9-2013

Inferring ongoing human activities based on recurrent self-organizing map trajectory

Qianru SUN
Singapore Management University, qianrusun@smu.edu.sg

Hong LIU

Citation

SUN, Qianru and LIU, Hong. Inferring ongoing human activities based on recurrent self-organizing map trajectory. (2013). *Proceedings of the 24th British Machine Vision Conference (BMVC 2013), Bristol, September 9-13*. 1-11.
Available at: https://ink.library.smu.edu.sg/sis_research/4466

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

Qianru Sun
qianrusun@sz.pku.edu.cn

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

Hong Liu
hongliu@pku.edu.cn

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

Abstract

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. While the existing problem is that typical bag-of-visual words (BoVW) focuses on the statistical distribution of features but ignores the inherent contexts in activity sequences, resulting in low discrimination when directly dealing with limited observations. To solve this problem, the Recurrent Self-Organizing Map (RSOM), which was designed to process sequential data, is novelly adopted in this paper for the high-level representation of ongoing human activities. The innovation lies that the currently 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. Two real-world datasets with markedly different characteristics, complex scenes and inter-class ambiguities, serve as sources of data for evaluation. Experimental results based on kNN classifiers confirm that our approach can infer ongoing human activities with high accuracies.

1 Introduction

Early recognition of human activity arises in a large amount of applications from human-machine interaction to video security. Taking a monitoring system as an example, if it can infer the ongoing destroying behavior and raise a timely alarm before it is done, it will be more meaningful than just identifying objects destroyed. In the field of human activity analysis, most previous methods are limited to the recognition of simple and single human actions [1, 2]. Additionally, most researchers used fully observed videos for training, making models unsuitable to infer unfinished activities [3, 4, 5, 6]. Therefore, the development of new methods to recognize ongoing activities from limited observation is rather necessary.

In recent years, though researchers have proposed some methods for inferring ongoing activities, there exist some limitations in different aspects. Lv and Nevatia modeled each human action as a hidden state sequence of body silhouettes and used HMM to infer intermediate status [7], which depended on the accurate detection of human body, thus suffered from the difficulties of multi-object and complex background in real-world videos. Ravichandran *et al.* realized a dynamic inference system using the histogram of single-frame optical flow for description [8]. It performed well to deal with simple actions, but it was incompetent to handle more complex activities (e.g. human interactions [9]) due to the disturbed optical flow. Ryoo used the local video features and proposed two variants of bag-of-visual words (BoVW) to predict human activity at high computational speed [10]. However, it obtained modest accuracies, which proved that BoVW could not capture enough discrimination from the limited data of unfinished activities. Most recently, Hoai and Torre addressed the early event detection by an active training process, for which Structural Output SVM was extended to anticipate the sequential data [11]. It emphasized the activity recognition not the representation. In summary, the specialized method for representing ongoing activities is relatively unexplored and remains to be an important bottleneck for early recognition. In this paper, the low-accuracy of unreasonable representations is what we attempt to improve, hence the goal is to extract sufficient information for the rich description of ongoing activities.

To reduce the effect of complex scenes, this paper utilizes spatio-temporal interest point (STIP) based features as the low-level video description [3, 5, 6]. Such features are directly extracted from videos, avoiding possible failures of body segmentation and target tracking in [7]. However, they are described in a high-dimension space and are time-consuming, which was evaluated in [12]. Methods of data clustering and dimensionality reduction are hence desirable. Referring to [10], the bag-of-visual words (BoVW) is the most common method to decrease the dimensionality of feature space for its simplicity and robustness. However, it ignores both the geometric information in body gestures and the time-varying information in activity sequences, resulting in the modest accuracies when dealing with limited observations [10]. Though previous researchers proposed to use the feature correlograms to add some contextual information to BoVW [13, 14], they focused on the statistic data using completed videos, thus are impractical for representing the unfinished activity in our case.

To encode the spatio-temporal contexts during video input, we introduce another algorithm for dimensionality reduction, called Recurrent Self-Organizing Map (RSOM) [15], which is a temporal variant of Self-Organizing Map (SOM) [16]. The SOM is quite popular in data mining applications, and its superiority lies that it can preserve the topology of data space, and well approximate the probability density distribution. In SOM, the chain of the best matching units (bmus) produces a trajectory on the map for any input sequence. This trajectory is conceptually different from the moving trajectory in [19], and was used for the visualization of speech signals [17] and process control [18]. Since the RSOM constitutes a direct extension of SOM, it successfully generalizes above properties, additionally it is more suitable to manage time-varying sequences. Another superiority of the RSOM trajectory for early pattern recognition is that the trajectory of new observation makes no affect on the previously generated trajectory, which can not be held in the overall-updated histogram of BoVW. For a trajectory of the long-range human activity, this characteristic means that the sub-actions are sequentially and independently encoded in it, which means a lot since in practice we have to use completed videos to train templates while use unfinished samples for tests. Considering these advantages, we employ the RSOM trajectories in conjunction with local features to keep track of human motion occurrences in section 2. To the best of our knowledge, we are the first to try RSOM for representing ongoing human activities.

After the RSOM is trained, an newly input feature sequence can be mapped to a structural trajectory that the inner-frame trajectory encodes the gesture information on one frame while the inter-frame trajectory contains the long-range temporal variation in the sequence. Therefore, the distance measurement should be carefully selected to reflect the underlying dissimilarity of these special patterns. Schemes such as HMM distance, Longest Common Sub-Sequence (LCSS), PCA+Euclidean, and Dynamic Time Warping (DTW) distance have been compared for normal trajectory measurement in [19]. The HMM distance needs to train a statistical model for each trajectory. Results in [19] show it very likely suffers from the over-fitting due to limited training data. LCSS concerns more about the similarity of data shapes, and also requires exhausting parameter settings [20]. PCA+Euclidean distance [21] is based on PCA decomposition, hence it needs equal-length sequences, which is impractical to the ever-changing trajectory. The DTW distance is an alignment based measurement, and has been successfully used in the clustering of sign language [22]. Considering the special structure of our RSOM trajectory, we utilize the DTW distance to measure the global cost during frame-by-frame warping. Besides, the distance between inner-frame trajectories is computed by Edit distance, which is the minimum number of single-character edits (insertion, deletion, substitution) required to change one word into the other [23]. The combination of DTW distance and Edit distance, named DTW-E distance, is introduced in section 3.

The main contribution of this paper lies in two aspects: the RSOM based representation for the representation of ongoing human activities – RSOM trajectory, and the DTW-E distance for pattern measurement of RSOM trajectories. To deal with the complex real-world scenes, the STIP based features are detected as the basic motion description and act as the input primitives of RSOM. The main idea behind our approach is to properly describe and measure the changing pattern during observation increases.

2 Recurrent Self-Organizing Map Trajectory

Each human activity can be regarded as a sequence of motion regions changing over time. In this paper, spatio-temporal interest point (STIP) is utilized to detect local regions for its recent popularity in [4, 10, 24, 27]. The advantages of STIP based local features are: 1) they avoid the pre-processes of background subtraction, body detection and tracking, which are inherently tough problems by themselves; 2) benefiting from their localized and unstructured nature, they are robust to scale change and body occlusion; 3) their density is flexible, hence can be stored and manipulated efficiently when they are sparsely distributed. As mentioned, the RSOM is designed to encode the before-after sequentiality of time series. In this section, we first introduce how to learn the RSOM in conjunction with the sequence of feature vectors, then describe how to transform newly input local features to a RSOM trajectory.

2.1 Learning Recurrent Self-Organizing Map

Since the RSOM constitutes a direct extension of SOM, the learning process starts with the introduction of 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 $\boldsymbol{x} = [x_1, x_2, ..., x_n]^T \in \Re^n$. The learning process that leads to self-organization on a map can be summarized as following steps,

(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:

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

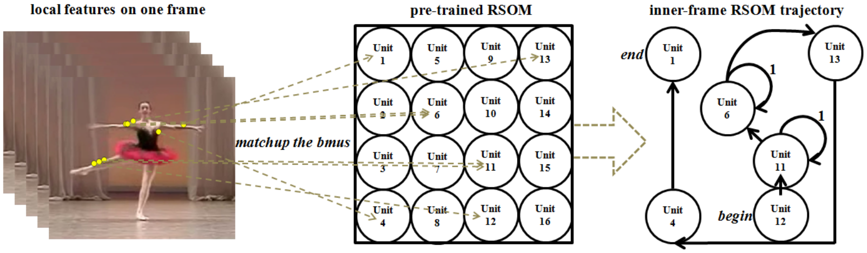


Figure 1: Illuminations of a pre-trained 4 × 4 RSOM and the assumed trajectory (12, 11, 11, 6, 6, 13, 4, 1) of local features (yellow points). Note that the bmus order in the trajectory is based on their image locations (detected order): from left to right, then from up to down.

(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.

Since SOM is not originally designed to accommodate the time series, its temporal extension RSOM is 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,

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

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

2.2 Dynamic generation of RSOM trajectory

In this section, 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 illustrates the key idea behind the proposed RSOM trajectory: local features are first detected as primitives on each frame then orderly mapped onto a pre-trained RSOM to generate a series of best matching units, i.e., inner-frame trajectory. Intuitively, the whole video trajectory, i.e., inter-frame trajectory, is the before-after concatenation of the inner-frame trajectories during frame-by-frame input. In this way, the sequential nature of the video is encoded, and the advantage of local feature to handle the noisy observation is also maintained in the generated trajectory.

3 DTW-E distance

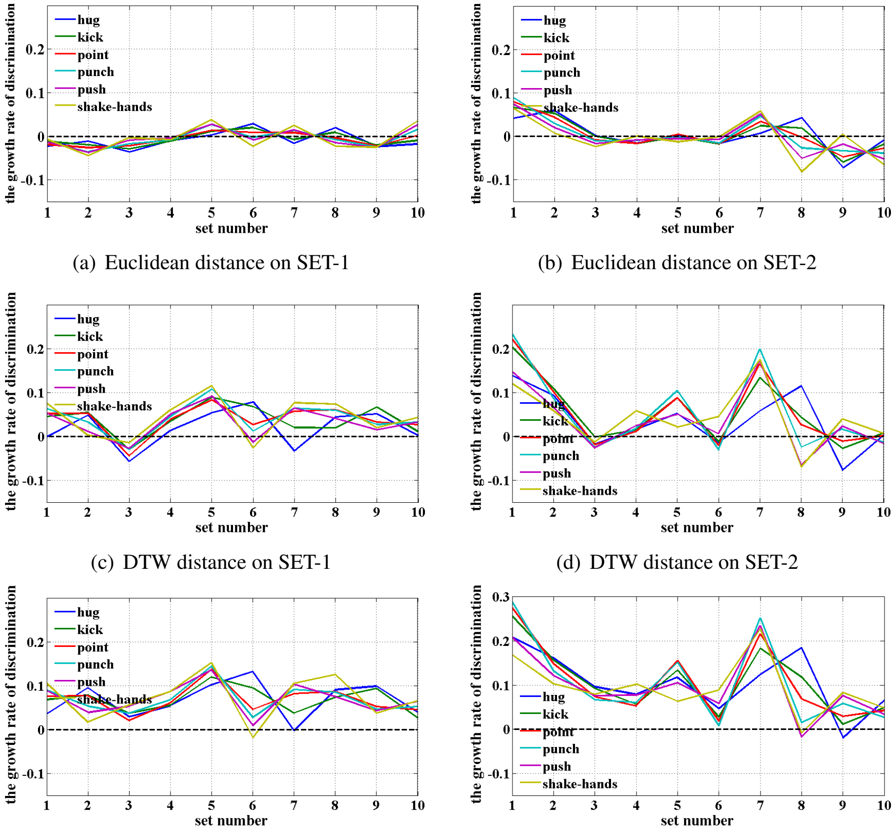
After the RSOM is learned, newly input activities can be represented as trajectories of best matching units (indexes). As discussed in section 1, the structure of RSOM trajectory is clear and special that each subset on one frame (inner-frame) contains the human shape information and the whole sequence (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.

Particularly, the basic idea behind DTW is to search the warping path between two time series that minimizes the warping cost. The warping cost thus can be used to measure the distance between the series. Since the time scale-variations are embodied in the inter-frame trajectory, the DTW distance is chosen as the distance to measure inter-frame trajectory. The Edit distance, which is more suitable to cope with the short and sparse series, is adopted to compute the distance between inner-frame trajectories. Given RSOM trajectories $\boldsymbol{Trj}_1$ and $\boldsymbol{Trj}_2$, the details of DTW-E distance $dist$ are given in Algorithm 1, where $length(\cdot)$ computes the length of a time series. Though the recursive computations of DTW and Edit distance are time-consuming to some extent, which could be made up by the development of high-speed computer or reasonable sparse sampling of motion features, the superiority of DTW-E distance is quite obvious in Figure 2, compared with the typical measurements.

Algorithm 1 Compute the DTW-E distance between two RSOM trajectories

```
INPUT: RSOM trajectory $\boldsymbol{Trj}_1 = \{\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n\}$, $\boldsymbol{Trj}_2 = \{\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_m\}$. $n, m$ are frame numbers. $\boldsymbol{a}_i, \boldsymbol{b}_i$ are the inner-frame trajectories.
OUTPUT: DTW-E distance $dist$ between $\boldsymbol{Trj}_1$ and $\boldsymbol{Trj}_2$
initialize $dist(i,1) \leftarrow infinity$ for all $i = 1$ to $n$
initialize $dist(1,j) \leftarrow infinity$ for all $j = 1$ to $m$
for $i = 2$ to $n$ do
    for $j = 2$ to $m$ do
        if $length(\boldsymbol{a}_i) = 0$ then $E(i,j) \leftarrow length(\boldsymbol{b}_j)$ end if
        if $length(\boldsymbol{b}_j) = 0$ then $E(i,j) \leftarrow length(\boldsymbol{a}_i)$ end if
        if $\boldsymbol{b}_j(length(\boldsymbol{b}_j) - 1) = \boldsymbol{a}_i(length(\boldsymbol{a}_i) - 1)$ then $cost_E \leftarrow 0$ else $cost_E \leftarrow 1$ end if
        $E(i,j) \leftarrow \min\{E(i-1,j)+1, E(i,j-1)+1, E(i-1,j-1)+cost_E\}$
        $dist(i,j) \leftarrow E(i,j) + \min\{dist(i-1,j), dist(i,j-1), dist(i-1,j-1)\}$
    end for
end for
return $dist$
```

The performance of trajectory measurement can be evaluated by the degree of enlarging inter-class distance and reducing inner-class distance. In this paper, to intuitionally show the superiority of our DTW-E distance, a criterion called “growth rate of discrimination” is defined. Assuming there are n sets, each set includes m human activities performed by one



(a) Euclidean distance on SET-1

(b) Euclidean distance on SET-2

(c) DTW distance on SET-1

(d) DTW distance on SET-2

(e) Our DTW-E distance on SET-1

(f) Our DTW-E distance on SET-2

Figure 2: Growth rates of discrimination on UT-Interaction [9]. Note that when computing the DTW distance (c)(d), Euclidean norm is used to compute the basic warping cost between a data pair.

actor, and $a_{i,j}$ is the ith activity in set j, the inter-class distance $inter_{i,j}$ of $a_{i,j}$ is defined as Eq. (7) and the inner-class distance $inter_{i,j}$ is defined as Eq. (8). Then, the growth rate of discrimination from inner-class to inter-class is computed as the $rate_{i,j}$ in Eq. (9), where $\|\cdot\|$ denotes the Euclidean norm.

The growth rates using three distances to measure the RSOM trajectories of UT-Interaction activities are shown in Figure 2. Normally, the rate is the bigger the better since it presents the ability to enlarge inter-class distance and/or reduce inner-class distance, which can be figured out in Eq. (9). Note that the positive rate means more reasonable than the negative one because inter-class activities statistically own bigger difference than that of inner-class activities. It can be concluded from Figure 2 that the DTW-E distance is more superior than typical Euclidean and DTW distance to measure RSOM trajectories.

$$inter_{i,j} = \frac{1}{n\cdot(m-1)} \sum_{l=1,l\neq i}^{m} \sum_{j=1}^{n} distance(a_{i,j}, a_{l,j}) \tag{7}$$

$$inner_{i,j} = \frac{1}{(n-1)\cdot m} \sum_{i=1}^{m} \sum_{k=1,k\neq j}^{n} distance(a_{i,j}, a_{i,k}) \tag{8}$$

$$rate_{i,j} = \frac{\|inter_{i,j} - inner_{i,j}\|}{\|inter_{i,j} + inner_{i,j}\|} \tag{9}$$

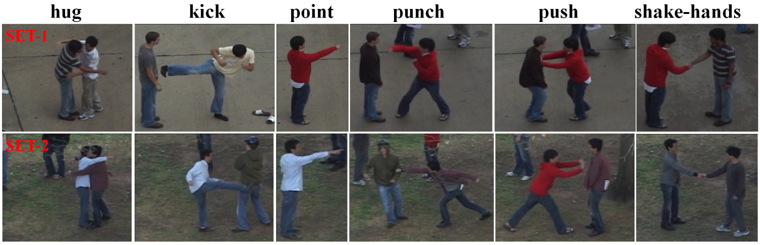


Figure 3: Snapshots of 6 interactions in two scenes of UT-Interaction dataset.

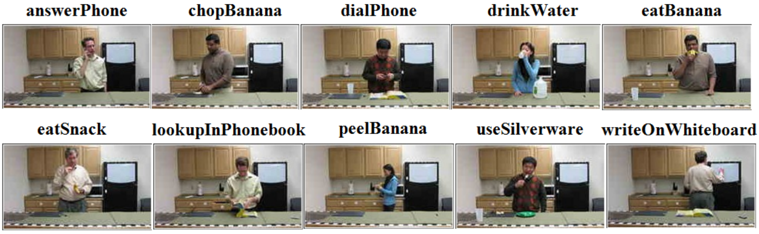


Figure 4: Snapshots of 10 daily activities on Rochester Activities dataset.

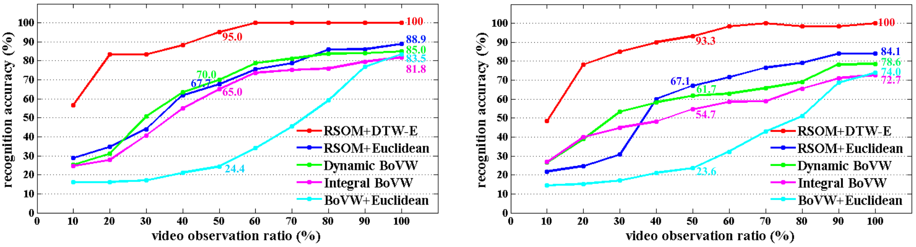
4 Experiments and Discussions

In this section, the proposed approach is evaluated and compared with related methods. The inference of ongoing activities is implemented on two recent activity datasets: UT-Interaction [9] and Rochester Activities dataset [25]. For testing, the inference process is conducted at ten observing ratios: 10%, 20%, ..., 100%.

Dataset: Experiments were implemented on the segmented version of UT-Interaction, which has been tested by the state-of-the-art works [10, 24]. It contains 6 classes: "hug", "kick", "point", "punch", "push" and "shake-hands". Except that "point" is a single action, each activity is respectively performed by 10 pairs of actors. Following previous works, videos are divided into two sets based on the filming scenes: SET-1 includes 60 videos taken on a parking lot with slightly different zoom rates and little camera jitter; other 60 video sequences, named SET-2, are taken on a lawn in a windy day with cluttered backgrounds: tree moves, passerby and more camera jitters. Main problems lie in the complex filming scenes and SET-2 is more difficult than SET-1, which can be figured out in Figure 3.

Rochester Activities dataset contains 10 classes of daily living activities, shown in Figure 4. Each activity is composed of a number of sub-actions. Videos are divided into 5 sets based on 5 actors of different shapes, genders and behavioral habits, and each set contains three-time repetitions of 10 classes. Different with UT-Interaction, the main difficulties in Rochester Activities are the inter-class activity ambiguities, e.g., eating a banana is similar to eating snacks, and turning pages in a telephone book seems having the same hand motions with peeling a banana.

Experimental settings: For accurate comparisons with [10], we follow its protocol to extract the spatio-temporal interest points introduced in [3]. For each frame, no more than 20 points are extracted. Then each $4 \times 4 \times 4$ point-centered cuboid is described as a 640-dimensional feature using 3DSIFT [5]. Note that other feature extractor and descriptor are also available as long as they provide the *xyt* location and feature value of the detected motion region. For representation, a 6×6 RSOM is randomly initialized, then trained during $T_{max} = 100$ iterations using the feature vectors of the one-third videos, randomly selected from the whole dataset. Other parameters are fixed as follows: initial learning rate $\alpha_i = 0.8$



(a) SET-1 (b) SET-2

Figure 5: Inference rates with respect to the video observing ratios on UT-Interaction dataset.

and final learning rate $\alpha_f = 0.001$, Gaussian kernel width $\sigma = 3$ and influence factor $\lambda = 0.7$. The kNN classifiers, where $k = 3$, are uniformly trained using fully observed videos for the test of any observing stage. On UT-Interaction, results are obtained by the method of leave-one-out cross validation on each "SET". On Rochester Activities, for each round, 120 videos taken by four actors are used for training, and the remaining 30 videos for testing.

To evaluate the superiority of the RSOM trajectory and its measurement DTW-E distance, three experimental settings are implemented: RSOM trajectory and DTW-E distance; RSOM trajectory and Euclidean distance (zero-padding to compensate different lengths); BoVW (using K-means algorithm) and Euclidean distance. Note that the size of codebook in K-means is set as 160 on UT-Interaction and 210 on Rochester Activities by "gap" statistic [26]. Since there are randomness in initializing the weights of RSOM units as well as the K-means clusters, each performance is reported as the average accuracy of 20 runs. The results of Dynamic BoVW and Integral BoVW in the state-of-the-art work [10] are included.

Results on UT-Interaction: The results of different methods are shown in Figure 5, where the crucial rates at 50% and 100% observing ratio are marked with three significant figures. It can be seen from the red/green/magenta curves that our approach significantly outperforms the Dynamic BoVW and Integral BoVW on both scenes. For example, on SET-1, our approach is able to make an inference with the accuracy higher than 80% after observing only the first 20% video contents, while the Dynamic BoVW must observe more than 70% to reach the same accuracy. It is valuable to note that the highest performances at the very beginning (10%) are reached by our approach, validating the powerful data mining strength of our approach with very limited observation. On one hand, this is attributed to the rich distinctiveness of activity contexts. On the other hand, it dose work that the redundant 90% trajectory in the trained template makes less false on the recognition of the tested 10% trajectory by our approach than using BoVW, in which the distribution histogram of 10% ratio is significantly changed in the 100% template. Besides, on SET-1, after 40% ratio, the blue curve is almost lower than the green one while it becomes higher on SET-2, proving the ability of RSOM trajectory to handle more complex videos. Comparing the red/blue/cyan curves, we can see both the RSOM trajectory and DTW-E distance contribute to the overall

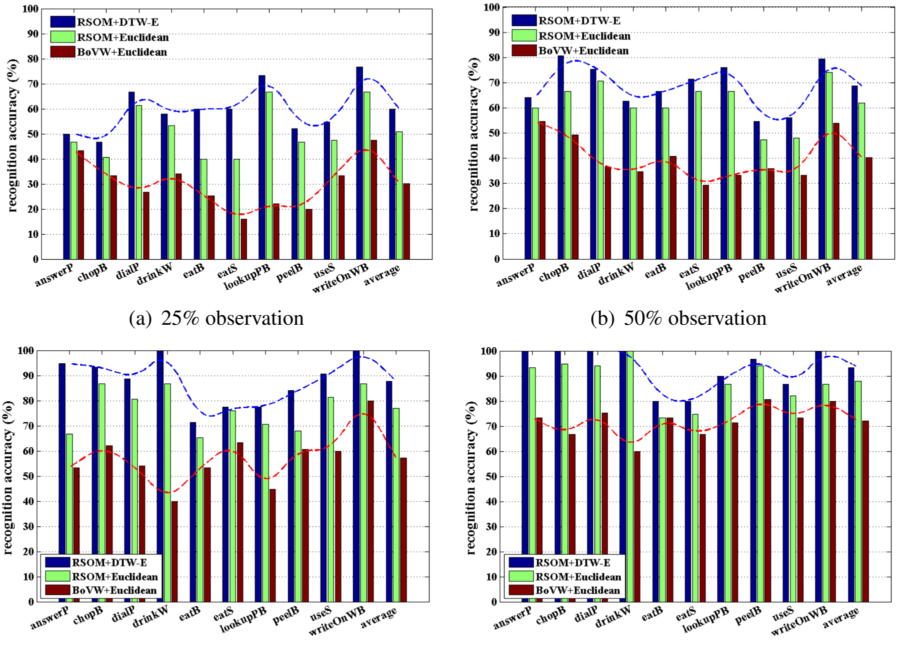
Table 1: Inference rates of 6 activities by cross-test on SET-1 and SET-2 at 10 observation stages. (%)

observation	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	average
hug	61.75	73.75	76.25	85.00	89.50	90.75	88.25	94.75	92.25	93.50	*84.58*
kick	45.25	65.00	71.50	86.25	84.75	86.25	93.25	93.25	98.50	95.75	*81.97*
point	50.50	57.25	78.50	85.25	85.50	90.25	91.50	92.75	95.00	93.75	*82.03*
punch	39.25	51.25	75.00	78.25	82.50	80.25	81.25	85.00	90.75	93.50	*75.70*
push	35.25	46.00	61.50	82.50	86.50	89.75	91.75	97.25	95.75	95.75	*78.20*
shake-hands	30.50	45.75	66.00	66.50	69.00	78.75	85.75	89.50	89.00	90.50	*71.12*
average	*43.75*	*56.50*	*71.46*	*80.63*	***82.96***	*86.00*	*88.62*	*92.08*	*93.54*	***93.79***	–

improvements. Particularly, the proposed DTW-E distance brings 11.1% improvements on SET-1 and 15.9% on SET-2 at 100% ratio, it hence validates that DTW-E can properly grasp the underlying semantics in RSOM trajectories, which are disturbed in complex scenes.

To test the scene portability of our approach, another experiment was implemented using the cross-test on SET-1 and SET-2. That is, in each test of SET-1, the classifiers are trained using the videos on SET-2, then videos on SET-1 are used for training in turn. Table 1 shows the total results of 6 activities at each observation ratio. The average rate of "shake-hands" (71.12%) is the lowest probably because the holding hands are too motionless to generate sufficient features. It is worth noting that there are some rates (marked with red) going down during observation increases because of the immediate occurrence of serious background interference. However, the average trend in the last row is steady increasing as expected, and the high accuracies prove the robustness of our approach to handle scene changing.

Results on Rochester Activities: In Figure 6, the superiority of our approach is obviously shown at four observing stages. It is attributed to the ability of RSOM to grasp sequential contexts for the disambiguation of complex activities. For example, at early stages, "write on a white board" contains turning to back with a writing brush, and both "look up a phone number" and "eat snacks" contain a back to front turn with an object in hand. It very likely leads to activity ambiguities when ignoring the before-after relationships among the local features of front/back moving. RSOM encodes this contextual information during its training process, hence greatly outperforms BoVW. Particularly, at 100% ratio, "write on a white board" is completely classified by our approach. In addition, the fitting curves of blue/red bars are plotted to show that our approach makes the highest improvements over BoVW clearly at the 25% ratio, validating our superiority for early recognition. Noting that our improvements over BoVW become smaller during observation increases because the overall-updated histogram of BoVW gradually approximates to the 100% histogram template, and our advantage of before-after trajectory independence becomes less obvious.



(a) 25% observation

(b) 50% observation

(c) 75% observation

(d) 100% observation

Figure 6: Inference rates with respect to 4 video observing ratios on Rochester Activities dataset.

5 Conclusions

To robustly infer ongoing human activities, the RSOM trajectory and its measurement, DTW-E distance, are novelly proposed in this paper. The motivation is to enable early recognition by encoding videos' contextual information to the greatest extent and using the alignment based measurement to properly reflect the structural semantics of RSOM trajectories. Experiments on the datasets, from human-human interaction to daily living activity, show that our approach often makes the efficient inference 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 than the histogram of BoVW.

Acknowledgement: This work is supported by National Natural Science Foundation of China(NSFC, No.60875050, 60675025), National High Technology Research and Development Program of China(863 Program, No.2006AA04Z247), Scientific and Technical Innovation Commission of Shenzhen Municipality (No.201005280682A, No.JCYJ2012061415223 4873, CXC201104210010A).

References

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri. Actions as Space-Time Shapes. In *Proc. ICCV*, pages 1395-1402, 2005.

[2] C. Schuldt, I. Laptev, B. Caputo. Recognizing Human Actions: A Local SVM Approach. In *Proc. ICPR*, pages 32-36, 2004.

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

[4] 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.

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

[6] A. Kl*ä*ser, M. Marszalek, C. Schmid. A Spatio-temporal Descriptor based on 3d-gradients. In *Proc. BMVC*, pages 995-1004, 2008.

[7] F. Lv, R. Nevatia. Single View Human Action Recognition Using Key Pose Matching and Viterbi Path Searching. In *Proc. CVPR*, pages 1-8, 2007.

[8] A. Ravichandran, G. Hager, R. Vidal. Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *Proc. CVPR*, pages 1932-1939, 2009.

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

[10] M. S. Ryoo. Human Activity Prediction: Early Recognition of Ongoing Activities from Streaming Videos. In *Proc. ICCV*, pages 1036-1043, 2011.

[11] M. Hoai, F. De la Torre, Max-Maigin Early Event Detection. In *Proc. CVPR*, pages 2863-2870, 2012.

[12] L. Shao, R. Mattivi. Feature detector and descriptor evaluation in human action recognition. In *Proc. CIVR*, pages 477-484, 2010.

[13] Qianru Sun, Hong Liu. Action Disambiguation Analysis Using Normalized Google-Like Distance Correlogram. In *Proc. ACCV'12*, Part III, *LNCS* 7726, pages 425-437, 2013.

[14] S. Savarese, A. DelPozo, J. C. Niebles, L. Fei-Fei. Spatial-Temporal correlatons for unsupervised action classification. In *Proc. WMVC*, pages 1-8, 2008.

[15] M. Varsta, Jos*é* del R. Mill*á*n, J. Heikkonen. A Recurrent Self-Orgnizing Map for Temporal Sequence Processing. *LNCS* 1327, pages 421-426, 1997.

[16] T. Kohenen. Self-Organizing Maps. vol. 30 of *Lecture Notes in Information Sciences*, Springer, 2nd Edn, 1997.

[17] L. Leinonen, J. Kangas, K. Torkkola, A. Juvas. Dysphonia detected by pattern recognition of spectral composition. *J. Speech and Hearing Res.*, 35: 287-295, 1992.

[18] V. Tryba, K. Goser. Self-Organizing Feature Maps for process control in chemistry. *Artifcial Neural Networks*, pages 847-852, 1991.

[19] Z. Zhang, K. Huang, T. Tan, Comparison of Similarity Measures for Trajectory Clustering in Outdoor Surveillance, In *Proc. ICPR*, vol. 3, pages 1135-1138, 2006.

[20] M. Vlachos, G. Kollios, D. Gunopulos. Discovering Similar Multidimensional Trajectories. In *Proc. ICDE*, pages 673-685, 2002.

[21] F. I. Bashir, A. A. Khokhar, D. Schonfeld. Segmented trajectory based indexing and retrieval of video data. In *Proc. ICIP*, pages 623-626, 2003.

[22] E. J. Keogh, M. J. Pazzani. Scaling up Dynamic Time Warping for Datamining Application. In *Proc. ACM SIGKDD*, pages 285-289, 2000.

[23] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707-710, 1966.

[24] F. Yuan, G. S. Xia, H. Sahbi, V. Prinet. Mid-level features and spatio-temporal context for activity recognition. *Pattern Recognition*, 45(12): 4182-4191, 2012.

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

[26] R. Tibshirani, G. Walther, and T. Hatie. Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2): 411-423, 2001.

[27] Qianru Sun, Hong Liu. Learning Spatio-Temporal Co-occurrence Correlograms for Efficient Human Action Classification. In *Proc. ICIP*, 2013.