Spatio-Temporal Self-Organizing Map Deep Network for Dynamic Object Detection from Videos

Yang Du, Chunfeng Yuan, Bing Li, Weiming Hu and Stephen Maybank
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, Beijing, China
duyang2014@ia.ac.cn, cfyuan@nlpr.ia.ac.cn, bli@nlpr.ia.ac.cn, wmhu@nlpr.ia.ac.cn, sjmaybank@dcs.bbk.ac.uk

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 2017 HONOLULU

Introduction:

➢ Modern detection algorithms are generally attained by background modeling. The difficulty of background modeling is to tackle background motions.

Motivations:

➢ We conclude that the motions of complex background mainly have two properties:

- Variation of the global background in the space. It is mainly caused by the zoom, translation, jitter, etc, of the camera. We refer to it as the spatial property of background motion.
- Variation of the local background in the time. It mainly indicates the dynamic elements in background and at different frames, such as river, fountain and bad weather. We refer to it as the temporal property of background motion.

Contributions:

➢ We propose a new STSOM and train it in two aspects, using the whole frames from spatial perspective and using the sequence of a pixel over time from temporal perspective. Then we propose a new method based on Bayesian parameter estimation to automatically learn the spatio-temporal threshold of background filtering. In order to further accurately model the complex background, we stack multiple STSOMs to form a deep network with a STSOM as a layer. The different parts of complex background are accurately modeled by different layers. The dynamic objects are detected by filtering out the background layer by layer and the segments are increasingly accurate with the deeper layer.

Self-Organizing Map:

➢ F(t): the t-th input

➢ w(t): weight vector

➢ c: the index of the winner node, which is the one with the weight vector w(t) that has the smallest Euclidean distance from F(t)

$$c = \underset{q}{argmin}\{||F(t) - w_q(t)||\}$$

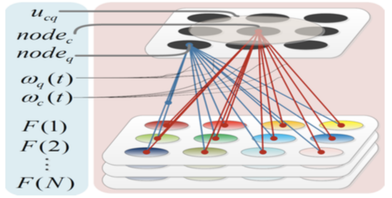


Figure 1. Structure of a self-organizing map.

➢ neighborhood function:

$$u_{cq} = \exp\left(-\frac{||q-c||}{2\sigma^2}\right)$$

➢ Learning rule:

$$w_q(t+1) = w_q(t) + u_{cq} * \alpha(t) * [F(t) - w_q(t)]$$

STSOM Deep Network for Dynamic Object Detection :

➢ A new frame of a video enters into this network and then its dynamic objects will be extracted layer by layer till the last output layer.

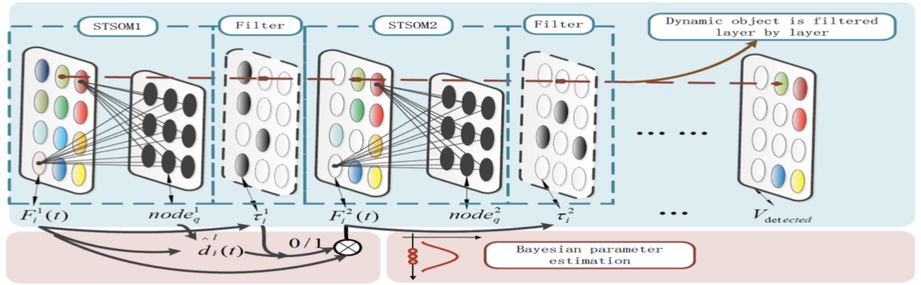


Figure 2. The architecture of the STSOM deep network.

➢ Spatial Weight Updating

$$D^l_{spatial.q}(t) = \sum_{i=1}^{P} d^l_{iq}(t)$$

$$q^*_{spatial}(t) = \underset{q}{argmin}\{D^l_{spatial.q}(t), q = 1, 2, ..., Q\}$$

$$w^l_{iq}(t+1) = u_q * (w^l_{iq}(t) + \alpha_{train} * (F^l_i(t) - w^l_{iq}(t))$$

➢ Temporal Weight Updating

$$q^*_{i.temporal}(t) = \underset{}{argmin}\{d^l_{iq}(t), q = 1, 2, ..., Q\}$$

➢ Forward Propagation

$$\hat{\mu}_i = \frac{N/\sigma_i^2}{N/\sigma_i^2 + 1/\hat{\delta}_i^2} F^l_{ave.i} + \frac{1/\hat{\delta}_i^2}{N/\sigma_i^2 + 1/\hat{\delta}_i^2} \hat{\gamma}_i$$

$$L(\hat{\mu}_i, \sigma_i^2) = \sum_{t=1}^{N} \ln \mathbf{N}(F^l_i(t)|\hat{\mu}_i, \sigma_i^2).$$

$$\hat{\mu}_i|_{\frac{\partial L}{\partial \hat{\mu}_i}=0} = \underset{\hat{\mu}_i}{argmax}\{L(\hat{\mu}_i, \sigma_i^2)\}$$

$$\tau^l_{spatial.i} = max\{D^l_{spatial.q}, q = 1, 2, ..., Q\}/P$$

$$\tau^l_{temporal.i} = max\{d^l_{iq}, q = 1, 2, ..., Q\}$$

$$d^l_{iq}(t) = ||(v^l_i(t)s^l_i(t)\cos(h^l_i(t)), v^l_i(t)s^l_i(t)\sin(h^l_i(t)), v^l_i(t)) - (w^l_{v_{iq}}(t)w^l_{s_{iq}}(t)\cos(w^l_{h_{iq}}(t)), w^l_{v_{iq}}(t)w^l_{s_{iq}}(t)\sin(w^l_{h_{iq}}(t)), w^l_{v_{iq}}(t))||_2^2.$$

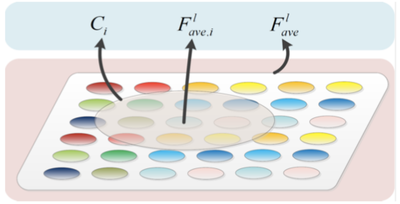


Figure 3. Estimation of γ_i and δ_i^2 with MLE.

$$\hat{\gamma}_i = 1/|C_i| \sum_{i \in C_i} F^l_{ave.i},$$

$$\hat{\delta}_i^2 = 1/|C_i| \sum_{i \in C_i} (F^l_{ave.i} - \hat{\gamma}_i)(F^l_{ave.i} - \hat{\gamma}_i)^T$$

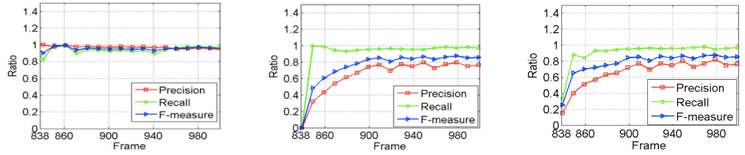
$$\tau^l_i = (\tau^l_{temporal.i} + \tau^l_{spatial.i})/2$$

$$\hat{d}^l_i(t) = \min\{\hat{d}^l_{i1}(t), \hat{d}^l_{i2}(t), ..., \hat{d}^l_{iQ}(t)\}$$

Evaluation on the CDnet 2014 Dataset :

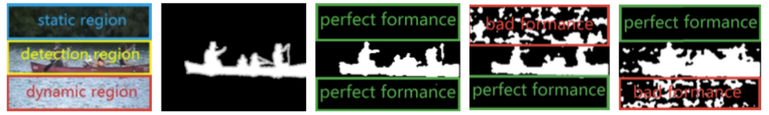
Method	$FM_{overall}$	FM_{BL}	FM_{CJ}	FM_{DB}	FM_{IOM}	FM_{SH}	FM_{TH}	FM_{BW}	FM_{LF}	FM_{NV}	FM_{PTZ}	FM_{TU}
STSOM	**0.816**	**0.957**	**0.888**	**0.923**	**0.835**	**0.910**	**0.848**	**0.892**	**0.812**	*0.563*	*0.575*	**0.800**
IUTIS-5	*0.771*	*0.956*	*0.833*	0.890	0.729	*0.908*	0.830	0.824	*0.774*	0.529	0.428	*0.783*
SharedModel	0.747	0.952	0.814	0.822	0.672	0.845	0.831	0.798	0.728	0.541	0.386	0.733
SuBSENSE	0.741	0.950	0.815	0.817	0.656	0.864	0.817	*0.861*	0.644	0.559	0.347	0.779
PAWCS	0.740	0.939	0.813	*0.893*	0.776	0.871	0.832	0.815	0.658	0.415	0.461	0.645
C-EFIC	0.730	0.930	0.824	0.562	0.622	0.845	*0.834*	0.786	0.680	**0.667**	**0.620**	0.627
MBS	0.728	0.928	0.836	0.791	0.756	0.826	0.819	0.798	0.635	0.515	0.552	0.585
FTSG	0.728	0.933	0.751	0.879	*0.789*	0.853	0.776	0.822	0.625	0.513	0.324	0.712
S-Subsense	0.717	0.948	0.807	0.815	0.601	0.865	0.685	0.859	0.651	0.534	0.339	0.751
SMSOM	-	0.944	0.732	0.675	-	-	0.793	-	-	-	-	-
SOBS	0.596	0.933	0.705	0.643	0.562	0.721	0.683	0.662	0.546	0.450	0.040	0.488
KDE	0.568	0.909	0.572	0.596	0.408	0.766	0.742	0.757	0.547	0.436	0.036	0.447
GMM	0.556	0.838	0.596	0.633	0.520	0.715	0.662	0.738	0.537	0.409	0.152	0.466

Evaluation of Pre-training and Fine-tuning :



(a) pre-training & fine-tuning (b) $\tau_i^l = 0.01$, without pre-training (e) $\tau_i^l = 0.01$, with pre-training

Evaluation of SSOM, TSOM and STSOM :



(a) original image (b) ground truth (c) with pre-training (e) TSOM result (f) SSOM result

Evaluation of Deep Network :

(g) result of 1st layer (h) result of 2nd layer (i) result of 4th layer (j) result of 5th layer (k) result of 6th layer (l) result of 7th layer