Inferring Ongoing Human Activities Based on Recurrent Self-Organizing Map Trajectory

Qianru Sun
qianrusun@sz.pku.edu.cn

Hong Liu
hongliu@pku.edu.cn

Engineering Lab on Intelligent Perception for Internet of Things(ELIP), Shenzhen Graduate School
Peking University, CHINA

Key Laboratory of Machine Perception(Ministry of Education)
Peking University, CHINA

Automatically inferring ongoing activities is to enable the early recognition of unfinished activities, which is quite meaningful for applications, such as online human-machine interaction and security monitoring. State-of-the-art methods use the spatio-temporal interest point (STIP) based features as the low-level video description to handle complex scenes [1, 2, 3]. While the existing problem is that typical bag-of-visual words (BoVW) focuses on feature distribution but ignores the inherent contexts in sequences, resulting in low discrimination when directly dealing with limited observations. To solve this problem, the Recurrent Self-Organizing Map (RSOM) [4], which was designed to process sequential data, is novelly adopted in this paper for the high-level representation of ongoing activities. The innovation lies that observed features and their spatio-temporal contexts are encoded in a trajectory of the pre-trained RSOM units. Additionally, a combination of Dynamic Time Warping (DTW) distance and Edit distance, named DTW-E, is specially proposed to measure the structural dissimilarity between RSOM trajectories.

RSOM Trajectory: Since the RSOM constitutes a direct extension of SOM, we start from SOM. SOM is to map the data from an input space $\boldsymbol{V}_I$ onto a lower dimensional space $\boldsymbol{V}_L$ (a map) in such way that the topological relationships in $\boldsymbol{V}_I$ are preserved and the SOM units approximate closely the probability density function of $\boldsymbol{V}_I$. Suppose each unit i in SOM is associated with a weight vector $\boldsymbol{w}_i = [w_{i1}, w_{i2}, ..., w_{in}]^T \in \Re^n$ with the same dimension as the input vector $\boldsymbol{x} = [x_1, x_2, ..., x_n]^T \in \Re^n$. Learning process that leads to self-organization on a map can be summarized as,

(i) The feature vector $\boldsymbol{x}(t)$ is input, then its best matching unit (*bmu*) on the map is found by computing the minimum distance as:

$$bmu = \arg\min_{i \in \boldsymbol{V}_L} \{\|\boldsymbol{x}(t) - \boldsymbol{w}_i(t)\|\} \tag{1}$$

(ii) The winner *bmu* and its neighbors on the map have their weights $\boldsymbol{w}_i(t)$ updated towards $\boldsymbol{x}(t)$ as:

$$\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) + \alpha(t) \cdot N_{bmu,i} \cdot \|\boldsymbol{x}(t) - \boldsymbol{w}_i(t)\| \tag{2}$$

where $\|\cdot\|$ denotes the Euclidean norm, $\alpha(t) = \alpha_i \cdot (\alpha_f / \alpha_i)^{T(i)/T_{max}} \in [0,1]$ is the learning rate, where the α_i and α_f denote the initial rate and final rate. $T(i) = \{1, 2, ..., T_{max}\}$ where T_{max} is iteration number. $N_{bmu,i}$ is called neighborhood function and defined over the units on the map. Typically, $N_{bmu,i} = \exp\{-\|r_{bmu} - r_i\|^2 / 2\sigma^2\}$, where $r_{bmu} \in \Re^2$ and $r_i \in \Re^2$ are the location vectors of unit *bmu* and i on the map, and σ defines the Gaussian kernel width.

SOM is not originally designed to accommodate the time series, its temporal extension RSOM is hence adopted here to learn the temporal contexts in activity sequences. It is to utilize both the feature vectors before $\boldsymbol{x}(t)$ and $\boldsymbol{x}(t)$ itself to search the best matching unit of $\boldsymbol{x}(t)$. This is done by associating the following recursive equation to each unit i to compute the difference vector $\boldsymbol{y}_i(t)$:

$$\boldsymbol{y}_i(t) = \lambda \cdot \|\boldsymbol{x}(t) - \boldsymbol{w}_i(t)\| + (1-\lambda) \cdot \boldsymbol{y}_i(t-1) \tag{3}$$

where $0 < \lambda < 1$ is a factor determining the influence of earlier difference vectors on the current $\boldsymbol{x}(t)$. When λ is close to 0, the system of Eq. (3) involves a heavy backward memory, whereas when λ is near to 1, Eq. (3) describes a slight memory. Now equations of RSOM for searching bmus and adapting weights are as follows,

$$bmu = \arg\min_{i \in \boldsymbol{V}_L} \{\|\boldsymbol{y}_i(t)\|\} \tag{4}$$

$$\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) + \alpha(t) \cdot N_{bmu,i} \cdot \|\boldsymbol{y}_i(t)\| \tag{5}$$

Then, we introduce how to map the input feature sequence to a time-varying trajectory of bmus. Supposing that an $M \times M$ map is learned after a fixed number of iterations using Eq. (3)(4)(5). For simplification, the one-dimensional value b of the original coordinate $bmu \in \Re^2$, i.e., $b = bmu(2) \times M + bmu(1) \in [1, M^2]$, is used as the location index in the final trajectory. During video input at time t, STIP based local features are first extracted on the current frame based on above equations. Then feature vectors search their bmus on the map, and compose an inner-frame trajectory $\boldsymbol{b}$. Finally, $\boldsymbol{b}$ is used to generate the inter-frame trajectory $\boldsymbol{Trj}$.

$$\boldsymbol{b}_f = \{b_k | k = 1, 2, ..., K(f)\}; \boldsymbol{Trj}(t) = \{\boldsymbol{b}_f | f = 1, 2, .., F(t)\} \tag{6}$$

where $K(f)$ is the number of local features on the fth frame and it changes with different frames. $F(t)$ is the frame number until time t. $\boldsymbol{Trj}(t)$ is thus used as a high-level representation of the current observed activity.

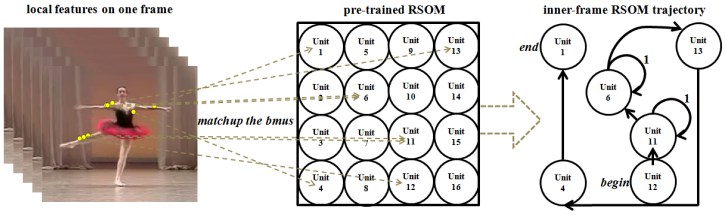

Figure 1: Illuminations of a pre-trained 4×4 RSOM and the assumed trajectory (12, 11, 11, 6, 6, 13, 4, 1) of local features (yellow points). Note that the bmus order in the trajectory is based on their image locations (detected order): from left to right, then from up to down.

Figure 2: Samples of UT-Interaction (a, b) and Rochester Activities dataset (c).

DTW-E distance: The structure of RSOM trajectory is clear and special that each subset on one frame (inner-frame) contains the human shape information and the whole sequence (inter-frame) contains the long-range temporal relationships. Therefore, how to reasonably measure the likelihood between RSOM trajectories for pattern classification arises another problem. To solve this, a hierarchical distance based on the combination of DTW distance and Edit distance, named DTW-E, is specially defined (implementation of this algorithm is described in our paper).

Experiments and Conclusions: Two real-world datasets (Figure 2) with different characteristics, complex scenes [5] and inter-class ambiguities [6], serve as sources of data for evaluation. Experimental results based on kNN classifiers confirm that our approach can infer ongoing human activities at any stage with high accuracies.

Our conclusion is that our approach can make efficient inferences even with complex scenes and inter-class ambiguities. It hence confirms the ability of RSOM trajectory to extract sufficient discrimination. Moreover, the RSOM trajectory with the advantage of before-after independence is proved more suitable to the recognition of unfinished patterns.

[1] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie. Behavior recognition via sparse spatio-temporal features. *VS-PETS*, pages 65-72, 2005.

[2] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proc. BMVC*, pages 124.1-124.11, 2009.

[3] P. Scovanner, S. Ali, M. Shah. A 3-Dimensional SIFT Descriptor and its Application to Action Recognition. *ACM Conf. Multimedia*, pages 357-360, 2007.

[4] M. Varsta, José del R. Millán, J. Heikkonen. A Recurrent Self-Orgnizing Map for Temporal Sequence Processing. *LNCS* 1327, pages 421-426, 1997.

[5] M. S. Ryoo, J. K. Aggarwal. UT-Interaction Dataset. *ICPR Contest on Semantic Description of Human Activities (SDHA)*, 2010.

[6] R. Messing, C. Pal, H. Kautz, Activity recognition using the velocity histories of tracked keypoints, In *Proc. ICCV*, pages 104-111, 2009.