A data-driven approach for tag refinement and localization in web videos

A data-driven approach for tag refinement and localization in web videos

Lamberto Ballan[a,c,*], Marco Bertini[a], Giuseppe Serra[b], Alberto Del Bimbo[a]

[a]*Media Integration and Communication Center (MICC), Università degli Studi di Firenze, Viale Morgagni 65, 50134 Firenze, Italy*
[b]*Dipartimento di Ingegneria "Enzo Ferrari", Università degli Studi di Modena e Reggio Emilia, Via Vignolese 905/b, 41125 Modena, Italy*
[c]*Computer Science Department, Stanford University, 353 Serra Mall, Stanford, CA 94305, United States*

Abstract

Tagging of visual content is becoming more and more widespread as web-based services and social networks have popularized tagging functionalities among their users. These user-generated tags are used to ease browsing and exploration of media collections, e.g. using tag clouds, or to retrieve multimedia content. However, not all media are equally tagged by users. Using the current systems is easy to tag a single photo, and even tagging a part of a photo, like a face, has become common in sites like Flickr and Facebook. On the other hand, tagging a video sequence is more complicated and time consuming, so that users just tag the overall content of a video. In this paper we present a method for automatic video annotation that increases the number of tags originally provided by users, and localizes them temporally, associating tags to keyframes. Our approach exploits collective knowledge embedded in user-generated tags and web sources, and visual similarity of keyframes and images uploaded to social sites like YouTube and Flickr, as well as web sources like Google and Bing. Given a keyframe, our method is able to select "on the fly" from these visual sources the training exemplars that should be the most relevant for this test sample, and proceeds to transfer labels across similar images. Compared to existing video tagging approaches that require training classifiers for each tag, our system has few parameters, is easy to implement and can deal with an open vocabulary scenario. We demonstrate the approach on tag refinement and localization on DUT-WEBV, a large dataset of web videos, and show state-of-the-art results.

Keywords: Video tagging, Web video, Tag refinement, Tag localization, Social media, Data-driven, Lazy learning

1. Introduction

Over the past recent years social media repositories such as Flickr and YouTube have become more and more popular, allowing users to upload, share and tag visual content. Tags provide contextual and semantic information which can be used to organize and facilitate media content search and access. The performance of current social image and video retrieval systems depends mainly on the availability and quality of tags. However, these are often imprecise, ambiguous and overly personalized [1]. Tags are also very few (typically three tags per image, on average) [2], and their use may change over time, following the creation of new folksonomies created by users. Another issue to be considered is the 'web-scale' of data, that calls for efficient and scalable annotation methods.

Many efforts have been done in the past few years in the area of content-based tag processing for social images [3, 4]. The main focus of these works has been put on three aspects: *tag relevance* (or *ranking*) [5], *tag refinement* (or *completion*) [6] and *tag-to-region localization* [7]. Among the others, nearest-neighbor based approaches have attracted much attention for image annotation [8, 9, 10, 11], tag relevance estimation [12] and tag refinement [13]. Here the key idea is that if different users label similar images with the same tags, these tags truly represent the actual visual content. So a simple voting procedure may be able to transfer annotations between similar images. This tag propagation can be seen as a lazy local learning method in which the generalization beyond the training data is deferred until test time. A nice property of this solution is that it naturally adapts to an open vocabulary scenario in which users may continuously add new labels to annotate the media content. In fact, a key limitation of the traditional methods in which classifiers are trained to label images with the concept represented within, is that the number of labels must be fixed in advance. More recently, some efforts have been made also to design methods to automatically assign the annotated labels at image level to those derived semantic regions [7, 14, 15]. A relevant example is the work of Yang *et al.* [14] in which the encoding ability of group sparse coding is reinforced with spatial correlations among regions.

The problem of *video tagging* so far has received less attention from the research community. Moreover, typically it has been considered the task of assigning tags to whole videos, rather than that of associating tags to single relevant keyframes or shots. Most of the recent works on web videos have addressed problems like: *i) near duplicate detection*, applied to IPR protection [16, 17] or to analyze the popularity of social videos [18]; *ii) video categorization*, e.g. addressing actions and events [19, 20], genres [21] or YouTube categories [22]. How-

*Corresponding author. Tel.: +39 055 2751395.
Email addresses: lamberto.ballan@unifi.it (Lamberto Ballan), marco.bertini@unifi.it (Marco Bertini), giuseppe.serra@unimore.it (Giuseppe Serra), alberto.delbimbo@unifi.it (Alberto Del Bimbo)

Figure 1: Example of video tag localization: *top)* YouTube video with its related tags; *bottom)* localization of tags in keyframes.

ever, the problem of video tagging "in the wild" remains open and it might have a great impact in many modern web applications.

In this paper, the proposed method aims at two goals: to extend and refine the video tags and, at the same time, associate the tags to the relevant keyframes that compose the video, as shown in Fig. 1. The first goal is related to the fact that the videos available on media sharing sites, like YouTube, have relatively few noisy tags that do not allow to annotate thoroughly the content of the whole video. Tackling this task can be viewed also as an application of image tag refinement to video keyframes [4, 6]. The second goal is related to the fact that tags describe the global content of a video, but they may be associated only to certain shots and not to others. Our approach takes inspiration from the recent success of nonparametric data-driven approaches [8, 23, 24, 25]. We build on the idea of nearest-neighbor voting for tag propagation, and we introduce a temporal smoothing strategy which exploits the continuity of a video. Compared to existing video tagging approaches in which classifiers are trained for each tag, our system has few parameters and does not require a fixed vocabulary. Although the basic idea has been previously used for image annotation, this is the first attempt to extend this idea to video annotation and tag localization.

Our contributions can be summarized as follows:

- We propose an automatic approach that locates the temporal positions of tags in videos at keyframe level. Our method is based on a lazy learning algorithm which is able to deal with a scenario in which there is no predefined set of tags.
- We show state-of-the-art results on DUT-WEBV, a large dataset for tag localization in web videos. Moreover, we report an extensive experimental validation about the use of different web sources (Flickr, Google, Bing) to enrich and reinforce the video annotation.
- We show how the proposed approach can be applied in a real-world scenario to perform open vocabulary tag annotation. To evaluate the results, we collected more than 5,000 frames from 40 YouTube videos and three individuals to manually verify the annotation.

2. Related work

Probably the most important effort in semantic video annotation is TRECVID [26], an evaluation campaign with the goal to promote progress in content-based retrieval from digital video archives. Recently, online videos have also attracted the attention of researchers [22, 27, 28, 29, 30], since millions of videos are available on the web and they include rich metadata such as title, comments and user tags.

2.1. Tags at the video-level

A vast amount of previous work has addressed the problem of online video tagging using a simple classification approach with multiple categories and classes. Siersdorfer *et al.* [31] proposed a method that combines visual analysis and content redundancy, strongly present in social sharing websites, to improve the quality of annotations associated to online videos. They first detect the duplication and overlap between two videos, and then propagate the video-level tags using automatic tagging rules. Similarly Zhao *et al.* [32] investigated techniques which allow annotation of web videos from a data-driven perspective. Their system implements a tag recommendation algorithm that uses the tagging behaviors in the pool of retrieved near-duplicate videos.

A strong effort has been made to design effective methods for harvesting images and videos from the web to learn models of actions or events and use this knowledge to automatically annotate new videos. This idea follows similar successful approaches for image classification [33, 34, 35] but it has been applied only for the particular case of single-label classification. To this end, a first attempt has been made by Ulges *et al.* [36] who proposed to train a concept detection system on web videos from portals such as YouTube. A similar idea is presented in [19] in which images collected from the web are used to learn representations of human actions and then this knowledge is used to automatically annotate actions in unconstrained videos. A main drawback of these works is that they require training classifiers for each label, and this procedure does not scale very well, especially on the web. Very recently, Kordumova *et al.* [37] have also studied the problem of training detectors from social media, considering both image and video sources, obtaining state-of-the-art results in TRECVID 2013 and concluding that tagged images are preferable over tagged videos.

2.2. Tags at the keyframe(or shot)-level

Several methods have recently been proposed for unsupervised spatio-temporal segmentation of unconstrained videos [38, 39, 40]. Hartmann *et al.* [39] presented an object segmentation system applied to a large set of weakly and noisily tagged videos. They formulate this problem as learning weakly supervised classifiers for a set of independent spatio-temporal video segments in which the object seeds are refined using graphcut. Although this method shows promising results, the proposed system requires a high computational effort to process videos at a large scale. Similarly, Tang *et al.* [40] have addressed keyframe segmentation in YouTube videos using a weakly supervised approach to segment semantic objects. The proposed method exploits negative video segments (i.e. those that are not related to the concept to be annotated) and their distance to the uncertain positive instances, based on the intuition that positive examples are less likely to be segments of the searched concept if they are near many negatives. Both these methods are able to classify each shot within the video either as coming from a particular concept (i.e. tag) or not, and they provide a rough tag-to-region assignment.

The specific task of tag localization, i.e. transferring tags from the whole video to the keyframe or shot level, has been addressed by a few different research groups. Wang *et al.* [41] proposed a method for event driven web video summarization by tag localization and key-shot mining. They first localize the tags that are associated with each video into its shots by adopting a multiple instance learning algorithm [42], treating a video as a bag and each shot as an instance. Then a set of keyshots are identified by performing near-duplicate keyframe detection. Zhu *et al.* [43] used a similar approach in which video tags are assigned to video shots analyzing the correlation between each shot and the videos in a corpus, using a variation of sparse group lasso. A strong effort in collecting a standard benchmark for video localization research has been recently done by Li *et al.* [44]. They released a public dataset designed for tag localization, composed by 1550 videos collected from YouTube with 31 concepts and providing precise time annotations for each concept. The authors provide also an annotation baseline obtained using multiple instance learning, following [42]. All of these techniques have been largely adopted training classifiers, but still strongly suffer the lack of comprehensive, large-scale training data.

An early version of the proposed approach was introduced in our preliminary conference papers [45, 27]. In this paper we made key modifications in the algorithm and obtained significant improvements in the results. Differently from our previous work we introduce multiple types of image sources for a more effective cross-media tag transfer; we design a vote weighting procedure based on visual similarity and the use of a temporal smoothing strategy which exploits the temporal continuity of a video; further, we show a better performance in terms both of precision and recall. Finally, large-scale experiments have been carried on using a new public dataset [44, 46], allowing fair comparisons w.r.t. other methods.

Variable	Meaning
V	collection of videos and metadata (titles, tags, descriptions, etc.)
v, f	a video from D and a keyframe within v
D	dictionary of tags to be used for annotation
T_v, T'_v	set of tags associated to the video v, prior and after the Tag refinement and localization procedure
S	set of images downloaded from Google, Bing and Flickr using T_v filtered by stopwords, dates, tags containing numbers, punctuations and symbols
t	a particular tag from D
S_t	set of images from S annotated with the tag t
I_i, T_i	an image from S and their tags
S_K, T_K	set of K image neighbors for a given keyframe f ($S_K \subseteq S$) and their tags
T_f	set of tags associated to the keyframe f
$f^{(k)}$	a keyframe at time k
$t^{(k)}$	a binary variable that defines whether the tag t is present in the keyframe $f^{(k)}$

Table 1: Summary of notations used in this paper.

3. Approach

The architecture of our system is schematically illustrated in Fig. 2 and our notation is defined in Tab. 1. Let us consider a corpus V composed of videos and metadata (e.g. titles, tags, descriptions). We further define D as a dictionary of tags to be used for annotation. Each video $v \in V$, with tags $T_v \subseteq D$, can be decomposed in different keyframes.

Online video annotation is performed in two stages: in the first stage a relevance measure of the video tags is computed for each keyframe, possibly eliminating tags that are not relevant; then new tags are added to the original list. Video keyframes can be obtained either from a video segmentation process or from a simple temporal frame subsampling scheme. Each keyframe of the video is annotated using a data-driven approach, meaning that (almost) no training takes place offline. Given a keyframe, our method retrieves images from several sources and proceeds to transfer labels across similar samples.

3.1. Retrieval set

Similarly to several other data-driven methods [8, 23, 24, 25, 47], we first find a training set of tagged images that will serve for label propagation. The tags $T_v = \{t_1, \ldots, t_l\}$ associated to a video v are filtered to eliminate stopwords, dates, tags containing numbers, punctuations and symbols. In addition, we also include the WordNet synonyms of all these labels to extend the initial set of tags [1]. This resulting list of tags is then used to download a set of images $S = \{I_1, \ldots, I_m\}$ from Google, Bing and Flickr. Following this procedure an image $I_i \in S$, retrieved using t_j as query, has the following set of tags $T_i = \{t_j, t'_1, \ldots, t'_z\}$

[1] To cope with the fact that WordNet synonyms may introduce a semantic drift, for these tags we downloaded a number of images equal to one third of the original set.

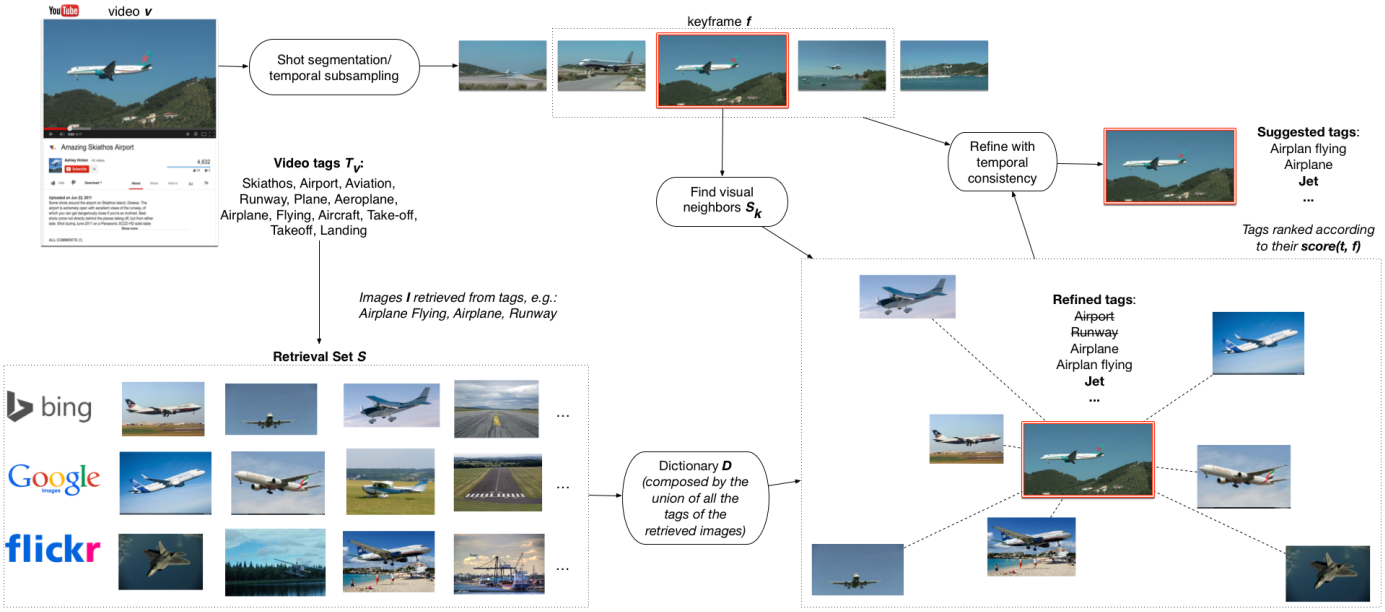

Figure 2: Overview of the proposed method.

if it has been obtained from Flickr or $T_i = \{t_j\}$ if it has been obtained from Google or Bing. It has to be noticed that in the latter case, only the query term has been collected as a label since the images do not contain any other additional tag. So let $D \supseteq T_v$ be the union of all the tags of the m images in S, after that they have been filtered with the same approach used for video tags (i.e. removing the stopwords, dates etc.). This set D is then used in the following steps to annotate "on the fly" the video.

Given the retrieval set S, for each keyframe f within the video v we find a (relatively) small set of K visual neighbors $S_K \subseteq S$. A good neighbor set will contain images that have similar scene types or objects (in our experiments we varied K from 150 to 300 images). In the attempt to indirectly capture this kind of similarity, we compute a 2000-d bag-of-visual-words descriptor, computed from densely sampled SIFT points. This descriptor can be efficiently used to find similar images using approximate search data structures by hierarchical k-means trees [48], in order to address scalability issues.

3.2. Tag localization and refinement

A simple approach to annotate a keyframe f is to consider only the tags belonging to the set of tags T_v that is associated to the video, computing their rank according to their relevance w.r.t. the keyframe to be annotated. This is a common procedure used for image tagging [8, 12]. However, this approach does not yield good results for the task of video annotation since the video tags may be associated only to certain keyframes and not to others. In fact, if we consider all the $t \in T_v$ for each keyframe, this procedure would simply result in a re-ranked list of the original video tags.

In order to solve this problem, we adopt the following approach: a tag t is kept in the list T_f, i.e. the set of tags associated to the keyframe f, only if it is present among the tags of the visual neighborhood (noted as T_K). Since the visual neighbors are images tagged by amateurs, such as Flickr users, or obtained from sources that can not be fully trusted, such as the images retrieved from Google or Bing, it is fundamental to evaluate the relevance of the tags that compose the lexicon. To this end, we build on the tag relevance algorithm for social image retrieval by Li *et al.* [12], and we present an effective framework to tackle the problem of tag localization and refinement in web videos.

The original *tag relevance* algorithm is based on the consideration that if different persons label visually similar images using the same tags, then these tags are more likely to reflect objective aspects of the visual content. Therefore it can be assumed that the more frequently the tag occurs in the neighborhood, the more relevant it might be. However, some frequently occurring tags are unlikely to be relevant for the majority of images. To consider this fact, given a keyframe f, the tag relevance score takes into account a prior term obtained by computing the ratio of cardinality of images tagged with t (denoted as S_t), to that of the entire retrieval set S:

$$tagRelevance(t, f, T_K) := \frac{1}{K} \sum_{i=1}^{K} R(t, T_i) - \frac{|S_t|}{|S|} \qquad (1)$$

where

$$R(t, T_i) = \begin{cases} 1 & \text{if } t \in T_i \\ 0 & otherwise \end{cases} \qquad (2)$$

where $|\cdot|$ is the cardinality of a set. Eq. 1 shows that more neighbor images labeled with the tag t imply larger tag relevance score. At the same time common frequent tags, that are less descriptive, are suppressed by the second term.

Differently from [12], the *tagRelevance* is not forced to be ≥ 1 and in case no visual neighbor is associated to t then it is

Algorithm 1: Tag refinement and localization

Input: A test video v with tags T_v.
Output: A set of keyframes $f \in v$ annotated with tags in D, The refined set T'_v at the video level.

Retrieve images from Google, Bing and Flickr for each $t \in T_v$ and let S be the retrieval set while D is the union of all the tags of the images in S;

for *each keyframe* $f \in v$ **do**
- Find K nearest visual neighbors of f from S;
- $tagRelevance(t, f, T_K) := \sum_{i=1}^{K} R(t, T_i) - Prior(t, D)$;
- Rank each candidate tag t by *tagRelevance* in descending order, and compute $score(t, f)$ (Eq. 4);
- Refine/compute the final $score(t, f)$ by exploiting temporal continuity (Eq. 5);

Define $T'_v := \bigcup_f T_f$ as the refined set of tags for v;

set to 0. This effectively allows to localize in time the original video tags.

The function $R(t, T_i)$ can be changed to account for the similarity between a keyframe and its visual neighbors. In our system we weight each vote with the inverse of the square of Euclidean distance between f and its neighbors:

$$R(t, T_i) = \begin{cases} \frac{1}{d(f, I_i)^2} & \text{if } t \in T_i \\ 0 & otherwise \end{cases} \quad (3)$$

where $d(f, I_i)$ is the Eucliden distance between feature vectors of the keyframe f and the image I_i. It has to be noticed that in case that a relevant tag is incorrectly eliminated in this phase, it may be recovered during the following stage of annotation.

Summarizing the above, the output of tag relevance estimation is a ranked list of tags for each keyframe f. In other words, $\forall t \in T_K$, the algorithm computes its tag relevance and a resulting rank position $rank_t$. Then, for each tag in T_f (as obtained from the previous steps), we compute the co-occurrence with all the tags in T_K. This results in a tag candidate list from which we select the tags that have a co-occurrence value that is above the average. For each candidate tag we then compute a suggestion score $score(t, T_f)$, according to the $Vote^+$ algorithm [2]. The final score is computed as follows:

$$score(t, f) = score(t, T_f) \cdot \frac{\lambda}{\lambda + (rank_t - 1)} \quad (4)$$

where λ is a damping parameter set to 20. We tuned λ on our training set by performing a parameter sweep and maximizing performance both in terms of precision and recall. The results obtained applying Eq. 4 are used to order all the candidate tags for the actual keyframe f, and the 5 most relevant tags are then selected. Finally, the union of all the tags selected at the keyframe level may be used to annotate the video at the global level (hereafter referred as to T'_v).

3.3. Temporal consistency

A main drawback of the procedure reported above is that the score computed using Eq. 4 does not account for the temporal aspects of a video. On the other hand, videos exhibit a strong temporal continuity in both visual content and semantics [49]. Thus we attempt to exploit this coherence by introducing a temporal smoothing to the relevance scores with respect to a tag. To this end, for each tag t and keyframe f, we re-evaluate the score function as reported below.

Let $f^{(k)}$ (or, for simplicity, f) be the actual keyframe at time k, and d the maximum temporal distance within which the keyframes are considered; thus $f^{(k-i)}$ refers to the nearby keyframe at a temporal distance i. The score is computed as follows:

$$score(t, f) = \sum_{i=-d}^{+d} w_i \cdot P(t^{(k)} = 1 | t^{(k-i)} = 1) \cdot score(t, f^{(k-i)}). \quad (5)$$

The term $score(t, f^{(k-i)})$ is the score obtained for the tag t and the keyframe that is temporally i keyframes apart from f, while w_i is a Gaussian weighting coefficient (which satisfies $\sum_i w_i = 1$).

The binary random variable $t^{(k)}$ is similarly defined to represent whether the tag t is present in the keyframe $f^{(k)}$. We then estimate the conditional probabilities between neighboring keyframes (for a tag at a time), from ground-truth annotations. These are computed as follows:

$$P(t^{(k)} = 1 | t^{(k-i)} = 1) = \frac{\#(t^{(k)} = 1, t^{(k-i)} = 1)}{\#(t^{(k-i)} = 1)} \quad (6)$$

where $\#(t^{(k-i)} = 1)$ is equivalent to the total numbers of relevant keyframes in the training dataset; $\#(t^{(k)} = 1, t^{(k-i)} = 1)$ is the total number that two keyframes are i frames apart and both relevant to the tag t.

We examine the contributions of changing the width of time window d in the experiments of Section 4.4. We finally summarize the procedure for tag refinement and localization by neighbor voting in Algorithm 1.

4. Experiments

Our proposed approach is a generic framework that can be used to annotate web videos and also to refine and localize their initial set of tags. To quantitatively evaluate the performance of our system, we present extensive experimental results for tag refinement and localization on a large public dataset.

4.1. DUT-WEBV dataset

Our experiments have been conducted on the DUT-WEBV dataset [44] which consists of a collection of web videos collected from YouTube by issuing 31 tags as queries. These tags, listed in Tab. 2, have been selected from LSCOM [50] and cover a wide range of semantic levels including *scenes*, *objects*, *events*, *people activities* and *sites*. There are 50 videos for each concept, but 2 videos are associated to two different tags, so that the total number of different videos is 1,458. For each video is

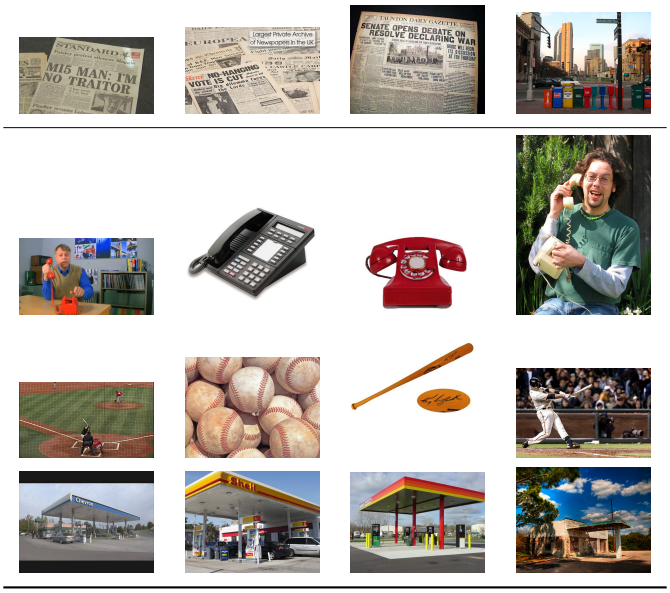

Figure 3: Example frames from YouTube videos, Google, Bing, Flickr images for the tag (top to bottom): *newspapers*, *telephones*, *baseball* and *gas station*.

provided also a ground truth that indicates the time extent in which a particular tag is present. In order to evaluate video annotation and tag refinement "in the wild", we have collected additional information with respect to the original dataset. In particular, for each video that is still available on YouTube, we have extracted the tags provided by the original users to complement title and description that are provided by the authors of the dataset. This effort allows to use the dataset also for generic video annotation and tag refinement research, and it is so an additional contribution of our work.

Our experimental setup follows the one proposed by the authors of the dataset, whose results are compared in Sect. 4.5. Video frames have been sub-sampled from each video every two seconds, following the experimental setup proposed by the authors of the dataset, obtaining 170,302 different frames. For each tag we have obtained images from web search engines, namely Google Images and Bing Images, and from a social network, i.e. Flickr. The overall number of images retrieved is 61,331. Considering all the video frames and downloaded images, the overall number of images in the dataset is thus 231,633, comparable to the dimension of NUS-WIDE (which is nowadays the largest common dataset used for social image retrieval and annotation). Some examples of the images retrieved from these web sources, as well as the corresponding keyframes from DUT-WEBV, are shown in Fig. 3.

4.2. Experiment 1: tag localization using only DUT-WEBV data

First of all, we present a tag localization baseline on the DUT-WEBV dataset relying only on the keyframes extracted from the web videos. The experimental setup used to build the image retrieval set follows the approach used in the baseline provided with the dataset [44]. So, given a particular tag t to be localized in a video v, we extract all the keyframes

Category	Tag	#frames *with tag*	#frames *total*
Events	airplane flying	2,217	5,241
	birthday	1,464	5,172
	explosion	2,050	3,870
	flood	2,216	4,083
	riot	4,462	6,582
Objects	cows	3,014	5,080
	food	1,773	6,576
	golf player	1,497	4,295
	newspapers	2,443	6,168
	suits	2,287	5,302
	telephones	2,720	5,587
	truck	2,382	6,171
Activities	baseball	2,459	3,991
	basketball	3,026	4,925
	cheering	2,788	6,605
	dancing	1,781	6,092
	handshaking	1,516	3,412
	interviews	4,217	7,206
	parade	3,445	5,756
	running	2,826	6,024
	singing	4,045	6,802
	soccer	3,204	4,747
	swimming	2,757	4,924
	walking	2,669	6,035
Scenes	beach	3,016	5,305
	forest	4,157	7,001
	mountain	2,735	6,394
Sites	aircraft cabin	2,593	5,110
	airport	4,187	6,538
	gas station	1,029	4,327
	highway	2,321	5,166
Total		*83,296*	*170,302*

Table 2: DUT-WEBV dataset: list of tags with their corresponding category, number of frames containing a particular tag/concept and total number of keyframes extracted from all the videos labeled with a particular tag.

of the other videos associated to t, and the keyframes of 10 randomly selected videos associated to other 10 randomly selected tags from T_v. Similarly to previous works [44, 46], we use *Precision@N* and *Recall@N* to evaluate results (i.e. precision/recall at top N ranked results).

In our experiments, the visual neighborhood S_K is obtained varying the number K of neighbors from 150 to 300. Tag relevance is computed using Eq. 4 and without weighting votes. These preliminary results are reported in Tab. 3. It can be observed that, as the number of visual neighbors increases so the performance slightly improves, both in terms of precision and recall. In the rest of the paper, if not mentioned otherwise, we fixed $K = 200$. We have conducted also similar experiments by weighting votes as reported in Eq. 3. Using this procedure we observed an improvement in recall of around 4% and a loss

#num. neigh.	*Precision*@1						*Recall*@1					
	events	objects	activities	scenes	sites	**Avg.**	events	objects	activities	scenes	sites	**Avg.**
150	59.4	52.9	60.7	70.8	50.5	**58.4**	54.4	46.3	41.4	55.5	66.6	**49.2**
200	59.8	54.2	59.8	70.9	52.4	**58.6**	53.3	48.7	40.6	55.9	68.8	**49.6**
250	57.7	53.4	60.9	70.4	53.1	**58.6**	53.2	48.4	42.4	54.4	69.1	**50.1**
300	59.0	54.7	62.3	69.5	53.3	**59.6**	54.8	48.6	42.9	55.1	71.4	**50.9**

Table 3: Results of tag localization using only DUT-WEBV data (Experiment 1).

in precision of more then 5%. The tag localization task is inherently more demanding in terms of precision since a tag that has not been recognized at a particular keyframe might be recovered in the forthcoming frames. So, in the following experiments, we only report the performance obtained using the original voting scheme (Eq. 2).

4.3. Experiment 2: tag localization using different web sources

In this experiment we evaluate the effect of using different sources to build the visual neighborhood. First of all we compare the results obtained with the previous baseline configuration (i.e. video only) with several combination of video and different web sources. Then we analyze the same configurations without the original video frames. Note that in these experiments the diversity of the images in the retrieval set grows, as well as the total number of tags in our dataset. The results are reported in Tab. 4; the first column indicates the sources used to create the neighborhood.

It can be observed that using all the available image sources provides the best precision result of 65.2%. In terms of precision any combination of video and additional source performs better than the same source alone, but it is interesting to notice that using all the social and web sources together (B+G+F) provides very good results, 62.3%. This is even better than using video alone, which achieves 58.6%, or any combination of video with a single additional image source [2] except when using Flickr. We believe that the results obtained using only web sources (B+G+F) are very interesting since this configuration might be the most useful in a "annotation in the wild" scenario, in which no previous video data are available. It has also to be noticed that this configuration provides higher results w.r.t. the "closed world" scenario in which only video data is used, on almost all the categories. In some cases, look for example at tags such as *highway*, *airport* and *airplane_flying*, the performance are significantly higher than in the baseline configuration. A comparison of the precision obtained for each individual tag with the most interesting configurations is shown in Fig. 4.

[2] We believe the main difference between the use of Google and Bing is due to technical reasons: these search engines do not provide an official API to download the images needed for the experiments, that were obtained from them through scraping. Bing apparently enforces stricter anti-scraping techniques that resulted in a more limited and less diverse set of images than Google.

Regarding recall results, the main difference is between using only video data which achieves 49.6% and any other combination which provides at most 29.9%. In case of using video alone, we rely on the training data provided in the original benchmark and this is obviously not possible in a real application of our system in which the set of tags is not known a priori. Moreover, the intra-class variation of the videos is not very high and this may facilitate too much the recall results.

Analysis of specific tags. To analyze more in depth the effect of using different image sources, the following Tab. 5 reports the results for a few tags that have large variations in terms of precision when using only one of the possible sources. Some of these variations can be motivated by the fact that images of some social sources, such as Flickr, have been created with a different intent. For example, on the one hand Flickr is often used by amateur photographers aiming at some aesthetics, and are therefore too different from videos that are documenting an object, like newspapers of gas stations. On the other hand Flickr users tend to represent objects like telephones or activities like baseball in their context, i.e. including persons using them or participating in the action, while Bing and Google tend to use more objective images. Examples of these differences are shown in Fig. 3.

4.4. Experiment 3: tag localization using temporal consistency

In this experiment we evaluate the effect of our temporal smoothing procedure (see Sect. 3.3) using the combination of parameters obtained from previous experiments that obtains the best precision, i.e. using all image sources and $K = 200$. In Tab. 6 we show the results obtained at varying width of keyframe time window, i.e. the value of d in Eq. 5. Keyframes have been temporally subsampled every 2 seconds, therefore if $d = 1$ the temporal extent of the video corresponds to 4 seconds.

The results show that considering temporal aspects is beneficial for the performance since it improves recall (around 4%) without reducing precision. Using larger temporal extents does not provide particular advantages since conditional probabilities of the presence of a concept at several seconds of distance are often not relevant. It has to be noticed that our temporal smoothing procedure has a negligible computational cost and so it gives great advantages with no drawbacks.

sources	*Precision*@1						*Recall*@1					
	events	objects	activities	scenes	sites	**Avg.**	events	objects	activities	scenes	sites	**Avg.**
V	59.8	54.2	59.8	70.9	52.4	**58.6**	53.3	48.7	40.6	55.9	68.8	**49.6**
V + B + G + F	66.3	58.1	66.0	77.0	64.7	**65.2**	27.3	28.4	20.7	43.7	24.3	**26.2**
V + B + G	69.9	54.1	68.2	76.7	62.3	**65.1**	25.3	26.1	18.3	41.8	26.4	**24.5**
V + B	65.3	56.7	60.9	72.5	56.9	**61.3**	25.9	32.3	23.1	51.1	35.0	**29.9**
V + G	81.2	53.1	54.5	62.2	72.6	**62.1**	48.1	44.0	5.2	27.5	43.1	**29.3**
V + F	70.4	54.7	64.7	77.1	54.7	**63.3**	25.2	29.2	20.4	34.1	18.8	**24.3**
B + G + F	66.6	54.2	60.3	77.1	66.0	**62.3**	24.9	29.4	13.5	35.1	18.8	**21.7**
B + G	63.7	54.1	51.4	76.2	64.4	**58.1**	25.6	24.4	11.5	37.6	20.8	**20.4**
B	47.1	54.7	57.2	66.1	21.6	**51.3**	4.4	23.6	12.3	38.5	1.4	**14.7**
G	64.8	52.9	54.8	75.7	61.4	**58.9**	31.8	22.4	11.0	28.8	23.0	**20.2**
F	51.2	43.1	56.4	71.6	44.5	**52.5**	19.0	27.1	15.1	8.9	4.9	**16.5**

Table 4: Results of tag localization using different combinations of image sources: DUT-WEBV frames (V), social images from Flickr (F) and web images from Google (G) and Bing (B). The visual neighborhood consists of 200 images.

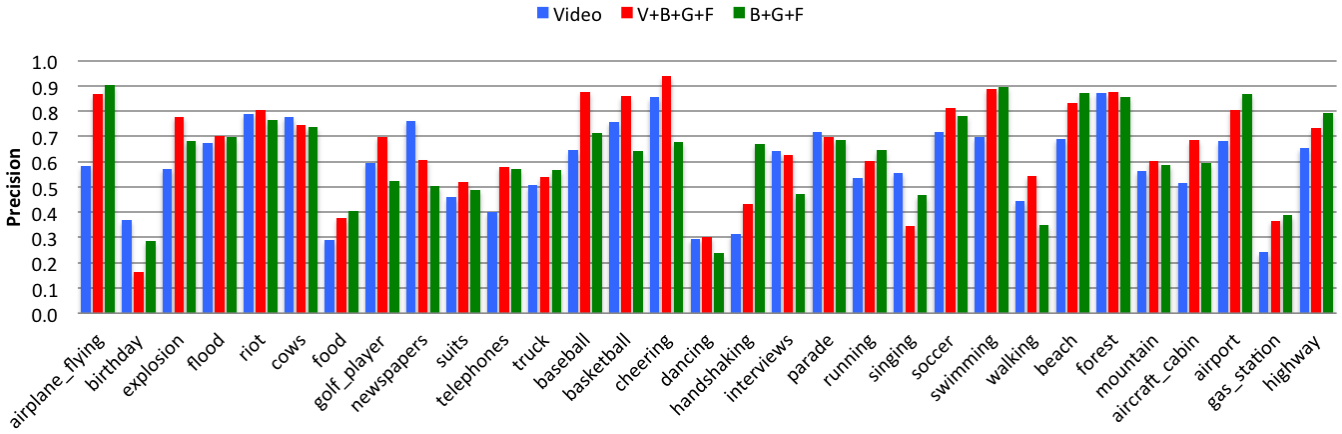

Figure 4: Precision rate broken down by tag for the most interesting combinations of image sources.

Tag	Video	Google	Bing	Flickr
newspapers	76.2	81.3	90.4	45.2
telephones	40.2	36.7	40.5	47.1
baseball	64.5	48.5	62.2	83.1
gas station	24.2	37.5	33.8	0.0

Table 5: *Precision*@1 for specific tags, when using different image sources.

4.5. Comparison with previous works

The tag localization method proposed by the authors of the dataset is MIL-BPNET [51]. The choice of this method is motivated by the fact that MIL has been used in other approaches for tag localization [42, 41, 44] and thus provides a sound baseline. In particular, given a tag t, the associated videos form the positive bags, while the others the negative bags. To reduce computational costs, for each tag, 10 negative tags are selected and for each negative tag 10 videos are randomly selected to create the negative bags. Performance is reported by the original authors as *Precision*@N, with a varying N that accounts for the number of video frames that contain each concept and the percentage of video parts that contains the concept.

We show in Tab. 7 a comparison of the results reported in [44] with our best combination of image sources (V+B+G+F) and temporal smoothing computed using $d = 3$. The table reports also figures of precision for two other methods as reported in [46]: the first one use kernel density estimation (KDE) to localize the most relevant frames for a given tag; the second one combines KDE with visual topic modeling (using LDA). For these latter methods, the original authors report results for a

	Precision@1						Recall@1					
d	events	objects	activities	scenes	sites	**Avg.**	events	objects	activities	scenes	sites	**Avg.**
0	66.3	58.1	66.0	77.0	64.7	**65.2**	27.3	28.4	20.7	43.7	24.3	**26.2**
1	64.4	57.9	66.5	77.0	66.9	**65.3**	31.2	29.8	22.5	48.3	26.7	**28.6**
2	64.9	57.7	66.0	76.6	67.0	**65.1**	32.4	29.9	22.9	49.9	27.4	**29.2**
3	64.8	57.8	66.1	76.6	67.5	**65.3**	32.6	29.9	23.0	49.9	27.7	**29.7**

Table 6: Results of tag localization with temporal consistency, using different time widths. Visual neighborhood obtained using $K = 200$ from DUT-WEBV frames, social and web images (Flickr, Google Images, Bing Images). If d is 0 then no temporal consistency is used.

subset of only ten tags.

Our proposed method obtains better results than MIL-BPNET for all tags but four, and overall performs better in all categories. On average, we outperform the baseline of 10%. Moreover, a comparison w.r.t. KDE and KDE+LDA shows that the proposed method obtains better results except for two tags. It has to be noticed that our results are reported as *Precision*@1 while these baselines were measured using *Precision*@*N* (with a large *N*), and so our improvements should be considered even more.

We compare also with a ConvNet-based classifier [52], trained using ImageNet 2010 metadata. Very recently, deep convolutional neural networks (CNN) have demonstrated state-of-the-art results for large-scale image classification and object detection [52, 53] and promising results on multilabel image annotation [54]. Similarly to our method, results are reported as *Precision*@1. As can be expected convolutional neural networks have a better performance in several classes, although in many others the results are comparable (e.g. basketball, soccer, gas station). On the other hand it has to be noted that even using a GPU implementation [3] the processing time is twice as slower than the proposed method, and that using ConvNets require an extremely large amount of manually annotated images.

4.6. Experiment 4: frame-level annotation "in the wild"

In this experiment we evaluate the performance of the annotation using an open vocabulary, performing tag localization and refinement. To this end we have selected 40 YouTube videos, for a total of 5,351 frames; visual neighborhoods have been built from Google, Bing and Flickr images, retrieved using the original video tags. The dictionary used for annotation is obtained by the union of all the tags of the retrieved images, and on average is composed by around 8,000 tags per video. With this approach it becomes possible to tag keyframes showing specific persons (e.g. TV hosts like Ellen DeGeneres), objects (e.g. Dodge Viper car or NBA Baller Beats videogame) or classes (e.g. marathon races). The annotation performance has been evaluated in terms of Precision@5 and Precision@10, through manual inspection of each annotated frame by three different persons, and averaging the results. Each annotator was

Category	Tag	**Our**	MIL [51]	KDE [46]	KDE + LDA [46]	ConvNet [52]
Events	airplane flying	84.3	72.6	72.0	72.9	-
	birthday	12.7	30.5	-	-	-
	explosion	82.0	65.0	-	-	-
	flood	69.6	55.0	58.3	63.1	-
	riot	78.8	69.3	-	-	-
	Avg.	**65.5**	**58.5**	**-**	**-**	**-**
Objects	cows	72.0	58.1	-	-	-
	food	37.3	41.6	-	-	-
	golf player	72.6	38.6	-	-	61.8
	newspapers	58.2	41.6	-	-	64.3
	suits	51.5	42.5	54.4	54.6	75.3
	telephones	59.7	53.4	58.1	58.4	72.2
	truck	53.3	52.1	-	-	77.1
	Avg.	**57.8**	**46.8**	**-**	**-**	**-**
Activities	baseball	91.2	66.9	-	-	95.4
	basketball	84.1	64.3	-	-	87.9
	cheering	96.3	58.2	-	-	-
	dancing	30.7	28.1	-	-	-
	handshaking	45.9	44.7	-	-	-
	interviews	71.2	61.8	65.6	69.6	-
	parade	67.8	69.4	-	-	-
	running	62.3	45.5	47.0	54.7	34.2
	singing	19.0	61.1	-	-	-
	soccer	82.6	76.3	71.4	79.7	83.8
	swimming	86.5	70.8	-	-	47.3
	walking	55.2	43.0	-	-	-
	Avg.	**66.1**	**57.5**	**-**	**-**	**-**
Scenes	beach	85.0	70.5	-	-	-
	forest	83.5	73.2	76.3	79.5	-
	mountain	61.6	57.4	53.9	58.6	-
	Avg.	**76.7**	**67.0**	**-**	**-**	**-**
Sites	aircraft cabin	75.4	51.9	-	-	-
	airport	80.9	70.1	-	-	-
	gas station	41.1	23.5	-	-	45.5
	highway	72.9	58.5	57.6	59.6	-
	Avg.	**67.6**	**51.0**	**-**	**-**	**-**
	Overall Avg.	**65.3**	**55.3**	**-**	**-**	**-**

Table 7: Comparison between our method and the MIL-BPNET [51] baseline in terms of precision. We report also the results of KDE and KDE+LDA [46] for a subset of nine tags as in the original paper.

[3] CCV – A Modern Computer Vision Library: `http://libccv.org`

requested to evaluate the relevance of each tag with respect to the visual content of each frame. Given the difficulty of this assessment this was performed after watching the whole video, and reading video title, description and list of original tags, so to understand the topics of each video and the content of the individual frames; frames were presented to the annotators following their order of appearance in the video. Results are reported in Tab. 8, comparing the results with a baseline that randomly selects tags, with a probability proportional to their frequency in the downloaded images. As can be expected the precision is lower than that of the other experiments, but this is due to the difficulty of multi-label annotation and to the very large vocabulary used to annotate each video.

Method	Precision@5	Precision@10
Random	6.1	4.5
Our	33.4	30.4

Table 8: Annotation "in the wild", using an open vocabulary. Comparison between our method and the random baseline.

4.7. Running time and system details

Finally, we provide a rough analysis about the computational requirements of our system. The Python implementation of the proposed algorithm annotates a video frame in about 0.17 seconds, of which 96% of the time is spent in computing the visual neighborhood, and $\sim$ 4% to compute tag localization and suggestion. The time required to compute temporal consistency is negligible. The average DUT-WEBV video is composed by around 110 keyframes (with a median of 98), requiring about 18.7 seconds to process it. This is mostly an unoptimized and un-parallelized implementation, and all our experiments are run on a single workstation with Xeon 2.67 GHz six core CPU and 48 GB RAM. As previously reported, for each image and keyframe we have computed a 2000-d bag-of-visual-words histogram obtained from densely sampled SIFT descriptors. Moreover, we used ANN and hierarchical k-Means trees [48] to speed up nearest neighbor search.

In order to promote further research on this topic, we provide all the additional annotation of the DUT-WEBV dataset to the public at large on our webpage `www.micc.unifi.it/vim`, as well as the visual features used in our experiments. We share also the images retrieved from the different web sources to build our retrieval set.

5. Conclusions

In this paper we have presented a tag refinement and localization approach based on lazy learning. Our system exploits collective knowledge embedded in user generated tags and visual similarity of keyframes and images uploaded to social sites like YouTube and Flickr, as well as web image sources like Google and Bing. We also improve our baseline algorithm with a temporal smoothing procedure which is able to exploit the strong temporal coherence which is normally present in a video.

We have demonstrated state-of-the-art results on the DUT-WEBV dataset and we have shown an extensive analysis of the contribution given by different web sources. We plan to extend this work with a large experimental campaign with an open set of tags (not only the ground truth labels provided in the original benchmark) in order to evaluate our system in a tag recommendation scenario.

Acknowledgments

This research was supported in part by a grant from the Tuscany Region, Italy, for the AQUIS-CH project (POR CRO FSE 2007-2013). L. Ballan acknowledges the support of a Marie Curie Individual Fellowship from the EU's Seventh Framework programme under grant agreement No. 623930.

References

[1] L. S. Kennedy, S.-F. Chang, I. V. Kozintsev, To search or to label? Predicting the performance of search-based automatic image classifiers, in: Proc. of ACM MIR, Santa Barbara, CA, USA, 2006, pp. 249–258.

[2] B. Sigurbjörnsson, R. van Zwol, Flickr tag recommendation based on collective knowledge, in: Proc. of WWW, Beijing, China, 2008, pp. 327–336.

[3] M. Wang, B. Ni, X.-S. Hua, T.-S. Chua, Assistive tagging: A survey of multimedia tagging with human-computer joint exploration, ACM Computing Surveys 44 (2012) 25:1–25:24.

[4] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. M. Snoek, A. Del Bimbo, Socializing the semantic gap: A comparative survey on image tag assignment, refinement and retrieval, arXiv preprint arXiv:1503.08248 (2015).

[5] D. Liu, X.-S. Hua, L. Yang, M. Wang, H.-J. Zhang, Tag ranking, in: Proc. of WWW, Madrid, Spain, 2009, pp. 351–360.

[6] D. Liu, X.-S. Hua, M. Wang, H.-J. Zhang, Image retagging, in: Proc. of ACM Multimedia, Firenze, Italy, 2010, pp. 491–500.

[7] Y. Wang, G. Mori, A discriminative latent model of image region and object tag correspondence, in: Proc. of NIPS, Vancouver, BC, Canada, 2010, pp. 2397–2405.

[8] A. Makadia, V. Pavlovic, S. Kumar, A new baseline for image annotation, in: Proc. of ECCV, Marseille, France, 2008, pp. 316–329.

[9] M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid, TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation, in: Proc. of ICCV, Kyoto, Japan, 2009, pp. 309–316.

[10] Y. Verma, C. V. Jawahar, Image annotation using metric learning in semantic neighbourhoods, in: Proc. of ECCV, Firenze, Italy, 2012, pp. 836–849.

[11] L. Ballan, T. Uricchio, L. Seidenari, A. Del Bimbo, A cross-media model for automatic image annotation, in: Proc. of ACM ICMR, Glasgow, UK, 2014, pp. 73–80.

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

[13] L. Ballan, M. Bertini, T. Uricchio, A. Del Bimbo, Data-driven approaches for social image and video tagging, Multimedia Tools and Applications 74 (2015) 1443–1468.

[14] Y. Yang, Y. Yang, Z. Huang, H. T. Shen, Tag localization with spatial correlations and joint group sparsity, in: Proc. of CVPR, Providence, RI, USA, 2011, pp. 881–888.

[15] X. Cao, X. Wei, Y. Han, Y. Yang, N. Sebe, A. Hauptmann, Unified dictionary learning and region tagging with hierarchical sparse representation, Computer Vision and Image Understanding 117 (2013) 934–946.

[16] M. Douze, H. Jégou, C. Schmid, P. Pérez, Compact video description for copy detection with precise temporal alignment, in: Proc. of ECCV, Heraklion, Greece, 2010, pp. 522–535.

[17] J. Song, Y. Yang, Z. Huang, H. T. Shen, J. Luo, Effective multiple feature hashing for large-scale near-duplicate video retrieval, IEEE Transactions on Multimedia 15 (2013) 1997–2008.

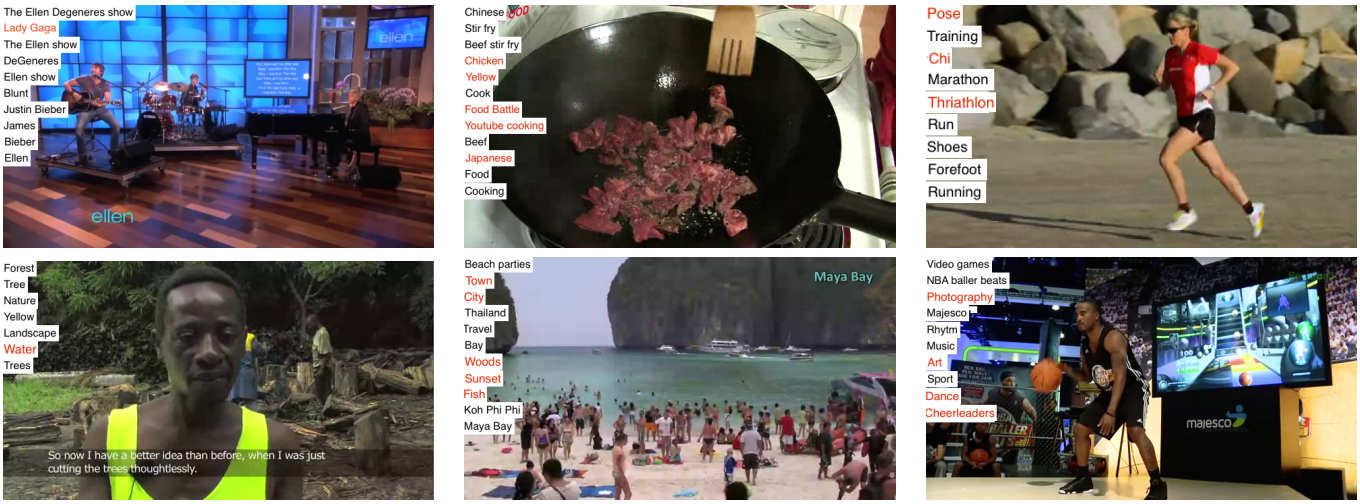

Figure 5: Annotation "in the wild" experiment: some example frames from a subset of YouTube videos of the original DUT-WEBV dataset.

[18] J. Liu, Y. Yang, Z. Huang, Y. Yang, H. T. Shen, On the influence propagation of web videos, IEEE Transactions on Knowledge and Data Engineering 26 (2014) 1961–1973.

[19] N. Ikizler-Cinbis, R. G. Cinbis, S. Sclaroff, Learning actions from the web, in: Proc. of ICCV, Kyoto, Japan, 2009, pp. 995–1002.

[20] A. Habibian, C. G. M. Snoek, Recommendations for recognizing video events by concept vocabularies, Computer Vision and Image Understanding 124 (2014) 110–122.

[21] J. Wu, M. Worring, Efficient genre-specific semantic video indexing, IEEE Transactions on Multimedia 14 (2012) 291–302.

[22] Z. Wang, M. Zhao, Y. Song, S. Kumar, B. Li, YouTubeCat: Learning to categorize wild web videos, in: Proc. of CVPR, San Francisco, CA, USA, 2010, pp. 879–886.

[23] J. Hays, A. A. Efros, Scene completion using millions of photographs, ACM Transactions on Graphics 26 (2007) 29–35.

[24] V. Ordonez, G. Kulkarni, T. L. Berg, Im2text: Describing images using 1 million captioned photographs, in: Proc. of NIPS, Granada, Spain, 2011, pp. 1143–1151.

[25] C. Liu, J. Yuen, A. Torralba, Nonparametric scene parsing via label transfer, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (2011) 2368–2382.

[26] A. F. Smeaton, P. Over, W. Kraaij, Evaluation campaigns and TRECVid, in: Proc. of ACM MIR, Santa Barbara, CA, USA, 2006, pp. 321–330.

[27] L. Ballan, M. Bertini, A. Del Bimbo, G. Serra, Enriching and localizing semantic tags in internet videos, in: Proc. of ACM Multimedia, Scottsdale, AZ, USA, 2011, pp. 1541–1544.

[28] W. Yang, G. Toderici, Discriminative tag learning on youtube videos with latent sub-tags, in: Proc. of CVPR, Colorado Springs, CO, USA, 2011, pp. 3217–3224.

[29] K. K. Reddy, M. Shah, Recognizing 50 human action categories of web videos, Machine Vision and Applications 44 (2013) 971–981.

[30] D. H. Nga, K. Yanai, Automatic extraction of relevant video shots of specific actions exploiting web data, Computer Vision and Image Understanding 118 (2014) 2–15.

[31] S. Siersdorfer, J. San Pedro, M. Sanderson, Automatic video tagging using content redundancy, in: Proc. of ACM SIGIR, Boston, MA, USA, 2009, pp. 395–402.

[32] W. Zhao, X. Wu, C.-W. Ngo, On the annotation of web videos by efficient near-duplicate search, IEEE Transactions on Multimedia 12 (2010) 448 – 461.

[33] R. Fergus, L. Fei-Fei, P. Perona, A. Zisserman, Learning object categories from Google's image search, in: Proc. of ICCV, Beijing, China, 2005, pp. 1816–1823.

[34] S. Vijayanarasimhan, K. Grauman, Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization, in: Proc. of CVPR, Anchorage, AK, USA, 2008, pp. 1–8.

[35] L.-J. Li, L. Fei-Fei, OPTIMOL: automatic online picture collection via incremental model learning, International Journal of Computer Vision 88 (2010) 147–168.

[36] A. Ulges, C. Schulze, M. Koch, T. M. Breuel, Learning automatic concept detectors from online video, Computer Vision and Image Understanding 114 (2010) 429–438.

[37] S. Kordumova, X. Li, C. G. M. Snoek, Best practices for learning video concept detectors from social media examples, Multimedia Tools and Applications 74 (2015) 1291–1315.

[38] M. Grundmann, V. Kwatra, M. Han, I. Essa, Efficient hierarchical graph-based video segmentation, in: Proc. of CVPR, San Francisco, CA, USA, 2010, pp. 2141–2148.

[39] G. Hartmann, M. Grundmann, J. Hoffman, D. Tsai, V. Kwatra, O. Madani, S. Vijayanarasimhan, I. Essa, J. Rehg, R. Sukthankar, Weakly supervised learning of object segmentations from web-scale video, in: Proc. of ECCV, VSM Workshop, Firenze, Italy, 2012, pp. 198–208.

[40] K. Tang, R. Sukthankar, J. Yagnik, L. Fei-Fei, Discriminative segment annotation in weakly labeled video, in: Proc. of CVPR, Portland, OR, USA, 2013, pp. 2483–2490.

[41] M. Wang, R. Hong, G. Li, Z.-J. Zha, S. Yan, T.-S. Chua, Event driven web video summarization by tag localization and key-shot identification, IEEE Transactions on Multimedia 14 (2012) 975–985.

[42] G. Li, M. Wang, Y.-T. Zheng, T.-S. Chua, ShotTagger: Tag location for internet videos, in: Proc. of ACM ICMR, Trento, Italy, 2011, pp. 1–8.

[43] X. Zhu, Z. Huang, J. Cui, H. T. Shen, Video-to-shot tag propagation by graph sparse group lasso, IEEE Transactions on Multimedia 15 (2013) 633–646.

[44] H. Li, L. Yi, Y. Guan, H. Zhang, DUT-WEBV: A benchmark dataset for performance evaluation of tag localization for web video, in: Proc. of MMM, Huangshan, China, 2013, pp. 305–315.

[45] L. Ballan, M. Bertini, A. Del Bimbo, M. Meoni, G. Serra, Tag suggestion and localization in user-generated videos based on social knowledge, in: Proc. of ACM Multimedia, WSM Workshop, Firenze, Italy, 2010, pp. 1–5.

[46] H. Li, L. Yi, B. Liu, Y. Wang, Localizing relevant frames in web videos using topic model and relevance filtering, Machine Vision and Applications 25 (2014) 1661–1670.

[47] D. Giordano, I. Kavasidis, S. Palazzo, S. Spampinato, Nonparametric label propagation using mutual local similarity in nearest neighbors, Computer Vision and Image Understanding 131 (2015) 116–127.

[48] M. Muja, D. G. Lowe, Scalable nearest neighbor algorithms for high dimensional data, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2014) 2227–2240.

[49] K.-H. Liu, M.-F. Weng, C.-Y. Tseng, Y.-Y. Chuang, M.-S. Chen, Association and temporal rule mining for post-filtering of semantic concept detection in video, IEEE Transactions on Multimedia 10 (2008) 240–251.

[50] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. S. Kennedy, A. Hauptmann, J. Curtis, Large-scale concept ontology for multimedia,

IEEE Multimedia 13 (2006) 86–91.
[51] M.-L. Zhang, Z.-H. Zhou, Improve multi-instance neural networks through feature selection, Neural Processing Letters 19 (2004) 1–10.
[52] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proc. of NIPS, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.
[53] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, International Journal of Computer Vision in press (2015) 1–1.
[54] Y. Gong, Y. Jia, T. Leung, A. Toshev, S. Ioffe, Deep convolutional ranking for multilabel image annotation, in: Proc. of ICLR, Banff, Canada, 2014, pp. 1–9.

Vitae

Lamberto Ballan is currently a postdoctoral researcher at Stanford University, supported by a prestigious Marie Curie Fellowship by the European Commission. He received the Laurea and Ph.D. degrees in computer engineering in 2006 and 2011, both from the University of Florence, Italy. He was a visiting scholar at Telecom ParisTech in 2010. He received the best paper award at the ACM Workshop on Social Media 2010, and is also the lead organizer of the Web-scale Vision and Social Media workshops at ECCV 2012 and CVPR 2014.

Marco Bertini is an assistant professor in the Department of Information Engineering at the University of Florence, Italy. His research interests include content-based indexing and retrieval of videos and semantic web technologies. He received the Laurea and Ph.D. degrees in electronic engineering from the University of Florence, Italy, in 1999 and 2004, respectively. He has been awarded the best paper award by the ACM Workshop on Social Media in 2010.

Giuseppe Serra is an assistant professor at the University of Modena and Reggio Emilia, Italy. He received the Laurea and Ph.D. degrees in computer engineering in 2006 and 2010, both from the University of Florence, Italy. He was also a visiting scholar at Carnegie Mellon University, USA, and at Telecom ParisTech, France. He has published around 40 publications in scientific journals and international conferences. He has been awarded the best paper award by the ACM Workshop on Social Media in 2010.

Alberto Del Bimbo is a professor of computer engineering at the University of Florence, Italy, where he is also the director of the Media Integration and Communication Center. He has published more then 300 publications in some of the most distinguished journals and conferences, and is the author of the monograph Visual Information Retrieval. He is an IAPR fellow and an associate editor of Multimedia Tools and Applications, Pattern Analysis and Applications, and International Journal of Image and Video Processing. He was general co-chair of ACM Multimedia 2010 and ECCV 2012.