© 2018 IJSRSET | Volume 4 | Issue 4 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Capitalizing The Collective Knowledge for Video Annotation Refinement using Dynamic Weighted Voting

Kirubai Dhanaraj, Rajkumar Kannan
Research Department of Computer Science Bishop Heber College, Tiruchirappalli, Tamil Nadu, India

ABSTRACT

Collective knowledge paves the way for most challenging task to be an interesting and improving efficiency in the field of multimedia annotations and retrievals. Automatic annotation is bridging the gap between low-level content and high-level semantic concepts. It has been an active research area in the field of multimedia retrieval, machine learning and social media environments. Even most automatic annotation approaches are often unsatisfactory, the annotation refinement has invited the attention of recent researchers. In this paper, a novel refinement algorithm is proposed using dynamic weighted voting based on mutual information. It leverages the collective knowledge of the social media like collection of videos, images, texts in the form of tags, and comments available online. The proposed algorithm invests collective knowledge to measure the relevance between the candidate annotations by assessing the probability and assigning a dynamic weights.

Keywords : Annotation Refinement, Collective knowledge, dynamic-weighted voting, SURF feature, Multimedia annotation

I. INTRODUCTION

Multimedia digital repository has enormous videos that are associated with user generated tags, comments and links. Automatically annotating a video can extract semantic features that are identified by incorporating the user generated tags, keywords, textual descriptions. These semantic features are limited in two ways: Firstly, semantic features cannot perceive the contextual meaning of the entire video, which can be understood only by humans. Secondly, semantic feature extraction may identify the number of persons and even their names using face recognition but the role played by them, mood and climate of the video content can only be felt and guessed by the humans. On the other hand social annotations generated by the user in social media can be incorporated with multimedia annotations for it reflects the user perception. These collective knowledge represents the user interest rather than the semantic concepts or semantic features.

Currently the performance of image and video retrieval system depends mainly on the availability and the quality of the tags. However existing studies show that tags are few, imprecise, ambiguous and overly personalized [1]. The automatic annotation approaches finds the concept similarity [2, 3, 4] between the image and labels or tags. The concept similarity is obtained by finding the visual similarity by low-level and semantic features of the image or video using the training samples. They cluster the similar concept images and propagate to new images of visually similar images [2, 3, 4].

Label propagation through graph construction [5, 6], neighbourhood propagation [4, 7, 8] for labelling the nearest neighbour, random walks [6, 7, 8] to find the neighbourhood through hierarchically [8, 9] dividing partitioning the graph. On the other hand computer vision techniques [10] like object recognition, and face recognition [10] are also increasing but they depend on semantic annotated accurate training samples. The labels or tags that are generated by the automatic annotation are not most relevant when comparing to the tags that user generates for the same image in the internet. Social tags are better than the automatic annotations for training the large-scale images samples [6, 7, 8]. Collaborative annotation approaches [2, 4] leverages these user tags to find the semantic relationship between the image content and the tags.

Collective knowledge generated by the user in social media tags to construct web-scale image graphs [5, 8] that represents semantically similar images, finding the tag relevance [2, 9] using semantic tag similarity and to improve the tag quality approaches like tag recommendation [2, 4], tag refinement [5], tag filtering [3, 7] are used. Collaborative tag depends on social user interest and their use the vocabulary by their choice. These user generated tags may not properly describe the content of the multimedia and sometimes they are irrelevant, negatively annotate, and have noisy tags. To refine the social annotations and to enhance the quality tag processing during tagging such as tag recommendation [4, 5] or after tagging such as tag refinement [3, 4] and re-ranking are the state-of-art approaches in social multimedia annotation and retrieval.

In this paper we propose a refinement algorithm that measures the relevance between integral images of the video and semantically similar images in social media. Rest of the paper is organized as follows: Section 2 describes the refinement algorithm based on collective knowledge using Hessian interest points. In section 3 the experimental results are explained. The section 4 draws the conclusion.

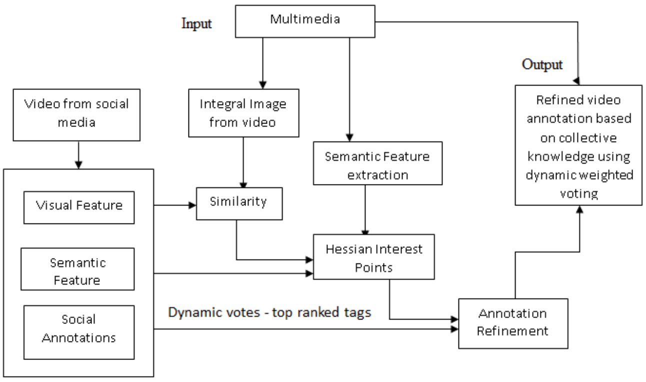

Input
Multimedia
Output
Video from social media
Integral Image from video
Semantic Feature extraction
Refined video annotation based on collective knowledge using dynamic weighted voting
Visual Feature
Similarity
Semantic Feature
Hessian Interest Points
Social Annotations
Dynamic votes - top ranked tags
Annotation Refinement

Figure 1. Overview of the proposed approach

II. VIDEO ANNOTATION REFINEMENT USING COLLECTIVE KNOWLEDGE

The human level of perception and understanding of images and videos depends on the context of the video, prior knowledge, imagination, details about specific task or person instead of object, region, and person. Motivated by the human perception and collective knowledge prevails in the social media, we propose a novel algorithm for video annotation refinement using dynamic weighted voting based on Hessian interest points. The proposed refinement algorithm in this paper aims at two goals: firstly, to annotate the video with interesting points at keyframe level known as integral image. Secondly, to refine these annotations with dynamic weighted votes computed by visual similarity between the integral image and online images.

The first goal relates to the fact that the whole video is not needed to be annotated rather the interest points in the keyframe exhibit the content. The second goal is related to the fact that the annotations for the integral keyframes in the same video has different relevance measure when compared with similar images online. To cope up these challenges, this paper proposes a novel algorithm for video annotation based on Hessian interest points and an interactive refinement algorithm based on dynamic weighted voting. An overview of the proposed approach is shown in figure 1.

A. VIDEO ANNOTATION ALGORITHM BASED ON HESSIAN INTERST POINTS

Video annotation is performed by adopting an improved relevance measure between interesting points of video keyframe and visually similar images from the web. The relevance score is computed by the probability of producing candidate annotation using frequency of the annotation terms and visual similarity of the candidate image to that of integral keyframes. The proposed annotation algorithm can be described as follows:

1. Extract keyframes kf from the video v.
2. Segment the keyframe kf into interest points as an integral image $I_{kf}(X)$ where $X = (x, y)^T$ stores the sum of all pixels in a rectangular area between origin and X.
3. Compute the feature vector for the integral images for each keyframe. $[d_1, \dots, d_{72}]$.
4. Let the $freq_{a,b}$ = the raw frequency of the integral image $I_{kf}(X)$ in the keyframe kf is given by
$$f_{a,b} = \frac{freq_{a,b}}{\max freq_{a,b}}$$
5. The term weight is calculated for each keyframe using $W_{a,b} = f_{a,b} * log\frac{N}{n_i}$, where N is the size of training image in the dataset and n_i is the number of images in which the integral image $I_{kf}(X)$ appears.
6. Compute the candidate annotation according to the relevance model in [2] $tagRelevance(W, I, kf) = n_w[N_f(I, Kf)] - Prior(W, kf)$
7. Assign the probability of each candidate annotation to the weight of annotation using [1]
$$P(y_{iw} = +1) = \sum_j \pi_{ij} P(y_{iw} = +1|j)$$
$$\pi_{ij} \geq 0 \wedge \sum\nolimits_j \pi_{ij} = 1$$

$$P(y_{iw} = +1) = \begin{cases} 1 - \epsilon & for \quad y_{iw} = +1 \\ \epsilon & otherwise. \end{cases}$$

Where $y_{iw} \in \{+1, -1\}$ indicates whether tag w is relevant or not for image i and π_{ij} is the weight of image j (from the visual neighbors) in respect to image i to be learned.

8. Transfer the top N highest weight tags.

B. ANNOTATION REFINEMENT ALGORITHM USING DYNAMIC WEIGHTED VOTING

Based on the fact that visually similar images has been semantically-related to each other when their weights or rank is high [8]. The visual consistency depends on visual feature selection on regions with similar images. To improve the accuracy of the annotation this paper proposes an annotation refinement algorithm using dynamic weighted voting with Hessian interest points. These interest points form an integral image for each keyframe that reflects close visual consistency between higher weighted nearest neighbour votes. In each integral image the SURF descriptor [4] describes an interest area with size 20s. The interest area is divided into 4 X 4 sub areas that is described by the values of a wavelet response in x and y directions. The interest area are weighted with a Gaussian center at the interest point to give robustness. The proposed refinement algorithm using dynamically calculated weights can be described as follows:

A) Segment the keyframe into integral images for every interest points
$$I(X) = \sum_{i=0}^{i<x} \sum_{j=0}^{j<y} I(x, y) \tag{1}$$
B) Construct the SURF feature vector for integral image with 72-dimensions.
C) Calculate the weight w with Gaussian kernel, using second order Gaussian kernels $\frac{\partial^2}{\partial y^2} g(\sigma)$ can be given as
$$\mathcal{H}(\mathrm{X}, \sigma) = \begin{bmatrix} L_{xx}(X, \sigma) & L_{xy}(X, \sigma) \\ L_{xy}(X, \sigma) & L_{yy}(X, \sigma) \end{bmatrix} \tag{2}$$
Where
$$L_{xx}(X, \sigma) = I(X) * \frac{\partial^2}{\partial x^2} g(\sigma)$$
$$L_{xy}(X, \sigma) = I(X) * \frac{\partial^2}{\partial xy} g(\sigma)$$
D) Construct graph G , whose vertex is corresponding to Integral image.

E) Compute weighted adjacency matrix W, whose edge weight $w_{i,j}$ is similarity and distance metric of node i and j corresponding to integral image x and y respectively.

F) Compare the unnormalized Laplacian L using

$$\nabla^2 L = tr(\mathcal{H}) = L_{xx}(X,\sigma) + L_{xx}(X,\sigma) \qquad (3)$$

G) Let $V \in R^{nXm}$ be the matrix containing the vectors $V_1, \ldots\ldots, V_m$ as columns.

H) Cluster the points $(Y_i)_{i=i,\ldots,n}$ in R^m with the k-means algorithm into clusters $C_1, \ldots \ldots, C_k$.

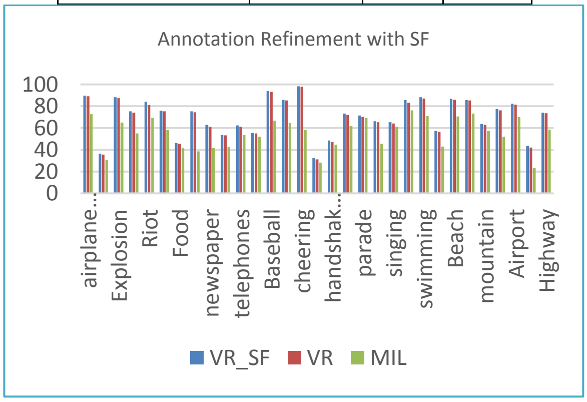
The Laplacian is the base of the hessian matrix (equation.3), and when calculating the determinant of the hessian matrix these values are available. It is a matter of storing the sign. The reason to store the sign of the Laplacian is that distinguishes between bright blobs on dark backgrounds and vice versa. It is only necessary to compare the full descriptor vector if they have same sign, which can lower the computational cost of matching.

III. EXPERIMENTAL RESULTS

To evaluate the proposed algorithm we experimented on the DUT-WEBV dataset [11] which consists of a collection of web videos collected from YouTube by issuing 31 tags as queries. The results obtained from our approach is compared with our previous method without SURF feature, and DUT-WEBV results. The results are shown in the table 1. The annotations obtained from refinement algorithm outperforms with the previous and baseline results.

Table 1. Comparison with our previous work and DUT-WEBV

Tag	**VR_SF**	VR	DUT
airplane flying	89.7	89.2	72.6
Birthday	36.5	35.5	30.5
Explosion	88.3	87.3	65
Flood	75.3	74.2	55
Riot	84.1	81.1	69.3
Cows	75.9	75.2	58.1
Food	46.2	45.5	41.6
golf player	75.4	74.3	38.6
Newspaper	62.8	61.2	41.6
Suits	53.9	53.2	42.5
telephones	62.4	61.2	53.4
Truck	55.7	54.9	52.1
Baseball	93.9	93.1	66.9
basketball	85.9	85.2	64.3
cheering	98.3	97.9	58.2
dancing	32.6	31.1	28.1
handshaking	48.4	47.2	44.7
interviews	73.2	72.1	61.8
parade	71.5	70.2	69.4
running	66.3	65.2	45.5
singing	65.3	64.2	61.1
soccer	85.7	83.3	76.3
swimming	88.2	87.2	70.8
walking	57.4	56.5	43
Beach	86.9	85.9	70.5
Forest	85.7	85.2	73.2
mountain	63.5	62.9	57.4
aircraft cabin	77.4	76.2	51.9
Airport	82.3	81.4	70.1
gas station	43.4	42.1	23.5
Highway	74.1	73.6	58.5



IV. CONCLUSION

The automatic video annotation refinement with dynamic weighted voting algorithm in this paper refines the annotations and improves the accuracy of annotation. It improves the performance of annotation regarding faster annotations when compared to other annotation methods. This can be combined with any retrieval platforms to improve the accuracy of annotation. In future, tag enrichment can be done to improve the informativeness which helps to study the affective influence of the multimedia content.

V. REFERENCES

[1]. M.Guillaumin,T.Mensink,J.Verbeek and C.Schmid.TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation.In Proc.Of ICCV,2009.

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

[3]. L.Ballan,T.Urricho,and A.Del Bimbo.2014.A cross-media Model for Automatic Image Annotation.In Proc.of ACM ICMR.73-80

[4]. Kirubai Dhanaraj,Rajkumar Kannan,Harnessing the Social Annotations for Tag Refinement in Cultural Multimedia,IJSRCEIT,2018,pp.1802-1808.

[5]. T.Uriccho,L.Ballan,M.Beritini,and A.Del Bimbo,"An evalution of nearest-neighbor methods for tag refinement",in Proc.of IEEE International conference on multimedia & Expo (ICME),San Jose,CA,USA,2013.

[6]. Emily Moxley,TaoMei,B.S.Manjunath,Video Annotation Through Search and Graph Reinforcement Mining,Published in IEEE Transaction on Multimedia Vol.12,No.3 April 2010 pp 184 – 193.

[7]. 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.

[8]. Z.Qian,P.Zhong,and R.Wang.2015.Tag refinement for user-contributed images via graph learning and nonnegative tensor factorization.IEEE Signal Processing Letters 22,9(2015),35-62.

[9]. Kirubai Dhanaraj,Rajkumar Kannan,A State-of-the are Review: A Survey on Multimedia Tagging Techniques,IJSRST Volume 4,Issue 5,pp 377-386,2018.

[10]. R.Kannan,G.Ghinea,S.Swaninathan,Salient region detection using patch level and region level image abstractions,2015,IEEE,Signal Processing Letters 22(6),pp 686-690.

[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.