Application Of Segmentation Of Object Video In Robot

GONG Heng

Chongqing College of Electronic Engineering, Chongqing, 401331, China

Gonghengxix75@163.com

Keywords: near duplicated video detection; segmentation of object image; video key frames

Abstract. In order to eliminate the redundancy of the video on visual information and improve the matching speed of video sub-sequence, an efficient solution is to extract a set key frame which is far less than an entire video frame to represent the video content. The common practice of extracting the video key frames is to split the video camera, and then extract the key frames in visual content which is most representative of the lens from each lens. The method based on shot segmentation has limitations in the two aspects. There is redundant visual information on the lens with similar content; When the query video is very short (less than 5s), the granularity of key frame extraction is difficult to ensure the accuracy of detection.

Introduction

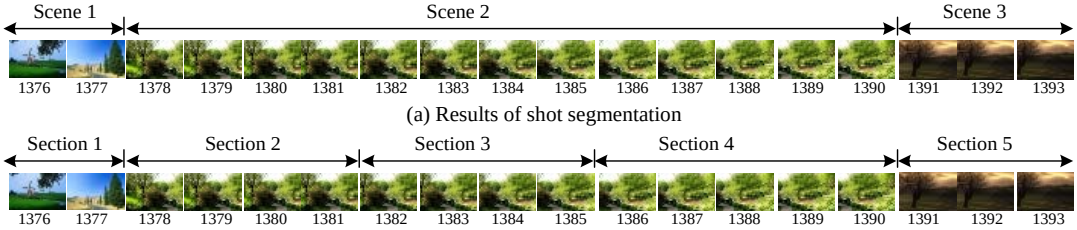
The important inner feature of video is its time characteristic, i.e. the video is composed of successive video frames in the direction of the time, thus content-based video retrieval or near repeated video judgment usually need to be compared the similarities of the video frames set to judge the similarity between two videos, although, in principle, content-based near duplicated video detection is basically the same with the content-based video retrieval[1]. However, in practical applications, there is a big difference in terms of implementation and existence. Content-based video retrieval pays more attention to the overall similarity of the query video and the goal video, that is, more of a "one to one" relationship, while for near duplicated video detection, its similarity retrieval may face more matches in the form. These matching forms can be summarized as “one-one”, “one-many” and “many-many”[2]. These three forms are described in Figure 1. Because the existing forms of near repeated video are entirely dependent on the specific application of the target video[3]. If the near duplicated video is operated by complex editing and modification, its detecting work will be a typical matching problem of video sub-sequence. In the complex in the complex near duplicated video detection, the length of the target sub-sequence, the existing location and the existing frequency are unknown; this form of matching task of video subsequence is much more complex than the ordinary video retrieval tasks[4,5].

Algorithm Realization

The proposed theory of extracting key frames adopts the adaptive dual-threshold shown in Figure 3; in the figure there are successive video frames with a period of time, in which C_{1f} means the first frame in segment 1; C_{1l} means the last frame in segment 1; C_{2f} represents the first frame in segment 2; T_h means the mutation threshold; T_l means the gradient threshold. The segment 1 and segment 2 in the video must satisfy any one of the following two conditions:

1)The similarity between the last frame in the video segment 1 and the first frame in the video segment 2 $sim(C_{1l}, C_{2f}) < T_h$; the similarity between the first frame in the video segment 1 and the first frame in the video segment 2 $sim(C_{1f}, C_{2f}) < T_l$. The continuous video frames can be divided into several visually similar video segments through the adaptive dual-threshold segmentation method, and three frames are extracted to represent each video segment. These three frames can be divided into the first frame, the key frame and the last frame, in which the key frame and the average frame (the means of all the frame features in this segment) can be replaced by the most similar frame. Key frame is used to match a video sequence, while the first frame and the last frame are mainly used for

precise positioning and auxiliary matching, which allocates contiguous ID to the divided video segments with the time direction. Figure 1 illustrates the difference between the method of proposed adaptive dual-threshold method eliminating the redundant video frames and the method of ordinary shot-based segmentation. Figure 1 (a) is the result obtained by the method using the shot segmentation; Figure 1 (b) segmentation results obtained by using the automatic dual-threshold to eliminate redundant video frames; in Figure 4, the distance (the distance between features based on the color) between the last frame(ID 1377) in the video segment 1 and the first frame(ID 1378) in the video segment 2 is greater than the mutation threshold T_h; the distance between the first frame(ID 1378) in the video segment 2 and the first frame(ID 1385) in the video segment 3 is greater than the mutation threshold T_l. Therefore, based on the proposed adaptive dual-threshold segmentation method, the lens 2 can be further divided segment 2 and segment 3, which has a smaller particle size. Therefore, it is called as sub-lens video or video segment in this article.



(a) Results of shot segmentation

(b) Segmentation results of adaptive dual-threshold

Figure 1: Results contrast of adaptive dual-threshold segmentation and the lens segmentation

After the matching results figure is built, the goal of matching the near duplicated video subsequence is to find the longest matching path in the Figure. So far, the matching problem of near duplicated video subsequence is converted into the problem of searching longest path in the matching results graph. Many classical algorithms can be used to find the shortest path of starting point of each node, such as Dijkstra, Bell man Ford, Floyd warshall and so on. Since what this thesis tying to find out is the longest path between any two nodes in graph, and thus the Floyd warshall algorithms is chosen. By using the Floyd_warshall algorithm, the longest path between any two points in the graph can be found out one time, and these paths determine the position of near duplicated video, but also determine the length of, and avoid exhaustive method to determine the position and length of near duplicated video, in which the detection efficiency is effectively improved.

As to any node in the figure, there is no path which is start with it, but maybe there is a path or several paths. node $M_{1,29}$,node $M_{1,76}$,node $M_{2,76}$ in figure have no path (the path is the node itself); while for node $M_{1,26}$, there are 16 paths. Wherein there are four longest path (L = 6), namely:

Route 1: $M_{1,26} \to M_{1,27} \to M_{1,28} \to M_{1,29}$;
$\to M_{1,30} \to M_{1,31} \to M_{1,32}$

Route 2: $M_{1,26} \to M_{1,27} \to M_{1,28} \to M_{1,29} \to M_{1,30} \to M_{1,31} \to M_{1,33}$;

Route 3: $M_{1,26} \to M_{1,27} \to M_{1,28} \to M_{1,29} \to M_{1,30} \to M_{1,32} \to M_{1,33}$;

Route 4: $M_{1,26} \to M_{1,27} \to M_{1,28} \to M_{1,29} \to M_{1,30} \to M_{1,32} \to M_{1,34}$;

These 16 matching paths are stacked in the time sequence, so when the final matching sequence is outputted, these laminated paths on times must be merged. Then according to the timestamp information corresponding to query video and the reference video of the matching paths' starting and final node, the location of these two near duplicated video is located. For the merging of these cascading paths, this thesis takes the following matching strategy to the longest path:

The matching strategy of the longest path: Given any two paths p_i and p_j, let's set the beginning point and final point of p_i respectively are $M_{a,b}$ and $M_{c,d}$. Then there are three possibilities of the two paths in the timing position relationship: divided, contained and intersected. The combined path is also divided into three cases:

1) Divided: if $c \le e$ and $d \le f$, or $g \le a$ and $h \le b$, then the path p_i is away from path p_j. At this point there is no lamination of these two paths in the video sequence, so the path p_i and path p_j are two mutually independent paths, and the combined paths are still two separate path p_i and

path p_j.

2) Contained: if $a \le e$, $b \le f$, $c \le g$ and $d \le h$, the path p_j contains the path p_i and the combined path is p_j. Conversely, if $a \ge e$, $b \ge f$, $c \ge g$ and $d \ge h$, the path p_i contains the path p_j and the combined path is p_i.

3) Intersected: If $e < a < g$, $f < b < h$, $c > g$ and $d > h$, the starting point $M_{a,b}$ of the path p_i is contained in the path p_j, while the final point $M_{c,d}$ of p_i falls after the end point $M_{g,h}$ in the path p_j, so the combined path is $p_j \to M_{c,d}$, i.e. node $M_{g,h}$ which is added after p_i is chosen as end point.

If $a < e < c$, $b < f < d$, $g > c$ and $h > d$, the starting point $M_{e,f}$ of the path p_j is contained in the path p_i, while the final point $M_{g,h}$ of p_j falls after the end point $M_{c,d}$ in the path p_i, so the combined path is $p_i \to M_{g,h}$, i.e. node $M_{g,h}$ which is added after p_i is chosen as end point.

After the matching path in the graph is merged through the longest path matching strategy, the discrete paths are ultimately got in the matching results figure, that is, there is no stacked path in time. For example, as for all paths in the matching results figure shown in Figure 6, when the longest path matching strategy is merged, the final matching path is as follows:

$M_{1,26} \to M_{1,27} \to M_{1,28} \to M_{1,29} \to M_{1,30} \to M_{1,32} \to M_{1,34}$

According to the final matching path, the frame set got in the query video sequence is $S_Q = \{1,2,3,4,5,7,8\}$ and the frame set got in the reference video sequence is $S_R = \{26,27,28,29,30,32,34\}$. Finally, according to the time-stamp information of the starting frame in S_Q and S_R, the position of the two videos sequences are located corresponding to the location in the video. Such as according to the corresponding time point of the first frame and the eighth frame in the query video, the location of the sub-sequence can be located in the in query sequence, and according to the corresponding time point of the 26th frame and the 34th frame in the reference video. Obviously, the proposed graph-based method can directly detect the situation of several near duplicated video presence existing in these two videos. Furthermore, in order to filter the matching noise brought by some short paths (such as some paths with the length as 2 or 3). In this paper, through the experimental study, Equation 2 can be used to measure the similarity between the sequences of two video:

$$sim(p_i) = \frac{\sum_{k=1}^{L} sim_k(M_{i,j})}{L} \log(1+L) \quad (1)$$

Wherein, L is the number of nodes in the path; $M_{i,j}$ is a node in the path; $sim(M_{i,j}) = sim(q_i, r_j)$.

Experimental Results

If there is an edge between two nodes in the matching figure, there are two conditions that must be simultaneously meted as following:

(1) The two nodes must be satisfied the consistency in the time direction;

(2) The time jump degree of two nodes is $\Delta t < \tau$ (τ is threshold of time jump degree)

The condition (1) represents the time direction of the query video represented by two nodes is consistent with the time direction of the target video graph, which may seems reasonable to do so, because the video sequence is a time s sequence and the time direction of the coping video and the copied video is the same. The direction of increasing time is directed edges between nodes direction. The condition (2) represents the jumps of the matching results represented by two nodes in the time direction can not exceed a certain threshold, otherwise there is no correlation between the two matching results. According to the above methods and conditions, there is obviously a directed acyclic graph in the matching results figure. The author also had a comparative experiment to obtain the most optimal time jump threshold as shown in Figure 2; T1-T10 in the figure are represent 10 different types of copied video; the experimental data shows that a relatively appropriate time jump threshold should be between 10 to 29 seconds. On the other hand, the author also makes the

statistics to the time length of the obtained video clips of the adaptive dual threshold method as shown in Figure 3. The time length of video segments shown by data in the figure mostly concentrates between 10 and 30 seconds, which also verified the reasonability of choosing threshold of time jump degree between 10 and 30 seconds.

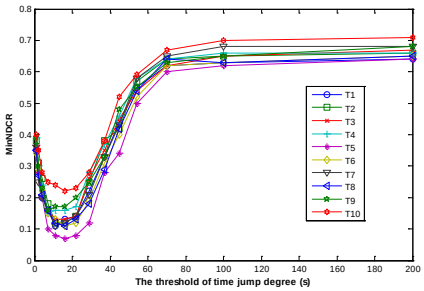


Figure 2: selection of threshold of time jump degree

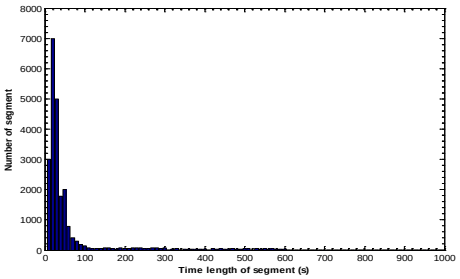


Figure 3: distribution of time length of segments after the video is divided by adaptive dual threshold method

Conclusion

the method based on shot segmentation has limitations in the two aspects. First,1) There is redundant visual information on the lens with similar content; 2) When the query video is very short (less than 5s), the granularity of key frame extraction is difficult to ensure the accuracy of detection. Therefore, the approach I picked up does not fully take the shot segmentation method to extract key frames, but rather follows the idea of dividing the video into the similar video clip on the visual content, and then extract fixing number of key frames on each video clip to represent the video clips, results shows the distinction between the proposed video clips segmentation and commonly used shot segmentation method.

References

[1]D. Xu, Z. Y. Feng, Y. Z. Li, et al. Fair Channel allocation and power control for uplink and downlink cognitive radio networks. IEEE., Workshop on mobile computing and emerging communication networks, 2011:591-596

[2]W. Q. Yao, Y. Wang, T. Wang. Joint optimization for downlink resource allocation in cognitive radio cellular networks. IEEE., 8th Annual IEEE consumer communications and networking conference, 2011:664-668

[3]S. H. Tang, M. C. Chen, Y. S. Sun, et al. A spectral efficient and fair user-centric spectrum allocation approach for downlink transmissions.IEEE., Globecom.,2011:1-6

[4]S. Li, Y. Geng, J. He, K. Pahlavan,Analysis of Three-dimensional Maximum Likelihood Algorithm for Capsule Endoscopy Localization, 2012 5th International Conference on Biomedical Engineering and Informatics (BMEI), Chongqing, China Oct. 2012 (page 721-725)

[5]Y. Geng, J. He, H. Deng and K. Pahlavan, Modeling the Effect of Human Body on TOA Ranging for Indoor Human Tracking with Wrist Mounted Sensor, 16th International Symposium on Wireless Personal Multimedia Communications (WPMC), Atlantic City, NJ, Jun. 2013.