International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

Harnessing the Social Annotations for Tag Refinement in Cultural Multimedia

Kirubai Dhanaraj*, Rajkumar Kannan
Department of Computer Science, Bishop Heber College, Tiruchirappalli, India

ABSTRACT

Videos are the source of social multimedia for the past few years and going to be the major source of all communications in the near future. On the other hand multimedia retrieval techniques lack in semantic context annotations for the video. Though the social media has numerous annotated tags and comments for similar image contents, it is not properly correlated with the context of the video retrieval techniques. In this paper we propose a method for video tag refinement and temporal localization for cultural multimedia. In this method the social annotations are exhibited to harness the temporal consistency of the video.

Keywords : Tag Refinement, Tag Localization, Temporal Consistency, Social Annotations, Multimedia Retrieval, SURF feature

I. INTRODUCTION

Cultural heritages are pride of any nation. Cultural memory institutions such as archives, museums, performing arts collections and libraries are more important repositories than ever. Social media is augmented with rich context such as user-provided tags, comments, geolocations, time, device metadata, and so on.

Videos are the source of social multimedia for the past few years and going to be the major source of all communications in the near future. On the other hand multimedia retrieval techniques lack in semantic context annotations for the content of the video. Though the social media has numerous annotated tags and comments for similar image contents, it is not properly correlated with the context of the video retrieval techniques. Features dealing with images will not thoroughly annotate the semantic content of the video. They need much understanding as human perception and views in form of social annotations.

In this paper we propose a method for video tag refinement and temporal localization for cultural videos based on social annotations. This method exploits the social annotation generated by the user for the videos to be explored in particular keyframe by tag relevance and visual consistency using SURF feature.

Some researchers are here which uses the social contexts to annotate and index multimedia. The performance of social image and video retrieval systems strictly depends on the availability and the quality of tags. However, these are often imprecise, ambiguous and overly personalized [1]. In the case of videos there is also another problem that the tags are not localized in the video frames. 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]. The main focus of these works has been put on three aspects: tag relevance (or ranking) [4], tag refinement (or completion) [5] and tag-to-region localization [6]. Among the others, nearest-neighbor based approaches have attracted much attention for image annotation [7,8], tag relevance estimation [9] and tag refinement [6]. The problem of video tagging so far has received less attention from the research community.

Table 1: Notations used in this paper

Variable	Meaning
V	Collection of videos and metadata (titles, tags, descriptions, etc.,)
D	Dictionary of tags to be used for annotation
v, kf	Video from D and a keyframe within v
T_v, T'_v	Set of tags associated to video v, prior and after tag refinement and localization process
I	Set of images downloaded from Facebook, Google, Bing and Flickr
I_t	Set of images from I annotated with the tag t
M_i, T_i	An image from I and their tags
T_{kf}	Set of tags associated to the keyframe kf
T_s	Set of S neighbor images for a given keyframe kf
H_{kf}	Integral Keyframe
$kf^{(tim)}$	A keyframe at a time

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 [9] or to analyze the popularity of social videos [18]; ii) video categorization, e.g. addressing actions and events [7, 13], genres [10] or YouTube categories [12]. However, the problem of video tagging *in the wild* remains open and it might have a great impact in many modern web applications.

The rest of the paper is organized as follows. The proposed method discussed in detail in section 2; experiments and results are presented in section 3. Finally the conclusions are drawn in section 4.

II. Approach

The framework of our system is schematically illustrated in Figure .1 and our notation is defined in Tab: 1. Consider a corpus V composed of videos and metadata (e.g: titles, descriptions, tags). Define D as a dictionary of tags to be used for annotation. Each video $v_i \in V$, with $T_v \subseteq D$, can be decomposed into different keyframes. Video tag refinement and localization is performed in two stages. In the first stage the video keyframes are extracted from the video using an automatic temporal segmentation tool. A relevance measure of the video tags is composed for each keyframe, after eliminating tags that are not relevant. In the second stage each keyframe of the video is annotated by retrieving images from online sources and proceeds to transfer labels across similar samples.

1.1 Retrieval Set

The collection of all tags $T_v = \{t_1 \dots t_i\}$ associated to a video V_i are filtered to eliminate stop words, dates, tags with numbers, punctuations and symbols. This resulting list of tags is used to retrieve a set of images $I = \{I_1, I_2, \dots\dots, I_m\}$from Google, Facebook and Flickr. By following this procedure an image$I_i \in I$, retrieved using t_j as query has the following set of tags$T_i = \{t_j, t'_1, \dots\dots, t'_z\}$if it has been obtained from Flickr or $T_i = \{t_j\}$ if it has been obtained from Google or Facebook. It is noticed that only the query term has been collected as a label since the images do not contain any other additional tag. So $D \supseteq T_v$ be the union of all the tags of the m images in I. This set D is then used in the following steps for tag localization on the video.

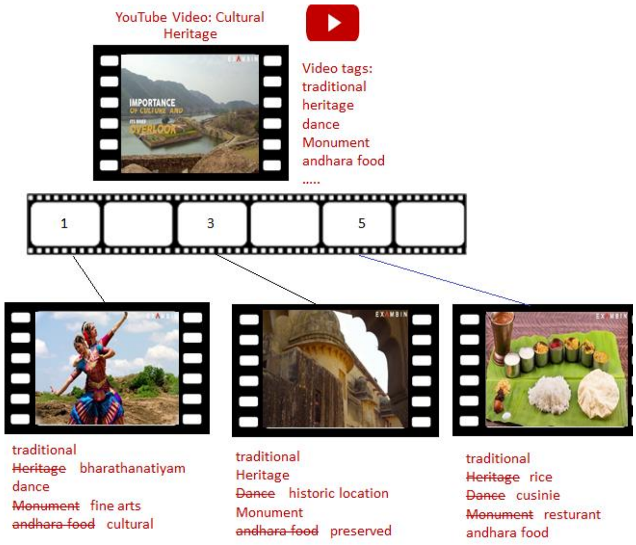


Figure 1 : Example of video localization: top) YouTube video with its tags; bottom) localization of tags in different keyframes.

Given the retrieval set I, for each keyframe kf within the video V finds a small set of k visual neighbors $I_k \subseteq I$. A good neighbor set will contain images that have similar scene types or objects (in our experiment - varied from 50 to 100 images). In the attempt to indirectly capture the similarity, we compute 200-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 [5], in order to address scalability issues. The SIFT feature examines several octaves and levels to detect features across scaling [Ref: surf_report.pdf]. The SURF algorithm uses progressively larger Gaussian kernels (eq.3) in the integral image to calculate the responses with arbitrary larger kernels.

1.2 Tag refinement and localization

A simple approach to annotate a keyframe kf is to consider only the tags belonging to the set of tag T_v that is associated to the video. Computing the tag relevance for each tag is to their rank their relevance w.r.t. the keyframe to be annotated. To solve this problem we adopt the following *tag-relevance* approach using visual neighborhood. Since visual neighborhood are the images tagged by social media users.

$$tag_{relevance(t,f,T_s)} = \frac{1}{S}\sum_{i=1}^{s} R(t,T_i) - \frac{|I_t|}{|I|} \qquad (1)$$

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

In our system we used a Hessian based blob detector to find the interest point in an integral keyframe. An Integral keyframe $H_{kf}(X)$ is a rectangular area between the origin and point X that stores the sum of all the pixels in a rectangular area $X = (x,y)^T$. Computing Integral keyframe $H_{kf}(X)$ for each keyframe in the video is as follows:

$$H_{kf}(X) = \sum_{i=0}^{i\le x}\sum_{j=0}^{j\le y} H(x,y), \ \ for\ all\ keyframes\ kf \qquad (3)$$

To find the weight of each vote between the integral keyframe and its neighbor the inverse of the square of Euclidean distance is used. The function $R(t,T_i)$ calculates the similarity between an integral keyframe and its visual neighbor. The $d(H_{kf}, M_i)$ is the Euclidean distance between the feature vectors of the integral keyframe H_{kf} and the image M_i.

$$R(t,T_i) = \begin{cases} \dfrac{1}{d(H_{kf},M_i)^{\,2}} & if\ \ t \in T_i \\ 0 \ \ otherwise \end{cases} \qquad (4)$$

In case a relevant is incorrectly eliminated, it may be recovered in the following stage of annotation. The algorithm computes tag relevance score and resulting rank position $rank_k$. Then, compute the co-occurrence for each tag in $T_{kf}$with all the tags in T_s. The tags that have a co-occurrence value above the average is selected from the resulting tag candidate list. Computing tag score for each candidate tag according to the $vote^+$ algorithm [2]. The tag score is computed as follows:

$$tag_{score}(t,kf) = tag_{score}(t,T_{kf}).\frac{\lambda}{\lambda+(rank_t-1)} \qquad (5)$$

where λ is a damping parameter set to 10. The performance in terms of precision and recall is maximized by the parameter λ. For all the candidate tags in actual keyframe kf, the above equation (5) is applied and the resulting 5 most relevant tags are selected. The union of all the tags selected for the keyframe level is used to annotate the video at the global level. These refined tags are referred as T'_v.

1.3 Tag Refinement with Temporal Consistency

Videos exhibit a strong temporal continuity in both visual content and semantics [13]. Harnessing this coherence is done by introducing a temporal smoothing to the tag_{score} function with respect to a tag. To compute the temporal continuity for each tag t and keyframekf, re-evaluate the tag_{score} function as below.

The keyframe kf at time $kf^{(tim)}$, and the dis the maximum temporal distance that the keyframes are considered; thus $kf^{(kf-i)}$ refers to the nearby keyframe at temporal distance i . The tag_{score} function is recomputed with temporal consistency as follows:

$$refinetag_{score}(t,kf) = \sum_{i=-dis}^{+dis} w_i \,.\, P\left(t^{(kf)} = 1 \middle| t^{(kf-i)} = 1\right) \qquad (6)$$

where w_i is a Gaussian weighting coefficient that satisfies $\sum_i w_i = 0.9$.

III. 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 qualitatively evaluate the performance of our system, we present experimental results for tag refinement and localization on public dataset and specifically with Indian cultural videos Dataset.

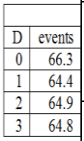
1.4 DUT-WEBV dataset

We have been conducted experiments on the DUT-WEBV dataset [11] which consists of a collection web videos collected from YouTube by issuing 31 tags as queries. It covers a wide range of semantic levels including *scenes, objects, events, people, activities and sites* and are in listed Tab. 2. There are 50 videos for each concept, and 2 videos are associated with two different tags. The total number of videos is 1,458. Our experimental setup follows the author of the dataset and the results are compared in section 3.4. Video frames are been extracted from video every two seconds, following

Table 2: DUT-WEBV dataset: list of tags with their corresponding category, number of frames containing a particular tag/concept and total no:of keyframes

Category	Tag	No.of Frames with Tag	No.of 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
	Newspaper	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	170,487

Table 3: Results of Tag Localization on DUT-WEBV dataset



D	events
0	66.3
1	64.4
2	64.9
3	64.8

the experimental setup by the authors of the dataset. It obtains 170,487 different frames. Images were obtained from different web sources namely Google Images, and from social network i.e., Flickr. The total number of all images retrieved is 58,832. Culminating all the video frames and the images retrieved from online is 229,319, which is comparable to the dimension of NUS_WIDE (the largest common dataset used for social image retrieval and annotation).

1.5 *Experiment 1: Tag Localization using DUT-WEBV Dataset*

Experiment 1 is conducted on the DUT-WEBV dataset relying only on the keyframes extracted from the web videos. This experimental setup is follows the approach used in the baseline provided with the dataset [5]. Given a particular tag t to be localized in a video v, extracts all the keyframes of the other videos associated to t and keyframes of 10 randomly selected videos associated to other 10 randomly selected tags from T_v. Similarly to the previous works [6], we use *Precision@N* and *Recall@N* to evaluate results (precision/recall at the top *N* ranked results).

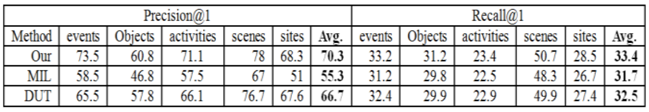
The tag relevance is computed using Eq.5 without weighting votes. The preliminary results are reported in Tab.4. We adopted the keyframes as an integral keyframes to improve the neighbor using SURF feature vector in Eq.3. As the visual neighbor increases we observation, that the performance also slightly improves, both in terms of precision and recall. Using weighting votes in Eq.4, the procedure obtain as improvement in recall around 3% and a loss in precision of more than 4%. The demand in tag localization task is in the terms of precision since a tag has not been recognized at a particular keyframe might be claimed in the following of the frames.

1.6 *Experiment 2: Tag Localization using Indian Cultural Video Dataset*

Our experiments have been conducted on Indian Cultural Videos that consists of a collection web videos collected from Facebook, YouTube, and Google by issuing tags for 4 categories as queries. These categories are listed in Tab. 5, covers a range of semantic levels including *dance, festivals and celebration, food and culture, Indian clothing.* There are 4 cultural category with 14 different videos for each video, total 56 videos. Video frames are automatically grouped by detected shots into semantically coherent temporal video fragments. It obtains 4,982 different keyframes from 56 videos. Images from search engine namely Google and Bing and from a social network Flickr are retrieved for the tags obtained from the video.

The collection of images retrieved is 5579. With this approach it become possible to tag keyframes showing specific dance (e.g. Bharathanatiyam, Bihu), festivals (eg. Christmas, Deepavali, Harvest Pongal), and food (Idli, Puri). The annotation performance has been evaluated in terms of Precision@5 and Precision@10, though manual inspection of each annotated frame by four different persons, and averaging the results. The obtained results are reported in Tab. 5, comparing the results with a baseline that randomly selects tags, with the probability proportional to their frequency in downloaded images.

Table 4(a): Results of tag localization using our method, DUT-WEBV, MIL-BPNET



	Precision@1						Recall@1					
Method	events	Objects	activities	scenes	sites	**Avg.**	events	Objects	activities	scenes	sites	**Avg.**
Our	73.5	60.8	71.1	78	68.3	**70.3**	33.2	31.2	23.4	50.7	28.5	**33.4**
MIL	58.5	46.8	57.5	67	51	**55.3**	31.2	29.8	22.5	48.3	26.7	**31.7**
DUT	65.5	57.8	66.1	76.7	67.6	**66.7**	32.4	29.9	22.9	49.9	27.4	**32.5**

1.7 *Comparison with previous works*

The authors of the dataset MIL-BPNET[5] proposed tag localization approaches that provides a reliable baseline. Our proposed method obtains better results than the base line method MIL-BPNET [5] for all tags

categories. On average our method outperforms the baseline with 8%. It is noticed that the results

Table 4(b): Comparison between our method and DUT-WEBV[11] and baseline MIL-BPNET[5].

Category	Tag	**Our**	DUT	MIL
Events	airplane flying	85.2	84.3	72.6
	Birthday	12.9	12.7	30.5
	Explosion	82.3	82.0	65.0
	Flood	71.2	69.6	55.0
	Riot	79.1	78.8	69.3
	Avg.	**66.1**	**65.5**	58.5
Objects	Cows	72.2	72.0	58.1
	Food	37.5	37.3	41.6
	golf player	73.4	72.6	38.6
	newspaper	59.6	58.2	41.6
	Suits	52.2	51.5	42.5
	telephones	60.2	59.7	53.4
	Truck	53.9	53.3	52.1
	Avg.	**58.4**	**57.8**	46.8
Activities	Baseball	92.1	91.2	66.9
	basketball	84.9	84.1	64.3
	cheering	97.4	96.3	58.2
	dancing	31.1	30.7	28.1
	handshaking	46.5	45.9	44.7
	interviews	71.9	71.2	61.8
	parade	68.3	67.8	69.4
	running	63.4	62.3	45.5
	singing	19.9	19.0	61.1
	soccer	83.3	82.6	76.3
	swimming	87.2	86.5	70.8
	walking	55.7	55.2	43.0
	Avg.	**66.8**	**66.1**	57.5
Scenes	Beach	85.6	85.0	70.5
	Forest	84.1	83.5	73.2
	mountain	62.1	61.6	57.4
	Avg.	**77.3**	**76.7**	67.0
Sites	aircraft cabin	75.9	75.4	51.9
	Airport	81.4	80.9	70.1
	gas station	41.8	41.1	23.5
	Highway	73.6	72.9	58.5
	Avg.	**68.2**	**67.6**	51.0
	Overall avg.	**67.4**	**66.7**	55.3

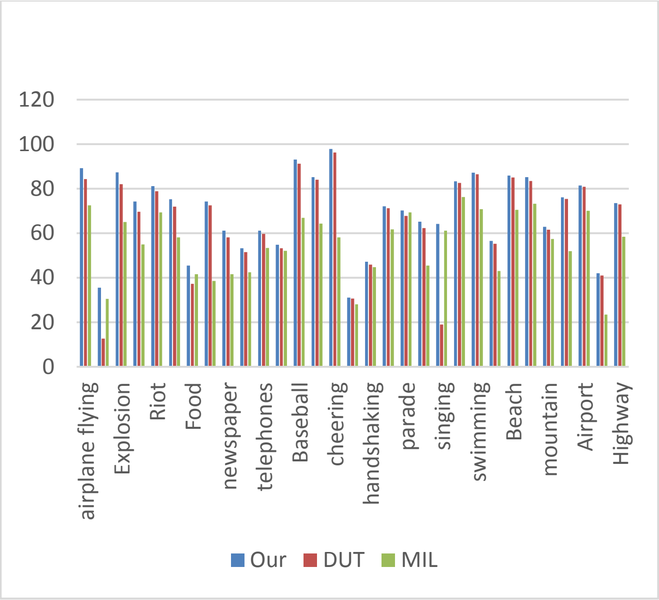


Table 5: Annotation "Indian Cultural Videos". Comparison between our method and the random baseline

Method	Precision@5	Precision@10
Random	6.3	4.7
Our	34.5	31.3

produced by our method are with *Precision@1* while the baselines were measured using *Precision@N,* and so our improvements should be considered even more. The results are reported in Tab. 4(a)

We also compare with a Lazy learning approach in DUT-WEBV[5] for tag refinement and localization using "in the wild" dataset. Similar to our method using only SIFT feature vector for neighbor voting with temporally subsampling the video for every 2 seconds. It can be noticed that using all the available image sources provides the best precision results of 65.7%. In terms of precision any combination of video and additional source performs better than the same source alone, but it is interesting to notice the good result of 62.5%in all social and web sources together. In recall the results has a main difference in using only video data by achieving 49.8% and any other combination that provides at most 29.4%. All these results are compared and listed in the Tab. 4(b).

IV.CONCLUSION

In this paper we have presented a tag refinement and localization approach based on Hessian based blob detector. Our system harnesses the collective knowledge embedded in user generated tags and visual similarity of the integral keyframes and images available in social media like Facebook and YouTube and web image sources like Google and Bing. This approach is improved from the baseline algorithm with temporal smoothing with an integral keyframe which is able to exploit the strong coherence normally present in the video.

V. 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. Sigurbj¨ornsson, R. van Zwol, Flickr tag recommendation based on collective knowledge, in: Proc. Of WWW, Beijing, China, 2008, pp. 327-336.

[3]. 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).

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

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

[6]. 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.

[7]. 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.

[8]. 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.

[9]. 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.

[10]. 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.

[11]. 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.

[12]. 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.

[13]. 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.