Surveillance Video Parsing with Single Frame Supervision

Si Liu[1], Changhu Wang[2], Ruihe Qian[1], Han Yu[1], Renda Bao[1], Yao Sun[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences,Beijing 100093,China

[2] Toutiao AI Lab

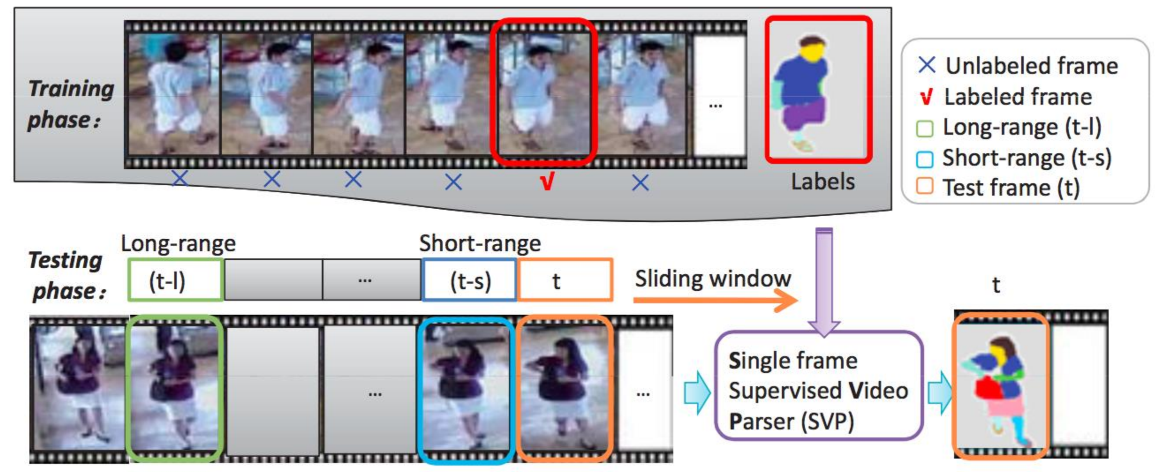
IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR July 21-26 HONOLULU 2017

Introduction :

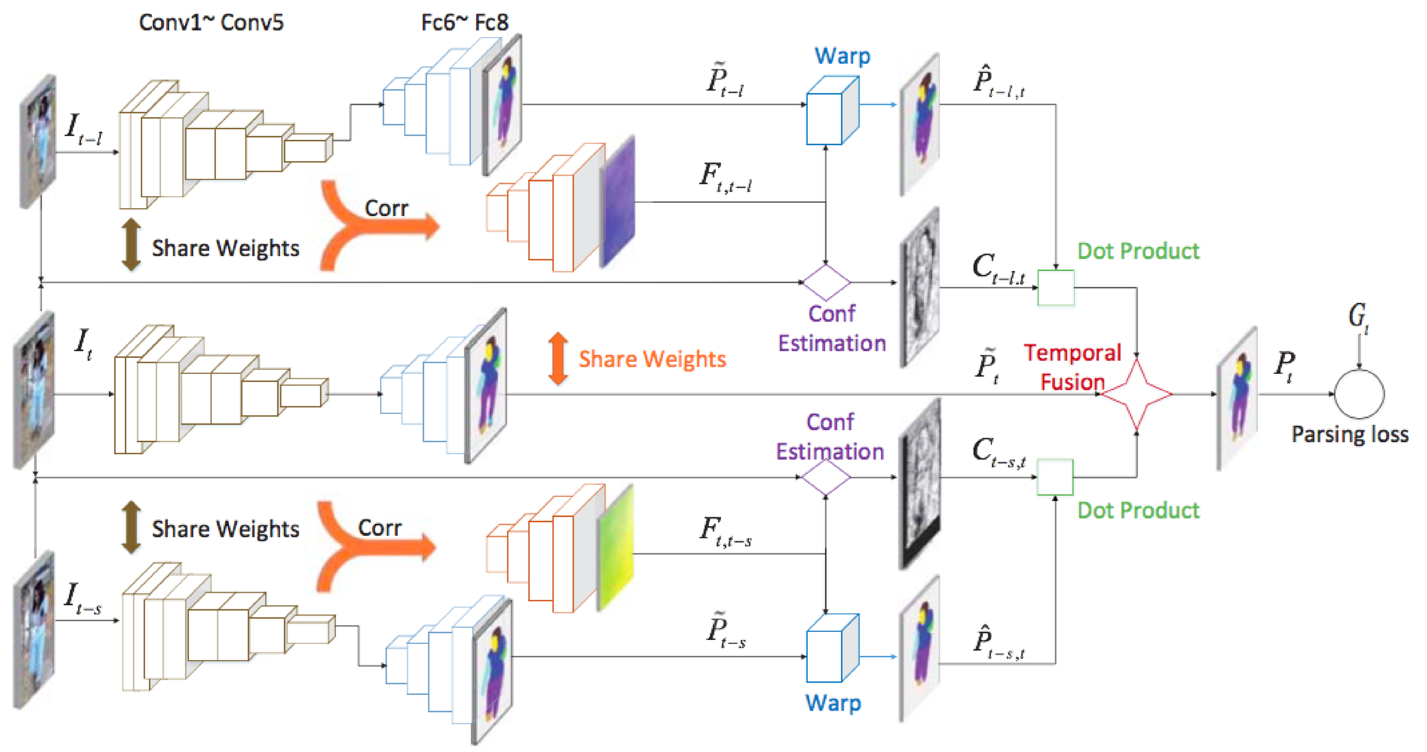
- We develop a Single frame Video Parsing (SVP) method which requires only one labeled frame per video in training stage to parse one Surveillance video.

- Function
  - Segment the video frames into several labels, e.g., face, pants, left-leg.
- Train & Test
  - During training, only a single frame per video is labeled(red box)
  - During testing, a parsing window is slided along the video. The parsing result(orange box) is determined by itself, the long-range frame(green box) and the short-range frame(blue box).



Figture1. The process diagram of Training and testing

- Network



Figture2. The proposed single frame supervised video paring (SVP) network. The network is trained end-to-end.

Approach :

- Contribution
  - Single Frame Supervision : the first attempt to segment the human parts in the surveillance video by labeling single frame per training video.
  - Good performance: the feature learning, pixelwise classification, correspondence mining and the temporal fusion are updated in a unified optimization process and collaboratively contribute to the parsing results.
  - Applicable : the proposed SVP framework is end-to-end and thus very applicable for real usage.

- Frame Parsing Sub-network
  - Video V = $\{I_1, \ldots, I_N\}$. The single labeled frame is I_N. The frame parsing sub-network produces the rough label maps for the triplet, donated as $\{\tilde{P}_{t-l}, \tilde{P}_{t-s}, \tilde{P}_t\}$.

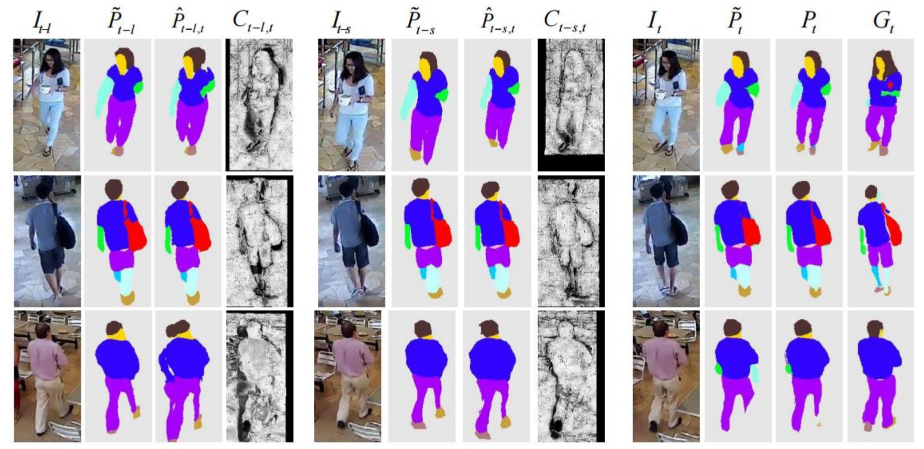
Figture3. 1,2 columns: the long-range frame, its the parsing result. 3,4 columns: the short-range frame, its the parsing result. 1,2 columns: the test frame, its the parsing result.

- Optical Flow Estimation Sub-network
  - Estimate the dense cor-respondence between adjacent frames on the fly.

$$F_{t,t-l} = o(I_t, I_{t-l}),$$

where o (a, b) is the operation of predicting the optical flow from a to b. $F_{t,t-s}$ is estimated similarly.

- Temporal Fusion Sub-network
  - Apply the obtained optical flow $F_{t,t-l}$ and $F_{t,t-s}$ to $\tilde{P}_{t-l}$ and $\tilde{P}_{t-s}$, producing $\hat{P}_{t-l}$ and $\hat{P}_{t-s}$. To alleviate the influence of imperfect optical flow, the pixel-wise flow confidence $C_{t-l,t}$ and $C_{t-s,t}$ are estimated.



Figture4. Step by step illustration of SVP. 1~4 columns: the long-range frame, its the parsing result, the warped parsing result and the confidence map. 5~8 columns: the short-range frame, its parsing result, the warped parsing result and the confidence map. 9~12 columns: test image, the rough parsing result, refined parsing result and ground truth parsing result.

Experiment:

Figture5. The test image, the groundtruth label, results of the EM-Adapt and SVP are shown sequentially

Table 1. Per-Class Comparison of F-1 scores with state-of-the-arts and several architectural variants of our model in Indoor dataset. (%).

Methods	bk	face	hair	U-clothes	L-arm	R-arm	pants	L-leg	R-leg	Dress	L-shoe	R-shoe	bag
PaperDoll [37]	92.62	57.16	58.22	62.52	19.96	14.99	52.47	25.43	20.7	9.92	20.66	24.41	14.32
ATR [15]	93.62	59.08	60.79	81.36	32.54	28.65	75.40	29.19	29.60	70.22	11.68	17.75	48.97
M-CNN [18]	93.40	53.94	59.12	75.53	24.46	20.51	78.46	36.15	21.92	43.61	14.53	18.79	53.43
Co-CNN [16]	94.06	64.64	73.53	81.54	26.82	31.66	77.13	25.47	34.11	76.08	15.42	20.57	46.91
FCN-8s [22]	94.80	71.35	74.90	79.53	33.55	32.29	81.89	36.57	33.98	43.53	33.03	31.50	43.66
DeepLab [4]	93.64	63.01	69.61	81.54	40.97	40.31	81.12	34.25	33.24	64.60	28.39	26.40	56.50
EM-Adapt [26]	93.46	66.54	70.54	77.72	42.95	42.20	82.19	39.42	37.19	63.22	33.18	31.68	53.00
SVP l	94.68	67.28	72.74	82.12	42.96	43.35	81.91	39.26	38.31	67.17	31.47	30.38	58.99
SVP s	94.65	66.27	73.48	83.12	45.17	44.89	82.72	38.62	38.43	66.04	30.93	31.46	58.81
SVP l+c	94.44	67.29	73.76	83.06	43.56	43.56	82.33	41.36	39.46	68.36	31.75	31.73	59.04
SVP s+c	94.64	67.62	74.13	83.48	45.13	**45.08**	83.21	39.89	40.11	68.17	31.15	32.27	58.75
SVP l+s	94.50	67.08	73.52	83.10	**45.51**	44.26	82.59	41.82	**42.31**	69.43	33.71	33.36	58.58
SVP l+s+c	**94.89**	**70.28**	**76.75**	**84.18**	44.79	43.29	**83.59**	**42.69**	40.30	**70.76**	**34.77**	**35.81**	**60.43**

Table 2. Per-Class Comparison of F-1 scores with state-of-the-arts and several architectural variants of our model in Outdoor dataset. (%).

Methods	bk	face	hair	U-clothes	L-arm	R-arm	pants	L-leg	R-leg	L-shoe	R-shoe	bag
FCN-8s [22]	92.00	62.64	65.58	78.64	28.73	28.97	79.69	38.88	9.08	32.04	30.56	29.45
DeepLab [4]	92.19	58.65	66.72	84.31	42.23	35.36	81.12	30.64	6.13	37.89	33.25	52.25
EM-Adapt [26]	92.68	60.84	67.17	84.78	41.28	33.61	**81.80**	**42.39**	7.28	39.54	32.20	**54.31**
SVP l	91.13	62.40	67.73	84.64	45.18	31.40	80.66	30.28	5.86	**40.32**	33.11	54.96
SVP s	92.51	64.25	67.14	84.99	45.28	32.14	79.71	32.31	18.49	37.24	31.45	51.58
SVP l+c	92.60	63.76	68.77	84.84	**45.83**	33.75	81.67	31.37	**19.06**	38.54	33.51	53.57
SVP s+c	**92.94**	64.40	69.93	**85.43**	44.44	31.86	81.65	35.88	18.22	37.48	33.36	54.23
SVP l+s	91.90	63.32	69.48	84.84	42.09	28.64	80.45	31.10	13.28	38.52	**35.52**	46.89
SVP l+s+c	92.27	**64.49**	**70.08**	85.38	39.94	**35.82**	80.83	30.39	13.14	37.95	34.54	50.38

project page:
http://liusi-group.com/projects/SVP

home page:
http://liusi-group.com