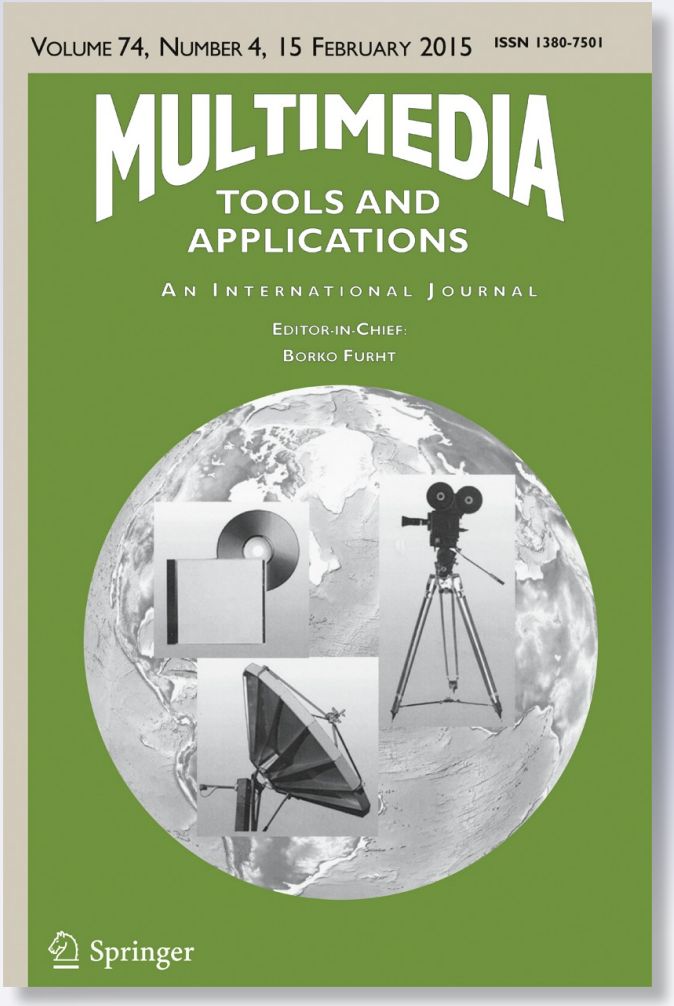
Data-driven approaches for social image and video tagging

Lamberto Ballan, Marco Bertini, Tiberio Uricchio & Alberto Del Bimbo

Multimedia Tools and Applications
An International Journal

ISSN 1380-7501
Volume 74
Number 4

Multimed Tools Appl (2015)
74:1443-1468
DOI 10.1007/s11042-014-1976-4



Springer

Multimed Tools Appl (2015) 74:1443–1468
DOI 10.1007/s11042-014-1976-4

Data-driven approaches for social image and video tagging

Lamberto Ballan · Marco Bertini · Tiberio Uricchio · Alberto Del Bimbo

Published online: 16 April 2014

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 c lustering and enable more effective semantic indexing and retrieval of visual data. However there is need to overcome the relatively low quality of these metadata: user produced tags and annotations are known to be ambiguous, imprecise and/or incomplete, excessively 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, considering also the temporal patterns of their usage, and discuss extensions to tag suggestion and localization in web video sequences.

Keywords Social media · Image tagging · Video tagging · Temporal analysis

L. Ballan · M. Bertini (✉) · T. Uricchio · A. Del Bimbo
Media Integration and Communication Center (MICC), Università degli Studi di Firenze, Firenze, Italy
e-mail: marco.bertini@gmail.com

L. Ballan
e-mail: lamberto.ballan@unifi.it

T. Uricchio
e-mail: tiberio.uricchio@unifi.it

A. Del Bimbo
e-mail: alberto.delbimbo@unifi.it

1 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) [13, 35]; *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 [35] 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 $t_j \in V_T$, where V_T is a vocabulary of tags; we can then define the main research problems that have been investigated as:

image auto-annotation:	assign tags to an image that has not been tagged;
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 $T_i = t_1, \ldots, t_k \in V_T$ are relevant for i and $tag(i,t) \forall t \in T_i$, the problem becomes to determine: $\text{suggestion}_M(i, T_i) : (I, \mathcal{P}(V_T)) \rightarrow \mathcal{P}(V_L) = \{l_1, l_2, \ldots, l_M\}$, 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: $\text{refine}_M(i, T_i) : (I, \mathcal{P}(V_T)) \rightarrow \mathcal{P}(V_L) = \{l_1, l_2, \ldots, l_M\}$. Figure 1 shows an example of tag refinement.
tag suggestion and localization:	in internet videos associates tags to specific shots. This problem can be viewed as *tag refinement* applied to keyframes, where each keyframe is associated to all the tags of the video. Figure 2 shows an example of tag suggestion and localization in videos.

Figure 3 shows a taxonomy of the most important works addressing these problems. The methods proposed can be divided in those based on statistical modeling techniques and data-driven approaches [25]. Given these definitions of the problems we can consider that tag refinement is the most general problem, while the others can be considered as

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*)

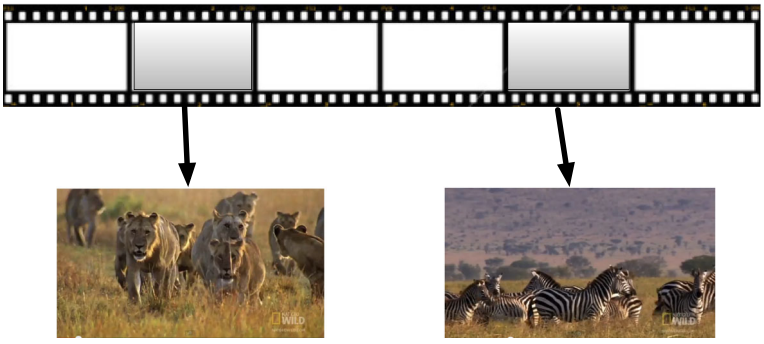
specializations, hence the rest of the paper will focus on tag refinement for images and tag suggestion and localization for videos. Considering this problem, the current state-of-the-art methods [24, 33, 43] – often based on matrix factorization approaches – require costly training procedures, that have to be redone periodically if a new set of images or terms are added to the system, thus making the approach impractical for large-scale processing or in social networks undergoing continuous evolution of image collections and tags. Recently, data-driven approaches have shown to be able to deal with these latter issues, and have been applied to tag ranking for social image retrieval, tag suggestion for social image annotation (considering also the case in which no tag is associated to an image) [9, 20, 28] and tag suggestion and localization in web videos [2, 19].

In this paper we present a review of state-of-the-art data-driven methods for image and video tagging, with a thorough comparison of nearest-neighbor approaches for tag refinement, in order to address the problem of large-scale collections, inherent with social media, and we provide an analysis of the temporal aspects of user tags in two standard social media datasets. We present also an adaptation of a data-driven approach for tag localization in video shots, a problem that can be recast to that of tag refinement applied to video key frames. The paper is organized as follows: related works are discussed in Section 2; a description of the nearest-neighbor methods that have been selected for application to tag-refinement is provided in Section 3; temporal analysis of tags is presented in Section 4; description of nearest-neighbor methods for video shot tagging is provided in Section 5; a description of the datasets used in the experiments is reported in Section 6, while experimental results are discussed in Section 7. Finally conclusions are drawn in Section 8.

2 Related works

Many researchers have addressed problems related to social media analysis and annotation. In this section we have selected the most relevant works dealing with images and video using different approaches and considering different problems, reporting for each type of approach and problem the most relevant works.

YouTube tags:
National Geographic, world's deadliest, deadliest, deadly, lions, zebra, lioness, hunt, pack, packs, attack, predation, predator, eat, kill, blood, animal



…, lions, ~~zebra~~, lioness, hunt, pack, packs, ...

…, ~~lions~~, zebra, ~~lioness~~, ~~hunt~~, ~~pack~~, ~~packs~~, ...

Fig. 2 Example of video tag localization: *top)* YouTube video with its related tags; *bottom)* localization of tags in shots

2.1 Images

The first attempt in the literature for image tag refinement is the RWR algorithm presented in [41]. In this work, Wang et al. performed belief propagation among tags within the Random Walk with Restart framework, to refine the imprecise original annotations. Random walk-based tag refinement step, following an initial probabilistic tag relevance estimation based on kernel density estimation (RWTR), has been proposed in D. Liu et al. [23] for tag ranking and image retrieval.

The problem of filtering out unreliable tags in social images has been considered by Kennedy et al. in [14], where it is shown that the tags used by different persons to annotate visually similar images are more related to visual content than the others. In the proposed approach 20 nearest neighbors of each processed image are considered and scalability is addressed using a learned low-dimensional image feature, and using the Map/Reduce framework to speed the exhaustive search.

Li et al. [20] have proposed a tag relevance measure for image retrieval based on the consideration, originally proposed in [14], that if different persons label visually similar

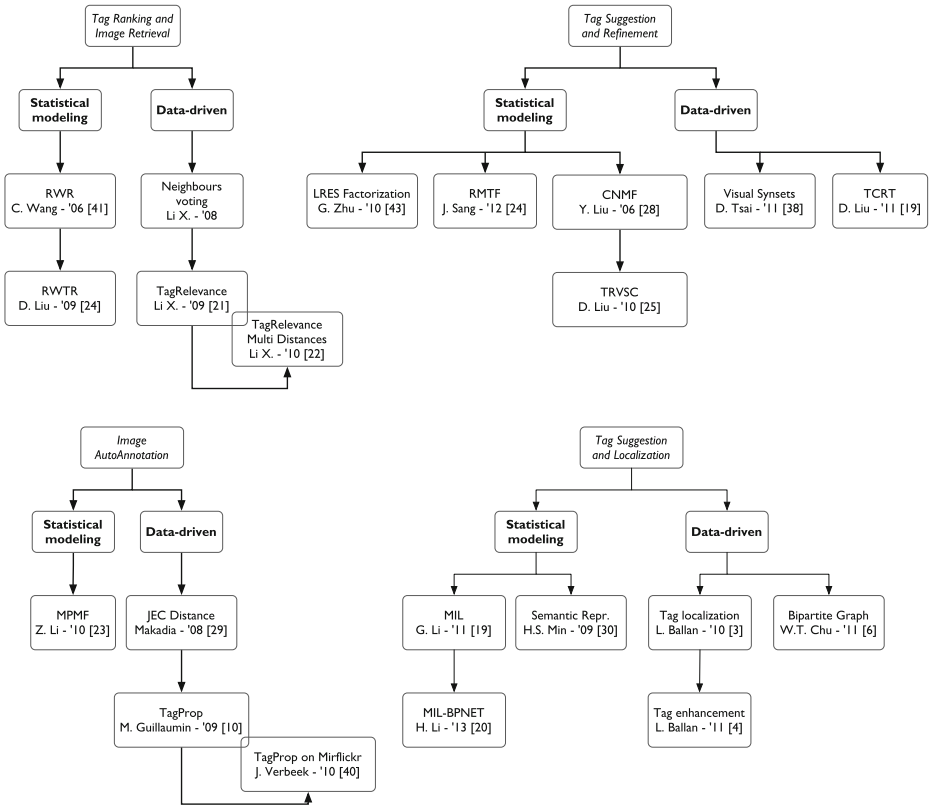


Fig. 3 Taxonomy of the most important works on social media annotation

images using the same tags, then these tags are more likely to reflect objective aspects of the visual content. Therefore it can be assumed that the more frequently the tag occurs in the neighbor set, the more relevant it might be. However, some frequently occurring tags are unlikely to be relevant to the majority of images. To account for this fact the proposed tag relevance measurement takes into account both the distribution of a tag t in the neighbor set for an image I and in the entire collection. The original method has been extended in [21] to fuse the outcomes of multiple tag relevance measures based on different visual features to compute image similarity.

Makadia et al. in [28] have proposed a baseline for image auto-annotation by using a simple method to transfer n tags to a test image from its visual neighborhood: similar images are ordered according to their similarity to the test image and the tags that are most frequent in a training set are 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] have proposed to learn a weighted nearest neighbor model, to automatically find the optimal combination of feature distances - e.g. local shape descriptors or global color histograms, to solve the task of image auto-annotation and tag relevance. Tags of a test image are based on neighbor rank or distance.

The assumption of consistency between visual and semantic similarity in social images is used by D. Liu et al. in [24] to formulate the tag refinement task as an optimization

framework, based on constrained non-negative matrix factorization (CNMF) [27] by Y. Liu et al., which tries to maximize the consistency while minimize the deviation the tags from initially provided by users. Considering that the consistency assumption is mainly applicable for content-related tags (see Fig. 1), a filtering procedure based on Wordnet is used to constrain the tagging vocabulary within content-related tags. Tag enrichment is done by considering tag synonyms and hypernyms. This method is usually referred in the literature as tag refinement based on visual and semantic consistency (TRVSC).

Tsai et al. [37] have proposed a structure named visual synset which is an organization of images which are visually-similar and semantically-related. Each visual synset correspond to a single prototypical visual concept with an associated set of weighted tags. Linear SVMs are then used to predict annotations to unseen images.

D. Liu et al. [26] have 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.

The method proposed by Zhu et al. in [43] is based on the assumptions that visually similar images are similarly tagged, that tags are often correlated and interact at the semantic level, that the semantic space spanned by all the tags can be approximated by a smaller subset of them and that user tags are accurate enough so that it can be assumed a condition of error sparsity for the image tag matrix. The problem of tag refinement is then cast into a decomposition of the user-provided tag matrix into a low-rank refined matrix and a sparse error matrix, and a convergence provable iterative procedure is proposed to accomplish the optimization. This tag refinement approach is referred as low-rank and error sparsity approximation (LRES).

A probabilistic approach, based on typical probabilistic matrix factorization (PMF) [32], is proposed by Z. Li et al. in [22] where they extend the formulation by jointly fusing different sources of correlation such as image-tag correlation, image similarity and tag correlation. Two sets of low dimensional latent factors are derived and used to predict newer annotations by reconstructing the image-tag correlations estimated.

Recently, Sang et al. [33] have 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, the task is cast into a ranking problem to determine which tag is more relevant for a user to describe an image than another tag. To this end is introduced a ternary semantics 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).

A characteristic that has received less attention, so far, is the temporal aspect of social media production. However, extracting time information from documents may improve several information retrieval applications such as hit-list clustering and exploratory search, as noted in [1]. In fact, several researchers have shown that the temporal information associated to search engine queries (e.g. frequency of query keywords over time) can be used to predict trends and behaviors related to economics (such as claims for unemployment benefits [4]) and medicine (such as flu epidemics [8]).

In [31] Rattenbury et al. have compared "burst" analysis techniques derived from signal processing against a novel method to identify social events in the associated social media, using the tags and geo-localization information of Flickr images. In [16] Kim et al. have proposed to use the temporal evolution of topics in social image collections to perform subtopic outbreak detection and to classify noisy social images. The authors used a non-parametric

approach in which images are represented using a similarity network, created using Sequential Monte Carlo, where images are the vertices and the edges connect the temporally related an visually similar images. Temporal dynamics of social image collections has been studied by Kim et al. in [15] to improve search relevance at query time, addressing both general and personalized interest searches. The authors propose a unified statistical model based on regularized multi-task regression on multivariate point process, in which an image stream is considered an instance of a process and a regression problem is formulated to learn the relations between image occurrence probabilities and temporal factors that influence them (e.g. seasons).

Analysis of the temporal evolution of social media collections have been proposed in [12] by Jin et al. to predict political success and product sales; regression-based and diffusion-based models have been adapted to account for a Flickr-based index, combining images' metadata and visual similarity, that models the popularity of politicians and products. The work presented by Kim et al. in [17] re-casts the problem of image retrieval re-ranking as a prediction of which images will be more likely to appear on the web at a future time point. Both collective group level and individual user level cases are considered, using a multivariate point process to model a stream of input images, and using a stochastic parametric model to solve the relations between the occurrences of the images and factors such as visual clusters, user descriptors and month of the image.

2.1.1 Visual features

Most of the approaches reported in this section rely on global features that have the advantage of being compact and require low computational costs. The authors of a commonly used dataset (see Section 6), NUS-WIDE-270K, provide precomputed descriptors composed by color moments, wavelet texture and edge histograms; these descriptors are used in papers that use this dataset like [43]. Similarly, the descriptors used in [20] and in [22] combine color correlogram, color moments and texture descriptors, while the descriptors used in [23] and [28] combine color moments and wavelet textures, or color histograms and wavelet textures, respectively. Tsai et al. [37] has added LBP to color histograms and wavelets.

Some works add local features to global descriptors. In [9] GIST and color histograms have been combined with SIFT and local color features. Global features like MPEG-7 Edge Histogram and local color descriptors (i.e. color SIFT) have been pre-computed also for MIRFlickr-25K [11]. Local features only have been used in [26], with SIFT descriptors and BoW.

2.2 Videos

Most of the recent works on internet videos have addressed problems like near duplicate detection [30], training concept detectors [38] or topic detection [34].

Ulges et al. [38] exploit YouTube videos in order to train concept detectors without using any manual annotation for the creation of ground truth data: using video tags as lexicon allows to scale the number of detected concepts, at the expense of detection performance, although training detectors with ground truth material prepared by experts in conjunction with social videos improves their performance.

Currently, only a few works have considered the problem of tag suggestion and localization in internet videos, i.e. associating video tags to specific shots: this problem can be recast as tag refinement applied to keyframes, where each keyframe is an image annotated with all the tags associated to the video. Ballan et al. [2] annotate automatically shots of

YouTube videos using Flickr images, with a variation of the tag relevance algorithm of [20] that, exploiting visual similarity of keyframes and images, can also add new tags that were not originally available in videos.

Localization of video tags is addressed by Li et al. in [18]; a multiple instance learning approach that considers semantic relatedness of co-occurring tags and temporal smoothness are used to model shots and videos.

Min et al. [29] annotate video shots with 34 concept detectors, using their results to build a semantic representation for each shot. The same detectors are applied to Flickr images and semantic similarity with video keyframes is used to suggest tags selected from those of the images.

Chu et al. [5] used Flickr images and the associated tags for tag localization, modeling relationship between keyframes in a video shot and candidate tags as a bipartite graph in which two disjoint sets of nodes (keyframes and tags), and each edge between nodes is associated with a weight calculated based on similarity between a pair of keyframe and tag, and tagging behaviors; best matching is used to determine the most appropriate tags to be associated with keyframes.

In [19] Li et al. have recently presented a dataset of 1550 YouTube videos with a ground-truth annotation and localization of 31 concepts, and results of tag localization performed using a baseline method based on multiple instance learning, based on the MIL-BPNET approach proposed in [42].

2.2.1 Visual features

Similarly to the papers dealing with images, the papers addressing videos have used global features like color and texture - i.e. color correlogram (computed in the HSV color space) and color moments, Tamura features [2] or MPEG-7 Edge Histogram Descriptor [3]. MPEG-7 color and texture descriptors have been used in [29]

Also local features have bee used: TOP-SIFT in [3] and SIFT with BoW in [19] and [5]. citeulges10 combines different feature extraction pipelines using SIFT and BoW, color histograms and textures and motion histograms.

3 Tag refinement using nearest neighbor methods

The basic idea of the nearest-neighbor methods is to select a set of visually similar images and then to select a set of relevant associated tags based on a tag transfer procedure. This type of methods has been typically applied to different tasks such as image auto-annotation and tag ranking/relevance. Considering a test image I 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), the methods selected are:

3.1 Simple label transfer: Makadia et al. [28]

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 has been originally tested on Corel5K, IAPR TC-12 and ESP datasets.

In our implementation 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 $\mathbf{f}_i = \left(f_i^1, \ldots, f_i^N\right)$, and σ^2 is set as the median value of all the distances.

3.2 Learning tag relevance from visual neighbors: Li et al. [20]

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 has been originally tested for image retrieval on a proprietary Flickr dataset with 20,000 manually checked images and for image auto-annotation using a subset of 331 images.

3.3 TagProp, discriminative metric learning in nearest neighbor models: Guillaumin et al. [9]

Using $y_{It} \in \{-1, +1\}$ to represent if tag t is relevant or not for the test image I, the probability of 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 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 according to each neighbor in the weighted sum, with $\pi_{II_i} \geq 0$ and $\sum_{N_k(I,K)} \pi_{II_i} = 1$. The objective is to maximize $\sum_{I,t} \ln p(y_{It})$.

The model can be used with rank-based or distance-based weighting. Furthermore, to compensate for varying frequencies of tags, a tag-specific sigmoid is used to scale the predictions, boosting the probability for rare tags and decrease that of frequent ones. 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 [39], 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.

4 Temporal evolution analysis

The correlation of the time series of the tags with Google searches (see Fig. 4) shows that for certain concepts web information sources may be beneficial to annotate social media.

To exploit the underlining time process and to be able to improve image annotation using temporal information, we need a way to evaluate quantitatively the possible correlation between sources. This let us analyze if a series can be estimated by another one and how a generalized model may describe the original time series. To this end we compute a correlation measure over two series. First of all we standardize all time series: given a time series $X = \{x_i : i \in D\}$, we compute $x_i = \frac{x_i - \overline{X}}{s}$, where $\overline{X}$ is the sample mean and s is the sample standard deviation. Even if sample mean and sample standard deviation are sensible to outliers, these can be removed thanks to a filtering and smoothing procedure described in Section 6.1.3. In our case X is the time series of user tags $t \in V_T$, while Y is the time series of the corresponding term in Google Trends. To evaluate the correlation between two time series, we choose to use the *sample Pearson correlation coefficient*, often denoted as r. Given two time series X and Y of n samples, r is defined as the ratio between covariance and the product of X variance and Y variance:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{Y})^2}} \tag{5}$$

which is defined in $[-1, 1]$. Values towards the positive or negative end reveal a strong correlation between the two time series, changing only in the sign. We can reformulate it as the mean of the products of the standard scores, which permits us to use standardized time series $\hat{x_i} = \frac{x_i - X}{s_X}$ and $\hat{y_i} = \frac{y_i - Y}{s_Y}$:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - X}{s_X}\right)\left(\frac{y_i - Y}{s_Y}\right) = \frac{1}{n-1}\sum_{i=1}^{n}\hat{x_i}\,\hat{y_i} \tag{6}$$

Given that the strength of correlation is not dependent on the direction or the sign, we also computed r-square. Unfortunately the interpretation of a correlation coefficient depends

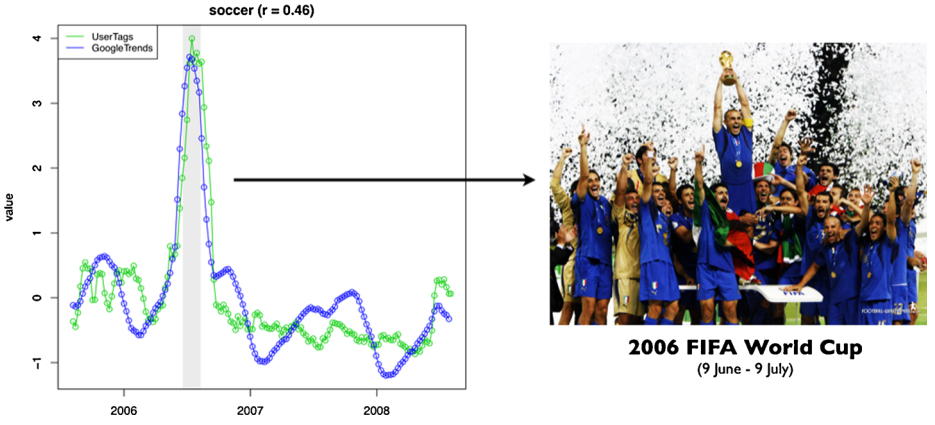


Fig. 4 Time series of Flickr user tags and Google searches for "soccer" in NUS-WIDE dataset

heavily on the context and purposes that can't be easily defined at this stage of work. However several works like [7] offered some guidelines which can be used to interpret our analysis, that are reported in Table 1.

5 Video tag localization using nearest neighbor methods

5.1 Video tag suggestion & localization: Ballan et al. [2]

The video tags $T_v = t_1, \ldots, t_l \in V_V$ are used as queries to retrieve images from Flickr, that are then used to build a visual neighborhood $N_k(I, K)$ for each shot keyframe I. The union of all the tags of the Flickr images of the neighborhood is the set of tags V_T associated with keyframe I. Then, tag relevance of these tags is computed as in Section 3.2; computing tag relevance on the set of tags of the whole video would not consider the fact that some tags describe the content of only certain shots, and would lead to simple re-ranking the same list of tags for all the shots; to avoid this problem the tag relevance algorithm is modified by computing relevance of only the tags that appear at least once in the visual neighborhood of a keyframe. Tag relevance is used to obtain the rank position of each tag $rank_t$ and a $Vote^+$ score according to [35], obtaining the suggestion score:

$$score(t, K) \cdot \frac{\lambda}{\lambda + (rank_t - 1)} \tag{7}$$

used to order the tags t be localized in the shot. Decision on the number of tags to be localized is based on the selection of a fixed number of ranked tags, ordered using the score.

5.2 Enriching and localizing semantic tags: Ballan et al. [3]

The model presented in the previous paragraph has been extended in [3] to compute weighted $tagRelevance$ based on visual similarity, and performing an initial video tag expansion using Wikipedia. Before the tag expansion step tags are filtered to eliminate those that are not very relevant, by computing a semantic relatedness score that consider tag presence in the video title, in the video neighborhood and tag co-occurrence. Selection of tags to be localized is based on selection of a minimum relevance score. Due to the lack of standard social video datasets for tag refinement, experiments have been performed on a new publicly available dataset.

6 Datasets

6.1 Image datasets

To demonstrate the effectiveness of nearest neighbor methods for image tag refinement in a real large-scale scenario, we performed thorough experiments on two large image datasets:

Table 1 Guidelines for sample Pearson correlation coefficient

Correlation	None	Small	Medium	Strong
Positive	0.0 to 0.09	0.1 to 0.3	0.3 to 0.5	0.5 to 1.0
Negative	−0.09 to 0.0	−0.3 to −0.1	−0.5 to −0.3	−1.0 to −0.5

MIRFlickr-25K [10] and NUS-WIDE-270K [6]. Both datasets have been collected from Flickr.

The MIRFlickr-25K dataset contains 25,000 images with 1,386 tags. The NUS-WIDE-270K dataset comprises a total of 269,648 images (provided as URLs) with 5,018 unique tags. In order to implement the method described in [20] (see Section 3.2) we had to download again the original data from Flickr for the NUS-WIDE-270K dataset, in order to obtain the users information that is not contained in the dataset; due to the fact that some of the original images of the NUS-WIDE-270K collection are not anymore available, we have been forced to use a subset of the 238,251 images that are still present on Flickr. Hereafter, we refer to this image collection as NUS-WIDE-240K.

Since the tags in the above two image collections are rather noisy and many of them are meaningless words, a pre-processing step was performed to filter out these tags. To this end we matched each tag with entries in Wordnet and only those tags with a corresponding item in Wordnet were retained, similarly to the approach used in [6]. Moreover, we removed the less frequent tags, whose occurrence numbers are below 50. The result of this pre-processing is that 219 and 684 unique tags were obtained in total for MIRFlickr-25K and NUS-WIDE-240K, respectively.

6.1.1 Temporal analysis

Since MIRFlickr-25K contains too few images to be useful for temporal analysis, we have substituted it with its superset MIRFlickr-1M [11], which contains 1 million images, selected by their Flickr interestingness score [10, 40]. Every image provided has full *Flickr metadata* which includes *taken* and *posted* timestamps, indicating when a photo was taken and when it was shared on Flickr. However, only about half of the images provide a valid "taken" timestamp, in particular only 584,892 are valid, as 330,454 have no timestamps and 84,654 have an invalid timestamp. Both MIRFlickr-1M and NUS-WIDE-240K have images that are unbalanced with respect to time, having very different number of images per date. The time interval of NUS-WIDE-240K goes from year 1900 (i.e. old photo scans) to 2009, concentrating most of the images between 2005–2008, while in MIRFlickr-1M images are concentrated around years 2007–2009.

Given that NUS-WIDE-240K has the biggest ground truth of the two datasets considered and that we are looking to discover the relations between tags and image content with respect to time, we choose to use it as the main reference. We use all the 81 manually checked tags as V_T set and consider four information sources which are different in the kind of underlining latent process:

- From NUS-WIDE-240K, for all images, we consider the V_T set of tags using the *manually validated* tags which constitute the entire ground truth; we refer to this source as *NUS-GT*.
- From NUS-WIDE-240K, for all images, we consider the V_T set of tags using the *user tags* (e.g. the tags provided by the respective Flickr users); we refer to this source as *NUS-TAGS*.
- From MIRFlickr-1M, for all images, we consider the V_T set of tags using the *user tags*; we refer to this source as *MIR-TAGS*.
- Beside image datasets, we also consider a source of temporal query information given by Google Trends. From Google Trends, we have downloaded all available query data for the V_T set of tags considered; we refer to this source as *GOO-TAGS*.

All sources are to be considered subject to different kinds of noise, in particular all images are highly unbalanced over time, resulting in days with hundreds of images and others with at most ten images. To reduce this effect, we choose to consider only the largest time span with at least 350 images per week. In addition the two image datasets differ in the time interval which has the most images. This forced us to use a reduced time interval that we choose as starting from 2005-06-01 and ending in 2008-08-01 for NUS-WIDE-240K (retaining 161,176 images from 179,128) and from 2007-01-01 to 2008-08-01 for MIRFlickr-1M (retaining 110,064 images from 531,670).

6.1.2 Visual features

For both these datasets, the visual similarity between images has been calculated using some simple visual descriptors. We started from the features provided by the authors of the NUS-WIDE dataset and, as in [43], for each image we have extracted a single 428-dimensional descriptor. This feature vector has been 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. These features have been computed for both the MIRFlickr-25K and NUS-WIDE-240K datasets, in order to have comparable results.

6.1.3 Temporal features

Given a set of images I, all taken in a set of dates D (as a daily interval), we denote as V_T the set of all tags used and U the set of all users. For every image $i \in I$ we denote $\text{tag}(i) \subseteq V_T$ the set of tags associated, $day(i) \in D$ the timestamp associated and $user(i) \in U$ the user who owns the image. We also consider two other time spans, a set of weeks W and a set of months M, easily computed by integrating over the interval of days considered. These can be thought as time series over the selected index set. For every set considered, we computed a set of features, as proposed in [17]:

– *Images per day*: the number of relevant images which are $taken$ in a day. More specifically, given a day $d \in D$, the number of images per day (IMD) is defined as

$$\text{IMD}(d) := |\{i \in I | day(i) = d\}| \tag{8}$$

  Similarly we also define a feature for the number of images per week (IMW) and per month (IMM).
– *Images per day for a tag*: the number of relevant images associated with a tag which are $taken$ in a day. More specifically, given a tag $t \in V_T$ and a day $d \in D$, the number of images with t per day (ITD) is defined as

$$\text{ITD}(t, d) := |\{i \in I | day(i) = d \wedge t \in tag(i)\}| \tag{9}$$

  Similarly we also define a feature per week (ITW) and per month (ITM).

However, a phenomenon associated with a social source is that of *batch tagging*: a user may decide to upload an entire album of photos and, instead of carefully tagging each photo, he could simply opt to tag each photo with the same tags (e.g. tag the album instead of every single photo). This may result in a kind of noise with respect to the normal use of tags in time. In addition, the features defined above are sensitive to this kind of noise, producing noisy peaks over single days. To produce a more meaningful analysis we decide to collapse all images that are batch tagged into a single entry. A set of images are considered *batch*

tagged if they are all uploaded by the same user on the same day and have the same set of tags. More specifically, given a user $\hat{u} \in U$, a day $\hat{d} \in D$ and a set of tags $\hat{t} \subseteq V_T$, a set of images $I_B = \{i_1, i_2, \ldots, i_k\}$ are considered *batch tagged* if $\text{tag}(i) = \hat{t}$, $\text{user}(i) = \hat{u}$, $\text{day}(i) = \hat{d}\ \forall i \in I_B$.

Flickr popularity model As described in [12], the selection of images of the two datasets is only a sample of all images in Flickr. In addition, the number of images over time in Flickr are mostly variable, based on the popularity of the site itself. This slow change over time can be modeled as a trend over all tags, independent from any particular query. Unfortunately, no statistics are released publicly and other sources such as Alexa[1] or Google Trends[2] are affected by the impact of news. Based on this preliminary analysis and supposing an uniform sampling in Flickr searches, we use the feature IMD to remove this background deviation by normalizing the ITD feature.

Given a tag $t \in V_T$ and a date $d \in D$ we compute:

$$\overline{ITD}(t, d) = \frac{ITD(t, d)}{IMD(d)} \tag{10}$$

This may also be considered as a frequentist probability distribution of tag t in day d with respect to all other tags considered, which is $p(t; d)$. Similarly we also compute $\overline{ITW}$ and $\overline{ITM}$ by considering a week and a month granularity, respectively. After collapsing all batch tagged images, the two datasets retain 179,128 images for NUS-WIDE-240K and 531,670 images for MIRFlickr-1M respectively. To make the time series patterns more clear, we computed a simple moving average over all time series, varying the windows size n from 2 to 10 weeks. For a day time series defined over a time span Ψ for a tag $t \in V_T$ is defined as:

$$ITD_n(t, d) = \frac{1}{n} \sum_{i=-n}^{n} \overline{ITD}(t, d + i) \quad \forall d \in \Psi \tag{11}$$

This has the effect to smooth the series, letting to visualize more clearly the trend. On the other hand, tags which have very sparse frequency tends to be worsened, so we adjusted the window size empirically, based on visualization clearness. The final time series are composed of 1,158 and 579 week samples respectively for NUS-WIDE-240K and MIRFlickr-1M.

6.2 Video dataset

The dataset[3] 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 3 h and 8 min and the number of detected shots is 4196. The number of tags per video varies from 8 to 22.

[1] Alexa Internet, Inc. http://www.alexa.com

[2] Google Trends. http://www.google.com/trends

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

Video tags are 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.

6.2.1 Visual features

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 [36], a compact image descriptor that combines interest points with visual words, designed for fast content-based image retrieval.

The Flickr images are clustered using k-means, to use the cluster centers as indexes for a fast approximate nearest neighbor search. For each keyframe of the video the nearest cluster center based on the visual similarity is retrieved. Images belonging to this cluster are considered as neighbors.

7 Experiments and results

7.1 Tag refinement evaluation framework

In order to measure the effectiveness of different tag refinement approaches, we evaluated the performance on the 18 tags in MIRFlickr-25K and the 81 tags in NUS-WIDE-240K where the ground-truth annotations have been provided by the respective authors of these datasets.[4] Following the most relevant previous works in the field [24, 26, 33, 41, 43], we report F-measure figures which have been widely used as evaluation metric of tag refinement. The F-measure is defined by $F = \frac{2RP}{(R+P)}$, where P is precision and R is recall.

The F-measure has been calculated to evaluate the refinement results for each tag, and then the overall results were usually obtained by averaging over the number of ground-truth annotations (i.e. classes) as a *macro-average*. Moreover, since both datasets are highly unbalanced, we show also the F-scores obtained by averaging over all the images as a *micro-average*. We believe that both *micro* and *macro* average F-scores are necessary to evaluate the performance of different tag refinement algorithms. The main reason is that because of the unbalance in the number of images per label, simple algorithms like Makadia et al. [28] tend to always predict the most common tags.

As previously done by most of the related works [26, 43], we report the overall results by retaining $m = 5$ tags per image. This is an important aspect since the performance are highly influenced by this number. For this reason, we report for both the datasets also some figures by varying m between 1 and 10. It has to be noticed that, 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.

Finally, we report also the figures for F-score *macro* while varying the m number of tags on both MIRFlicr-25K and NUS-WIDE-240K datasets.

[4] Source code and dataset metadata are available from http://www.micc.unifi.it/uricchio/

7.1.1 Evaluation of tag refinement on MIRFlickr-25K

To evaluate the effectiveness of the proposed methods, we compare the following four algorithms:

- Baseline, the original tags provided by the users (UT);
- Simple Label Transfer (SLT) [28], described in Section 3.1; as shown in Fig. 5 the best results are obtained using $K = 500$ neighbors;
- Learning Tag Relevance from Visual Neighbors (TR) [20], described in Section 3.2; again, see Fig. 5, the best results are obtained using $K = 500$ visual neighbors;
- TagProp, Discriminative Metric Learning in Nearest Neighbor Models (TP) [9], described in Section 3.3; the best results are obtained by defining the weights of the model directly as a function of the distance.

We performed two sets of experiments. The first one has been conducted on the entire dataset (i.e. 25,000 images) and the results are shown in Table 2. The second one has been conducted using 15,000 images as training set and 10,000 images as test set. Therefore, the results reported in Table 3 refer to the F-scores obtained on the test set (as averages among 10 random train/test splits). It has to be noticed that in this second set of experiments, the performance drop - about 5 % for each method - is due to the smaller number of visual neighbors available for the tag propagation.

In general, the Tag Relevance algorithm by Li et al. [20] guarantees superior performance with respect to the Simple Label Transfer algorithm by Makadia et al. [28] (e.g. 0.27 vs 0.26 on the MIRFlickr-25K full dataset, see Table 2). TagProp shows very similar results (e.g. 0.20 vs 0.19, as reported in Table 3) but it requires more computational costs and a learning phase, that does not allow to apply it to the full dataset. Regarding other methods recently presented in the literature, we report in Table 4 the most relevant previous results.

These results demonstrate that nearest-neighbor methods, when applied to tag refinement, give comparable results to more complex state-of-the-art approaches, despite their simplicity and low computational cost. Complex and computationally intensive algorithms such as TRVSC [24] and LRES [43] give an improvement in performance of about 2 percent, but require re-training if the datasets change. The recent results by Liu et al. [26],

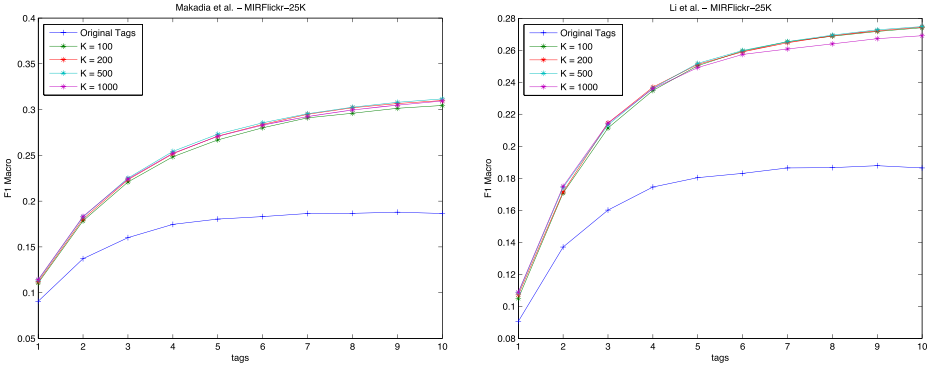


Fig. 5 F-score results (y axis) on the MIRFlickr-25K dataset with (**a**) the Simple Label Transfer algorithm [28], (**b**) the Tag Relevance Learning algorithm [20]. These results are obtained by varying the number of visual neighbors (K) and the number m of retained tags per image (x axis)

Table 2 Average performances of different algorithms for tag refinement on MIRFlickr-25K (full dataset)

	UT	SLT [28]	TR [20]
F-score *macro*	0.18	0.26	0.27
F-score *micro*	0.06	0.14	0.13

obtained using different visual features (i.e. 500-d BoW of SIFT descriptors), confirm the same trend.

7.1.2 Evaluation of tag refinement on NUS-WIDE-240K

We have done similar experiments on the NUS-WIDE-240K dataset, using the same parameters and the same experimental methodology. Again, we performed two sets of experiments. The first one has been conducted on the entire dataset (i.e. 238,251 images) and the results are shown in Table 5. The second one has been conducted using 158,834 images as training set and the remaining 79,417 as test set. In this case, the results are reported in Table 6. The variation of performance due to changes in the number of visual neighbors K and number of retained tags m per image is similar to that reported in Fig. 5 for MIRFlickr-25K.

The experiments on the NUS-WIDE-240K dataset confirm that the TR algorithm of Li et al. [20] gives the best results, both in terms of F-score macro and micro-average figures. It is more difficult to compare our results with the previous works since, in the case of the NUS-WIDE dataset, the previous works often use a subset of the full dataset (often due to the large-scale nature of this dataset) and some undocumented/non-standard experimental procedures. Zhu et al. [43] reported in their paper some results on the NUS-WIDE-270K dataset. Their pre-processing step on the tags vocabulary results in 521 tags (instead of our 684 tags). Their results are lower than the others reported by us and by the other works in the literature; their baseline UT is 0.269 while in our case is 0.35 (see Table 5) and so their results are not comparable to us; our results is more similar to those reported by Liu et al. [26] (UT=0.45) and Sang et al. [33] (UT=0.477). But both [26] and [33] used subsets of the NUS-WIDE-270K dataset, due to the inapplicability of their methods for such a huge number of images. In particular, Liu et al. [26] used a subset of only 24,300 images, while Sang et al. [33] used a subset of 124,099 images (about half of our NUS-WIDE-240K). Sang et al. have used also the same features of us but they have reported results obtained with $m = 10$ tags per image. On their dataset, they have obtained 0.475 with the RWR [41] method, 0.49 with TRVSC [24], 0.523 with LR [43], and 0.571 with their best algorithm.

Also in the case of a large-scale dataset such as NUS-WIDE-240K, nearest-neighbor based methods show competitive performance. Moreover, an important aspect that is clear from the other previous works is that this kind of approaches (i.e. matrix factorization and graph-based methods) suffer in a large-scale scenario. This fact enforces the interest in nearest-neighbor methods for tag refinement.

Table 3 Average performances of different algorithms for tag refinement on MIRFlickr-25K (test set)

	UT	SLT [28]	TR [20]	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 4 F-score performances of other algorithms for tag refinement on MIRFlickr-25K, as reported in the literature

	UT	RWTR [41]	TRVSC [24]	LRES [43]
Zhu et al. [43]	0.22	0.34	0.41	0.42
Liu et al. [26]	0.2	0.31	0.37	–

7.1.3 Dependency of precision on number of tags suggested

In a final experiment we have evaluated the F-score *macro* while varying the number of suggested tags, using both MIRFlickr-25K and NUS-WIDE-240K datasets. Figure 6 shows the best combination in terms of F-score *macro* for the train/test split, while Fig. 7 shows the results obtained using the full datasets.

7.2 Temporal analysis

In the following we will consider both the presence of the tags that have been added by the users that uploaded the images to Flickr (referring to them as "user tags") and the tags that have been manually checked by the creators of NUS-WIDE as referring to visual content of images (referring to them as "ground-truth" tags), to account for the fact that tags are often ambiguous and personalized [13, 35], and do not necessarily reflect the visual content of the image. As an example consider Fig. 8, showing the temporal usage of the tags "snow" and "soccer" in NUS-WIDE, along with the respective Google searches, as obtained from Google Trends. It can be observed that the peak in usage of the "soccer" tag - associated with the 2006 FIFA World Cup - reflects that in Google Trends, but the peak is much less pronounced in the ground truth tags; this indicates that for this tag the relationship between tag and image may exist because of how people react to social events, rather than uploading photos depicting that event on Flickr. On the other hand the peaks of both user and ground truth "snow" tag are corresponding to that of Google Trends: in this case the relationship may exist because it is more likely that people take pictures of snow scenes during winter, and this concept is less related to social aspects than to visual content of these images.

7.2.1 Qualitative analysis

Considering time series composed of the frequencies of image tags (either user or ground-truth) and Google searches obtained from Google Trends, it is possible to observe that they exhibit the presence of different components, that may appear mixed together:

trend: long term variation, that can be increasing, decreasing or also stable (see Fig. 9 left). Terms such as "computer" or "military" have this pattern;

cyclical variation: repeated but not periodic variations. Tags like "sports" or "flags" have this pattern;

Table 5 Average performances of different algorithms for tag refinement on NUS-WIDE-240K (full dataset)

	UT	SLT [28]	TR [20]
F-score *macro*	0.35	0.36	0.44
F-score *micro*	0.11	0.18	0.23

Table 6 Average performances of different algorithms for tag refinement on NUS-WIDE-240K (test set)

	UT	SLT [28]	TR [20]	TP [9]
F-score *macro*	0.35	0.36	0.45	0.44
F-score *micro*	0.11	0.18	0.22	0.21

seasonal variation: periodic variations, e.g. due to concepts associated with some regular event (see Fig. 9 center). Concepts related to seasons show this behavior, like "garden", "snow", "beach" or "frost";

irregular variation: random irregular variations, e.g. due to the sudden emergence of a topic (see Fig. 9 right), that appears as a burst of activity. Concepts that exhibit this pattern are related to social or natural events like "soccer", "earthquake" and "protest".

7.2.2 Correlation analysis

Figure 10 reports the outcome of correlation analysis of NUS-TAGS with NUS-GT, NUS-TAGS with GOO-TAGS and NUS-GT with MIR-TAGS. In particular it can be observed that the correlation of NUS-TAGS and NUS-GT has a vast majority of "Medium" and "Strong" values, while the correlation between user tags and Google searches is overall weaker and can be useful for a selected number of tags. The correlation between NUS-GT and MIR-TAGS has a large number of "Medium" and "Strong" values, suggesting that the temporal information of NUS-WIDE can be used in MIRFlickr-1M.

Correlation analysis of NUS-TAGS with GOO-TAGS, followed by averaging of r-square values over tags classes, determined by assigning each tag to the nearest Wordnet class – see Fig. 11 left - shows that Plant, Event, Phenomenon and Action obtain the higher values. A second group of categories comprises Artifact, Person+Group, Animal, Object and Time. In general, the categories that obtain the best performances are benefitting from tags whose time series show seasonal behaviors (e.g. "snow", "frost", "grass", "leaf") or have a "burst" behavior associated with specific social events (e.g. "soccer", "protest", "earthquake").

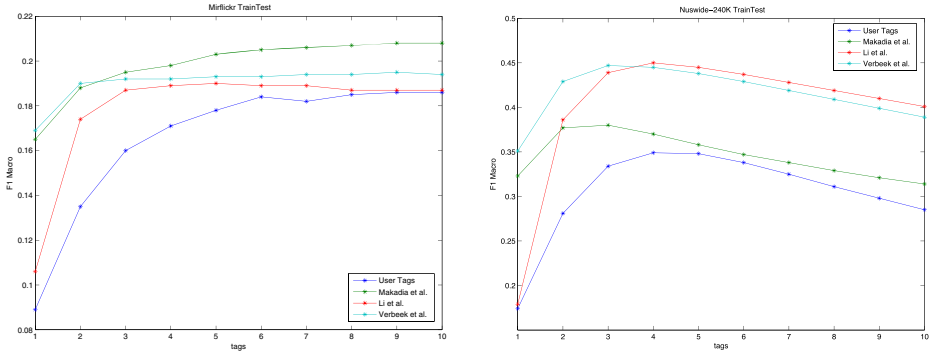


Fig. 6 F-score *macro* results (y axis) on the MIRFlickr-25K (*left*) and NUS-WIDE-204K train/test datasets (*right*) with user tags, Makadia et al. (Simple Label Transfer algorithm [28]), Li et al. (Tag Relevance Learning algorithm [20]), Verbeek et al. (TagProp algorithm [9]). These results are obtained by varying the number *m* of retained tags per image (x axis)

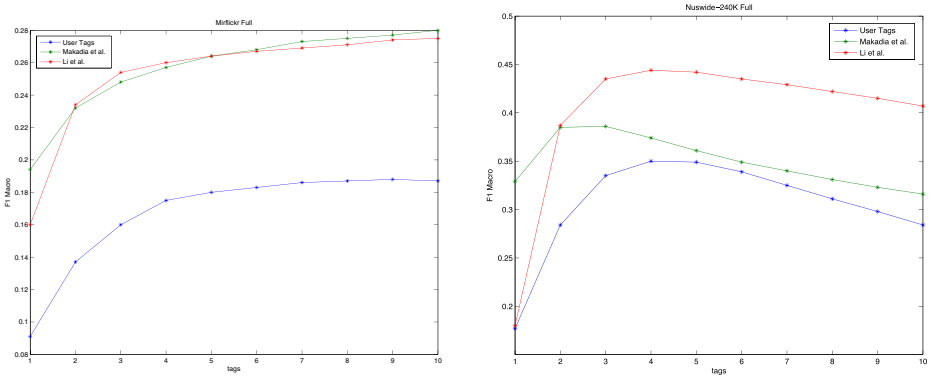


Fig. 7 F-score *macro* results (y axis) on the MIRFlickr-25K (*left*) and NUS-WIDE-204K full datasets (*right*) with user tags, Makadia et al. (Simple Label Transfer algorithm [28]), Li et al. (Tag Relevance Learning algorithm [20]). These results are obtained by varying the number m of retained tags per image (x axis)

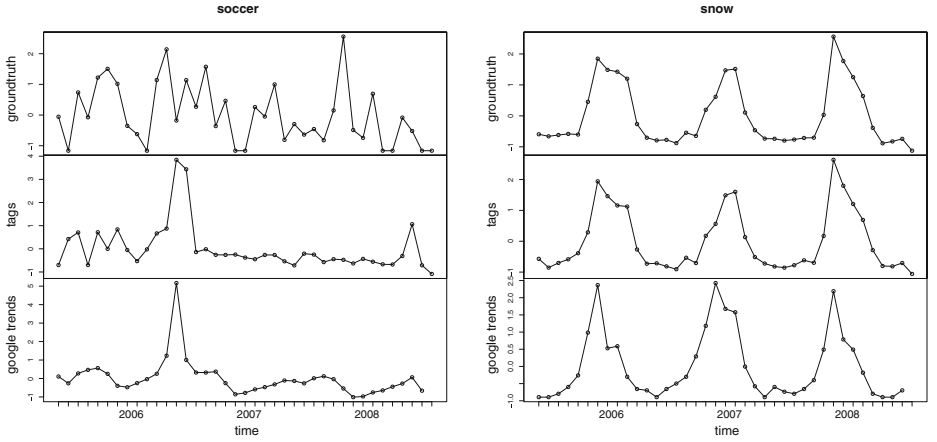


Fig. 8 *left)* frequency of "soccer" in NUS-GT, NUS-TAGS and GOO-TAGS: the peak of Google Trends and user tags in the summer of 2006 are related to the World Soccer Championship; *right)* frequency of "snow" in NUS-GT, NUS-TAGS and GOO-TAGS: the peaks are associated with winter seasons. Tag frequencies have been normalized by the number of images of the same day

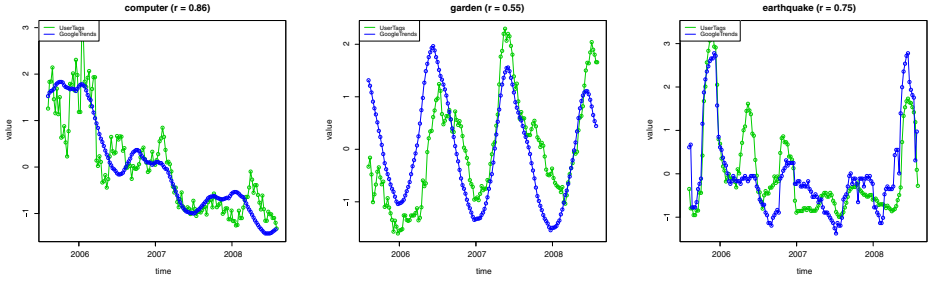


Fig. 9 Time series patterns of NUS-TAGS and GOO-TAGS, averaged over 10 weeks. *left)* trend (computer); *center)* seasonal (garden); *right)* episodic (earthquake: peaks correspond to earthquakes in China and Pakistan)

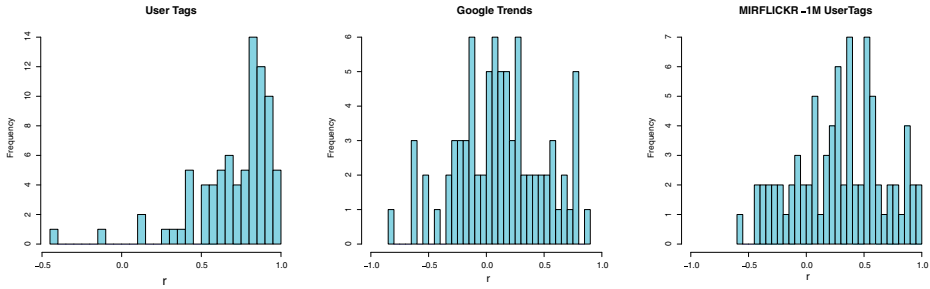


Fig. 10 *left)* r values computed between NUS-TAGS and NUS-GT; *center)* r values computed between NUS-TAGS and GOO-TAGS; *right)* r values computed between NUS-GT and MIR-TAGS

Correlation analysis of NUS-GT with GOO-TAGS (Fig. 11 right) shows that Plant and Phenomenon categories maintain their position among the best performing classes, because of the tags that exhibit a seasonal pattern. Instead the correlation of Event and Action categories is lower because the ground-truth tags that have an episodic pattern like "soccer", "protest" and "earthquake" have a lower correlation. This is due to the fact that these tags are employed by users also when the content of the image is not visually related to the described event.

7.3 Evaluation of video tag localization

The performance of [3] is measured in terms of accuracy: i.e. ratio between the number of tags correctly suggested and the total number of suggested tags. For each tag, resulting from the filtering and expansion process, the system downloads the first 15 Flickr images ranked according the "relevance" criterion provided by the Flickr API. Table 7 reports, for different relevance threshold scores, the accuracy and the mean number of correctly suggested tags for shot. The overall performance of the system is promising. We can observe that the mean accuracy on the entire dataset increases until score equals to seven and slightly decreases for higher scores, remaining close to 0.9; while the mean number of suggested tags correctly

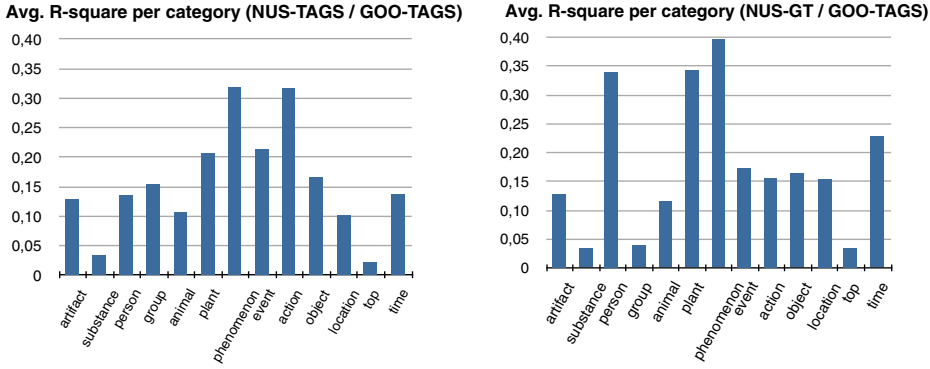


Fig. 11 NUS-WIDE dataset: r-square averages for tags classes. *left)* NUS-TAGS correlation with GOO-TAGS; *right)* NUS-GT correlation with GOO-TAGS

Table 7 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

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

decreases significantly for high scores (e.g. when requiring a threshold above 5). From the experimental results we can also note that some 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" we saw that 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.

7.4 Discussion and interpretation

Data driven approaches, as shown in Section 7.1.1, compare favorably with respect to more complex state-of-the-art approaches, requiring much less computation. Tag Relevance [20] and TagProp [9] show a better performance than Simple Lable Transfer [28] and this fact is visible using also the *F-micro* score and not only the *F-macro* score. TagProp has a slightly better performance than Tag Relevance, but it requires a training step that may not always be desirable. An advantage of Tag Relevance is that it can be easily adapted to video domain, as shown in Section 5. An important benefit of using data-driven approaches is visible from the results reported in Section 7.1.2, in which this class of methods has been tested on the larger NUS-WIDE-240K dataset, while the competing approaches like [26] and [33] have been applied to subsets only, obtaining results that are just marginally better. Finally, the results of the temporal analysis of user tags reported in Section 7.2 suggest that adding this contextual information could improve the annotation results.

8 Conclusion

We reviewed the state of the art approaches to automatic annotation of social media. In particular we analysed nearest neighbor methods since they have shown good recognition performance, and they are also suitable for large-scale recognition problems. We have presented a comparison of tag refinement methods for social images using standard datasets, presenting also a temporal analysis of the use of tags with respect to their presence in other social signals like Google Trends, and showing how this type of analysis could be beneficial for a certain number of classes of tags. We have also presented some extensions of nearest-neighbor methods for tag refinement to the problem of tag suggestion and localization in web videos, showing how these methods are flexible and can be adapted to different use cases.

References

1. Alonso O, Gertz M, Baeza-Yates R (2007) On the value of temporal information in information retrieval. SIGIR Forum 41(2):35–41
2. Ballan L, Bertini M, Del Bimbo A, Meoni M, Serra G (2010) Tag suggestion and localization in user-generated videos based on social knowledge. In: Proceedings of ACM SIGMM Workshop on Social Media (WSM). Firenze
3. Ballan L, Bertini M, Del Bimbo A, Serra G (2011) Enriching and localizing semantic tags in internet videos. In: Proceedings of ACM international conference on multimedia (ACM MM), pp 1541–1544. doi:10.1145/2072298.2072060
4. Choi H, Varian H (2011) Predicting the present with Google Trends. Tech. rep., Google
5. Chu WT, Li CJ (2011) Tag suggestion and localization for web videos by bipartite graph matching. In: Proceedings of ACM SIGMM Workshop on Social Media (WSM). New York
6. Chua TS, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) NUS-WIDE: a real-world web image database from National University of Singapore. In: Proceedings of ACM CIVR
7. Cohen J (1988) Statistical power analysis for the behavioral sciences. Routledge Academic
8. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009) Detecting influenza epidemics using search engine query data. Nature 457(7232):1012–1014
9. Guillaumin M, Mensink T, Verbeek J, Schmid C (2009) Tagprop: discriminative metric learning in nearest neighbor models for image auto-annotation. In: Proceedings of ICCV
10. Huiskes MJ, Lew MS (2008), The MIR Flickr retrieval evaluation. In: Proceeding of ACM MIR
11. Huiskes MJ, Thomee B, Lew MS (2010) New trends and ideas in visual concept detection: the MIR Flickr retrieval evaluation initiative. In: Proceedings of ACM MIR, pp 527–536
12. Jin X, Gallagher A, Cao L, Luo J, Han J (2010) The wisdom of social multimedia: using Flickr for prediction and forecast. In: Proceedings of ACM MM, pp 1235–1244
13. Kennedy LS, Chang SF, Kozintsev IV (2006) To search or to label? Predicting the performance of search-based automatic image classifiers. In: Proceedings of ACM MIR
14. Kennedy LS, Slaney M, Weinberger K (2009) Reliable tags using image similarity: mining specificity and expertise from large-scale multimedia databases. In: Proceedings of ACM-MM Workshop on Web-Scale Multimedia Corpus. Beijing
15. Kim G, Xing EP (2013) Time-sensitive web image ranking and retrieval via dynamic multi-task regression. In: Proceedings of ACM WSDM, pp 163–172
16. Kim G, Xing EP, Torralba A (2010) Modeling and analysis of dynamic behaviors of web image collections. In: Proceedings of ECCV, pp 85–98
17. Kim G, Fei-Fei L, Xing EP (2012) Web image prediction using multivariate point processes. In: Proceedings of ACM SIGKDD, pp 1068–1076
18. Li G, Wang M, Zheng YT, Chua TS (2011) ShotTagger: tag location for internet videos. In: Proceedings of ACM ICMR
19. Li H, Yi L, Guan Y, Zhang H (2013) DUT-WEBV: a benchmark dataset for performance evaluation of tag localization for web video. In: Proceedings of MMM
20. Li X, Snoek CGM, Worring M (2009) Learning social tag relevance by neighbor voting. IEEE Trans Multimed 11(7):1310–1322

21. Li X, Snoek CGM, Worring M (2010a) Unsupervised multi-feature tag relevance learning for social image retrieval. In: Proceedings of ACM CIVR
22. Li Z, Liu J, Zhu X, Liu T, Lu H (2010b) Image annotation using multi-correlation probabilistic matrix factorization. In: Proceedings of the international conference on multimedia, MM'10. ACM, New York, pp 11871190
23. Liu D, Hua XS, Yang L, Wang M, Zhang HJ (2009) Tag ranking. In: Proceedings of WWW
24. Liu D, Hua XS, Wang M, Zhang HJ (2010) Image retagging. In: Proceedings of ACM multimedia
25. Liu D, Hua XS, Zhang HJ (2011a) Content-based tag processing for internet social images. Multimed Tools Appl 51(1):723–738
26. Liu D, Yan S, Hua XS, Zhang HJ (2011b) Image retagging using collaborative tag propagation. IEEE Trans Multimed 13(4):702–712
27. Liu Y, Jin R, Yang L (2006) 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, p 421
28. Makadia A, Pavlovic V, Kumar S (2008) A new baseline for image annotation. In: Proceedings of ECCV
29. Min HS, Choi J, De Neve W, Ro YM, Plataniotis KN (2009) Semantic annotation of personal video content using an image folksonomy. In: Proceedings of IEEE ICIP
30. Paisitkriangkrai S, Mei T, Zhang J, Hua XS (2010) Scalable clip-based near-duplicate video detection with ordinal measure. In: Proceedings of ACM CIVR
31. Rattenbury T, Good N, Naaman M (2007) Towards automatic extraction of event and place semantics from flickr tags. In: Proceedings of ACM SIGIR, pp 103–110
32. Salakhutdinov R, Mnih A (2008) Probabilistic matrix factorization. Adv. Neural Info Process Syst 20:1257–1264
33. Sang J, Xu C, Liu J (2012) User-aware image tag refinement via ternary semantic analysis. IEEE Trans Multimed 14(3):883–895
34. Shao J, Yin W, Ma S, Zhuang Y (2010) Topic discovery of web video using star-structured k-partite graph. In: Proceedings of ACM multimedia
35. Sigurbjörnsson B, van Zwol R (2008) Flickr tag recommendation based on collective knowledge. In: Proceedings of WWW, pp 327–336
36. Thomee B, Bakker EM, Lew MS (2010) TOP-SURF: a visual words toolkit. In: Proceedings of ACM multimedia. doi:10.1145/1873951.1874250
37. Tsai D, Jing Y, Liu Y, Rowley HA, Ioffe S, Rehg JM (2011) Large-scale image annotation using visual synset. In: 2011 IEEE International conference on computer vision (ICCV). IEEE, pp 611–618
38. Ulges A, Schulze C, Koch M, Breuel TM (2010) Learning automatic concept detectors from online video. Comp Vision Image Underst 114(4):429–438
39. Verbeek J, Guillaumin M, Mensink T, Schmid C (2010) Image annotation with TagProp on the MIRFLICKR set. In: Proceedings of ACM MIR
40. von Ahn L, Dabbish L (2004) Labeling images with a computer game. In: Proceedings of ACM CHI
41. Wang C, Jing F, Zhang L, Zhang HJ (2007) Content-based image annotation refinement. In: Proceedings of CVPR
42. Zhang ML, Zhou ZH (2004) Improve multi-instance neural networks through feature selection. Neural Process Lett 19(1):1–10. doi:10.1023/B:NEPL.0000016836.03614.9f
43. Zhu G, Yan S, Ma Y (2010) Image tag refinement towards low-rank. In: Proceedings of ACM multimedia

Lamberto Ballan received the Laurea degree (M.S. equivalent) in computer engineering in 2006 and the Ph.D. degree in computer engineering, multimedia and telecommunication in 2011, both from the University of Florence, Italy. Currently he is a postdoctoral researcher at the Media Integration and Communication Center, University of Florence. He was a visiting scholar at the Signal and Image Processing department at Telecom Paristech/ ENST, Paris, in 2010. His research interests lie at the intersection of Multimedia, Computer Vision and Pattern Recognition. His work was conducted in the context of several EU and national projects, and his results have led to more than 30 publications in international journals and conferences, mainly in multimedia and image analysis. He has been awarded the best paper award by the ACM-SIGMM Workshop on Social Media in 2010. He coorganized the 1st Int'l Workshop on Web-scale Vision and Social Media in conjunction with ECCV 2012.

Marco Bertini is an assistant professor in the Department of Ingegneria dell'Informazione at the University of Florence, Italy. His research interests include content-based indexing and retrieval of videos and Semantic Web technologies. Bertini has a PhD in electronic engineering from the University of Florence. Contact him at marco.bertini@unifi.it. He has been awarded the best paper award by the ACM-SIGMM Workshop on Social Media in 2010. He co-organized the 1st Intl Workshop on Web-scale Vision and Social Media in conjunction with ECCV 2012.

Tiberio Uricchio is a Ph.D student at the Dipartimento di Ingegneria dell'Informazione at the University of Florence. He received the MS degree in computer engineering from the University of Florence, Italy, in 2012 with a thesis titled "Learning to tag images". His research interests focus on image processing, machine learning and pattern recognition, in particular automatic image and video annotation.

Alberto Del Bimbo is currently a Full Professor of computer engineering with the University of Florence, Florence, Italy, where he is also the Director of the Master's Program in Multimedia and the Director of the Media Integration and Communication Center. He has authored or co-authored more than 300 papers in scientific journals and international conferences, and is the author of the monograph Visual Information Retrieval. His current research interests include pattern recognition, multimedia information retrieval, computer vision, and human-computer interaction. Prof. Del Bimbo is an IAPR fellow and an Associate Editor of Multimedia Tools and Applications, Pattern Analysis and Applications, the Journal of Visual Languages and Computing, the International Journal of Image and Video Processing, and the International Journal of Multimedia Information Retrieval. He was an Associate Editor of Pattern Recognition, the IEEE Transactions on Multimedia, and the IEEE Transactions on Pattern Analysis and Machine Intelligence. He was a General Co-Chair of the ACM Multimedia in 2010 and the 12th European Conference on Computer Vision in 2012.