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

Do Hang Nga\textsuperscript{1,a)} Keiji Yanai\textsuperscript{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\textsuperscript{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].

However, in case of human actions, since they 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).

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\textsuperscript{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\textsuperscript{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:

\begin{align*}
RS_k &= \alpha \times SM^* \times SC^* \times RT_k + (1-\alpha) \ p \ \ (1) \\
RT_{k+1} &= (SC')^* \times RS_k \ \ (2)
\end{align*}

$RS$ 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 $SM_{ij}$ measures 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.

\textsuperscript{1}Faculty of Informatics, The University of Electro-Communications, Chofugaoka 1–5–1, Chofu-shi, Tokyo, 182–8585 Japan

\textsuperscript{a)} dohang@mm.cs.uec.ac.jp

\textsuperscript{b)} yanai@cs.uec.ac.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 $\alpha$ and damping vector $p$ into shot ranking. Damping factor $\alpha$ 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 our 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 $\alpha$ is chosen as 0.8 for practice. Damping vector $p$ 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).

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


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