VisualTextualRank: A Video Shot Ranking Method Using Visual Similarity and Tag Co-occurrence

Do Hang Nga[1,a)] Keiji YANAI[1,b)]

1. Introduction and Objective

The explosive growth of video sharing websites makes it easier for researchers of action recognition field to construct action shot database. However, video shot retrieval for specific actions still encounters many difficulties including computation cost, noise, diversity of keywords as well as human actions and so on. Here, a video shot is a part of a video which refers to a set of consecutive frames representing a specific scene.

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 ranked to the top of search results. VisualRank can also be applied to video shot ranking as in [3], [4].

Howerver, in case of human actions, since they are too diverse, their corresponding video shots are not always visually similar even 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).

Fig. 1 An example of two video shots with tag lists of their videos which are retrieved by YouTube with "blow candle" keyword. Since some relevant words such as "birthday" and "cake" are tagged to both videos, 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.

In this paper, we propose a novel ranking method, VisualTextualRank, which extends VisualRank [1]. 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.*

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 co-ranks 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.

VTR is an extension of VisualRank [1] with idea inspired by [2]. In [2], tags and videos are also co-ranked using their correlation to refine their relevance with 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. 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 SCi, k measures textual relevance score between the video of shot i and tag

[1] Faculty of Informatics, The University of Eletro-Communications, Chofugaoka 1–5–1, Chofu-shi, Tokyo, 182–8585 Japan
a) dohang@mm.cs.uec.ac.jp
b) yanai@cs.uec.co.jp

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. RT is initially defined as a uniform vector. 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. Damping vector p can be a uniform vector or a nonuniform vector.

3. Experiments

3.1 Implementation Details and Dataset

We chose the system of automatically extracting from tagged Web videos video shots corresponding to specific actions proposed in [3] to validate our ranking method since it provides a large-scale shot database which is suitable for our purpose and it is easy to implement. The system in [3] consists of two main steps: video ranking and shot ranking. At the shot ranking step, they apply VisualRank to rank shots from top ranked videos. We adopt our method to this step and compare the performance of their system with VisualRank and with our ranking method.

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. As for the calculation of SM, we use the same method as described in [3]. Relevance of a video to a tag is measured in the similar way as represented in [3] using the tag database constructed in advance. Note that here shots are obtained by segmenting selected videos but filtered by their length and tags are tags of selected videos but filtered based on their occurrence frequencies. Following [3], we select only shots which last longer than one second and shorter than one minute. To avoid using personal and subjective tags, we choose tags which appear at least five times over selected videos.

Damping factor α is chosen as 0.8 for practice. Damping vector is defined following the best results obtained in [3]. That means 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 practice.

We conduct experiments with the human action database of [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).

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**

3.2 Experimental Results

Experimental results are shown in Table 1. 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.

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%. 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", "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.

4. 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.

References

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

[3] Nga, D. H. and Yanai, K.: Automatic Construction of an Action Video Shot Database using Web Videos, *Proc. of IEEE International Conference on Computer Vision* (2011).

[4] Nga, D. H. and Yanai, K.: Automatic Collection of Web Video Shots Corresponding to Specific Actions using Web Images, *Proc. of CVPR Workshop on Large-Scale Video Search and Mining (LSVSM)* (2012).