Synopsis of Video Files Using Neural Networks: Component Analysis

Georgi Kostadinov(✉)

New Bulgarian University, 21 Montevideo str, 1618 Sofia, Bulgaria
grgkostadinov@gmail.com

Abstract. The following paper provides a detailed analysis of the components of a novel framework for generating synopsis of CCTV videos. A synopsis is a video file obtained by overlaying the main objects from a source video on a single scene. This allows for a file length reduction and optimization of the storage of such files. This paper extends the presented work based on convolutional neural networks by discussing the effect that such algorithms may have on the final synopsis result which in turn helps in understanding how they can be further improved for this task. For the purposes of the component analysis presented in this paper, specialized datasets and metrics were selected to quantify the quality of the algorithms of the video synopsis framework.

Keywords: Video synopsis · Convolutional neural networks · Machine learning · Object localization · Multiple object tracking · Feature extraction · Background segmentation · Person re-identification

1 Introduction

1.1 Current Challenges and Opportunities

Recent years have witnessed an explosive growth of video surveillance technologies. Constant video monitoring is required in order for governments and owners of private properties to ensure the security and safety of their people and assets. However, the effectiveness of analysis and storage of raw videos remain a challenge. In most of the cases in video surveillance, the static repetitive frames can have a length of hours whilSe the segments that contain useful information are no more than several seconds. An emerging solution called video synopsis is developed to cope with these challenges. It enables the reduction of hours of video footage in minutes showing the most important events while at the same time retains the quality of the source video.

In [1], an end-to-end framework is proposed, that can generate a synopsis of CCTV video footage of pedestrians using convolutional neural networks (CNN). A "synopsis" of a video file is an output file containing only the main moving objects from the source video file, placed together on the extracted background. The output file generated during the process is much smaller in size compared to the source file, making its storage cost-effective, and at the same time shorter in length, making the manual analysis and review of such CCTV footage a much easier process.

T. Zlateva and R. Goleva (Eds.): CSECS 2022, LNICST 450, pp. 11–23, 2022.
https://doi.org/10.1007/978-3-031-17292-2_2

1.2 Classification of Video Synopsis Methods

Over the years numerous methods for video condensation have been developed. They can be classified broadly in two categories: frame-based and object-based approaches.

Frame-based approaches are extracting the key frames from the source video either by skipping several frames or analysing each of them individually using certain criteria and extracting only those with the highest importance. The extracted frames are then blended to create the output video summarization. Earlier methods are based on video skimming approach [2], which skip several low interest frames. More advanced methods [3] analyse the structured motion of the frames to extract the key ones. Although frame-based approaches are simple and fast condensation methods, they result in unrealistic artifacts due to the frame skipping and are ineffective with highly dynamic scenes with little to no repetitive frames.

Modern video condensation methods are object-based. Such methods extract sequences of the main moving objects in the source video called *tubes* that are shifted along temporarily and placed together on the extracted background. This allows for the simultaneous visualization of all moving objects from the source video allowing for higher condensation ratio than the frame-based approaches. For example, both [4] and [5] are using estimation algorithms where object's positions are chronologically rearranged with no trajectory analysis. Whereas [6] is clustering the objects trajectories via event-based trajectory kinematics descriptors extracted from each object.

The rest of the paper is structured as follows. Section 2 discusses the details of the synopsis framework for videos with pedestrians presented in [1] as well as the importance of understanding the effect each component has on the final output. Section 3 analyses the background extraction method, whereas Sect. 4 measures the quality of the pedestrian localization, tracking, and pedestrian re-identification algorithms. Finally, in Sect. 5 conclusions are drawn, and future work is presented.

2 Video Synopsis Framework Overview

In [1] a novel framework for creating a synopsis of CCTV footage of pedestrians is presented. The framework can be divided into five main components: (1) extracting the background using mixture of Gaussian models (GMM) [7]; (2) localization of pedestrians using the convolutional neural network *You Only Look Once v3* (YOLOv3) [8]; (3) extraction of their visual features for accurate re-identification via the CNN for person re-identification *Omni-Scale Feature Learning for Person Re-Identification* (OSNet) [9]; (4) tracking them in the source video scenes using *Deep Simple Online and Realtime Tracking* (DeepSORT) [10] and the visual features from (3); and (5) generating the synopsis file for the joint visualization of the tracked identities from (4) on the extracted background from (1). The structure of the framework is visualized on Fig. 1.

The synopsis is generated via two main processes – *analysis* and *generation*. During the *analysis*, each frame is fed into the object localization algorithm YOLOv3 and the pedestrians are localized and segmented. For each pedestrian a structure determining its spatial-temporal data is created that includes the position and the time of occurrence in the original video. Moreover, using OSNet the visual features for each pedestrian localized on the scene are extracted. These visual features are later used as part of the

DeepSORT algorithm to re-identify the pedestrians in subsequent frames based on the accumulated state of the previous frames. This way the object tubes or in the context of pedestrians – their *identities*, are created. Each identity is a sequence of the occurrences of a given pedestrian from the source video. In parallel, the background is extracted using the algorithm from [7]. The output from the analysis step is both the extracted background and tracked identities. The *generation* process takes the tracked identities, filters any whose occurrence is low, and uses two control parameters to render the identities on the extracted background – m for the maximum number of identities to render at once and k for the number of frames to wait before rendering a new identity on the scene. The final output of the process is the rendered synopsis file where each identity is temporarily shifted along for their joint visualization.

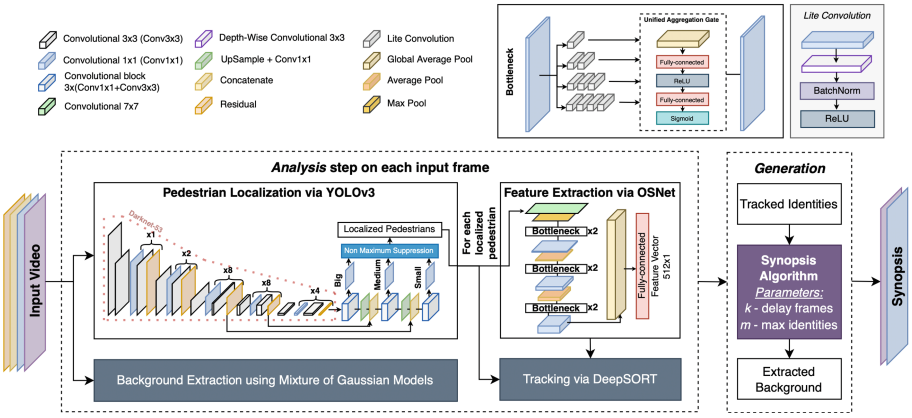

Fig. 1. Overview of the presented video synopsis framework in [1]. Two convolutional neural networks are used – YOLOv3 [8] for localization of the pedestrians and OSNet [9] for the extraction of their visual features. The localized pedestrians are then tracked via DeepSORT [10] and put on the extracted using mixture of Gaussian models [7] background to create the synopsis.

Comparing the presented framework in [1] with prior art [4–6] shows that the frame reduction is much higher while maintaining the same visual quality as in the source video. The comparison results achieved in [1] on a CCTV video of a hall room from [5] are presented in Table 1 (the video synopsis framework in [1] is marked as **VSF**) and the synopsis output is visualized on Fig. 2.

Table 1. Comparison with prior art on the hall video dataset [5]. *SynopsisO* is the frame reduction rate or the ratio between the number of synopsis frames F_o and number of source frames.

Metric	Huang et al. [4]	Huang et al. [5]	Wang et al. [6]	**VSF** [1]
F_o	14379	11271	8814	**1566**
SynopsisO	0.785	0.831	0.868	**0.977**

Fig. 2. Visualization of the hall dataset [5] synopsis.

As seen from the results in Table 1, the framework presented in [1] achieves state-of-the-art frame reduction rate. However, being a multi-component solution, no further analysis has been presented in [1] to understand the effect that each component may have on the final synopsis file, and which are the situations where they work sub-optimally. To do so, each component must be examined individually using specialized datasets and metrics for each task to get a quantitative measure of its accuracy and the effect on the final output. Moreover, the achieved results can be used as a base to determine how these algorithms could be further improved.

This paper presents a methodology of such analysis for each task and to the extent of the conducted research, no prior works, that discuss the effect each component used in a multi-component object-based video synopsis algorithm may have, exist.

3 Analysis of the Background Extraction

The algorithm for the adaptive Gaussian mixture model presented in [7] and shown on Fig. 1, is an important component of the video synopsis framework. A mixture of Gaussian models is used to model the multimodal background image of the source video. The algorithm is adaptive, which means that it iteratively updates the background mixture model to cope with illumination changes and scene noise. The extracted background is then used as a base where the tracked identities will be later rendered. This means that the algorithm must work as precisely as possible in segmenting the background from the foreground. Failing to do so will result in a lesser quality of the final synopsis output. A quantitative assessment of the quality of the algorithm must be generated to understand the effect that such algorithms have on the final output.

3.1 Control Dataset

For the overall assessment of the algorithm, a specialized dataset for motion detection called CDnet 2014 is used [11]. The goal of CDnet is to provide a balanced dataset with multiple scenarios that are common in motion recognition algorithms. The dataset includes 53 video files with a total of ~158,000 frame annotations. In addition, the video files are divided into 11 categories, each of which represents a different type of challenge – from easy to segment videos, to more challenging scenes with bad weather or night vision. CDnet's annotations are hand-made images for each frame of a video, which on a pixel level annotate the moving objects on the stage.

The official metrics presented in [11] were used to generate the results and rank them with other methods: *Recall* (Re), *Specificity* (Sp), *False Positive Rate* (FPR), *False Negative Rate* (FNR), *Percentage of Wrong Classifications* (PWC), *Precision* (Pr) and *F1-Score* (F1). For their specific descriptions, refer to [11].

3.2 Results

Measurements for the algorithm used in the synopsis framework [1] were obtained by calculating the metrics presented by CDnet for each category from the dataset. Then an average of these per-category metrics was calculated to obtain the overall results. The final results are presented in Table 2 and marked as **VSF**. They are also compared with the results of six other background segmentation methods, calculated in the same way. The data of the other methods is taken from the public results page of CDnet.

Table 2. Comparison of the background extraction algorithm used in [1] with other methods. Ordered in descending order of *F1*. Arrows are an indicator of low or high optimal values.

Methodology	*Re* ↑	*Sp* ↑	*FPR* ↓	*FNR* ↓	*PWC* ↓	*Pr* ↑	*F*1 ↑
FTSG [12]	0.77	**0.99**	**0.01**	0.23	**1.38**	**0.80**	**0.73**
SuBSENSE [13]	**0.81**	**0.99**	**0.01**	**0.19**	1.84	0.75	**0.73**
CwisarDH [14]	0.66	**0.99**	**0.01**	0.34	1.53	0.77	0.68
VSF [1]	0.64	0.98	0.02	0.36	3.32	0.68	0.58
CP3-online [15]	0.72	0.97	0.03	0.28	3.43	0.56	0.58
GMM [7]	0.68	0.98	0.03	0.32	3.77	0.60	0.57
Euclidean [16]	0.68	0.94	0.06	0.32	6.54	0.55	0.52

The presented methods in Table 2 show both more classical algorithms for background segmentation such as Euclidean distance [16] and GMM [7], and recent developments such as FTSG [12] and SuBSENSE [13]. [16] is a simplified method for background segmentation, which makes a direct comparison of pixels in two consecutive frames. This is followed by many errors, thus higher percentage of PWC (classification errors). GMM Stauffer-Grimson [7] is the adaptive Gaussian mixtures model of Stauffer and Grimson, that the algorithm in [1] is based on. In CP3-online [15], instead of modelling each pixel individually, the colour distribution of the pixels is modelled with strong spatial correlation. According to the authors, this type of spatial model copes with abrupt changes in scene's lighting.

The algorithm in the video synopsis framework (VSF) [1], which is based on [7] shows an improvement in accuracy and a lower error rate than the original work. This is due to the additional steps to post-process the resulting segmentation by blurring and morphological transformations to remove smaller regions of pixels that have been misclassified as foreground by the algorithm. Figure 3 shows an example segmentation of a video frame from CDnet, as well as the result of the additional processing that is performed by VSF.

Newer methodologies significantly improve F1-Score, precision (Pr) and reduce error rate (PWC). CwisarDH [14] uses a neural network that, for any set of previous pixel values in the image, approximates corresponding values of background or foreground components. The main disadvantage of this approach is the need for the neural network to be trained with pre-segmented frames. In [13], colour and local binary signs of similarity

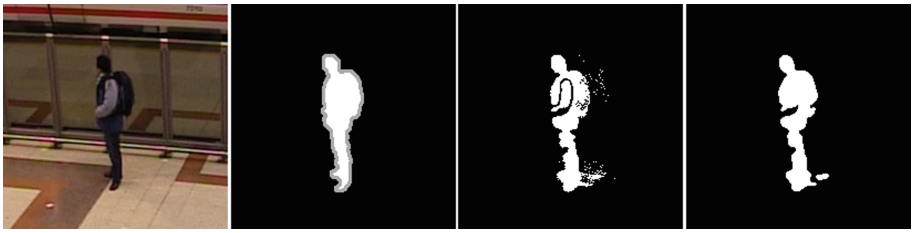

Fig. 3. Sample segmented frame via VSF [1]. Left to right: video frame snippet from CDnet [11]; CDnet frame annotations; GMM [7] output before processing; final result after blur processing, threshold filtering and morphological transformation.

are used to give approximations at the pixel level. In addition, it automatically adjusts the parameters and adapts locally according to the dynamics of the scene. FTSG [12] is a three-step algorithm: (1) recognizing moving objects using a pixel energy flux tensor, and a GMM [7] algorithm; (2) combining the recognized scene movements; and (3) removing artifacts obtained from stagnation of moving objects in one place. The results show that, despite the improvement in F1-Score compared to the baseline GMM model, the algorithm in [1] is far behind state-of-the-art methods that are either using neural networks or based on more complex frame analysis. However, to better understand where the algorithm in [1] lacks the necessary accuracy, Table 3 presents a more thorough analysis for each dataset category showing the F1-Score of the algorithm.

Table 3. Results of the background extraction from [1] for the CDnet dataset categories [11].

Video	Description	$F1$
badWeath	Different weather conditions, 4 videos	0.79
baseline	Easy to segment foreground, 4 videos	0.79
shadow	Dynamic scene shadow, 6 videos	0.75
dynamic	Dynamic movement on the background, 6 videos	0.68
cameraJ	Unstable CCTV footage, 4 videos	0.64
thermal	Infrared thermal sensor videos, 5 videos	0.62
turbule	Infrared videos with turbulence due to hot air, 4 videos	0.60
lowFram	Low FPS videos, 4 videos	0.50
intermi	With objects that are static for long periods of time, 6 videos	0.42
nightVi	Videos captured in low-light conditions, 6 videos	0.40
PTZ	From cameras with pan, tilt, and zoom control, 4 videos	0.22

The results in Table 3 indicate that the hardest category is the one from PTZ cameras due to the pan, tilt, and zoom movements of the camera. This result is expected as the algorithm used in the framework which is based on GMM [7] relies on the fact that the scene composition will be mostly static. The lack of lighting in the *nightVi* category leads

to much higher number of pixels being recognized as a background, whereas the result for the *intermi* category shows the inability of the algorithm to adapt if an object has stayed static for a longer period of time. Moreover, the lower frame rate in the *lowFram* category creates situations where the algorithm does not adapt fast enough. Sample background visualizations of the worst performing categories are shown on Fig. 4. The PTZ video background visualized is corrupted due to the pan and tilt movements of the camera. Meanwhile, the background for both *nightVi* and *lowFram* videos show artefacts due to either the low light conditions or the lower frame updates.

Fig. 4. Sample extracted backgrounds by [1] for the worst performing CDnet [11] categories.

The results on the CDnet dataset show that a more robust methodology needs to be developed in [1] to be able to cope with low-light conditions and adapt quickly to sudden scene changes. This will increase both the quality of the backgrounds extracted in such conditions and the overall quality of the final synopsis output.

4 Analysis of the Pedestrian Localization, Extraction, and Tracking Components

As visualized in Fig. 1, the algorithms for pedestrian localization, re-identification, and tracking are core to the video synopsis framework. If the localization algorithm does not recognize a pedestrian in a sequence of frames, it will be also missing in the synopsis video. On the other hand, if the feature extraction algorithm for re-identification does not extract robust-enough features, the tracking algorithm won't be able to correctly identify the identities on the scene. To understand the impact of these algorithms to the final output, a control dataset of CCTV videos of pedestrians is required that contains both annotations of their positions within each frame as well as their identities.

4.1 Control Dataset

Based on the defined requirements, the datasets from the MOTChallenge [17] competition were selected. Being a specialized dataset for multi-object tracking algorithms, MOTChallenge consists of several video files that capture pedestrians in busy public places with frequent, partial, or complete occlusions from other objects. Videos differ in shooting angle, scene brightness, subject size, camera sensor quality. The competition consists of several different tracking tests. For testing the localization, tracking, and re-identification algorithms, the test dataset MOT17 [18] has been selected.

The MOT17 dataset consists of a total of 7 videos that have public results from three different localization algorithms – DPM [19], Faster R-CNN [20] and SDP [21]. Moreover, it has additional detailed annotations for each frame that can be used to measure the quality of the tracking algorithm. One such annotation contains an identifier of the object, a region with its exact position in the frame, a value between 0 and 1, indicating how occluded the object is, as well as the type of object. Sample video frames for each of the 7 videos from MOT17 are visualized on Fig. 5.

Fig. 5. Sample frames from the MOT17 dataset [18].

MOTChallenge also includes official metrics for calculating the performance of an algorithm on their test sets. For measuring the accuracy of the tracking and re-identification algorithms, the following metrics from MOTChallenge were used: *Multiple Object Tracking Accuracy* (MOTA), *Multiple Object Tracking Precision* (MOTP). And for measuring the precision of the object localization algorithm *Recall* (Re), *Precision* (Pr), *F1-Score* (F1) were selected. For their specific descriptions, refer to [17].

4.2 Results for Pedestrian Localization with YOLOv3

YOLOv3 [8] consists of a single feed-forward 53-layer convolutional neural network Darknet-53 [8] as well as three scaled detection layers for detecting small, medium, and large objects. The algorithm considers the object localization task as a regression problem and in a single feed-forward pass can make predictions both for the regions of the objects within the image as well as their class affiliation. A visualization of its architecture is presented on Fig. 1. At the final stage, the three scaled detection layers are fused together and post-processed with a Non-maximum Suppression [22] algorithm to produce the localization output. Being a one-stage object detector, YOLOv3 is a very efficient and fast algorithm, making it an extremely good choice for real-time video processing applications.

The YOLOv3 algorithm in the video synopsis framework (VSF) is trained on the *person* category from two datasets – *Common Objects in Context (COCO)* [23] and *Pascal Visual Objects Classes (VOC)* [24]. The final dataset consists of 66,109 images for training and *4,786* for validation. Training on *15,000* iterations with learning rate of *0.001* achieved a validation accuracy of *93%*. However, measurement of the real-world performance is required in order to understand how the algorithm behaves in different occlusion situations or scene changes, all of which are present in the MOT17 dataset.

To test the localization algorithm, the metrics *Re*, *Pr*, and *F1* were calculated on each video of the MOT17 dataset and then averaged to get the overall results for the whole dataset. Moreover, public results for the algorithms DPM [19], Faster R-CNN [20], and SDP [21] were used to compare the results of the YOLOv3 algorithm used in the framework. DPM is a model that combines different parts of pedestrian recognition objects, but due to the hand-built descriptors it is inaccurate in abrupt changes in scene lighting or in recognizing pedestrians from a distance. The other two algorithms – Faster R-CNN and SDP, are using two neural networks to localize objects – one for region proposals and the other for classification of these regions. The difference between the two algorithms is that SDP is using an additional algorithm to clean the regions so that the bounding boxes approximations are more accurate.

Table 4 presents the results for each method. The YOLOv3 model used in the video synopsis framework is labelled as **VSF**. The frames per second (FPS) values have been measured on the same hardware – NVIDIA GTX 1080Ti, for SDP, VSF, and Faster R-CNN. The DPM algorithm was tested on a 3.4 GHz Intel CPU.

Table 4. Results of the YOLOv3 [8] localization algorithm compared with the results of the methods used in MOT17 [18]. Ordered by F1-Score. Arrows indicate low or high optimal values.

Methodology	*Re* ↑	*Pr* ↑	*F*1 ↑	*FPS* ↑
SDP [21]	**0.682**	0.954	**0.795**	4
VSF [1]	0.624	0.916	0.742	**78**
Faster R-CNN [20]	0.520	**0.958**	0.674	5.6
DPM [19]	0.314	0.935	0.470	30

Based on the results presented on the table and in terms of F1-Score, YOLOv3 proves to be the most optimal method for the video synopsis framework out of these four, as it achieves 19.5 times faster frame rate compared to SDP for 5.3% lower *F1*. However, further training of the algorithm is required to surpass the localization accuracy of SDP.

4.3 Results for Pedestrian Tracking with DeepSORT

DeepSORT [10] is tracking-by-detection algorithm that uses Kalman filtering for tracking multiple objects localized in new frames based on how their state changed in past frames and a convolutional neural network for object re-identification. The DeepSORT algorithm in the video synopsis framework is also responsible for creating the identities within a single video, making it an important component that has a direct impact on the final synopsis. The inability to correctly identify the pedestrians, can lead to less identities created, thus missing important frames from the source video.

For testing the tracking algorithm, the *MOTP* and *MOTA* metrics were calculated. Outputs from DPM, SDP, and Faster R-CNN were also used as tracked object regions to compare how the tracking algorithm's performance changes with methods other than the YOLOv3 used in VSF. As seen from the results in Table 5 the YOLOv3 algorithm,

when paired with DeepSORT, is both less precise and accurate than SDP and Faster R-CNN. This is since YOLOv3 approximates the regions in a single feed-forward pass instead in a dedicated neural network as in Faster R-CNN and SDP. This leads to less precise and varying regions, therefore lower tracking precision.

Table 5. Results of the DeepSORT [10] algorithm. Arrows indicate low or high optimal values.

Methodology	*MOTP* ↑	*MOTA* ↑
SDP [21]	0.832	**0.643**
VSF [1]	0.798	0.558
Faster R-CNN [20]	**0.883**	0.493
DPM [19]	0.776	0.290

4.4 Results for Pedestrian Re-identification with OSNet

An essential component of a tracking algorithm is the ability to re-identify the same object in a new video frame. The most common way is by creating an associative matrix, whose elements determine a score whether a tracked object correlates to a newly localized object. In SORT [25] such score is calculated using the Mahalanobis distance between the object states from the Kalman filter and the objects from the localization algorithm. SORT does not use any additional data such as object's visual similarity. However, this leads to an issue where the identities of two objects can be misidentified or switched by the tracking algorithm when their paths cross on the same scene. DeepSORT improves SORT and mitigates this issue by adding a second distance metric to the final score in the associative matrix – an Euclidean distance between the visual features of a given localized object and those of the last 100 tracked objects. These visual features are encoded using a convolutional neural network.

The convolutional neural network used to extract the visual features of the pedestrians in the video synopsis framework from [1] as shown in Fig. 1 is OSNet [9]. OSNet is a specialized CNN for pedestrian re-identification that uses residual blocks composed of convolutional streams for detecting features in different spatial scales – from small, local features (shoes, glasses) to more global, bigger features (size, age, clothing). It also uses a novel unified aggregation gate [9] that fuses the different-scaled features together with the input-dependent weights to learn spatial correlations. The final output is a 512-D feature vector that is used for re-identification as part of DeepSORT. The OSNet model used as part of the framework is trained on three datasets for person re-identification: *Market1501* [26], *DukeMTMC* [27], and *CUHK03* [28].

To test the accuracy of the OSNet model, the *MOTA* metric is calculated for SORT and compared with the same for the video synopsis framework (VSF) using DeepSORT and OSNet from Table 5. Table 6 presents these results with an additional column IDSW which is the number of times there was an identity switch. The results indicate that the accuracy improvement of using OSNet is 3.7% and, at the same time, this leads to

39.4% less identity switches. Figure 6 visualizes these results. Although they show a major improvement in decreasing the identity switches, the total number for [1] is still too high. Improvements to the algorithm will be required to decrease them even further.

Table 6. Comparison of the results on the MOT17 [18] dataset for the VSF [1] (that uses DeepSORT [10] and OSNet [9]) and SORT [25]. Arrows indicate low or high optimal values.

Methodology	*MOTA* ↑	*IDSW* ↓
VSF [1]	**0.558**	**941**
SORT [25]	0.521	1554

Fig. 6. Frames from MOT17 [18]. Top row are results using SORT [25], bottom row – using VSF [1] (DeepSORT [10] and OSNet [9]). VSF correctly identifies the pedestrian in red rectangle with an identity 3, whereas SORT incorrectly assigns him three different identities – 48, 65, 75.

5 Conclusions

This paper reviewed and analysed the components of the video synopsis framework presented in [1]. Results were generated on specialized datasets, namely CDNet [11] for evaluation of the background extraction algorithm and MOTChallenge [17] for the pedestrian localization, tracking, and re-identification algorithms. Moreover, the algorithms in [1] were compared with other methodologies. Results show that, despite being optimal for the framework, cases exist where the algorithms are not precise enough – there are still high number of identity switches from the tracking algorithm and the background subtraction algorithm could be further improved to handle more challenging scene situations. Future work will include incremental improvements to the algorithms increasing their accuracy in order to have higher quality synopsis videos.

References

1. Kostadinov, G.: Synopsis of video files using neural networks. In: Proceedings of the 23rd EANN 2022. https://doi.org/10.1007/978-3-031-08223-8_16

2. Smith, M.A., Kanade, T.: Video skimming and characterization through the combination of image and language understanding. In: Proceedings CAIVD, pp. 61–70 (1998)
3. Fu, W., Wang, J., Gui, L., Lu, H., Ma, S.: Online video synopsis of structured motion. Neurocomputing **135**, 155–162 (2014)
4. Huang, C.R., Chen, H.C., Chung, P.C.: Online surveillance video synopsis. In: IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1843–1846. IEEE (2012)
5. Huang, C.R., Chung, P.C.J., Yang, D.K., Chen, H.C., Huang, G.J.: Maximum a posteriori probability estimation for online surveillance video synopsis. IEEE Trans. Circuits Syst. Video Technol. **24**(8), 1417–1429 (2014)
6. Wang, W.C., Chung, P.C., Huang, C.R., Huang, W.Y.: Event based surveillance video synopsis using trajectory kinematics descriptors. In: Fifteenth IAPR International Conference on Machine Vision Applications, pp. 250–253 (2017)
7. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In CVPR (1999)
8. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv preprint arXiv: arXiv:1804.02767 (2018)
9. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: ICCV, pp. 3702–3712 (2019)
10. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: ICIP, pp. 3645–3649 (2017)
11. Wang, Y., Jodoin, P.-M., Porikli, F., Konrad, J., Benezeth, Y., Ishwar, P.: Cdnet 2014: an expanded change detection benchmark dataset. In: CVPR, pp. 387–394 (2014)
12. Wang, R., Bunyak, F., Seetharaman, G., Palaniappan, K.: Static and moving object detection using flux tensor with split Gaussian models. In: CVPR, pp. 414–418 (2014)
13. St-Charles, P.-L., Bilodeau, G.-A., Bergevin, R.: Flexible background subtraction with self-balanced local sensitivity. In: CVPR, pp. 408–413 (2014)
14. De Gregorio, M., Giordano, M.: Change detection with weightless neural networks. In: CVPR, pp. 403–407 (2014)
15. Liang, D., Kaneko, S.: Improvements and experiments of a compact statistical background model. arXiv preprint arXiv:1405.6275 (2014)
16. Benezeth, Y., Jodoin, P.-M., Emile, B., Laurent, H., Rosenberger, C.: Comparative study of background subtraction algorithms. J. Electron Imaging **19** (2010)
17. Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.: Motchallenge 2015: towards a benchmark for multi-target tracking. arXiv preprint arXiv:1504.01942 (2015)
18. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: a benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
19. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. TPAMI **32**(9), 1627–1645 (2009)
20. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. In: ANIPS, pp. 91–99 (2015)
21. Yang, F., Choi, W., Lin, Y.: Exploit all the layers: fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers. In: CVPR (2016)
22. Neubeck, A., Van Gool, L.: Efficient non-maximum suppression. In: ICPR (2006)
23. Lin, T.Y., et al.: Microsoft coco: common objects in context. In: ECCV, pp. 740–755 (2014)
24. Hoiem, D., Divvala, S.K., Hays, J.H.: Pascal VOC 2008 challenge. World Literature Today, p. 24 (2009)
25. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: ICIP, pp. 3464–3468 (2016)
26. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: ICCV, pp. 1116–1124 (2015)

27. Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In ECCV, pp. 17–35. (2016)
28. Li, W., Zhao, R., Xiao, T., Wang, X:. Deepreid: deep filter pairing neural network for person re-identification. In: CVPR, pp. 152–159 (2014)