A NEW FRAMEWORK FOR STUDYING TUBES REARRANGEMENT STRATEGIES IN SURVEILLANCE VIDEO SYNOPSIS

Giovanna Pappalardo, Dario Allegra, Filippo Stanco and Sebastiano Battiato

Department of Mathematics and Computer Science, University of Catania
Viale Andrea Doria, 6 - 95125 Catania, Italy
giovanna.pappalardo1@unict.it, {allegra, fstanco, battiato}@dmi.unict.it

ABSTRACT

The manual review of raw surveillance video is a time consuming task which can be optimized by using a Video Synopsis (VS) algorithm. The aim of such approaches is to condense a long video into shorter one to allow a quicker review of surveillance data. However, VS is a complex problem. A typical object-based VS algorithm requires three main modules to perform the following tasks: object detection and tracking, tubes rearrangement, condensed video generation. Although the aforementioned three steps are equally critical, we realized that the core of Video Synopsis lies in the tubes rearrangement. This led us to propose an original approach to tackle the problem of tubes rearrangement. To this aim, we first introduce a new toolbox to generate a proper testing dataset, which allows to bypass the lack of public databases including proper annotated videos for testing synopsis approaches. Additionally, we propose an improvement of a tubes arrangement algorithm based on graph colouring and we prove its validity on our generated dataset. For a proper comparison, we show that our algorithm also outperforms the original one on UA-DETRAC public dataset.

***Index Terms*—** Video Synopsis, Videos Surveillance, Synthetic dataset, Tubes rearrangement, Graph colouring.

1. INTRODUCTION

In recent years, the amount of video data is explosively increased with the expansion of surveillance cameras located in lot of environments such as banks, airports, petrol stations, private homes, shops and so on. Most of these cameras record 24/7 hours per day, so that the acquired videos cannot be entirely reviewed by human operator. Although some Computer Vision algorithms allow to automatically detect a set of simple activities [1, 2], the manual review can be mandatory, especially during crime investigation. Nevertheless, manual research of interesting activities, in an extremely large volume of video data, is time-consuming and highly inefficient because of the high redundant content. Video Synopsis, introduced in 1996 [3], aims to make shorter the original recording for taking a full view of the video content efficiently.

1.1. Related Works

Video Synopsis algorithms are usually classified in two main categories: frame-based and object-based. Frame-based approaches treat the frame as fundamental unity of videos, which means it cannot further decomposed. Basic approaches, in this category, perform a time compression and skip any frames to make the view faster [4]. Other frame-based strategies work by extracting a set of key frames or some short clips according to certain criteria (e.g., remove frames with no activities) [5, 6, 7]. However, these methods result in a lose of video dynamic. On the other hand, object-based approaches can produce a more dynamic output video; their aim, is to extract objects from the original video and convert them to temporal domain. Then, they reduce spatio-temporal redundancies but may cause many collisions between objects [8, 9, 10]. In this context, the set of object's positions in every frame of the video is called tube. Hence, there is one and only one tube for each object. The core of these approaches is to efficiently assigning a new start time to each object i.e., rearranging tubes. Outside the presented categories, other methods introduce some optimization like clustering of objects with similar activities [11] and scaling down objects [12]. In 2017, He et al. [13, 14] proposed a method where original video tubes are used to build a graph structure which models all the potential collisions. In order to minimize objects collisions, they employ a graph colouring algorithm which assigns a new time label to each tube.

1.2. Motivation and Contributions

Object-based synopsis preserves the activities in the input video and produce an output video in which most of the spatio-temporal redundancies are reduced (e.g., frames with no objects). Moreover, the moving objects in the video are temporally shifted to condense as many objects as possible in the same frame (Fig. 1). A comprehensive object-based VS algorithm consists of three main components which make the problem relatively complex: (1) a detector and a tracker for the objects captured in the video (i.e., tubes extraction) (2) a tubes rearrangement algorithm which plays the key role

978-1-5386-6249-6/19/$31.00 ©2019 IEEE

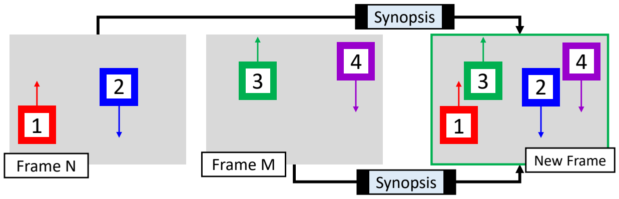

Fig. 1. Example of object-based Video Synopsis.

in VS. Condensation degree and collision avoidance mainly depends on this component; (3) an algorithm for final video generation. It should be created a clear and not confusing merging of rearranged tubes and background.

Our paper, focuses on the core of an object-based Video Synopsis algorithm, namely tubes reorganization; we do not address detection and tracking, because they are different problems that can be solved independently and for which numerous solutions have been proposed in the last decades [15, 16, 17, 18]. Moreover, in most of VS related works, detection and tracking are performed with basic approaches (e.g, background subtraction, frame difference [8, 12, 19]). However, for studying and testing VS methods, properly annotated datasets for object tracking and detection are required. Unfortunately, annotating objects in each frame of long videos involves a great human effort; also, for a proper evaluation of VS algorithm, one needs videos with many objects which appear at different time, so that the algorithm could condense the objects in the same frames (Fig. 1). The difficulties to get many annotated videos with this properties, result in a lack of public datasets suitable for VS study. For instance, He et al. [13] tested their approach on just 8 videos. Moreover, Li et al. [20] remark that in challenging scenario (e.g., very crowed video), the existing Video Synopsis algorithms do not perform very well (5 videos used for testing). Hence, annotated videos of these complex situations can be very useful for properly testing new solutions.

For this reason, our first contribution is a toolbox for random generation of synthetic annotated video where duration and objects density can be chosen by researchers. The proposed toolbox allows the researchers to generate new datasets in order to test their own approaches under many different and challenging conditions (e.g., [20]). The second contribution, is a new tubes rearrangement approach based on graph colouring. Inspired by the works of He et al. [13, 14], we propose a new solution which is more effective and efficient. However, we also test the proposed method on a subset of UA-DETRAC Benchmark Suite in order to show that we get similar results and to prove that tubes arrangement problem can be successfully studied with synthetic videos. Our findings reveal the proposed dataset generator and the proposed tubes rearrangements strategy are effective and robust.

The rest of the paper is organized as follows: Section 2, details the proposed toolbox and the new algorithm of tubes rearrangement; Section 3 presents a quantitative evaluation of the proposed algorithm on synthetic dataset and UA-DETRAC one. Conclusions are summarized in Section 4.

2. TOOLBOX AND PROPOSED METHOD

The contributions of this work is two-fold:

- a new toolbox to generate synthetic annotated videos of tracked bounding boxes;
- a new object-based Video Synopsis strategy based on graph colouring approach.

The first contribution gives researchers the opportunity to test different Video Synopsis strategies by focusing on tubes rearrangement step. This choice was motivated by the fact that most of the datasets used for testing VS approaches have been originally built to address detection and tracking problem. This makes these datasets not always ideal and suitable for VS studies. For example, in many literature annotated datasets, there are videos where some objects remain in the scene from the first frame to the last one. Consequently, tubes length for such objects is the same as the video length and hence, there is no way to rearrange them and condense the original video. The second contribution consists of a new effective and efficient tubes rearrangement strategy which also allows to benchmark the synthetic dataset.

2.1. Toolbox

The toolbox was designed to generate a synthetic dataset for testing VS approaches in many different cases; it allows to create a video by choosing the following parameters:

- resolution and video length l (in frames) at given frame rate (e.g., 25fps);
- probability p that in a certain frame a new object enters into the scene;
- objects movement direction and speed;

Let p the chosen probability and l the video length in frames, then the total number of objects in the video will be approximately $p * l$. For each frame, a new object randomly appears with probability p, whereas nothing happens with probability $(1 - p)$. Then, a new spawn location is randomly chosen according to movement direction. It also is possible to set unidirectional or bidirectional mode. In unidirectional mode, all the objects move in accordance with the user chosen direction. In bidirectional mode, the scene is divided in two parts along the chosen orientation and in each part the object move in opposite directions. The output is a text file which includes a list of position for each pair object-frame. Finally, we provide a tool to generate the synthetic video by reading the text file. Final generated video shows a scene with a static background and a set of moving bounding boxes. This allows to generate videos where synthetic objects enter in the scene and follow their paths according to a model which describes a realistic behaviour. Specifically, in this work we simulate a top-view traffic video where the vehicles are moving from the bottom to the top of the scene; the spawn location is randomly selected along the x axis, while the y position is fixed to 0. To promote the research on the field the toolbox is publicly available[1].

2.2. Video Synopsis Algorithm

In this section we first present the original work of He et al. [13]. However, for the sake of conciseness, we describe the main concepts; the reader is referred to [13] for a comprehensive description. Then, we describe our proposal to make the graph colouring more efficient. Finally, we detail the approach used to improve the effectiveness.

2.2.1. Overview of the Original Algorithm

In [13], He et al. used graph colouring algorithm to perform Video Synopsis. Each tube can be represented as a set of triplets (x, y, t) where (x, y) is the position in the space while t is the frame. A projection along t can show the potential collision between tubes. For each tube extracted from the original video, the potential collision points are identified to create a Potential Collision Graph (PCG). PCG consists of two sets V and E i.e., the set of nodes and the set of edges. He et al. introduce two kinds of nodes called main node (m-node) and sub node (s-node). An m-node represents a tube, while an s-node a potential collision. If a tube presents a potential collision, then a parent-child relationship between the related m-node and s-node is created. Also, they distinguish two kinds of collisions: intersection and overlapping. An intersection occurs when, after the projection, two tubes share each other a limited part. In this case, an edge is created between the s-nodes in the m-nodes related to the intersected tubes. On the contrary, an overlapping occurs when two tubes share each other a large part. In the latter case, two s-nodes are created in both the m-nodes involved in the overlapping. To distinguish between intersection and overlapping they set an overlapping threshold th. The s-nodes couple stands for the overlapping start time and overlapping end time. An edge is created between the two s-nodes related to the overlapping start time as well as the two s-nodes related to the overlapping end time. Finally, an edge is created between the s-nodes couple. If a tube does not present potential collision a so-called isolated m-node is created. He et al. propose to employ $L(q)$-colouring algorithm [21], to assign a colour label to each s-node and isolated m-node in PCG. Then, the labels assigned to the s-nodes are used to compute the labels for the related parent m-nodes. At the end, m-node final label stands for the new time assigned to the related tubes in synopsis. The value q in $L(q)$-colouring algorithm indicates the minimum difference between colours of two adjacent nodes. According to [13], q influences the time distance between two objects in the synopsis and, consequently, the artifacts due to the objects overlaps.

2.2.2. Colouring Algorithm Optimization

Our first improvement focuses on the PCG colouring strategy. When an overlapping between two m-nodes (tubes) m and m' occurs, the original algorithm to assign a colour to an s-node v_i in m, searches for a label l which satisfies the following condition:

$$(l - l'_i)(l - t_{ij} - l'_j) > 0 \tag{1}$$

where l'_i is the label of the s-node v'_i in m', connected with v_i; t_{ij} is the time difference in frame between the overlapping start/end time related to v_i and overlapping end/start time related to v_j; l'_j is the label of the s-node v'_j in m', connected with v_j. In the original algorithm, l is found by iterating a loop and increase its value by one for each iteration. Our proposal is to compute l by solving the second degree inequality in (1):

$$l = \frac{-(t_{ij} - l'_j - l'_i) + \sqrt{((t_{ij} - l'_j - l'_i)^2 - 4(l'_i l'_j - l'_j t_{ij}))}}{2} \tag{2}$$

Only the positive results of (2) are considered.

2.2.3. Condensation Optimization

A Video Synopsis algorithm should produce a high condensed video with a low number of collisions. We propose a reorganization step in PCG colouring which exploits graph connected components to improve condensation rate without collision avoidance drops. Our insight comes from the fact that if two m-nodes m and n belonged to different connected components, no potential collisions occur between the respective tubes. Hence, we can treat these groups separately to operate a further optimization. The first step is to assign to each m-node a time label by using colour values of its s-nodes. Let m_i a generic m-node; $s_{i,j}$ an s-node related to overlap start or to the intersection; $l_{i,j}$ the label for the s-node $s_{i,j}$; t_i the time when the tube related to m_i starts; $t_{i,j}$ the time related to $s_{i,j}$; then we assign to m_i the label l_i computed as follow:

$$l_i = max(1, \min_j \{l_{i,j} - (t_{i,j} - t_i)\}) \tag{3}$$

This guarantees the minimum positive time label for each m-node. As second step, we extract the N connected components from PCG to create a family of sets C. The generic set C_k, where $k = 1...N$, includes all the m-nodes in the k-th connected component, namely all the m-nodes related to the

[1] iplab.dmi.unict.it/videoSynopsis

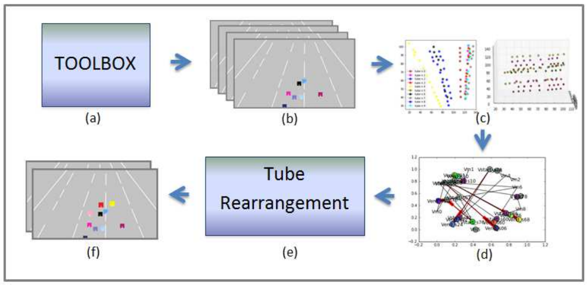

Fig. 2. Overview of the proposed framework. (a) Use the toolbox and choose video features; (b) generation of the synthetic videos; (c) projection of tubes in spatial domain; (d) PCG building (e) tube rearrangement by coloring and connected components analysis; (f) video synopsis generation.

objects with mutual potential collisions. The $\mid C_k \mid$ m-nodes m_i in C_k are sorted according to the following rule: $l_i \leq l_{i+1}$ for $i = 1, ..., \mid C_k \mid$-1. Then, the final time $l^f{}_i$ for all the m-nodes in C_k is computed as follow:

$$l^f{}_i = l_1 + (q * (i-1)) \quad for \quad i = 1, ..., \mid C_k \mid \tag{4}$$

Where the value q is the same used in colouring algorithms. In a nutshell, we found the object in each cluster C_k, which first occurs in the condensed video; then, we force a minimum time distance q between two consecutive objects involved in a potential collision to get a better condenseness. Low values of q improve condenseness, but introduce objects overlaps in the final video. A proper value for q should be chosen according to video resolution, objects size and speed.

3. EXPERIMENTAL RESULTS

We tested our approach on a synthetic dataset randomly generated through our toolbox. To this aim, we generated 40 videos of different length and different number of objects. Videos size, in frames, falls in 2646-38812, while the probability to generate a new object falls in $0.002 - 0.010$; the number of objects falls in $16 - 79$. We have chosen a resolution of 960×540 and a direction parallel to y axis to simulate a traffic camera. However, we also performed a comparison on real videos taken by UA-DETRAC Benchmark Suite [22, 23], which consists of traffic scenes. Unfortunately, most of the UA-DETRAC videos are short or include vehicles which do not leave the scene. This makes impossible to condense such videos, because tubes length cannot be changed. Hence we selected 10 videos where the synopsis is feasible.

To quantitatively evaluate the proposed approach, we use the same metrics employed in [13]. These metrics, allow to measure Frame Condensation Ratio (FR), Frame Compact Rate (CR) and Overlap Ratio (OR). FR gives information about the video length reduction operates by Video Synopsis. CR is used to evaluate if the foreground in condensed video is rearranged closely. A better Video Synopsis shows high CR value. Finally, OR indicates the collision degree of the object, hence it should be as lower as possible. Metrics formulas can be found in [13]. Experiments have been conducted with $q = 15$ and $th = 10$. The value of q has been chosen according to the object size and object speed [13]. Low values for q involve better compactness (i.e., CR and FR) but promote the overlapping artifacts (i.e, OR); viceversa, high values of q reduce the overlaps but penalize the compactness. We run experiments for each video and then we computed the ratio between results obtained with the proposed approach and with the one proposed by He et al. [13] (e.g., $FR_{ratio} = FR_{our}/FR_{He}$). However, we present average results. In Table 1, we report the comparison on both the datasets, synthetic and UA-DETRAC. It can be seen that our method outperforms the state-of-art in terms of FR, CR, OR and it is even faster. Specifically, our synopsis for synthetic videos is about 25% shorter; we also achieve better compactness (about 50% more) and object overlapping reduction of 15%. Regarding the execution time, our algorithm requires 5% less than the original one. Experiments on UA-DETRAC dataset show similar results; this proves our synthetic dataset can be successfully used for testing tubes rearrangement strategies. The main difference on UA-DETRAC videos, is our algorithm performs twice faster (50% less). It depends from the fact that UA-DETRAC videos present more chaotic situations in which the solution for equation (1) is larger. Hence, the proposed closed form in (2) makes the algorithm faster.

Table 1. Results on synthetic and UA-DETRAC datasets.

Dataset	FR ratio	CR Ratio	OR Ratio
Synthetic	0.7637	1.4822	0.7241
UA-DETRAC	0.9013	1.3794	0.7168

4. CONCLUSION

In this work, we explore object-based Video Synopsis and propose to address tubes rearrangements stage separately. Hence, we introduce a new toolbox to generate synthetic videos and the related annotation files bypassing detection and tracking steps. Additionally, we propose an improvement of the method in [13] and use a generated synthetic dataset to prove our proposal achieves better results. As further validation, we test our algorithm in real videos taken from UA-DETRAC dataset. In future works, we aim to extend toolbox functionalities by introducing curve trajectories and perspective transformation. Furthermore, we are investigating new strategies based on collision graph that can modify the length of specific tubes, namely objects speed.

5. REFERENCES

[1] K. A. Joshi and D. G. Thakore, "A survey on moving object detection and tracking in video surveillance system," *International Journal of Soft Computing and Engineering*, vol. 2, no. 3, pp. 44–48, 2012.

[2] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Fall detection from human shape and motion history using video surveillance," in *International Conference on Advanced Information Networking and Applications Workshops*. IEEE, 2007, vol. 2, pp. 875–880.

[3] M. Irani, P. A. Anandan, J. Bergen, R. Kumar, and S. Hsu, "Efficient representations of video sequences and their applications," *Signal Processing: Image Communication*, vol. 8, no. 4, pp. 327–351, 1996.

[4] B.M. Wildemuth, G. Marchionini, Meng Yang, G. Geisler, T. Wilkens, A. Hughes, and R. Gruss, "How fast is too fast? evaluating fast forward surrogates for digital video," in *Joint Conference on Digital Libraries*, 2003.

[5] X. Zhu, X. Wu, J. Fan, A. K. Elmagarmid, and W. G. Aref, "Exploring video content structure for hierarchical summarization," *Multimedia Systems*, vol. 10, no. 2, pp. 98–115, 2004.

[6] T. Liu, X. Zhang, J. Feng, and K. T. Lo, "Shot reconstruction degree: a novel criterion for key frame selection," *Pattern Recognition Letters*, vol. 25, no. 12, pp. 1451–1457, 2004.

[7] Y. F. Ma and H. J. Zhang, "A model of motion attention for video skimming," in *International Conference on Image Processing*, 2002.

[8] A. Rav-Acha, Y. Pritch, and S. Peleg, "Making a long video short: Dynamic video synopsis," in *Computer Vision and Pattern Recognition*, 2006, vol. 1, pp. 435–441.

[9] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg, "Webcam synopsis: Peeking around the world," in *International Conference on Computer Vision*, 2007, pp. 1–8.

[10] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological video synopsis and indexing," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 30, no. 11, pp. 1971–1984, 2008.

[11] Y. Pritch, S. Ratovitch, A. Hendel, and S. Peleg, "Clustered synopsis of surveillance video," in *Advanced Video and Signal Based Surveillance*, 2009, pp. 195–200.

[12] X. Li, Z. Wang, and X. Lu, "Surveillance video synopsis via scaling down objects," *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 740–755, 2016.

[13] Y. He, C. Gao, N. Sang, Z. Qu, and J. Han, "Graph coloring based surveillance video synopsis," *Neurocomputing*, vol. 225, pp. 64–79, 2017.

[14] Y. He, Z. Qu, C. Gao, and N. Sang, "Fast online video synopsis based on potential collision graph," *Signal Processing Letters*, vol. 24, no. 1, pp. 22–26, 2017.

[15] S. Battiato, G. Gallo, G. Puglisi, and S. Scellato, "SIFT features tracking for video stabilization," in *International Conference on Image Analysis and Processing*, 2007, pp. 825–830.

[16] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M. M. Cheng, S. L. Hicks, and P. H. Torr, "Struck: Structured output tracking with kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2096–2109, 2016.

[17] Z. He, S. Yi, Y. M. Cheung, X. You, and Y. Y. Tang, "Robust object tracking via key patch sparse representation," *IEEE Transactions on Cybernetics*, vol. 47, no. 2, pp. 354–364, 2017.

[18] S. Battiato, G. M. Farinella, A. Furnari, G. Puglisi, A. Snijders, and J. Spiekstra, "An integrated system for vehicle tracking and classification," *Expert Systems with Applications*, vol. 42, no. 21, pp. 7263–7275, 2015.

[19] T. Yao, M. Xiao, C. Ma, C. Shen, and P. Li, "Object based video synopsis," in *Workshop on Advanced Research and Technology in Industry Applications*, 2014.

[20] X. Li, Z. Wang, and X. Lu, "Video synopsis in complex situations," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3798–3812, 2018.

[21] J. R. Griggs and R. K. Yeh, "Labelling graphs with a condition at distance 2," *Journal on Discrete Mathematics*, vol. 5, no. 4, pp. 586–595, 1992.

[22] L. Wen, D. Du, Z. Cai, Z. Lei, M-C. Chang, H. Qi, J. Lim, M-H. Yang, and S. Lyu, "UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking," *arXiv preprint arXiv:1511.04136*, 2015.

[23] S. Lyu, M-C. Chang, D. Du, L. Wen, H. Qi, Y. Li, Y. Wei, L. Ke, T. Hu, M. Del Coco, et al., "UA-DETRAC 2017: Report of AVSS2017 & IWT4S challenge on advanced traffic monitoring," in *International Conference on Advanced Video and Signal Based Surveillance*, 2017, pp. 1–7.