RSVP: A Real-Time Surveillance Video Parsing System with Single Frame Supervision

Han Yu*
SKLOIS, IIE, CAS
School of Cyber Security, UCAS
yuhan@iie.ac.cn

Guanghui Ren*
SKLOIS, IIE, CAS
School of Cyber Security, UCAS
sundrops.ren@gmail.com

Ruihe Qian
SKLOIS, IIE, CAS
School of Cyber Security, UCAS
qianruihe@iie.ac.cn

Yao Sun
SKLOIS, IIE, CAS
sunyao@iie.ac.cn

Changhu Wang
Toutiao AI Lab
wangchanghu@toutiao.com

Hanqing Lu
IA, CAS
luhq@nlpr.ia.ac.cn

Si Liu†
SKLOIS, IIE, CAS
Jiangsu Key Laboratory of Big Data Analysis Technology, NUIST
Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, NUIST
liusi@iie.ac.cn

ABSTRACT

In this demo, we present a real-time surveillance video parsing (RSVP) system to parse surveillance videos. Surveillance video parsing, which aims to segment the video frames into several labels, e.g., face, pants, left-legs, has wide applications, especially in security filed. However, it is very tedious and time-consuming to annotate all the frames in a video. We design a RSVP system to parse the surveillance videos in real-time. The RSVP system requires only one labeled frame in training stage. The RSVP system jointly considers the segmentation of preceding frames when parsing one particular frame within the video. The RSVP system is proved to be effective and efficient in real applications.

CCS CONCEPTS

• **Computing methodologies** → **Pattern Analysis and Machine Intelligence**;

KEYWORDS

human parsing, deep learning, surveillance video parsing system

ACM Reference format:
Han Yu, Guanghui Ren, Ruihe Qian, Yao Sun, Changhu Wang, Hanqing Lu, and Si Liu. 2017. RSVP: A Real-Time Surveillance Video Parsing System with Single Frame Supervision. In *Proceedings of ACM Multimedia conference, Mountain View, CA USA, October 2017 (MM'17),* 3 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Both authors contributed equally to the paper
†corresponding author

MM'17, October 2017, Mountain View, CA USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Due to the wide range of applications, such as person reidentification and person attribute prediction, human parsing is growing rapidly in recent years. Insufficient labeled data always lead to over-fitting, especially in the deep learning based methods. However, using the context information is not common in most existing human parsing methods due to the lack of temporal information. But a large quantity of context information can be exacted from surveillance videos. Inspired by this, a system making full use of context information in surveillance videos is designed in this demo.

In this demonstration, we propose a RSVP system to parse surveillance videos in real-time and get outstanding parsing results over state-of-the-arts. The RSVP system leverages the very limited labeled images and the large amount of unlabeled images [3] with online estimated dense correspondences among video frames. The RSVP system can also solve the over-fittting problem caused by insufficient labeled data which is common in the deep learning based methods.

The RSVP system is consisted of three modules: video capturing module, video parsing module and result displaying module. The framework of this system is shown in Figure 1. In the video capturing module, a camera is used to capture a video. After acquiring the video frames, the video parsing module outputs the segmentation results of these video frames. This module uses Single frame Video Parsing(SVP) network [3] which takes the temporal relation between video frames into consideration. Finally, the result displaying module will display the captured video and its corresponding parsing results simultaneously in real-time.

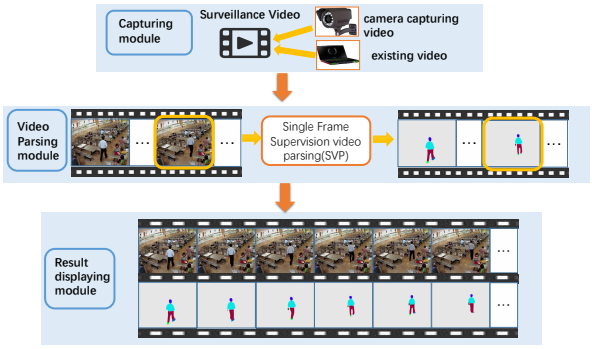

Figure 1: The framework of RSVP system.

To extend the application of this system, parsing existing video files is also supported in RSVP system. While parsing existing video files, the selected video file will be fed into the video parsing module directly to produce the parsing results. Then the parsing results and the video file will be displayed in real-time in the result displaying module.

2 TECHNOLOGY

The training and testing phase of SVP network [3] proposed in RSVP system is as follows: During training, only a single frame per video is labeled , while others are unlabeled. The SVP network is learned from the extremely sparsely labeled videos. During testing, a parsing window is slided along the video. The parsing result of testing frame I_t is determined by three frames, that is, the testing frame I_t itself, the preceding long-range frame I_{t-l} and the preceding short-range frame I_{t-s} .

The SVP network includes three sub-networks, that is: the frame parsing sub-network, the optical flow sub-network and the temporal fusion sub-network. There are four stages to build up the SVP network. First, we train the optical flow sub-network to generate the optical flow. Second, we train the frame parsing sub-network and the temporal fusion sub-network together using the optical flow estimated in the first step. Third, we fix the Conv1~Conv5 layers of optical flow sub-network by those of frame parsing sub-network and only fine-tune the layers unique to optical flow. Now the two sub-networks share convolutional layers. Finally, keeping Conv1~Conv5 layers fixed, we fine-tune the unique layers of frame parsing and temporal fusion sub-networks [3]. Details follow.

On-line optical flow generation: To acquire the pixel-wise correspondences between frames in real-time. Most state-of-the-art optical flow estimation methods, such as EpicFlow [4] etc, suffer from relatively slow speed. Because the video parsing task requires extensive online optical flow computation, a real-time and accurate optical flow estimation component called optical flow sub-network is proposed. The optical flow sub-network is based on [2], and is trained with flying chairs dataset. In the following training process and the final system opration, this sub-network is integrated to SVP to generate optical flow on-line.

Temporal fusion: In order to combine the temporal context of frames, a temporal fusion sub-network is trained to fuse three parsing results to output the final refined parsing result of the current testing frame. During fusing, the optical flow generated by optical flow sub-network will be used to warp parsing results of preceding frames to the parsing result of the current testing frame. After warping, parsing results of different related frames will be fused via a temporal fusion layer with several 1×1 filters to produce the final parsing result of the current testing frame.

Figure 2: The demonstration of parsing video.

Variable sampling interval: The sampling interval of the video parsing sub-network is adjestable. To achieve best performance for different needs, the proper sampling intervals can be various.

Parameters sharing: All the three sub-networks share the parameters of Conv1~Conv5 layers, which output features for operation of following sub-networks. This design reduces the amounts of computation and makes the real-time parsing possiable. It also improves the robustness of extracted features, which makes the segmentation results more accurate and stable.

3 USER INTERFACE

While working, there are two input options in the user interface for users to select: existing video files and videoes captured by camera. While selecting parsing an existing video file or a video captured by camera in real-time, as shown in Figure 2, the video will be displayed in the left panel. And the output of video parsing module will be demonstrated in the right panel. When users operate RSVP system in real world, the low latency and high accuracy will be experienced.

The options include using different models to extract features and whether using the temporal context or not. By turning these options on, users could experience the significant improvement on video parsing brought by our real-time RSVP system.

4 CONCLUSION

In this demo, we showcase a system to parse surveillance videos in real-time which utilizes insufficient labeled data in model training phase. Based on our SVP network proposed in RSVP system, we obtain an accurate, computationally efficient and stable parsor to parse surveillance videos. In future, we plan to apply this sytem to parse other kind of videos, such as urban scene videos [1]. There is a supplementary file in PowerPoint format attached in this demo.

5 ACKNOWLEDGMENT

This work was supported by the Open Project Program of the Jiangsu Key Laboratory of Big Data Analysis Technology and National Natural Science Foundation of China (No.61572493, Grant U1536203).

REFERENCES

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. *CVPR* (2016).

[2] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. 2015. FlowNet: Learning Optical Flow with Convolutional Networks. *CoRR* (2015).

[3] Si Liu, Changhu Wang, Ruihe Qian, Han Yu, Renda Bao, and Yao Sun. 2017. Surveillance Video Parsing with Single Frame Supervision. *CVPR* (2017).

[4] Jérôme Revaud, Philippe Weinzaepfel, Zaïd Harchaoui, and Cordelia Schmid. 2015. EpicFlow: Edge-preserving interpolation of correspondences for optical flow. *CVPR* (2015).