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. However, 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 efficiently ranks the visual similarity between shots calculated via bipartite graph which represents the correlation between 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 efficient 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...
The proposed co-ranking method can be represented by following iterative processes:

\[
RS_k = \alpha \times SM^* \times SC^* \times RT_k + (1 - \alpha) p \tag{1}
\]

\[
RT_{k+1} = (SC^*)^* \times RS_k \tag{2}
\]

\(RS\) 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 \(n_s \times n_s\). \(SC\) 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 \(n_t \times n_t\) 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 \(\alpha\) as a damping vector to give more weight to ratings of the previous step. Damping factor \(\alpha\) 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 "key- word" and "title" fields. The co-occurrence frequencies among tags are then exploited to build a tag database which provides us tag co-occurrence 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
The 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. It 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. \( k \) equaled 1000 in our experiments.

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. 

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

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. \( k \) equaled 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.

<table>
<thead>
<tr>
<th>Action</th>
<th>Prec@10</th>
<th>Prec@50</th>
<th>Prec@100</th>
<th>Recall@10</th>
<th>Recall@50</th>
<th>Recall@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>paint+picture</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>dance+flamenco</td>
<td>30</td>
<td>35</td>
<td>40</td>
<td>25</td>
<td>30</td>
<td>35</td>
</tr>
<tr>
<td>ride+bicycle</td>
<td>98</td>
<td>99</td>
<td>99</td>
<td>98</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td>do+yoga</td>
<td>90</td>
<td>95</td>
<td>90</td>
<td>85</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>wash+dishes</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>wash+face</td>
<td>98</td>
<td>98</td>
<td>98</td>
<td>98</td>
<td>98</td>
<td>98</td>
</tr>
<tr>
<td>row+boat</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>draw+eyebrows</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>paint+wall</td>
<td>90</td>
<td>95</td>
<td>90</td>
<td>90</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td>climb+tree</td>
<td>70</td>
<td>70</td>
<td>70</td>
<td>70</td>
<td>70</td>
<td>70</td>
</tr>
</tbody>
</table>

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 retrieval system proposed in [3] adopting VisualRank and VisualTextualRank proposed in this paper. Experimental results are shown in Table 1.
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. The results show that our ranking method helps to boost more relevant shot to the top.

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”, “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.

6. REFERENCES