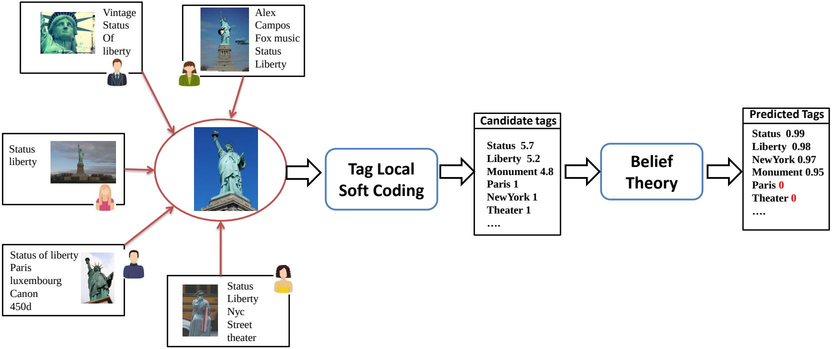


Related works

- [Makadia et al.2008] developed the *Joint Equal Contribution*: a combination of multiple features and distance metrics,
- [Li et al.2009] learns *tag relevancy* by accumulating votes from visually similar neighbors,
- [Wang et al.2009] proposed to build a *normalized histogram of tags* from k-nearest neighbor images,
- [Guillaumin et al.2009] proposed the *tag propagation* method to annotate a input image by propagating the tags of the weighted nearest neighbors,

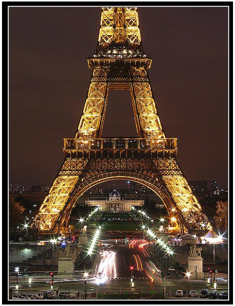
The proposed method

- Create a list of candidate tags from visual neighbors,
- Use them as *pieces of evidence* to provide final tags.



Probabilistic Tag Modeling

- Exploiting external knowledge for tag modeling,
- Contextual similarity based on tag social relatedness in Flickr.



Tags:

Eiffel Tower

cat | Paris | Eiffel Tower | monument

$$\begin{pmatrix} 0 & \cancel{0} & \ldots & 1 & \ldots & \cancel{0} \\ & 0.89 & & & & 0.78 \end{pmatrix}$$

Contextual similarity based on Flickr

- An adaptation of the TF-IDF model to the social space to compute social tag relatedness [Popescu and Grefenstette2011]:

$$\mathbf{S}(i,j) = \text{users}(\mathbf{t}_i, \mathbf{t}_j) \times \log\left(\frac{\text{users}_{\text{collection}}}{\text{users}_{\text{collection}(\mathbf{t}_j)}}\right),$$

- A Flickr normalized model for tags:

$$\mathbf{w}_i = [w_{i,1}, w_{i,2}, ..., w_{i,K}]^T, w_{i,j} = \frac{\mathbf{S}(i,j)}{max\{\mathbf{S}(i,k), k = 1, ..., K\}}.$$

$$\text{sim}_{\text{contextual}}(\mathbf{t}_i, \mathbf{t}_j) = \frac{\mathbf{w}_i^T \mathbf{w}_j}{||\mathbf{w}_i||||\mathbf{w}_j||}.$$

Finding candidate tags

- Let I be an untagged image and $\mathcal{N} = \{I^1, \ldots I^k\}$ the set of its nearest neighbors,
- Local Soft Coding for each tag,

$$z_{p,q} = \begin{cases} \text{sim}_{\text{contextual}}(\mathbf{t}_p^r, \mathbf{b}_q) & \text{if } \mathbf{b}_q \in \mathcal{N}_M(\mathbf{t}_p^r)\,, \\ 0 & \text{otherwise,} \end{cases}$$

Tags: cat | Sport | Tesnnis | challenge

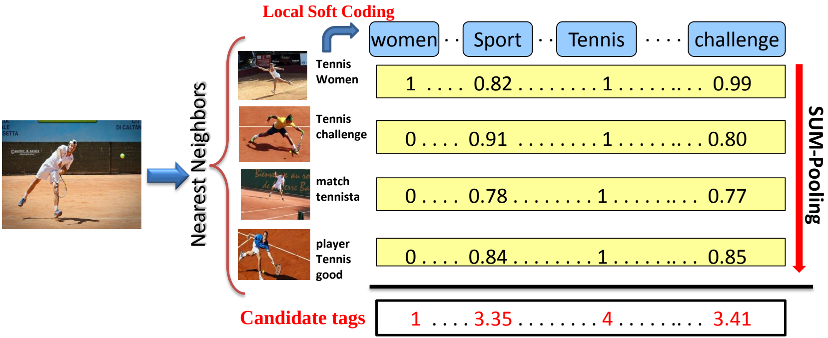
$$\begin{array}{l} \text{Tennis} \\ \text{Challenger} \\ \text{Tennista} \end{array} \begin{pmatrix} 0 \ldots\ldots 0.82 \ldots\ldots 1 \ldots\ldots\ldots 0.64 \\ 0 \ldots\ldots 0.6 \ldots\ldots 0.5 \ldots\ldots\ldots 0.99 \\ 0 \ldots\ldots 0.58 \ldots\ldots 0.98 \ldots\ldots\ldots 0.7 \end{pmatrix}$$

Max-Pooling

Final signature 0 . . . . 0.82 . . . . 1 . . . . 0.99

Finding candidate tags

- For each neighbor, a tag-signature is obtained based on Local Soft Coding,
- A sum-pooling across the *k* nearest tag-signatures to obtain "candidate tags",



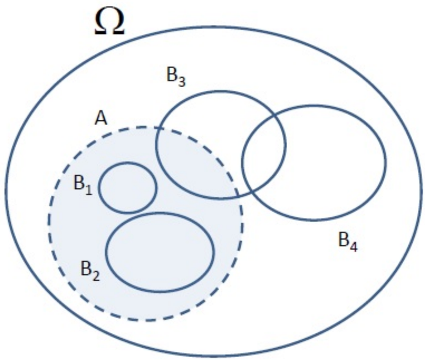
Belief Theory

- Evidence and Dempster-Shafer theory,
- Take into account uncertainties and imprecisions,
- Ω the *frame of discernment*: set of all hypothesis in a domain,
- A basic belief assignment (BBA) is a function m:

$$m : 2^{\Omega} \rightarrow [0, 1], \sum_{A \in 2^{\Omega}} m(A) = 1 \tag{1}$$

- $m(A)$: measure of the belief *exactly* committed to A,

Belief Theory

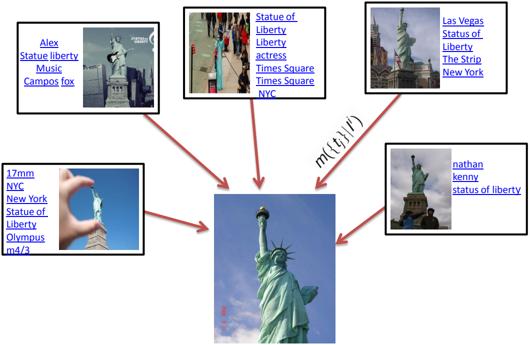


$$Bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B)$$

$$Pl(A) = \sum_{A \cap B \neq \emptyset} m(B)$$

$$m_1 \oplus m_2 = \begin{cases} \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B) m_2(C)}, & \forall\ A \subseteq \Omega,\ A \neq \emptyset \\ 0 & if\ \ A = \emptyset \end{cases}$$

Predicting Final tags

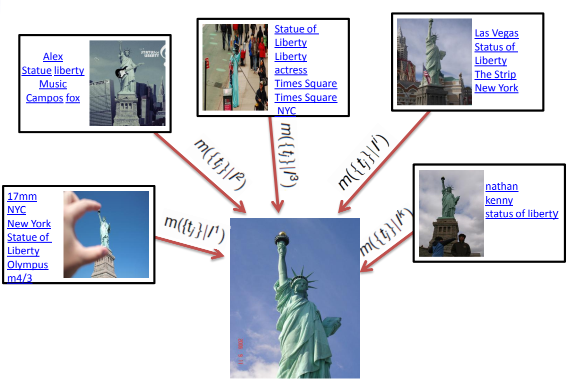


- $\Omega = \{t_1, ..., t_n\}$ the set of "candidate tags",
- $\mathcal{N} = \{I^1, \ldots I^k\}$ the set of k nearest neighbors of I,
- (I^i, t_j): pieces of evidence regarding the relevance of t_j,
- Strength of evidence decreases with $d(I, I^i)$,

$$m(\{t_j\}|I^i) = \alpha\phi_j(d^i), \quad m(\Omega|I^i) = 1 - \alpha\phi_j(d^i)$$

$$\phi_j(d) = exp(-\gamma_j d^2)$$

Predicting Final tags



$$m(\{t_j\}) = \frac{1}{K}(1 - \prod_{i \in \mathcal{N}_j}(1 - \alpha\phi_j(d^i))) \prod_{l \neq j} \prod_{i \in \mathcal{N}_l}(1 - \alpha\phi_l(d^i))$$

Image classification

- PASCAL VOC 2007 dataset
  - ≈10k images (5k for training and 5k for test),
  - 20 object classes.
- MIR Flickr dataset
  - 18k images (10k for training and 8k for test),
  - 99 concepts.

Table : Number and proportion of untagged images.

Dataset	# untagged Train	# untagged Test
Pascal VOC 2007 (prop. total)	1917 (38.3%)	1847 (37.3%)
MIR Flickr (prop. total)	812 (10.1%)	930 (9.3%)

Image classification

- Our method remains stable and more effective while varying the neighborhood size.

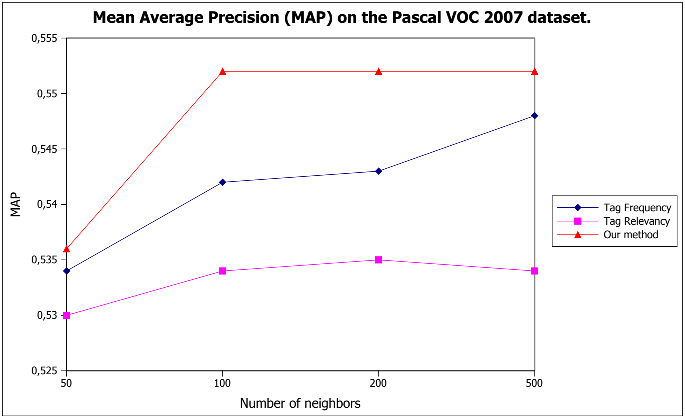


Image classification

Table : Classification performances on PASCAL VOC 07 in terms of MAP

Method	Textual	Multimodal
Tag Relevancy [Li et al.2009]	0.534	0.668
Tag Frequency [Wang et al.2009]	0.542	0.676
Our method	**0.552**	**0.684**

Table : Classification performances on MIR Flickr in terms of MAP

Method	Textual	Multimodal
Tag Relevancy [Li et al.2009]	0.337	0.412
Tag Frequency [Wang et al.2009]	0.343	0.417
Our method	**0.37**	**0.440**

Tag suggestion

- The dataset of [Sigurbjörnsson and van Zwol2008]: 331 images manually annotated,
- Our dataset[1]: 241 images manually re-annotated.




Initial ground truth	2005, february	2006 , costume october	2006, Asia, chinese, City travel	2006
Our ground truth	Music, concert Live, Show, Lights, night	People, portrait Makeup, Girl	Baby, sleeping, Bicycle, man Market, Asia	Girl, music Party, Food

[1] http://elm.eeng.dcu.ie/~hlborgne/tagcompletion.html

Tag suggestion

- We select the top 5 tags as final suggestion for each untagged image.

Method	Average Precision@5
Tag Relevancy [Li et al.2009]	0,349
Tag Frequency [Wang et al.2009]	0,387
Our method	**0,413**

Table : Comparison of our system to the state-of-the art methods on the tag suggestion task.

Conclusions and perspectives

- A novel approach for tag suggestion based on local soft coding and belief theory,
- Scheme to tackle with **imprecision** and **uncertainty** that are inherent to this type of information in a social media context,
- Results show the **competitive performances** of the proposed method on both tag suggestion and image classification,
- For tag suggestion, we manually annotated 241 queries to **propose a new benchmark** to the community,
- Exploit other visual signatures to search for neighbors.

References I

Guillaumin, M., Mensink, T., Verbeek, J., and Schmid, C. 2009.
Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation.
In *International Conference on Computer Vision*, pages 309–316.

Li, X., Snoek, C. G. M., and Worring, M. 2009.
Learning social tag relevance by neighbor voting.
IEEE Transactions on Multimedia, 11(7):1310–1322.

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

References II

Popescu, A. and Grefenstette, G. 2011.
Social media driven image retrieval.
In *ACM International Conference on Multimedia Retrieval (ICMR)*, pages 33:1–33:8.

Sigurbjörnsson, B. and van Zwol, R. 2008.
Flickr tag recommendation based on collective knowledge.
In *WWW '08*, pages 327–336, New York, NY, USA. ACM.

Wang, G., Hoiem, D., and Forsyth, D. A. 2009.
Building text features for object image classification.
In *CVPR*.

Thank you for your Attention.