Large-scale Web Video Shot Ranking Based on Visual Features and Tag Co-occurrence

Do Hang Nga Keiji Yanai

Department of Informatics, The University of Electro-Communications
Chofu, Tokyo 182-8585 JAPAN
{dohang,yanai}@mm.cs.uec.ac.jp

ABSTRACT

In this paper, we propose a novel ranking method, VisualTextualRank, which extends [1] and [2]. Our method is based on random walk over bipartite graph to integrate visual information of video shots and tag information of Web videos effectively. Note that instead of treating the textual information as an additional feature for shot ranking, we explore the mutual reinforcement between shots and textual information of their corresponding videos to improve shot ranking. We apply our proposed method to the system of extracting automatically relevant video shots of specific actions from Web videos [3]. Based on our experimental results, we demonstrate that our ranking method can improve the performance of video shot retrieval.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Design, Experimentation

Keywords

Web Video, Ranking, Bipartite Graph

1. INTRODUCTION AND RELATED WORKS

Given an action keyword such as “running” or “watching television” provided by a user, an online video search engine is supposed to rank the Web videos in its database according to their relevance to the query and then return the most relevant ones at the top. Howerver, in general, the relevant videos account for only a small portion of the videos retrieved by conventional video search engines. Moreover, even relevant videos may contain many unrelated video shots since Web videos belong to UGC (User Generated Content) and are created for various purposes. Here, a video shot is a part of a video which refers to a set of consecutive frames representing a specific scene. Video shot retrieval for specific actions must encounter many difficulties including computation cost, noise, diversity of keywords as well as human actions and so on.

In case of image retrieval, the popular Google Image search engine adopts a ranking method called VisualRank [1] which exploits the visual link structure between images. According to VisualRank, images found to share the most visual characteristics with the group at large shall be determined as the most relevant ones and brought to the top of search results. VisualRank was applied in our previous works [3, 4]. With VisualRank, we succeeded in retrieving relevant video shots for many action categories. However, the problem is that, by applying VisualRank, solely visual relationships between shots are explored, thus we obtained at the top only video shots which have similar appearances. This causes the loss of variety in the results. Especially, in some cases, most of those top ranked video shots do not correspond to the given action keywords even though they are visually related.

Since human actions are too diverse, their corresponding video shots are not always visually similar even if they are semantically related. The change in camera view or the way how people perform the action may cause visual differences. Our intuition is that, two video shots which belong to two videos tagged with related keywords may represent the same action even if they do not hold the same visual features (See Fig.1). Hence, shot ranking should additionally consider tag information. Besides, tags are supposed to be more efficiently adopted if their relevance is evaluated considering not only their intra-relationships but also their correlation with video shots. For example, if we find that a video shot is important, or in other words, related to the given action keyword, so that the tags of the video have high chance to be important as well. And vice versa, if a tag was found as being relevant to the keyword, it is highly probable that the videos annotated with it are also relevant.

As efforts on tag ranking considering their relevance, Yang et al. [5] proposed a method to evaluate tag relevance scores of all tags based on tag co-occurrence statistics. Dong et al. [6] proposed a method to measure tag relevance scores by combining the probabilistic relevance score estimation and random-walk based refinement. Especially, Liu et al. [2] presented a Web video topic discovery and tracking method via bipartite graph which represents the correlation between videos and their tags. Actually, our idea is inspired by their work. However, the objective as well as the methodology of their work are different from ours. While they tried to find relevant videos of a topic, we aim to detect relevant video shots of a specific action. In terms of methodology, they used only textual information, while we use both textual and visual features.

In this paper, we propose a novel ranking method, VisualTextualRank, which extends [1] and [2]. Our method is

ACM Multimedia'13 October 21-25, 2013, Barcelona, Spain.

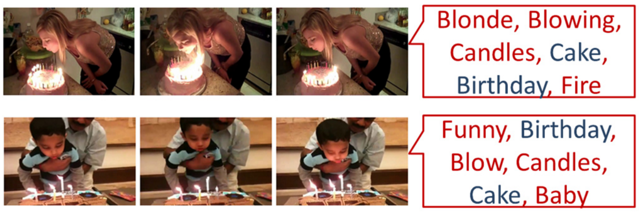

Figure 1: An example of Web video retrieval result. This figure shows two video shots together with tag lists of their videos which are retrieved by YouTube with "blow candle" keyword. We can see that some relevant words such as "birthday" and "cake" are tagged to both videos. Thus we can presume that these two video shots are semantically related to each other and relevant to "blow candle" even though they are not visually similar.

based on a random walk over bipartite graph to integrate visual information of video shots and tag information of Web videos effectively. Note that instead of treating the textual information as an additional feature for shot ranking, we explore the mutual reinforcement between shots and textual information of their corresponding videos to improve shot ranking. We apply our proposed method to the system of extracting automatically relevant video shots of specific actions from Web videos [3]. Based on our experimental results, we demonstrate that our ranking method can improve the performance of video shot retrieval. Our contribution is a co-analysis of visual links among video shots along with textual link between videos and their tags and its application to the learning of semantic similarities of video shots.

The reminder of this paper is organized as follows: In the next section we shall introduce our proposed method. In Section 3, we describe how we apply it to the system of video shot extraction. We present conducted experiments and show their results in Section 4. Conclusion and further discussion are presented in Section 5.

2. PROPOSED METHOD

The basic idea of VisualTextualRank (abbreviated as VTR) is that, the relevant tags are used to annotate relevant videos; the relevant video shots are from videos annotated with relevant tags and visually similar to each other. Thus VTR coranks tags and video shots so that at each iterative ranking step, ranks of shots are refined using their visual similarities as well as their relevance with corresponding tags, and then, ranks of tags are updated based on their relevance with video shots in conjunction with refined ranking positions of video shots. Figure 2 sketches the idea of VTR.

VTR is an extension of VisualRank [1] with idea motivated by [2]. In [2], tags and videos are also co-ranked using their correlation to refine their relevance with a specific topic. However, unlike our work, in [2], relevance of the whole video, not every scene in it, is evaluated and visual features of videos are totally ignored. On the other hand, VisualRank exploits only a visual linkage between images and does not take textual information into account. In VisualRank, the ranking position of the image which looks similar to many images with high ranking position becomes higher after iterative processing. Our proposed VTR employs both visual and textual features of Web videos to explore the mutual reinforcement across video shots and tags.

The proposed co-ranking method can be represented by following iterative processes:

$$\mathbf{RS_k} = \alpha \times SM^* \times SC^* \times \mathbf{RT_k} + (\mathbf{1}-\alpha)\ \mathbf{p} \quad (1)$$

$$\mathbf{RT_{k+1}} = (SC')^* \times \mathbf{RS_k} \quad (2)$$

RS and RT are vectors which represent rank positions of shots and tags, respectively. Let the number of shots be n_s and the number of tags be n_t, the dimension of RS will be $n_s \times 1$ and the dimension of RT will be $n_t \times 1$. SM refers to shot-shot similarity matrix where $SM_{i,j}$ means visual similarity score between shot i and shot j; SM^* is its column-normalized matrix with size as $n_s \times n_s$. SC represents shot-tag similarity matrix where $SC_{i,k}$ measures textual relevance score between the video of shot i and tag k; SC^* is its $n_s \times n_t$ column-normalized matrix. SC' refers to the transposed matrix of SC which represents tag-shot similarity matrix and SC'^* is its column-normalized matrix. Note that since the textual features, here refer to tag co-occurrence, are considered as being noisier than content-based features,we rank video shots first and use their refined ranking positions to update ranks of tags.

RT is initially defined as a uniform vector. At each ranking step, after ranking positions of video shots are updated based on their visual similarities and their correlation with tags following Eq.1, video shots cast their votes for tags through Eq.2. Thus relevant shots will cast important votes for tags which are strongly connected with them. And then at the next iterative step, those tags again help boost ranking positions for video shots which are tight linked with them. Gradually, video shots and tags with few important votes will go to the bottom.

Following VisualRank, we also introduce damping factor α and damping vector p into shot ranking. Damping factor α has been found empirically as holding minor impact on global ordering in ranking results [1, 3]. $\alpha \geq 0.8$ is often chosen for practice. Damping vector p can be a uniform vector or a nonuniform vector. For example, we can use a nonuniform damping vector as in [4] to bias shots which are visually related to relevant images during ranking computation.

3. APPLICATION AND IMPLEMENTATION DETAILS

We apply our proposed ranking method to our system of automatically extracting from tagged Web videos video shots corresponding to specific actions proposed in [3]. Our system consists of two main steps: video ranking and shot ranking. At the shot ranking step, in [3] we applied VisualRank to rank shots from top ranked videos. In this paper, we adopt our ranking method to this step to compare the effectiveness of our method and VisualRank.

The implementation details are described as follows. At first, we collect video IDs and tags for at most 1000 Web videos of search results of the action keyword via YouTube API. Tags here refer to words in retrieved values for "keyword" and "title" fields. The co-occurrence frequencies among tags are then exploited to build a tag database which provides us tag co-occurence based inter- and intra-relationships between tags and videos. Note that while in [3], this tag database is only adopted to measure relevance between videos and the action keyword, here it is also employed to calculate relevance between videos and tags.

In video ranking, videos are ranked in the descending order of their tag relevance scores with the given keyword. Tag relevance score of a video V to a word t is calculated by the

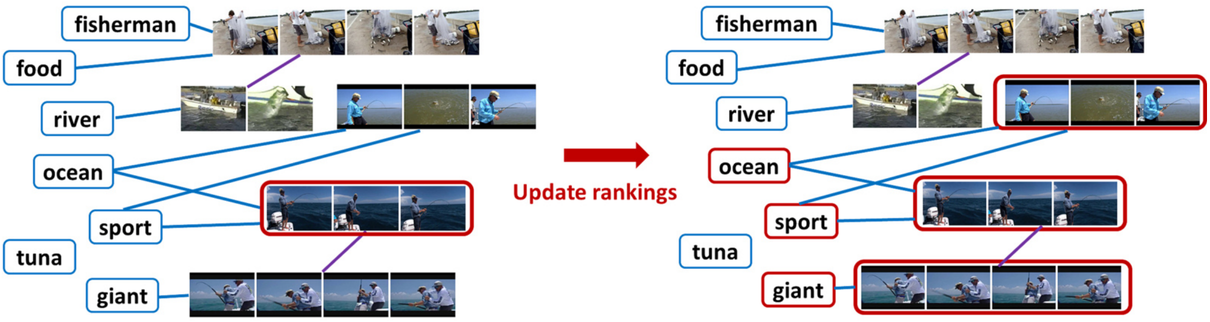

Figure 2: Illustration of VisualTextualRank by an example of "catch fish" action. Blue links represent relevance between video shots and tags. Purple links refer to visual relationships between shots. Objects marked with red bounding box are considered as being important. Assume that at first we found one important shots as shown at the left of this figure. It will cast its vote for shots and tags which are strongly linked with it. And then at the next step of ranking process, those shots and tags again cast their votes for objects which are tight connected with them. Finally, we can obtain relevant objects of "catch fish" as seen at the left of this figure.

following equation.

$$Sc_t(V) \quad = \quad \frac{1}{|\mathcal{T}|}\sum_{t_i \in \mathcal{T}} \log_2 F(t, t_i) \qquad (3)$$

$F(t, t_i)$ is the number of videos tagged with both word t and word t_i so it represents the co-occurrence frequency of t and t_i. $\mathcal{T}$ is a set of the words other than t which are tagged to V and frequently co-appear with t. The maximum size of $\mathcal{T}$ is set to 10. As the result of the video ranking step, the videos tagged with many words which have high co-occurence frequency with the action keyword will be ranked to the top.

Only the top ranked videos are selected to be downloaded and used since they are considered as being more related to the action keywords than the remains. These videos are then segmented into video shots based on their RGB histograms. Finally, shot rankings are iteratively calculated by applying our proposed method, VisualTextualRank.

To calculate ranking positions of shots in VisualTextualRank, we must construct shot-shot similarity matrix SM and shot-tag similarity matrix SC as shown in Eq.1. The similarity between two shots is calculated by histogram section of their bags of spatio-temporal features. We use the same method of spatio-temporal feature extraction described in [3]. Relevance of a video to a tag is measured in the similar way as represented in Eq.3 using the tag database constructed in advance. Note that here shots are obtained by segmenting selected videos but filtered according to their length and tags are tags of selected videos but filtered based on their occurrence frequencies. We select only shots which last longer than one second and shorter than one minute. To avoid using personal tags, we choose tags which appear at least five times over selected videos.

Damping factor is set to 0.8 and damping vector is defined following the best results obtained in [3]. That means damping vector here is a nonuniform vector derived from tag relevance of the videos to the keyword as follows.

$$p_i \quad = \quad \begin{cases} 1/k & (i \leq k) \\ 0 & (i > k) \end{cases} \qquad (4)$$

As shown above, damping vector is defined by giving uniform bias values to the elements corresponding to the top k shots regarding tag relevance of their videos to the keyword. k equals 1000 in our experiments.

Table 1: Experimental results. VR and VTR refer to performance of video shot retrieval system adopting VisualRank and proposed VisualTextualRank respectively.

Action	VR	VTR
soccer+dribble	100	100
fold+origami	96	99
crochet+hat	95	97
arrange+flower	94	96
paint+picture	88	87
boxing	86	84
jump+parachute	82	63
jump+trampoline	82	92
do+exercise	79	61
do+aerobics	78	79
do+yoga	77	70
surf+wave	75	73
shoot+arrow	73	81
massage+leg	72	78
fix+tire	67	77
batting	66	61
basketball+dribble	64	87
blow-dry+hair	64	59
knit+sweater	64	68
ride+bicycle	62	70
curl+bicep	58	59
shoot+ball	58	58
tie+shoelace	57	73
laugh	50	54
dive+sea	49	41
harvest+rice	49	46
ski	49	60
iron+clothes	47	48
twist+crunch	47	32
dance+flamenco	45	53
dance+hiphop	43	68
eat+ramen	42	47
dance+tango	41	41
play+trumpet	41	59
AVG. (1-34)	**65.9**	**68.3**

Action	VR	VTR
play+drum	40	45
skate	37	42
swim+crawl	36	49
cut+hair	35	42
run+marathon	35	43
count+money	33	58
paint+wall	33	32
shoot+football	33	29
draw+eyebrows	32	32
fieldhockey+dribble	32	68
hit+golfball	32	70
lunge	32	27
play+piano	32	34
row+boat	32	23
sing	32	65
chat+friend	31	52
clean+floor	31	38
cut+onion	31	24
shave+mustache	31	30
pick+lock	30	28
plaster+wall	30	55
blow+candle	29	44
wash+face	29	24
walking+street	29	46
brush+teeth	28	27
catch+fish	28	59
drive+car	28	40
plant+flower	28	24
play+guitar	28	41
lift+weight	27	51
raise+leg	27	40
hang+wallpaper	26	46
jump+rope	26	49
AVG. (35-67)	**31.0**	**41.7**

Action	VR	VTR
climb+tree	24	24
ride+horse	24	15
roll+makizushi	24	36
sew+button	24	46
fry+tempura	23	12
slap+face	20	45
read+book	19	21
squat	19	34
row+dumbell	16	30
wash+clothes	15	29
wash+dishes	15	39
comb+hair	14	26
drink+coffee	14	16
swim+breaststroke	13	18
cry	12	12
eat+sushi	12	23
serve+teniss	11	27
tie+necktie	11	28
boil+egg	9	11
head+ball	9	16
swim+backstroke	9	9
take+medicine	8	7
serve+volleyball	7	40
swim+butterfly	7	9
bake+bread	6	8
cook+rice	6	11
grill+fish	5	13
jog	5	6
slice+apple	5	16
peel+apple	5	14
bowl+ball	4	4
smile	4	6
kiss	2	3
AVG. (68-100)	**12.2**	**19.8**
AVG. (ALL)	**36.6**	**43.5**

4. EXPERIMENTS

We conduct experiments on our human action database described in [3]. This database consists of 100 action categories. Each category has 2000 video shots on average. Precision is defined as the percentage of relevant video shots in the top ranked 100 shots (Prec@100). We consider action with precision higher than 40% as "succeeded action", action with precision lower than or equal to 40% but higher than 25% as "acceptable action" and the remain as "failed action". The results reported in [3] are: 34 succeeded, 33 acceptable, 34 failed. We want to compare the performance of our video shot retrival system proposed in [3] adopting VisualRank and VisualTextualRank proposed in this paper. Experimental results are shown in Table 1.

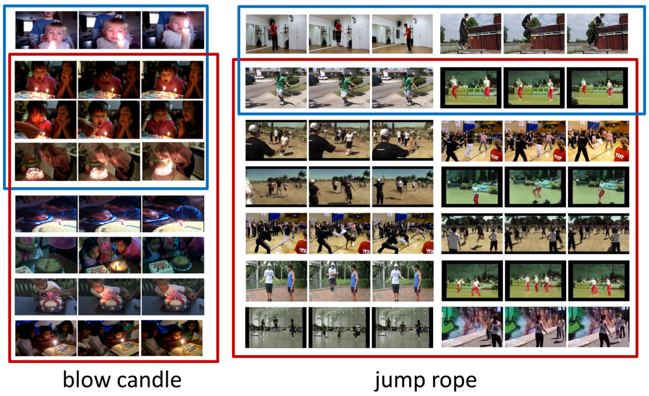

Figure 3: Relevant shots among top ranked 15 shots. Bounded by blue boxes and red boxes are respectively results obtained by applying VisualRank and our proposed VisualTexutualRank. The results show that our ranking method helps to boost more relevant shot to the top.

Experimental results demonstrate that by adopting our proposed ranking method, more relevant shots are brought to the top. In terms of overall performance, VTR improves the average precision by approximately 7%. Fig.3 shows some examples of detected relevant shots by applying VisualRank and our VisualTextualRank.

Table 1 shows that VTR enhances video shot retrieval system on most of categories. Especially, precision is boosted greatly in cases such as "hit+golfball", "dance+hiphop", "plaster+wall", "blow+candle", "jump+rope", "catch+fish", "play+guitar", "wash+dishes" and "slap+face". The acceptable group is the most significantly improved. By applying proposed VTR, the number of succeeded actions increases from 34 to 51 and the number of failed ones decrease from 34 to 23.

Not only precision, VTR also improves the performance of the shot retrieval system in the sense that it increases the variety of final results. Since VisualRank employs only visual features, visually similar elements are often ranked to the top. On the other hand, VTR additionally exploits the correlation between videos and tags so that not only visually similar video shots but also video shots having strong textual links with relevant shots also have chances to be ranked high as well (See Fig.4).

5. CONCLUSION

In this paper, we propose a novel graph based ranking method, VisualTextualRank, which performs co-ranking of video shots and tags employing both visual links between video shots along with textual links between videos and their tags. We apply VTR to the system of extracting automatically relevant video shots for specific human actions. The effectiveness of proposed VTR was validated by experimental results.

As for future work, we intend to improve our VTR by introducing more visual information such as appearance of objects. Most of human actions are associated with particular objects. For example, in general, "eat" scenes include "tableware" or in scenes of "type" action, "keyboard" is supposed to be seen. Hence we expect that by applying human-object interaction models proposed by [7] or [8], our VTR can achieve better performance.

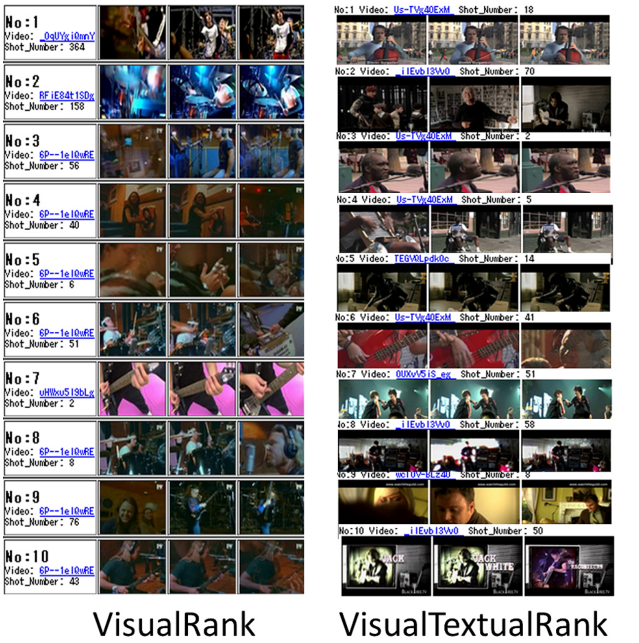

Figure 4: Diversity of results obtained by video shot retrieval system with VisualRank (right) and with VisualTextualRank (left). The category here is "play+guitar". In the original framework [3], more than half of top 10 shots are from the same video with ID "6P–1elQwRE" since they are visually similar. On the other hand, applying VisualTextualRank can select relevant shots from different videos since it regards not only visual similarities but also textual similarities of the shots.

6. REFERENCES

[1] Y. Jing and S. Baluja, "Visualrank: Applying pagerank to large-scale image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1870–1890, 2008.

[2] L. Liu, L. Sun, Y. Rui, Y. Shi, and S. Yang, "Web video topic discovery and tracking via bipartite graph reinforcement model," in *Proc. of the ACM International World Wide Web Conference*, pp. 1009–1018, 2008.

[3] D. H. Nga and K. Yanai, "Automatic construction of an action video shot database using web videos," in *Proc. of IEEE International Conference on Computer Vision*, 2011.

[4] D. H. Nga and K. Yanai, "Automatic collection of web video shots corresponding to specific actions using web images," in *Proc. of CVPR Workshop on Large-Scale Video Search and Mining (LSVSM)*, 2012.

[5] Q. Yang, X. Chen, and G. Wang, "Web 2.0 dictionary," in *Proc. of ACM International Conference on Image and Video Retrieval*, pp. 591–600, 2008.

[6] J. Deng, W. Dong, R. Socher, J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 6 2009.

[7] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 17–24, 2010.

[8] A. Gupta, A. Kembhavi, and L.S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1775–1789, 2009.