Social Media Annotation

Lamberto Ballan, Marco Bertini, Tiberio Uricchio, Alberto Del Bimbo
Media Integration and Communication Center (MICC)
Università degli Studi di Firenze, Italy
{lamberto.ballan|marco.bertini|tiberio.uricchio|alberto.delbimbo}@unifi.it

***Abstract*—The large success of online social platforms for creation, sharing and tagging of user-generated media has lead to a strong interest by the multimedia and computer vision communities in research on methods and techniques for annotating and searching social media. Visual content similarity, geo-tags and tag co-occurrence, together with social connections and comments, can be exploited to perform tag suggestion as well as to perform content classification and clustering and enable more effective semantic indexing and retrieval of visual data. However there is need to countervail the relatively low quality of these metadata user produced tags and annotations are known to be ambiguous, imprecise and/or incomplete, overly personalized and limited - and at the same time take into account the 'web-scale' quantity of media and the fact that social network users continuously add new images and create new terms. We will review the state of the art approaches to automatic annotation and tag refinement for social images and discuss extensions to tag suggestion and localization in web video sequences.**

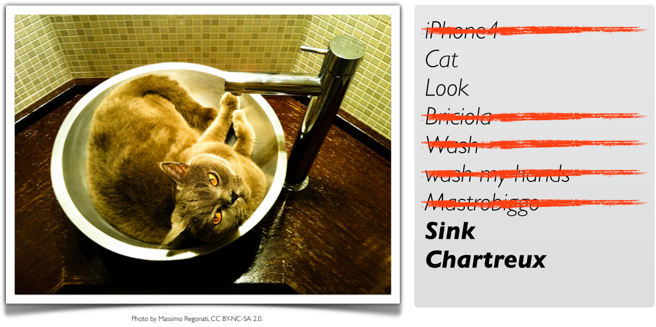


Photo by Massimo Regonati, CC BY-NC-SA 2.0.

Fig. 1. Example of tag refinement: some tags are not relevant with respect to image content (strike-through), some tags describing content should be added (bold).

I. Introduction

The success of online social platforms that let users share, rate, comment and tag media motivates social image analysis, annotation and retrieval as important research topics for the multimedia community. In fact, the availability of huge quantities of user-generated information, including media, social connections, multimodal content and descriptions, location and comments in various forms (ranking, votes, likes) and associated metadata are considered valuable resources for improving the results of tasks such as semantic indexing and retrieval. However, this wealth of media content and metadata poses several challenges: *i)* the relatively low quality of these metadata – i.e. tags and annotations are known to be ambiguous, overly personalized, and limited (typically an image is associated with only one-three tags) [1], [2]; *ii)* the 'web-scale' quantity of media; *iii)* in a social network, users continuously add images and create new terms given the freedom of tagging. So folksonomies and changing ontologies are a challenging issue to extract valuable information; *iv)* tags may be unrelated to visual content: among the most common Flickr tags analyzed in [2] there are "2006", "2005" and "2004".

To provide a more formal description of the problem, let us consider a corpus Φ composed of images and metadata, an image $i \in I$, with tags $t_j \in V_T$; we can then define the main research paths that have been addressed as:

image auto-annotation assign tags to an image that has not been tagged before;

tag (re-)ranking assign the right order or weight to each tag associated to an image, i.e. determine r so that: $r(i,t) : (I, V_T) \rightarrow \mathbb{R}$, where $r(i,t_1) > r(i,t_2)$ if t_1 is relevant for i, while t_2 is not and $r(i_1,t) > r(i_2,t)$ if the tag is relevant for the first image and not for the second - considering users $u \in U$, personalized ranking becomes: $r(u,i,t) : (U, I, V_T) \rightarrow \mathbb{R}$;

tag suggestion suggest new tags that are appropriate to the image content. Existing tags, are assumed as appropriate. Considering that the tags $T_i = t_1, \dots, t_k \in V_T$ are relevant for i and $tag(i,t) \forall t \in T_i$, the problem becomes to determine: suggestion$_M(i, T_i) : (I, \mathcal{P}(V_T)) \rightarrow \mathcal{P}(V_L) = \{l_1, l_2, \dots, l_M\}$, where $\mathcal{P}$ is the power set operator and $V_T \subseteq V_L$;

tag refinement refine existing tags by dropping out inappropriate tags and adding new / missing tags: refine$_M(i, T_i) : (I, \mathcal{P}(V_T)) \rightarrow \mathcal{P}(V_L) = \{l_1, l_2, \dots, l_M\}$. Fig. 1 shows an example of tag refinement.

We can consider tag refinement as the most general problem and the others as specializations. The vast majority of recent tag refinement methods have addressed images, while very few worked with videos.

Tag suggestion and localization (see Fig. 2) in internet videos is the task of associating tags to specific shots. In general terms, this problem can be viewed as image tag refinement applied to keyframes, where each keyframe is annotated with all the tags associated to the video.

In the following we will focus on tag refinement for social media, review the state of the art methods for image and video refinement, and present performance figures of the nearest neighbor methods for image and video tag refinement on several datasets. Finally we discuss adaptation of the data-driven approach for tag localization in video shots.

The paper is organized as follows: related works are discussed in Sect. II; a description of data-driven tag-refinement methods is provided in Sect. III; a description of the datasets

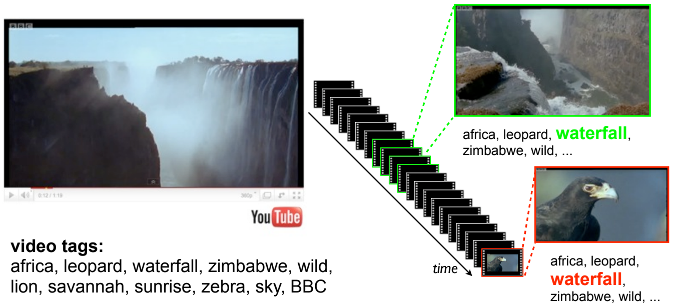

You Tube
video tags:
africa, leopard, waterfall, zimbabwe, wild,
lion, savannah, sunrise, zebra, sky, BBC
time
africa, leopard, waterfall,
zimbabwe, wild, ...
africa, leopard,
waterfall,
zimbabwe, wild, ...

Fig. 2. *Left*: Example of a YouTube video with its related tags. *Right*: localization of tags in shots.

used in the experiments is reported in Sect. IV, while experimental results are discussed in Sect. V. Finally conclusions are drawn in Sect. VI.

II. Principal Contributions and State of the Art

Methods for image tag refinement can be distinguished into two broad categories, respectively following statistical modeling or data-driven approaches [3]. Generally, the statistical modeling techniques [4]–[6] achieve the highest performance, but have the drawback that learning must be applied periodically as new images or terms are added, which is someway impractical in large-scale continuously evolving collections. Data-driven approaches have shown to have easier application. They have been successfully used for tag ranking in social image retrieval, tag suggestion for automatic image annotation [7]–[9] and tag suggestion and localization in web videos [10], [11].

A. Image Data

Figure 3 shows a taxonomy of the most important and recent contributions on social media annotation. Several works explicitly addressed the subject of tag ranking in the framework of social image retrieval.

The seminal paper by [12] was the first to suggest tag refinement as a processing step to improve the imprecise original annotations of social media. He proposed to perform belief propagation among tags using the Random Walk with Restart framework (RWR). This method was adopted by D. Liu *et al.* [13] for tag ranking and image retrieval (RWTR). In their approach, a random walk-based tag refinement step was applied after an initial probabilistic tag relevance estimation based on kernel density estimation.

Kennedy *et al.* in [14] addressed the problem of filtering out unreliable tags in social images. They demonstrated that tags used by different persons to annotate visually similar images are more related to visual content than the others. In their approach the collected the 20 nearest neighbors of each image. Scalability was addressed using a (learned) low-dimensional image feature, and the Map/Reduce framework to speed up search.

Li *et al.* [8] proposed a tag relevance measure for image retrieval in social networks based on the Kennedy's observations. They assumed that the more frequently a tag occurs in the neighbor set, the more relevant it is likely to be for the description of the image content. However, frequently occurring tags are unlikely to be relevant to the majority of images. Due to this, their tag relevance measure takes into account both the distribution of a tag in the neighbor set of an image and its distribution in the entire collection. The original method was extended in [15] where the outputs of several tag relevance measures based on different visual features were used to compute image similarity.

Similar ideas have inspired other authors for the task of image auto-annotation. Makadia *et al.* in [7] proposed a baseline method for image auto-annotation by to transfer tags to an image from its visual neighborhood. Similar images were ordered according to their similarity to the test image and the most frequent tags were assigned starting from the most similar image, until a specified number of them has been reached. The method is comprised of a composite image distance measure (JEC - Joint Equal Contribution - or Lasso) for nearest neighbor ranking. Guillaumin *et al.* [9], [16] proposed to learn a weighted nearest neighbor model, that provides a tag relevance measure to support image auto-annotation. The model finds the optimal combination of feature distances (e.g. local shape descriptors or global color histograms) automatically. Tag relevance was calculated on the basis of a neighbor rank or distance.

Consistency between visual and semantic similarity in social images was assumed by D. Liu *et al.* in [4] to formulate tag refinement as an optimization task. Their method was referred as tag refinement based on visual and semantic consistency (TRVSC). They used constrained non-negative matrix factorization (CNMF) by Y. Liu *et al.* [17], maximizing consistency while minimizing the deviation from the the tags initially provided by users. Since consistency is mainly referred to content-related tags (see Fig. 1), a filtering procedure based on Wordnet was used to constrain the tagging vocabulary to content-related tags only. Tag enrichment was done by considering tag synonyms and hypernyms.

Tsai *et al.* [18] proposed visual synset as a structure of visually-similar and semantically-related images. Each visual synset corresponds to a single prototypical visual concept with a set of weighted tags associated to it. Based on visual synsets, linear SVMs were then used to predict annotations to unseen images.

D. Liu *et al.* [19] proposed an expansion to the single graph multi label learning algorithms by learning a tag-specific visual vocabulary. Every annotation gets a correlation graph which is used to propagate the information by reflecting the particular relationship among images with respect to the specific tag.

Zhu *et al.* in [5] proposed a tag refinement approach which is referred as low-rank and error sparsity approximation (LRES). That method is based on several assumptions: visually similar images are similarly tagged; tags are often correlated and interact at the semantic level; the semantic space spanned by all the tags can be approximated by a smaller subset of them; user tags are sufficiently accurate so that the image tag matrix has error sparsity condition. Following these assumptions tag refinement was cast into the problem of decomposing the user-provided tag matrix into a low-rank refined matrix and a sparse error matrix. A convergence provable iterative procedure was proposed to perform the optimization.

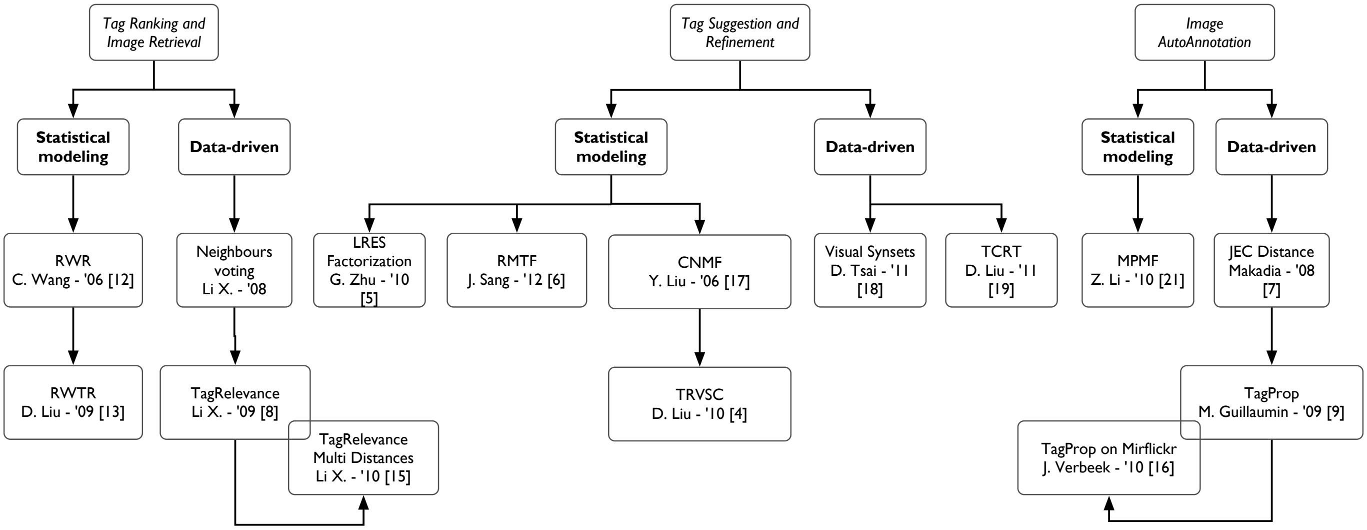

Tag Ranking and Image Retrieval
Statistical modeling
Data-driven
RWR C. Wang - '06 [12]
Neighbours voting Li X. - '08
RWTR D. Liu - '09 [13]
TagRelevance Li X. - '09 [8]
TagRelevance Multi Distances Li X. - '10 [15]
Tag Suggestion and Refinement
Statistical modeling
Data-driven
LRES Factorization G. Zhu - '10 [5]
RMTF J. Sang - '12 [6]
CNMF Y. Liu - '06 [17]
TRVSC D. Liu - '10 [4]
Visual Synsets D. Tsai - '11 [18]
TCRT D. Liu - '11 [19]
Image AutoAnnotation
Statistical modeling
Data-driven
MPMF Z. Li - '10 [21]
JEC Distance Makadia - '08 [7]
TagProp M. Guillaumin - '09 [9]
TagProp on Mirflickr J. Verbeek - '10 [16]

Fig. 3. Taxonomy of the most important works on social media annotation.

A probabilistic approach, based on typical probabilistic matrix factorization (PMF) [20], was proposed by Z. Li *et al.* in [21]. They extended the original formulation by fusing different sources of correlation, namely image-tag correlation, image similarity and tag correlation. Two sets of low dimensional latent factors were derived and used to predict new annotations by reconstructing the image-tag correlations estimated.

More recently, Sang *et al.* [6] proposed to jointly model the ternary relations between users, tags and images employing tensor factorization and using Tucker decomposition for the latent factor inference (RMTF). Since the traditional factorization models used in recommendation and collaborative filtering systems cannot fully account for missing and noisy tags, tag suggestion was cast into a ranking problem to determine which tag is more relevant for a user to describe a given image. To this end they introduced and managed a ternary semantic for tags, that can be positive (those assigned by the users), negative (tags that are dissimilar and that rarely occur together with positive tags) and neutral (all the other tags).

B. Video Data

Many recent works on internet videos have focused on near duplicate detection [22], training concept detectors [23] or topic detection [24] that are all relevant to tag suggestion. Ulges *et al.* [23] have exploited YouTube videos in order to train concept detectors with no manual annotation for the creation of ground truth data. In this way they used video tags as a lexicon and could scale on the number of concepts detected. Training detectors with ground truth prepared by experts in conjunction with social videos could improve their performance.

Ballan *et al.* [10] suggested a method to automatically annotate shots of YouTube videos using Flickr images. They observed that computation of tag relevance on the tags of the whole video would not consider the fact that some tags only refer to specific shots, and would result into re-ranking the same list of tags for all the shots. According to this they employed a variation of the tag relevance algorithm of [8] and used visual similarity of keyframes and images to add new tags that were not originally present in the video. This model was extended in [25] to compute weighted $tagRelevance$ based on visual similarity, and performing an initial video tag expansion using Wikipedia.

Recently H. Li *et al.* [11] presented a dataset of 1550 YouTube videos with ground-truth annotation and localization of 31 concepts. They performed tag localization using a multiple instance learning baseline method, based on the MIL-BPNET approach of [26].

Localization of video tags was addressed also by G. Li *et al.* in [27]. They used a multiple instance learning approach that considers semantic relatedness of co-occurring tags. Temporal smoothness was assumed to model shots and video.

Min *et al.* [28] applied 34 concept detectors to video keyframes to build a semantic representation for video shots. The same detectors were applied to Flickr images and the semantic similarity between the concepts detected in the video shots and the Flickr images was used to suggest tags from Flickr images.

Chu *et al.* [29] used Flickr images and their tags for tag localization. They modeled the relationship between keyframes in a video shot and candidate tags as a bipartite graph. In the graph keyframes and tags were two disjoint sets of nodes, and each edge between nodes was associated with a weight calculated based on the similarity between a pair of keyframe and tag, and tagging behaviors. Best matching was used to determine the most appropriate tags to associate to video keyframes.

III. Data-driven Tag Refinement

Data-driven Tag Refinement grounds 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 to this end. This method has been successfully applied to different tasks such as image auto-annotation and tag ranking/relevance.

In the following we discuss in detail the most relevant data driven approaches. We indicate with I a test image and a set of K visually similar images $N_k(I,K) = \{I_1, I_2, \ldots, I_K\}$, ordered according to their increasing distance (where I_1 is the nearest image and I_K is the farthest). In Sect. V we will provide a comparative evaluation of these methods.

A. Simple Label Transfer: Makadia et al. [7]

Considering $N_k(I,K)$, the label transfer procedure is:

1) Rank the tags of I_1 according to their frequency in the training set. We denote this set as S_1.
2) Transfer the highest n ranking tags of I_1. If I_1 has at least n tags, the algorithm terminates.
3) Rank the tags of neighbors I_2 through I_K (excluding $|S_1|$) according to the co-occurrence in the training set with the tags transferred in step 2 (S_1) and according to the local frequency.
4) Transfer the highest n - $|S_1|$ ranking tags from step 3.

The method was originally tested on Corel5K, IAPR TC-12 and ESP datasets. In our implementation of the method the distance between images is computed as:

$$d(I_i, I_k) = \frac{e^{||\mathbf{f}_i - \mathbf{f}_k||}}{\sigma^2} \tag{1}$$

where I_i is the visual neighbor in the i position, with N features $\mathbf{f}_i = (f_i^1, \ldots, f_i^N)$, and σ^2 is set as the median value of all the distances.

B. Learning Tag Relevance from Visual Neighbors: Li et al. [8]

Tag relevance measure of a tag t for an image I considering its the neighbor set K is:

$$tagRelevance(t, I, K) := n_t[N_k(I,K)] - Prior(t, K) \tag{2}$$

where n_t is an operator counting the occurrences of t in the neighborhood $N_k(I,K)$ of K similar images, and $Prior(t,K)$ is the occurrence frequency of t in the entire collection. In order to reduce user bias, only one image per different user is considered when computing the visual neighborhood. The method was been originally applied to image retrieval on a subset of the Flickr dataset with 20,000 manually checked images and to image auto-annotation using a 331 images subset.

C. TagProp, Discriminative Metric Learning in Nearest Neighbor Models: Guillaumin et al. [9]

The probability of a tag for being relevant given a neighborhood of K images $N_k(I,K) = \{I_1, I_2, \ldots, I_K\}$ is:

$$p(y_{It} = +1) = \sum_{N_k(I,K)} \pi_{II_i} p(y_{It} = +1 | N_k(I,K)) \tag{3}$$

$$p(y_{It} = +1 | N_k(I,K)) = \begin{cases} 1 - \epsilon & \text{for } y_{It} = +1, \\ \epsilon & \text{otherwise} \end{cases} \tag{4}$$

where:

- $y_{It} \in \{-1, +1\}$ indicates whether tag t is relevant or not for the test image I;
- π_{II_i} is the weight of a training image I_i of the neighborhood $N_k(I,K)$;
- $p(y_{It} = +1 | N_k(I,K))$ is the prediction of tag t.

The objective is to maximize $\sum_{I,t} \ln p(y_{It})$.

This model can be used with rank-based or distance-based weighting. To compensate the frequencies of tags, a tag-specific sigmoid is used to scale the predictions, boosting the probability for rare tags and lowering the probability of the frequent tags. Image tags have been used for model learning. The method has been initially experimented on Corel5K, IAPR TC-12 and ESP datasets. More recently it has also been tested on MIRFlickr-25K [16], using two sets of manually annotated concepts with different degrees of relevance, and a train/test split of the dataset that is different from the one proposed by the creators of the dataset.

D. Video Tag Suggestion & Localization: Ballan et al. [10]

Video tags $T_v = \{t_1, \ldots, t_l\} \in V_V$ were used as queries to retrieve images from Flickr. Visual neighborhoods $N_k(I,K)$ of keyframes are created from the Flickr images retrieved. The set of tags V_T of the images in the neighborhood is associated with the keyframe. Then, tag relevance of these tags is computed as in Sect. III-B. Due to the fact that public social datasets for video tag refinement are not available, we created a new dataset from the YouTube service, and used this dataset for the experiments.

IV. Datasets

A. Image Datasets

Effectiveness of nearest neighbor methods for image tag refinement in large-scale scenarios, was verified on two large image datasets with ground-truth annotations extracted from Flickr and publicly available: MIRFlickr-25K (18 tags) [30] and NUS-WIDE-240K (81 tags). The MIRFlickr-25K dataset contains 25,000 images with 1,386 tags. The NUS-WIDE-240K is a subset of the NUS-WIDE-270K dataset [31]. It contains only 238,251 of the 269,648 images (provided as URLs) originally present that are still present in Flickr. This dataset was created in order to make it possible to implement the method of [8] that required to download again the original data from Flickr for the NUS-WIDE-270K dataset to obtain user information.

Since the tags in the above two image collections are noisy and in large part meaningless, a pre-processing step was performed to filter out non useful tags. To this end, similarly to [31] only the tags with a corresponding item in Wordnet were retained. Moreover, we removed the less frequent tags, with occurrence below 50. As a result 219 and 684 unique tags were respectively obtained for MIRFlickr-25K and NUS-WIDE-240K.

For both the MIRFlickr-25K and NUS-WIDE-240K datasets, the visual similarity between images was calculated using simple visual descriptors. Starting from the features given in the NUS-WIDE dataset, as in [5], for each image we extracted a single 428-dimensional descriptor. This feature was obtained as the early-fusion of a 225-d block-wise color moment features generated from 5-by-5 fixed partition on image, a 128-d wavelet texture features, and a 75-d edge distribution histogram features.

B. Video Datasets

The dataset[1] is composed by four randomly selected YouTube videos for each of the 15 categories (*Auto & Vehicles, Comedy, Education, Entertainment, Film and Animation, Gaming, Howto & Style, Music, News and Politics, Nonprofits & Activism, Pets & Animals, Science & Technology, Sports, Travel & Events*). The total duration of videos is three hours and eight minutes and the number of detected shots is 4196. The number of tags per video varies from 8 to 22. Video tags were filtered to eliminate stopwords, dates and numbers. To select Flickr images the set of video tags is expanded considering their co-occurrrence of the related YouTube videos and the anchors of Wikipedia articles titled as these tags.

To compute visual similarity between keyframes K and Flickr images I we use a 370-dimensional feature vectors that includes local and global features. This feature vector is composed by a 50 dimensional color correlogram computed in the HSV color space, a 80 dimensional vector for the MPEG-7 Edge Histogram Descriptor and a 240 dimension vector for the TOP-SIFT descriptor. This latter descriptor is a variation of TOP-SURF [32], a compact image descriptor that combines interest points with visual words, designed for fast content-based image retrieval.

Flickr images were clustered using k-means, and the cluster centers were used as indexes to fasten the approximate nearest neighbor search. For each video keyframe the nearest cluster center is retrieved based on visual similarity. Images of the cluster are considered as neighbors of the keyframe.

V. Experiments and Discussion

A. Tag Refinement Evaluation Framework

Following [4]–[6], [12], [19], to evaluate tag refinement for each tag we report the F-measure figures, computed as *macro-average* F-scores by averaging the F-score over the number of ground-truth annotations. However, in order to provide meaningful scores, we must consider the fact that the average number of images per label is different in the two datasets: on average, each image of the MIRFlickr-25K dataset contains 1.3 tags, while in the NUS-WIDE-240K dataset there are 4 tags per image. To account for this fact, and avoid that algorithms like Makadia *et al.* [7] always predict the most common tags, we also compute the the *micro-average* F-scores by averaging F-score over all the images. As [5], [19], we only keep $m = 5$ tags per image. We also report figures by varying m between 1 and 10.

B. Evaluation of Tag Refinement on MIRFlickr-25K

On this dataset we compared the following methods:

- Baseline, the original tags provided by the users (UT);
- Simple Label Transfer (SLT) [7], described in Sect. III-A; as shown in Fig. 4 the best results are obtained using $K = 500$ neighbors;
- Learning Tag Relevance from Visual Neighbors (TR) [8], described in Sect. III-B; again, see Fig. 4, the best results are obtained using $K = 500$ visual neighbors;
- TagProp, Discriminative Metric Learning in Nearest Neighbor Models (TP) [9], described in Sect. III-C; the best results are obtained by defining the weights of the model directly as a function of the distance.

	UT	SLT [7]	TR [8]
F-score *macro*	0.18	0.26	0.27
F-score *micro*	0.06	0.14	0.13

TABLE I. Average performances of different algorithms for tag refinement on MIRFlickr-25K (full dataset).

	UT	SLT [7]	TR [8]	TP [9]
F-score *macro*	0.18	0.20	0.19	0.20
F-score *micro*	0.06	0.11	0.11	0.11

TABLE II. Average performances of different algorithms for tag refinement on MIRFlickr-25K (test set).

Two sets of experiments were performed. The first was applied to the entire dataset of 25,000 images (results shown in Table I). The second was conducted using 15,000 images as training set and 10,000 images as test set (results reported in Table II as averages of 10 random train/test splits).

It is possible to notice that the algorithm by Li *et al.* [8] has superior performance with respect to the Simple Label Transfer algorithm by Makadia *et al.* [7] (e.g. 0.27 vs 0.26 on the MIRFlickr-25K full dataset. TagProp shows very similar results (e.g. 0.20 vs 0.19, as reported in Table II) on the smaller dataset but requires more computational effort and a learning phase is necessary. It must be observed that with the second dataset performance drops of about 5% due to the smaller number of visual neighbors available for the tag propagation.

Table III reports the F-measure of the other methods as appeared in the literature. A comparison of these results demonstrates that, despite their simplicity and low computational cost, nearest-neighbor methods have comparable performance for tag refinement with respect to more complex state-of-the-art approaches, Complex and computationally intensive algorithms such as TRVSC [4] and LRES [5] only provide performance improvement of about 2 percent, and require re-training if the dataset has substantial variations. The recent results by Liu *et al.* [19] with different visual features (i.e. 500-d BoW of SIFT descriptors), confirm this fact.

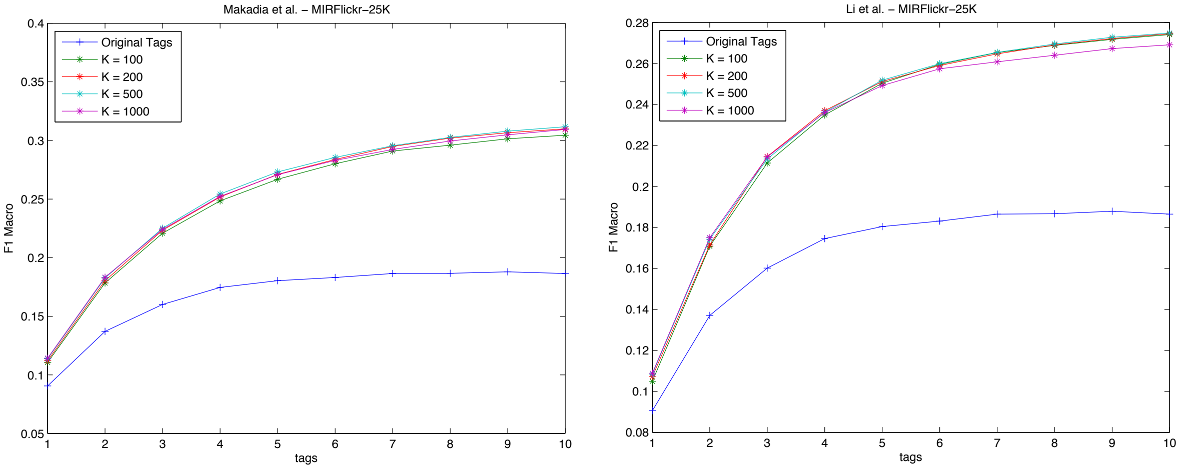
	UT	RWTR [12]	TRVSC [4]	LRES [5]
Zhu *et al.* [5]	0.22	0.34	0.41	0.42
Liu *et al.* [19]	0.2	0.31	0.37	-

TABLE III. F-score performances of other algorithms for tag refinement on MIRFlickr-25K, as reported in the literature.

C. Evaluation of Tag Refinement on NUS-WIDE-240K

We performed similar experiments on the NUS-WIDE-240K dataset. The first experiment was applied to the entire

[1] Available on request at: http://www.micc.unifi.it/ballan/research/tag-webvideos/



(a) (b)

Fig. 4. F-score results (y axis) on the MIRFlickr-25K dataset with (a) the Simple Label Transfer algorithm [7], (b) the Tag Relevance Learning algorithm [8]. These results are obtained by varying the number of visual neighbors (K) and the number *m* of retained tags per image (x axis).

dataset of 238,251 images (results shown in Table IV). The second was conducted using 158,834 images as training set and the remaining 79,417 as test set (results reported in Table V as averages of 10 random train/test splits). The variation of performance due to the number of visual neighbors and retained tags per image is similar to that reported in Fig. 4 for MIRFlickr-25K.

The experiments on the NUS-WIDE-240K dataset confirm that the algorithm of Li *et al.* [8] provides the best results in terms of F-score macro and micro-average figures. Comparison with previous works is complicated by the fact that most of them used a subset of the full dataset and undocumented/non-standard experimental procedures. In particular, Zhu *et al.* [5] reported figures lower than those verified by us and lower with respect to those of the other works in the literature. They performed a pre-processing on the tags vocabulary and reduced the number of unique tags to 521 tags (instead of 684 tags of our tests). Their baseline UT is 0.269 instead of 0.35 of our case. All in all they cannot be compared with our results.

Liu *et al.* [19] (UT=0.45) and Sang *et al.* [6] (UT=0.477) used subsets of the NUS-WIDE-270K dataset due to the fact that their methods could not be applied to such huge dataset: Liu *et al.* used a subset of 24,300 images; Sang *et al.* used a subset of 124,099 images. Sang *et al.* used the same features of ours but reported figures only with $m = 10$ tags per image. On their dataset, they measured 0.475 F-measure with the RWR [12] method, 0.49 F-measure with TRVSC [4], 0.523 F-measure with LR [5], and 0.571 F-measure with the best implementation of their method.

	UT	SLT [7]	TR [8]
F-score *macro*	0.35	0.37	0.44
F-score *micro*	0.11	0.18	0.23

TABLE IV. AVERAGE PERFORMANCES OF DIFFERENT ALGORITHMS FOR TAG REFINEMENT ON NUS-WIDE-240K (FULL DATASET).

	UT	SLT [7]	TR [8]	TP [9]
F-score *macro*	0.35	0.36	0.45	0.44
F-score *micro*	0.11	0.18	0.22	0.21

TABLE V. AVERAGE PERFORMANCES OF DIFFERENT ALGORITHMS FOR TAG REFINEMENT ON NUS-WIDE-240K (TEST SET).

These results indicate that also in the case of a large-scale dataset such as NUS-WIDE-240K, nearest-neighbor based methods have competitive performance. An important point is that while matrix factorization and graph-based methods suffer in a large-scale scenario, nearest-neighbor based methods appear to be able to perform tag refinement with reasonable performance.

D. Evaluation of Video Tag Localization

We have measured tag refinement performance in terms of accuracy, defined as the ratio between the number of tags correctly suggested and the total number of suggested tags. For each tag obtained from tag filtering and expansion the system downloads the first 15 Flickr images ranked according the "relevance" criterion given by the Flickr API.

Table VI reports, for different relevance threshold scores, the accuracy and the mean number of correctly suggested tags for shot. We can observe that the average accuracy on the entire dataset increases until the score equals to seven and slightly decreases for higher scores, remaining close to 0.9. The average number of correct tags suggested decreases for high scores (for threshold above 5). It is also evident that some video categories are more tractable than the others. In the "Auto & Vehicle" and "Travel & Events" categories, the extracted Flickr images are very relevant and similar to the shots analysed. This can be seen from the number of suggested tags which is quite large. In "Film & Animation" it is difficult to retrieve Flickr images similar to trailer scenes of feature films. "Howto & Style" collects very diverse content that is hard to be correctly annotated.

VI. CONCLUSION

We reviewed the state of the art approaches to automatic annotation and tag refinement for social images and discussed some extensions to tag suggestion and localization in web videos. In particular we analysed nearest neighbor methods

YouTube category	$\tau_{relevance}$=1		$\tau_{relevance}$=3		$\tau_{relevance}$=5		$\tau_{relevance}$=7		$\tau_{relevance}$=11	
	Acc.	Tags	Acc.	Tags	Acc.	Tags	Acc.	Tags	Acc.	Tags
Auto & Vehicles	0.41	10.99	0.65	4.09	0.78	2.13	0.86	1.36	0.93	0.66
Comedy	0.58	5.49	0.85	2.68	0.95	1.68	0.92	0.89	0.77	0.16
Education	0.49	3.97	0.62	1.83	0.76	0.84	0.72	0.39	0.69	0.11
Entertainment	0.60	4.46	0.84	2.98	0.99	1.94	1	0.89	1	0.03
Film & Animation	0.54	2.16	0.93	1.28	0.99	0.59	1	0.19	1	0.01
Gaming	0.47	3.85	0.85	2.13	0.93	0.97	0.99	0.60	1	0.2
Howto & Style	0.39	3.91	0.61	2.02	0.69	1.04	0.71	0.45	0,71	0.31
Music	0.39	2.48	0.69	0.48	1	0.10	1	0.012	1	0.06
News & Politics	0.62	5.32	0.87	2.40	0.97	1.04	1	0.46	1	0.04
No-profit & Activism	0.61	2.62	0.93	1	0.98	0.42	1	0.17	1	0.04
People & Blogs	0.40	5.70	0.67	2.74	0.79	1.22	0.82	0.58	0.50	0.15
Pets & Animals	0.56	4.83	0.75	2.28	0.86	1.04	0.85	0.55	0.94	0.23
Science & Technology	0.44	4.80	0.64	1.67	0.81	0.84	0.89	0.44	0.87	0.16
Sport	0.41	4.49	0.74	2.63	0.82	1.39	0.92	0.62	0.94	0.14
Travel & Events	0.61	12.57	0.79	7.34	0.87	4.21	0.91	2.45	0.98	1.18
Average	0.50	5.18	0.76	2.50	0.88	1.30	0.91	0.67	0.90	0.23

TABLE VI. RESULTS FOR TAG LOCALIZATION AND SUGGESTION FOR EACH YOUTUBE CATEGORY, IN TERMS OF ACCURACY AND AVERAGE NUMBER OF CORRECTLY ADDED TAGS, AS $\tau_{relevance}$ VARIES.

since they have shown good recognition performance, and they are also suitable for large-scale recognition problems.

REFERENCES

[1] L. S. Kennedy, S.-F. Chang, and I. V. Kozintsev, "To search or to label? Predicting the performance of search-based automatic image classifiers," in *Proc. of ACM MIR*, 2006.

[2] B. Sigurbjörnsson and R. van Zwol, "Flickr tag recommendation based on collective knowledge," in *Proc. of WWW*, 2008.

[3] D. Liu, X.-S. Hua, and H.-J. Zhang, "Content-based tag processing for internet social images," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 723–738, 2011.

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

[5] 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.

[6] J. Sang, C. Xu, and J. Liu, "User-aware image tag refinement via ternary semantic analysis," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 883–895, 2012.

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

[8] X. Li, C. G. M. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1310–1322, 2009.

[9] 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.

[10] L. Ballan, M. Bertini, A. Del Bimbo, M. Meoni, and G. Serra, "Tag suggestion and localization in user-generated videos based on social knowledge," in *Proc. of ACM SIGMM Workshop on Social Media (WSM)*, Firenze, Italy, 2010.

[11] H. Li, L. Yi, Y. Guan, and H. Zhang, "Dut-webv: A benchmark dataset for performance evaluation of tag localization for web video," in *Proc. of MMM*, 2013.

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

[13] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang, "Tag ranking," in *Proc. of WWW*, 2009.

[14] L. S. Kennedy, M. Slaney, and K. Weinberger, "Reliable tags using image similarity: mining specificity and expertise from large-scale multimedia databases," in *Proc. of ACM-MM Workshop on Web-Scale Multimedia Corpus*, Beijing, China, 2009.

[15] X. Li, C. G. M. Snoek, and M. Worring, "Unsupervised multi-feature tag relevance learning for social image retrieval," in *Proc. of ACM CIVR*, 2010.

[16] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid, "Image annotation with tagprop on the mirflickr set," in *Proc. of ACM MIR*, 2010.

[17] Y. Liu, R. Jin, and L. Yang, "Semi-supervised multi-label learning by constrained non-negative matrix factorization," in *AAAI-06: Proceedings of the Ninth National Conference on Artificial Intelligence*, vol. 21. AAAI Press, 2006, p. 421.

[18] D. Tsai, Y. Jing, Y. Liu, H. A. Rowley, S. Ioffe, and J. M. Rehg, "Large-scale image annotation using visual synset," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 611–618.

[19] D. Liu, S. Yan, X.-S. Hua, and H.-J. Zhang, "Image retagging using collaborative tag propagation," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 702–712, 2011.

[20] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," *Advances in neural information processing systems*, vol. 20, pp. 1257–1264, 2008.

[21] Z. Li, J. Liu, X. Zhu, T. Liu, and H. Lu, "Image annotation using multi-correlation probabilistic matrix factorization," in *Proceedings of the international conference on Multimedia*, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 1187–1190.

[22] S. Paisitkriangkrai, T. Mei, J. Zhang, and X.-S. Hua, "Scalable clip-based near-duplicate video detection with ordinal measure," in *Proc. of ACM CIVR*, 2010.

[23] A. Ulges, C. Schulze, M. Koch, and T. M. Breuel, "Learning automatic concept detectors from online video," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 429–438, 2010.

[24] J. Shao, W. Yin, S. Ma, and Y. Zhuang, "Topic discovery of web video using star-structured k-partite graph," in *Proc. of ACM Multimedia*, 2010.

[25] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra, "Enriching and localizing semantic tags in internet videos," in *Proc. of ACM International Conference on Multimedia (ACM MM)*, 2011, pp. 1541–1544.

[26] M.-L. Zhang and Z.-H. Zhou, "Improve multi-instance neural networks through feature selection," *Neural Processing Letters*, vol. 19, no. 1, pp. 1–10, 2004.

[27] G. Li, M. Wang, Y.-T. Zheng, and T.-S. Chua, "ShotTagger: Tag location for internet videos," in *Proc. of ACM ICMR*, 2011.

[28] H.-S. Min, J. Choi, W. De Neve, Y. M. Ro, and K. N. Plataniotis, "Semantic annotation of personal video content using an image folksonomy," in *Proc. of IEEE ICIP*, 2009.

[29] W.-T. Chu, C.-J. Li, and Y.-K. Chou, "Tag suggestion and localization for web videos by bipartite graph matching," in *Proc. of ACM SIGMM Workshop on Social Media (WSM)*, New York, NY, USA, 2011.

[30] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proc. of ACM MIR*, 2008.

[31] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. of ACM CIVR*, 2009.

[32] B. Thomee, E. M. Bakker, and M. S. Lew, "TOP-SURF: a visual words toolkit," in *Proc. of ACM Multimedia*, 2010.