21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Key Observation Selection for Effective Video Synopsis

Xiaobin Zhu, Jing Liu, Jinqiao Wang, Hanqing Lu
NLPR, Institute of Automation, Chinese Academy of Sciences
{*xbzhu,jliu,jqwang, luhq*}*@nlpr.ia.ac.cn*

Abstract

Millions of video surveillance cameras distribute around the world, and capture tremendous number of video data endlessly. Video browsing by frame is time consuming and inefficient, since needless information is abundant in the raw videos. Video synopsis is an effective way to solve this problem by producing a short video abstraction, while keeping the essential activities of the original video. However, traditional video synopsis only eliminates redundancy in spatial and temporal domain, while neglects redundancy in content domain. However, too many observations will make synopsis video confusing and degrade synopsis efficiency. In this paper, we present a novel video synopsis method based on key observation selection. Key observation selection is conducted for activity to eliminate content redundancy. We have demonstrated the effectiveness of our approach on real surveillance videos.

1. Introduction

With the development of imaging techniques and storage ability, surveillance videos for 24-hours everyday are produced in real world. However, browsing and indexing the large amount of raw videos is a time-consuming and even impossible task for us. Therefore, how to obtain a compressed video abstraction become a hot topic in related fields.

There are mainly two kinds of techniques in video abstraction, namely key-frame extraction and video skimming. In the former [10][2], the key frames are selected randomly, or selected according to some importance criteria, from the original video. Key frame representation could largely save video browsing time, but it neglects the dynamic aspect of video. The latter [4], also called moving-image abstract, attempt to extract video segments from the original video to obtain a shorter video, which is more coherent and expressive compared with those derived from the key frame based technique. However, it is likely for people to spend large amounts of time browsing video segments with little information. Video synopsis [8][6][7][5][3] breaks the previous framework, and makes the video abstraction shorter than the original one by displaying the moving objects from different periods simultaneously. They create a synopsis that combines objects which have appeared at different times. Video synopsis can eliminate redundancy in spatial-temporal domain, and give a summary of video to viewer. For example, a five minutes video synopsis may be created from ten hours raw video, while keeping the dynamic of aspect of video. However, traditional video synopsis concentrates on eliminating redundancy in spatial-temporal domain, while neglecting the redundancy in content domain. Too many observations can degrade the efficiency and subjective comprehension in video synopsis.

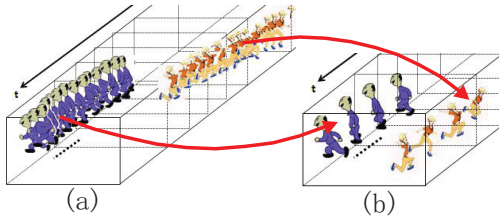


Figure 1. (a) Original video; (b) Key observation selection based synopsis video.

In this paper, we present a novel key observation selection based video synopsis method. The intuitive analysis of key observation selection can be seen in Fig. 1. The original input video contain a man and a girl at two different times, namely two objects, as showed in Fig. 1 (a). We refer to the space-time sequences of an object as a tube, in which all the observations belong to one object As we can see in Fig. 1 (a), numerous observations may exist in tubes, causing great content redundancy. Through key observation selection, we can obtain a more compact and comprehensive synopsis video than conventional method, as showed in Fig. 1 (b). The

main contributions of our work are: 1) A novel multiple kernel similarity method is adopted in selecting key observations; 2) An modified energy lost function combined with key observation selection is encoded into the synopsis framework. 3) Key observation selection in the tube of each object aims to sample the representatives in video data, for eliminating content redundancy, and resulting in higher compression ratio compared with state-of-art video synopsis method;

2. Observation selection

The method mentioned in [6] enables video synopsis more convenient for browsing by changing the spatial-temporal relationships among different objects. However, there may be numerous observations belong to one object. In typical scenarios, adjacent objects may be similar in action and appearance. In addition, too many observations will tend to cause collision, and negative impaction on subjective effect. So we select a reduced number of key observations, which can well represent the original activity, according to the change of action and appearance, to promote the efficiency of video synopsis.

In [9], k-means clustering method is adopted in selecting a pre-defined number of key actions. However, the number of key actions can't be fixed for different objects even in the same scenario. So we adopt a data-driven method in selecting key observations. We try to sample the objects in each tube to maximize the representation of the object moving process. The observations which have significant action or shape change are deemed as key ones, according to our criteria. Different from [2], we extract key observations from every object instead of input video. In order to explain our algorithm explicitly and effectively, we adopt the definition described in [6]. Every tube can be denoted as temporal duration $[t_b^s, t_b^e]$, but different from [6], observations are included in this duration, as $[t_b^s, t_b^{(s+1)}, ..., t_b^{(e-1)}, t_b^e]$.

Let's O_i and O_j denote two observations belong to one tube. We adopt a multiple kernel based similarity measurement to select key observations.

Distance kernel. This kernel is used to measure the spatial uniform of observations in tube, and defined as:

$$D_p(O_i, O_j) = exp(-dist(O_i, O_j)^2/\sigma_p^2) \quad (1)$$

where $dist(O_i, O_j)$ can be set as the simple $L1$ distance in Euclidean space for the x position of two observations.

Motion kernel. If two observations have similar action direction, they should be similar in motion space. This is defined as:

$$D_m(O_i, O_j) = exp(-sin(\theta(O_i, O_j))^2/\sigma_m^2) \quad (2)$$

where $\theta(O_i, O_j)$ measures the motion angle between the two observations.

Appearance kernel. This is defined as:

$$D_a(O_i, O_j) = exp(-KL(h(O_i), h(O_j))^2/\sigma_a^2) \quad (3)$$

where $KL(h(O_i), (O_j))$ measures the KL-divergence between two appearance distributions of observations($h(.)$ is the color histogram in HSV space).

There are multiple ways to associate these kernels together, but which is best is an open problem in machine learning field. In our setting, we adopt the linearly combine the three and find it effective.

$$\begin{aligned} SIM(O_i, O_j) =& \lambda_1 D_p(O_i, O_j) + \lambda_2 D_m(O_i, O_j) \\ &+ \lambda_3 D_a(O_i, O_j) \end{aligned} \quad (4)$$

The $SIM(O_i, O_j)$ is the selection criteria in our method. λ_1, λ_2 and $\lambda_3(\lambda_1 + \lambda_2 + \lambda_3 = 1, \lambda_1, \lambda_2, \lambda_3 > 0)$ are three weight parameters, which are learned from our ground-truth data for particular scenario in advance. Observations of objects's entering in and leaving out of camera scope should be regarded as key ones. Then we adopt the distance criteria $SIM(O_k^{last_k}, O_k^i)$ between observation k belong to last defined key observation and current observation in one tube, if the similarity is smaller than a pre-determined threshold, then we select the current observation as key observation. We use to define key observations as 1 and non-key ones 0 as in every tube. And observations with 1 are extracted to form a new tube, and the time of observations in new tube is organized in continuous mode. The detailed algorithm is summarized in $Algorithm1$.

Algorithm 1 Selecting key observations in tubes

Input:N,the number of tubes
Output:N new tubes $tnew$ consist of key observations
Data:$[t_1^s, t_1^{(s+1)}, ..., t_1^e]$,...,$[t_N^s, t_N^{(s+1)}, ..., t_N^e]$; T_F, similarity threshold.
for $k = 1; k < N + 1; k++$ **do**
  t_1^s, $y\,(k, s) \leftarrow 1$, $tnew_1^s = t_1^1$
  t_1^e, $y\,(k, e) \leftarrow 1$
  $p = 1$
  for $i = s + 1; i < e; i++$ **do**
    t_1^i, $y\,(k, i) \leftarrow 0$
    if $SIM(O_k^{last_k}, O_k^i) < T\_F$ **then**
      $y\,(k, i) \leftarrow 1$; $last_k \leftarrow i$
      $tnew_1^{(s+p)} = t_1^i$, p++
    end if
  end for
end for

3. Key observation selection based video synopsis

Different from traditional methods of abstraction, the temporal relationship of objects in synopsis will be changed in order to obtain higher compressive video abstraction, which can display objects appearing in different periods of original video simultaneously. It eliminates the spatial-temporal redundancy of original video. In [8], Rav-Acha et al., proposed the object-based synopsis method for surveillance video. Afterwards, Pritch et al. construct a framework of webcam video synopsis [6]. First, it extracts the object from tubes(the 3D space-time representation of each object), then formulates an energy function composed of activity lost cost E_a, background consistency cost E_s , time consistency cost E_t, and occlusion cost E_c.

Following the above method, we also introduce concepts of collision and time consistency cost. In addition, we combine the key observation selection with video synopsis generation, looking for a temporal mapping M and T_F that minimize the object function in (5). The energy function we formulate is as follows:

$$E = \arg\min_{\forall M, T\_F} E(M, T\_F) \tag{5}$$

$$\begin{aligned} E(M, T\_F) = &\sum_{b_n \in B} (E_a(\hat{b_n}) + E_k(\hat{b_n}) + E_s(\hat{b_n})) \\ &+ \sum_{b_n, b'_n \in B} (\alpha E_t(\hat{b_n}, \hat{b'_n}) + \beta E_c(\hat{b_n}, \hat{b'_n})) \end{aligned} \tag{6}$$

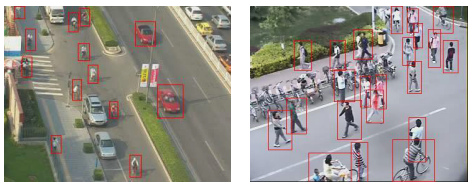
Where b_n and b'_n represent two key observation based new tubes selected according to $Algorithm1$ with threshold T_F, $\hat{b_n}$ and $\hat{b'_n}$ are two key observation based new tubes mapped into video synopsis. $E_s(\hat{b_n})$ is tube and background consistency cost, measuring the cost of stitching objects to the time-lapsed background. $E_t(\hat{b_n}, \hat{b'_n})$ is time consistency cost, preserving the chronological order of objects. $E_c(\hat{b_n}, \hat{b'_n})$ is collision cost, penalizing for the spatial-temporal overlaps among objects. α and β are two empirical parameters set by user. $E_k(\hat{b_n})$ is observation selection cost, penalizing for the lose of observations in key observation selection.

$$E_k(\hat{b_n}) = \sum_{t \in \hat{t_b} - t_{\hat{b_n}}} \sum_{x,y} \chi(x, y, t) \tag{7}$$

Where $t \in \hat{t_b} - t_{\hat{b_n}}$ denotes the observations discarded during key observation selection for the tubes appear in synopsis video, and $\chi(x, y, t)$ is characteristic functions representing appearance of tube, and defined as:

$$\chi_b(x, y, t) = \begin{cases} \parallel I(x, y, t) - B(x, y, t) \parallel & t \in t_b \\ 0 & otherwise \end{cases} \tag{8}$$

Where $B(x, y, t)$ is a pixel in the background image, $I(x, y, t)$ is the respective pixel in the image, and t_b is the time in which this object exists. $E_a(\hat{b_n})$ is activity lost cost, penalizing for the lose of observations during key observation selection. If key observation based new tubes haven't mapped into synopsis video, then $E_k(\hat{b_n}) + E_a(\hat{b_n})$ is activity lost cost for original tube. If key observation based new tubes appear in synopsis video, then $E_a(\hat{b_n})$ equals to 0, and only lose of observations in key observation selection exists. In our method, T_F is set from $[0.2, 0.4]$, with a step 0.02. Finally, the energy function is minimized by using simulated annealing algorithm for every T_F. And we select the smallest energy lost as best arrangement for synopsis. After we achieve the best arrangement of tubes, we stitch projected tubes into background image using Poisson Editing [1] to generate final synopsis video.



(a) Dataset1 (b) Dataset2

Figure 2. Two examples of synopsis videos.

4. Experiments

4.1 Synopsis for single camera

In this section, we employ two videos(Resolution $320 * 240$, 15 FPS) to evaluate the objective performances of our method, generating two synopsis videos. In Fig. 2, two representative frames from two synopsis videos, are displayed. We also compare our key observation selection based method (denoted as $Proposed$) with traditional method without key observation selection (denoted as $Method1$) [6]. The detailed results are displayed in Tab. 1 and Tab. 2. The energy lost of E_a for $Proposed$ includes E_a and E_k, while only E_a keeps for $Method1$. From Tab. 1 we can conclude, our method obtains 5.4% compression rate, while causing 3518 energy lost. And $Mehod1$ obtains 7.7% compression rate, while causing 4942 energy lost. From Tab. 2 we can conclude, our method obtains 3.9% compression rate, while causing 4587 energy lost. And $Mehod1$ obtains 5.8% compression rate, while causing 6001 energy lost. Obviously, our method can achieve higher compression rate, while causing a lower energy lost. In addition, from the subjective viewpoint, our method can generate more comprehensive and pleasing synopsis video.

Dataset1	Energy Cost				Frame Number	
	E_a	E_s	E_t	E_c	Original	Synopsis
Proposed	556	411	447	2104	12045	648
Method1	206	432	485	3819	12045	927

Table 1. Comparison results for dataset1.

Dataset2	Energy Cost				Frame Number	
	E_a	E_s	E_t	E_c	Original	Synopsis
Proposed	681	732	581	2593	13710	531
Method1	397	791	476	4337	13710	801

Table 2. Comparison results for dataset2.

4.2 Synopsis for camera-network

We employ video sequence($Dataset3$) captured by our camera-network equipment, which is located in challenging outdoor scenarios, describing a scene simultaneously recorded by two cameras located at different viewpoints with overlapping field of view. Two videos are all with resolution $320 * 240$ and frame rate 15. Fig. 3, is one representative frame of synopsis video. The camera-network video synopsis can provide overall dynamic of object, enabling video retrieval and browsing more efficient. We also compare our method with $Method1$ using the same setting in $Sec.4.1$. The detailed results are displayed in Tab. 3. We also can conclude our method has better performance than $Method1$.

5. Conclusion

In this paper, a novel key observation selection based video synopsis method is presented and discussed. A novel data-driven multiple kernel similarity is adopted in key observation selection. Our method can greatly eliminate the redundancy in content domain, and promote the efficiency of video synopsis. However, although we introduce spatial uniform item in key observation selection, our method still can cause jumping effect for objects in synopsis video. A good solution is to change the position of observations in individual local area, and combine this process with lost cost minimization, to further keep entirely spatial uniform of objects. In all, this technology enjoys a promising surveillance-oriented application especially in video searching and retrieval.

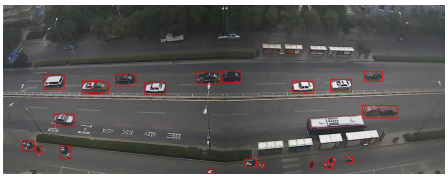


Figure 3. Representative frame of synopsis for dataset3.

Dataset3	Energy Cost				Frame Number	
	E_t	E_s	E_c	E_l	Original	Synopsis
Proposed	366	540	2593	694	9570	507
Method1	337	612	4937	1464	9570	846

Table 3. Detail synopsis information for Dataset3.

6. Acknowledgement

This work was supported by the National Natural Science Foundation of China 973 Program (Project No. 2010CB327905) and the National Natural Science Foundation of China (Grant No. 60903146, 60905008, 0835002).

References

[1] M. Gangnet, P. Perez, and A. Blake. Poisson image editing. In *ACM SIGGRAPH*, pages 313–318, 2003.
[2] C. Kim and J.Wang. An integrated scheme for object-based video abstraction. In *ACMMM*, pages 303–311, 2000.
[3] T. Li, T. Mei, I. Kweon, and X. Hua. Videom̂ Multi-video synopsis. In *ICDM Workshops*, pages 854–861, 2008.
[4] C. Ngo, Y. Ma, and H. Zhang. Automatic video summarization by graph modeling. In *ICCV*, pages 104–109, 2003.
[5] Y. Pritch, S. Ratovitch, A. Hendel, and S. Peleg. Clustered synopsis of surveillance video. In *AVSS*, pages 195–200, 2009.
[6] Y. Pritch, A. Rav-Ach, and S. Peleg. Webcam synopsis: Peeking around the world. In *ICCV*, pages 1–8, 2007.
[7] Y. Pritch, A. Rav-Ach, and S. Peleg. Nonchronological video synopsis and indexing. *TPAMI*, 30(11):1971–1984, 2008.
[8] A. Rav-Ach, Y. Pritch, and S. Peleg. Making a long video short: Dynamic video synopsis. In *CVPR*, pages 435–441, 2006.
[9] Z. Tian, J. Xue, X. Lan, C. Li, and N. Zheng. Key object-based static video summarization. In *ACMMM*, pages 1301–1304, 2011.
[10] X. Zhu, X.Wu, J. Fan, A. Elmagarmid, and W. Aref. Exploring video content structure for heirarchical summarization. *Multimedia Syst.*, 10(2):98–115, 2004.