


Signal, Image, and Video Processing

# Signal, Image and Video Processing



 Springer

# Visual trajectory analysis via Replicated Softmax-based models

Xiaogang Chen · Qixiang Ye · Jialing Zou · Ce Li ·  
Yanting Cui · Jianbin Jiao

Received: 3 January 2014 / Revised: 28 April 2014 / Accepted: 28 June 2014 / Published online: 15 July 2014  
© Springer-Verlag London 2014

**Abstract** In this paper, we apply the Replicated Softmax model to visual trajectory representation and analysis problems. We propose to represent trajectories with “bag-of-word” features from a spatially distributed codebook and then use a Replicated Softmax model to characterize trajectories with latent topic units. By stacking an additional label layer or representation layers, the Replicated Softmax model is enhanced with discrimination and generalization capability. Experiments on trajectory classification and trajectory route analysis are conducted to demonstrate the effectiveness of the proposed model.

**Keywords** Replicated Softmax model · Trajectory analysis · Abnormal classification · Trajectory route analysis

## 1 Introduction

Trajectory representation and analysis has been one of the most important research topics with applications in many visual surveillance-related tasks, such as abnormal behavior detection, video indexing and semantic scene understanding [1–5]. Many features extraction and machine learning techniques have been engaged in this topic.

Existing works about trajectory representation can be generalized into two categories: the “fine” and the “coarse” methods. The “fine” methods, which are often used in behavior detection and video indexing, represent trajectories with precise shape information. For example, in [1], Sillito and Fisher proposed to use cubic B-spline curves to approxi-

mate trajectories. In the method, the parameters of control points are extracted as feature representation, with which Gaussian mixture models (GMMs) [1] and sparse reconstruction analysis (SRA) [2] are used to classify trajectories. Besides cubic B-spline curves, Sillito and Fisher [6] explored haar wavelet coefficients, discrete Fourier Transform (DCT) and Chebyshev Polynomial Coefficients for parametric trajectory representation. To better capture trajectory change points for trajectory retrieval, Dyana and Das [7] proposed to use Gabor filters and spectral analysis. Despite of the simplicity of above-reviewed methods, most of them extract features from trajectory shapes, which are difficult to be obtained in noisy environments with an off-the-shelf visual tracking algorithm.

In contrast, the “coarse” methods represent trajectory in a simple way and pay more attention to the analysis approaches. Wang et al. [8] constructed trajectory similarity measure with trajectory point positions, then cluster trajectories into object and activity groups based on the measure. Junejo [9] proposed to use trajectories to train a scene path model and conduct anomaly detection with dynamic Bayesian networks. Recent works about semantic region analysis [5, 10, 11] introduced the notion from the natural language process community and cast the video scene into a codebook (dictionary). A trajectory is treated as a document generated from the dictionary, and the “bag-of-word” (BOW) representation is obtained by quantizing the occupation in the spatial-distributed cells and motion directions. The “BOW” representation is then clustered by a topic model [5, 10, 11] to learn semantic regions. Intuitively, the “BOW” features adopt coarse grid cells and discarded the temporal relationship among words, which limit their usage in problems that model trajectory in a fine manner, and the employed non-parametric Bayesian methods and topic model do work well in discriminative classification tasks.

X. Chen · Q. Ye · J. Zou · C. Li · Y. Cui · J. Jiao (✉)  
School of Electronics, Electrical and Communication Engineering,  
University of Chinese Academy of Sciences, Huairou District,  
Huaibei Village, A-2 Building, Beijing, China  
e-mail: jiaojb@ucas.ac.cn

Recent research about representation learning [12], also known as deep learning, has demonstrated that by stacking multiple layers of nonlinear transformation model like Restricted Boltzmann Machines (RBM) into a deep multi-layer model [13], one can obtain a compact representation of data in the top layer. The deep learning methods have achieved great success in computer vision research. And the most recently proposed Replicated Softmax model [14], a variant of RBM and “undirected topic model,” outperforms Latent Dirichlet Allocation (LDA) [15] in document modeling [16]. This inspires us to engage the deep learning methods in the trajectory analysis tasks. We show that it can achieve the state-of-the-art performance in abnormal trajectory detection and classification tasks. We also show that the model can work as well as a topic model in semantic scene modeling.

The rest of paper is organized as follows. In Sect. 2, the RBM and its variant Replicated Softmax model are introduced. In Sect. 3, experiments about trajectory analysis and classification are presented and analyzed, followed by conclusion and discussion in Sect. 4.

## 2 Methodology

The Replicated Softmax (RS) is a variant of the RBM. Therefore, we first introduce the RBM briefly and then extend it to the RS model and show its application in trajectory analysis.

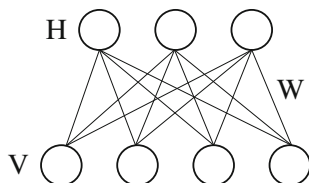
### 2.1 Restricted Boltzmann Machines

RBM, initially known as Harmonium [17], is a bipartite undirected graphical model, as shown in Fig. 1, it models a set of visible and hidden units through an energy function.

Specifically, for binary visible units  $V \in \{0, 1\}^N$  and hidden units  $H \in \{0, 1\}^M$ , where  $N$  and  $M$  are number of visible and hidden units; the energy function is defined as:

$$E(V, H) = -VWH - a^T V - b^T H, \quad (1)$$

where  $W$  is the weight matrix of connecting weights between visible and hidden units, and  $a$  and  $b$  are bias terms for visible and hidden units, respectively. Given the energy function, one can define a joint probability distribution over the bipartite as the free energy



**Fig. 1** Restricted Boltzmann Machines

$$P(V, H) = \frac{1}{Z(\theta)} e^{-E(V, H)}, \quad (2)$$

where  $Z(\theta)$  is the partition function, and  $\theta = \{W, a, b\}$  are model parameters. Since there is no connection among intra level units, the conditional distribution can be derived as

$$P(h_j = 1|v) = \sigma \left( b_j + \sum_{i=1}^N w_{i,j} v_i \right), \quad (3)$$

$$P(v_i = 1|h) = \sigma \left( a_i + \sum_{j=1}^M w_{i,j} h_j \right), \quad (4)$$

where  $\sigma(x) = 1/(1 + e^{-x})$  is a sigmoid function. Given a set of observations, the training procedure of the RBM is to maximize the probability assigned to it, as:

$$\max P(V) = \frac{1}{Z} \sum_h e^{-E(V, h)}. \quad (5)$$

For convenience, the optimization objective can use the log probability instead, and the derivative with respect to parameter  $\theta$  is:

$$\frac{\partial \log p(V)}{\partial \theta} = \frac{1}{T} \sum_t E_{H|V_t} \left[ \frac{\partial E(V_t, H)}{\partial \theta} \right] - E_{V, H} \left[ \frac{\partial E(V, H)}{\partial \theta} \right]. \quad (6)$$

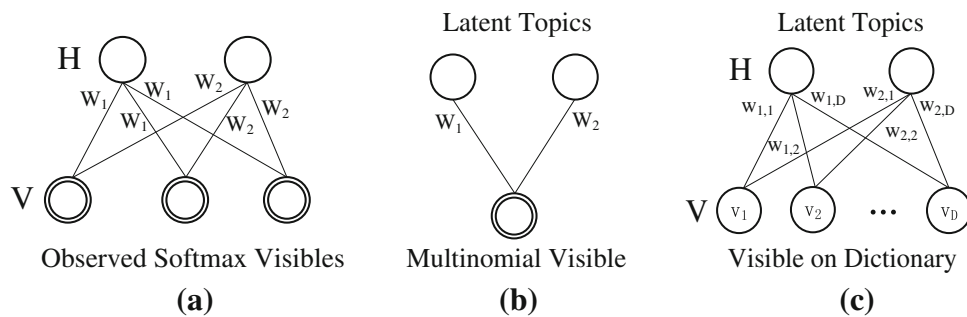
Since the second term of the gradient is intractable, Hinton [18] proposed contrastive divergence to approximate the expectation with a finite step Gibbs sampling, which turns out to be very efficient in RBM model training.

### 2.2 Replicated Softmax model

Replicated Softmax (RS) model is first proposed by Hinton and Salakhutdinov [14] as a topic model to model discrete word count data for document analysis. To understand the RS model, first change the RBM visible units from binary to softmax units, with which, the conditional distribution on visible units (4) become:

$$P(v_i = 1|h) = \frac{\exp(a_i + \sum_{j=1}^M w_{i,j} h_j)}{\sum_{i'=1}^N \exp(a_{i'} + \sum_{j=1}^M w_{i',j} h_j)} \quad (7)$$

In the softmax unit, only one of the  $N$  visible units can be activated, so it can be used to model a specific word in document, i.e., for a specific position in a document, choose only one word in the dictionary to fill that position. (The softmax unit can also be used to model class label for classification task, which will be introduced in the later section). Based on this interpretation, for a document of  $K$  words, we can model the document with  $K$  RBMs, each for a specific word in the document. Further, if we ignore the order of



**Fig. 2** Three representations of the Replicated Softmax model. **a** Model of document that contains three words [14]. **b** Replace the  $K$  softmax units in the left with a single multinomial unit. For a document has  $K$  words, the visible is sampled  $K$  times [14]. These two representations explain the generating process of a document. Notice that  $W_j$  is

a weight vector whose dimension is equal to the dictionary size  $D$ . This is slightly different from that in RBM. **c** A document is a word count observation on a dictionary and the hidden units are connected to the dictionary elements, which might be easier to understand the inference procedure

words in the document, the parameter can be shared with all  $K$  RBMs, so the  $K$  groups of softmax units can be assembled to a multinomial unit to construct the RS model. In this sense, the RS model can be seen as a special type of RBM in which the hidden units remain binary and the visible unit is multinomial. Figure 2 shows different interpretations of the RS model.

Specifically, for a document containing  $K$  words, considering a word dictionary of size  $D$ ,  $V \in R^D$  is the visible unit and  $H \in \{0, 1\}^M$  the binary hidden unit, the energy of state  $(V, H)$  is defined as:

$$E(V, H) = - \sum_{j=1}^M \sum_{d=1}^D v_d w_{j,d} h_j - \sum_{d=1}^D b_d v_d - K \sum_{j=1}^M a_j h_j, \quad (8)$$

where  $\sum v_d = K$ ,  $v_d$  is the word count for the  $d$ -th element of dictionary. As RBM, the conditional distribution can be derived from the following energy function

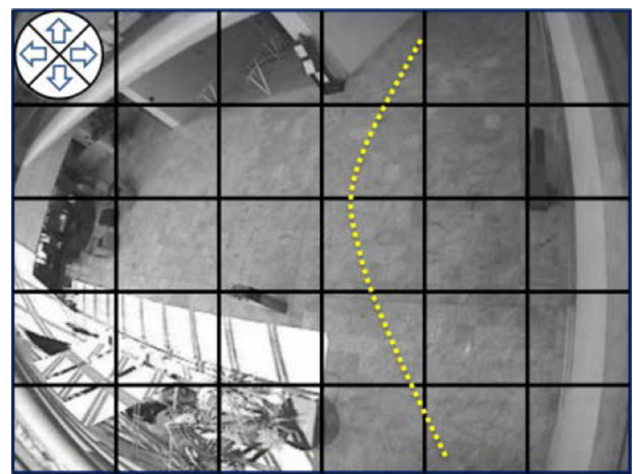
$$P(h_j = 1 | V) = \sigma \left( \sum_{d=1}^D w_{j,d} v_d + K a_j \right), \quad (9)$$

$$P(v_d = 1 | H) = \frac{\exp \left( \sum_{j=1}^M w_{j,d} h_j + b_d \right)}{\sum_{d'=1}^D \exp \left( \sum_{j=1}^M w_{j,d'} h_j + b_{d'} \right)}. \quad (10)$$

Noticing that the hidden unit bias term is multiplied with the document length variable  $K$ , this is important in modeling different length of documents. And the visible conditional distribution has the softmax form.

### 2.3 Replicated Softmax model for trajectory analysis

To employ the RS model in trajectory analysis, we first construct the BOW trajectory representation. As Fig. 3 show, the video scene is divided into grid regions, in each, the motion



**Fig. 3** The construction of bag-of-words representation

directions are partitioned into different bins, with which the dictionary is built. A trajectory that originally obtained as serial discrete points is then projected into dictionary, and the occupation of which are collected to form the final feature representation. After these, the features are trained with the RS model.

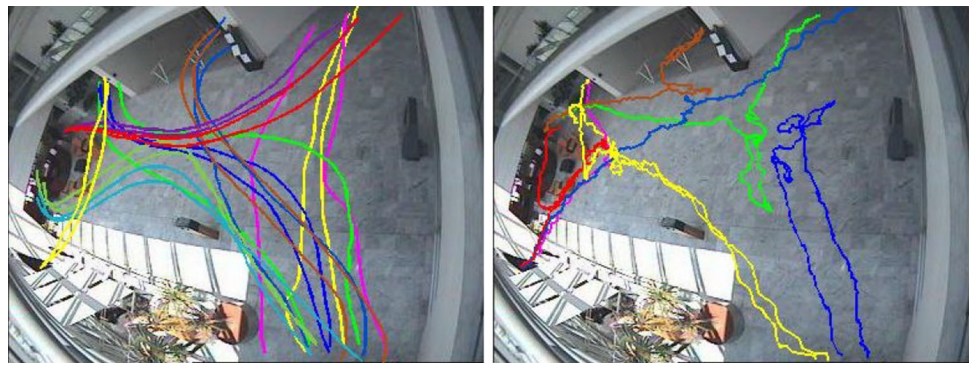
## 3 Experiments

We conduct two groups of experiments with distinct purposes, one is to show how the Replicated Softmax model performs discriminative trajectory classification problem, and the other is to demonstrate its usage in trajectory route analysis.

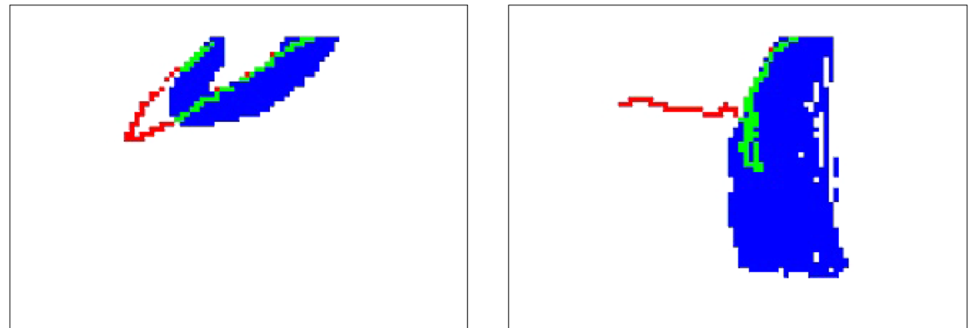
### 3.1 Trajectory classification

The first experiment is conducted on the abnormal behavior detection and classification task, which shows how the proposed method outperforms other methods. We test the

**Fig. 4** Examples of normal and abnormal trajectories. *Left* is normal and *right* is abnormal. (Best viewed in color)



**Fig. 5** Examples of dictionary occupation



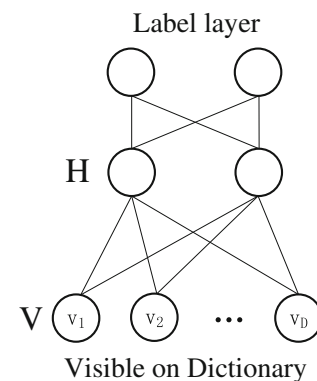
algorithm on the CAVIAR Lobby dataset [1]. The training set consists of 22 categories of behaviors, each of which has 100 simulated trajectories. The test set contains 40 trajectories, of which 19 are abnormal and 21 are normal behaviors, Fig. 4 shows some of the trajectory examples. The task can be split into a detection stage and a classification stage. The detection stage is designed to distinguish abnormal behaviors from normal ones, while the classification stage is used to classify the normal behaviors into specific categories. We use a single-layer RS model with the label layer stack on the top for discriminative training, as shown in Fig. 6. This is based on the observation that a single-layer structure that requires much less training time than multiple layers is sufficient to obtain good results.

We use Detection Accuracy (DACC) and correct classification rate (CCR) defined in [2] as measurements to quantitatively evaluate the trajectory classification performance. DACC and CCR are defined as

$$\text{DACC} = \frac{\text{True Positive Number} + \text{True Negative Number}}{\text{Number of total test trajectories}} \quad (11)$$

$$\text{CCR} = \frac{\text{Number of correctly classified behavior}}{\text{Number of total test trajectories}}. \quad (12)$$

To obtain the BOW representation, we first divide the scene with  $6 \times 6$  pixels rectangles and quantize moving directions into 4 bins. The dictionary size of a  $384 \times 288$  resolution video scene is therefore  $64 \times 48 \times 4$ .



**Fig. 6** Classification structure

At the detection stage, for a testing sample, we iteratively compare its dictionary occupation rate with each category of training behaviors. The dictionary occupation rate is defined as the percent of a testing sample's dictionary elements possession shared with specific category behaviors. As show in Fig. 5, blue regions indicate the occupation of a specific category of behaviors, and red curves are the occupied regions of an abnormal behavior. The green curves are shared by above two. For example, if a test sample has 100 nonzero dictionary entities, and 70 of them appear in a category's dictionary occupation, then the occupation rate with this category is set to 70 % (Fig. 6).

If one of the occupation rates is larger than a preset threshold, the testing trajectory is classified as a normal behavior; otherwise, it is an abnormal behavior. It reaches 90 % DACC rate which is comparable to the SRA method [2], which has



**Table 1** Correct classification rate of the CAVIAR dataset

Correct classification rate (CCR)	
Sparse reconstruction analysis (SRA) [2]	70.07 $\pm$ 6.13 %
Replicated Softmax (100 hidden)	74.29 $\pm$ 5.43 %
Replicated Softmax (200 hidden)	<b>75.71 <math>\pm</math> 5.81 %</b>
Replicated Softmax (300 hidden)	71.42 $\pm$ 3.85 %
Bag-of-Word Feature+SVM	66.67 %
Replicated Softmax Feature+SVM	75.24 $\pm$ 1.90 %

The best result is shown in bold

90.42  $\pm$  3.85 % DACC rate. This shows that the detection strategy is simple but effective.

At the trajectory classification stage, testing samples are directly fed into the RS model to obtain the classification results. It should be noticed that there is only one parameter in the single-layer RS model: the number of hidden units, which is highly relevant to the performance. With less hidden units, the model might be unable to capture characteristics of different behaviors. With more hidden units, the training data could be insufficient for training, introducing the under-fitting problem. In experiments, we test the performance with 100, 200 and 300 hidden units. Considering that the RS training might stuck at the local minimum, we conduct the 10 trails experiments. The average results are given in Table 1.

In Table 1, it can be seen that the increasing of hidden unit number does not always improve the classification performance. With 200 hidden layers, the performance reaches the highest. However, with 300 hidden units, the CCR falls, which might implies an under-fitting model. In this case, more training samples might help to solve the under-fitting problem.

It can also be seen that our proposed method has a 5.7 % CCR performance improvement over the SRA-based method. This can be seen as a significant improvement over the state of the art. In addition, as the classification in RS model is straight forward, the computation efficiency is much higher than that of SRA, which uses an online optimization in classification. This is crucial for systems of real-time performance requirements.

We also conduct two experiments to show how the RS model can improve the BOW representation, as listed at the last two lines in Table 1. It can be seen that using original BOW features with an SVM classifier can achieve 66.67 % CCR. However, if we extract the hidden units activations of

RS model as features, the mean value of CCR is boosted to 75.24 % in 10 trails experiment, which indicates that the Replicated Softmax model does capture the discriminative information among examples.

### 3.2 Trajectory route analysis

The trajectory route analysis is to discover the commonly paths taken by the objects in a specific scene, these include the object's entrance and exit area, also moving paths. We conduct the trajectory route analysis experiment on Grand Central Station Dataset as [10] with the same parameter setting.

The 720  $\times$  480 resolution video scene is cast into 72  $\times$  48 of size 10  $\times$  10 rectangle cells. The moving orientation is divided into up, down left and right four directions, thus forming a dictionary with 13,824 words.

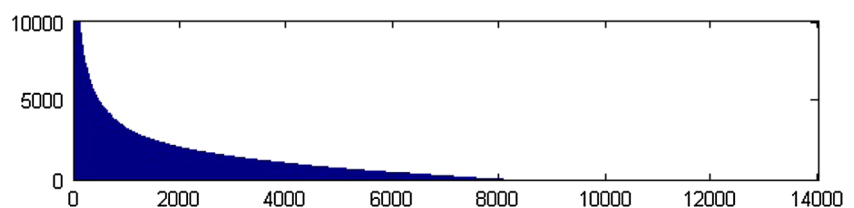
In this construction, the training features might not cover all dictionary elements. There are about 5,000 elements of zero coefficients, as show in Fig. 7, which means that these elements do not appear in the trajectory set. Since the zero coefficient elements would not be captured in model training, we choose the most frequent 9,000 words to present trajectories to accelerate the training process.

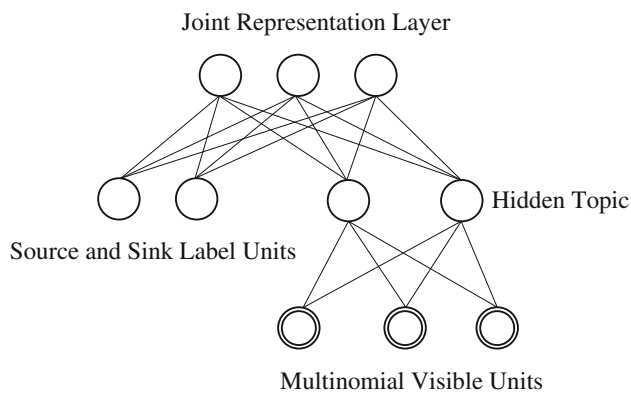
In [10], Zhou et al. added a source and sink node in topic model to denote the entrance and exit. With this, a topic is presented with two sets of parameters. One is for source and sink labels and the other for word distribution. However, the Replicated Softmax model only captures the latent topic-word relations. To overcome this shortage, we propose to use a structure of Multimodal Deep Boltzmann Machine [19], as show in Fig. 8. On the left side, we use simple softmax units to present source and sink labels, and the right side is a RBM with Replicated Softmax visible units and binary hidden units. Based on these, we stack an additional layer of binary units on the top of two sides to capture their joint distribution.

With an additional layer to enrich model's representation capability, however, parameter learning becomes more difficult. We employ the approximate learning method [19] that uses mean-field inference to estimate data-dependent expectations and MCMC approximation in model's expectation to maximize a variational likelihood bound.

Given a trained model, inferring the joint representation can be achieved by conducting a Gibbs sampling from

**Fig. 7** Word frequency. About 5,000 words have zero coefficients

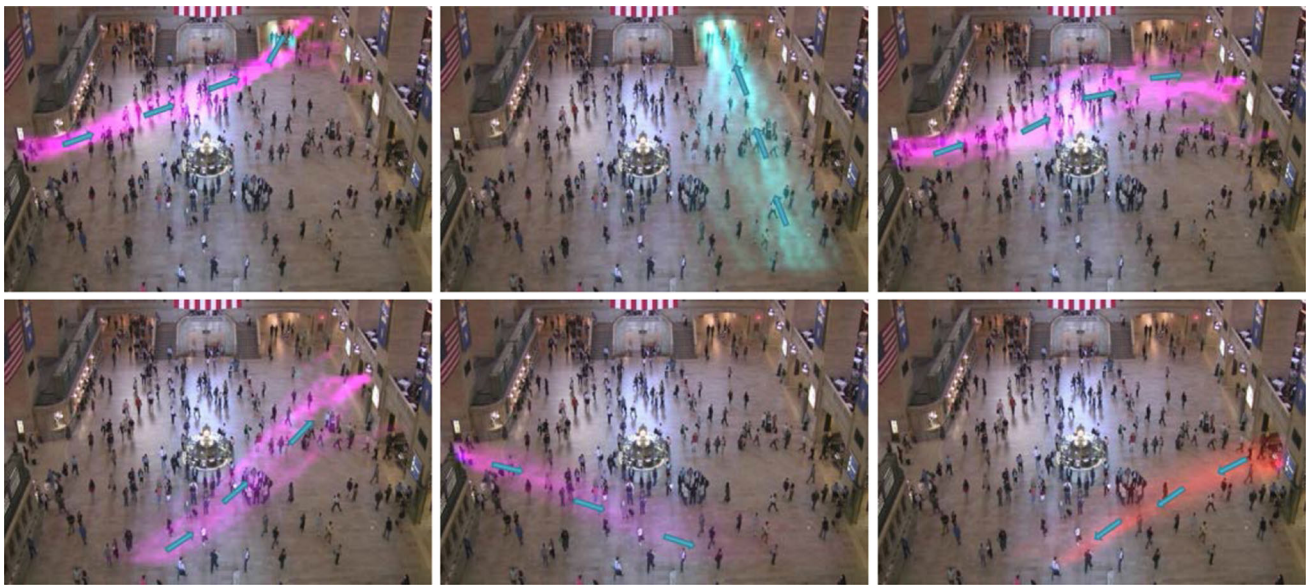




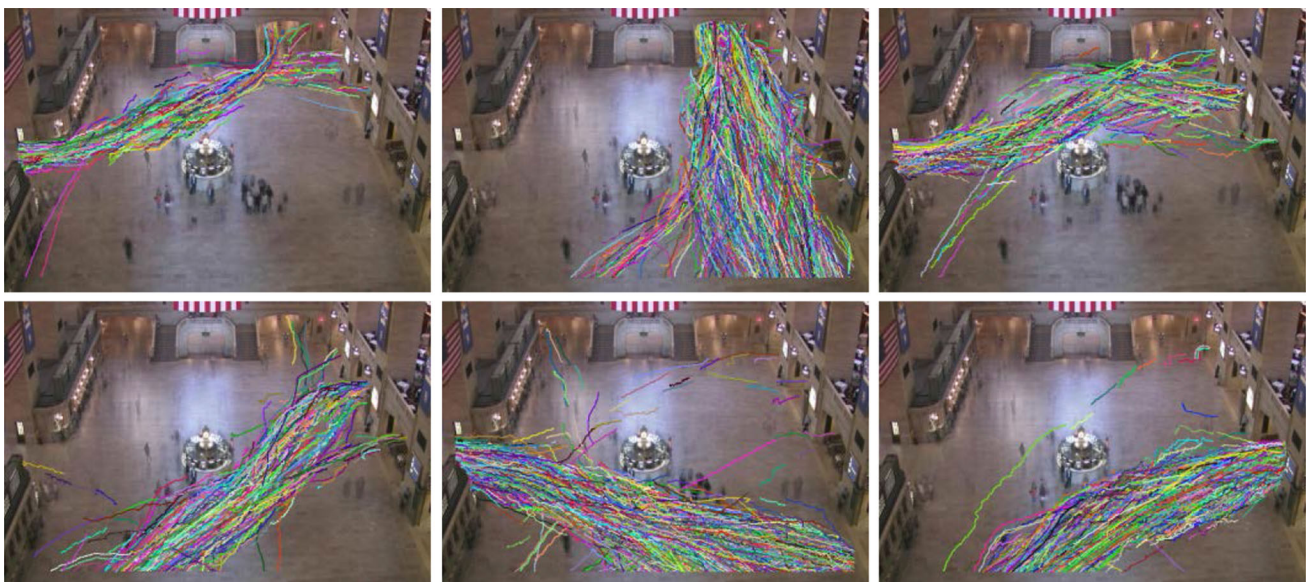
**Fig. 8** Multimodal deep Boltzmann machine for semantic scene learning

$P(h_{\text{Joint}}|V_{RS}, V_{SS})$  or  $P(h_{\text{Joint}}|V_{RS})$ . The latter is for the case when source and sink labels are not provided. A faster alternative of the inference is to use a mean-field approximation, for more details and explanation, we refer to [19]. In [10], Zhou et al. proposed to use a spanning tree to assemble tracklets into long trajectories; however, there might be some trajectories that miss source or sink labels after the assembling. This is well handled in the proposed DBM model.

Since the Replicated Softmax is an undirected graphical topic model, the hidden units in the DBM did not have a clear “topic” meaning as a directed model. Nevertheless, it is pointed out that the hidden representation can be recognized as latent topic features, and they are sufficient to characterize trajectories [20]. One can manage to obtain the semantic



**Fig. 9** Representative semantic regions



**Fig. 10** Representative trajectory clusters



**Fig. 11** Examples of failure cases



word distribution that represents the meaningful route patterns. Figure 9 shows some of the representative semantic regions, and Fig. 10 shows their corresponding clustered trajectories.

It can be seen that the proposed model can capture meaningful motion patterns with specific entrances and exits, along with spatial-distributed routes. With the learned structure, trajectories can be correctly clustered for further processing, e.g., classification and entrance/exit prediction. Though the deep learning method shows good modeling capability, it requires some trivial parameter fine turning. The failure of parameter fine turning could introduce failure cases, as shown in Fig. 11. In these cases, trajectories with different sources or sinks are grouped into one cluster. This could be introduced by two reasons. One is the inappropriate parameter setting, which includes number of hidden units, training iterations, and mini-batch of the stochastic gradient descent. The other is the small number of training samples. It is observed that a specific semantic category that has small amount of samples could be merged by adjacent ones.

#### 4 Conclusion

In this paper, we propose to use deep learning methods in trajectory representation and analysis. Based on the document type BOW representation, a Replicated Softmax (RS) model is used to characterize trajectories. Two distinct tasks, abnormal behavior analysis and trajectory route analysis, are conducted for validation. We explore the Replicated Softmax model in trajectory analysis, and experiments show that

it can achieve comparable performance with state-of-the-art methods. Experiments also show that the model can work as well as the popular topic model in trajectory route analysis. The good performance on two different tasks and datasets demonstrate the broad applicability of the method.

In the future, it is required to test the proposed methods in related tasks and more challenging datasets. In addition, the automatic parameter setting problem should be considered [21].

**Acknowledgments** This work was supported in Part by National Basic Research Program of China (973 Program) with Nos. 2011CB706900, 2010CB731800, and National Science Foundation of China with Nos. 61039003, 61271433 and 61202323.

#### References

1. Sillito, R.R., Fisher, R.B.: Semi-supervised learning for anomalous trajectory detection. In: Proceedings of the British Machine Vision Conference, Leeds, pp. 1035–1044 (2008)
2. Li, C., Han, Z., Ye, Q., Jiao, J.: Visual abnormal behavior detection based on trajectory sparse reconstruction analysis. *Neuralcomputing* **119**, 94–100 (2013)
3. Jung, Y.K., Lee, K.W., Ho, Y.S.: Content-based event retrieval using semantic scene interpretation for automated traffic surveillance. *Intell. Transp. Syst. IEEE Trans.* **2**(3), 151–163 (2001)
4. Bashir, F., Khokhar, A., Schonfeld, D.: Real-time motion trajectory-based indexing and retrieval of video sequences. *Multimed. IEEE Trans.* **9**(1), 58–65 (2007)
5. Wang, X., Ma, K.T., Ng, G.W., Grimson, E.L.: Trajectory analysis and semantic region modeling using nonparametric hierarchical Bayesian models. *Int. J. Comput. Vis.* **95**(3), 287–312 (2011)
6. Sillito, R.R., Fisher, R.B.: Parametric trajectory representations for behaviour classification. In: Proceedings of the British Machine Vision Conference, London, pp. 227–238 (2009)



7. Dyana, A., Das, S.: Trajectory representation using Gabor features for motion-based video retrieval. *Pattern Recognit. Lett.* **30**(10), 877–892 (2009)
8. Wang, X.G., Tieu, K., Grimson, E.: Learning semantic scene models by trajectory analysis. In: *Proceedings of the European Conference on Computer Vision*, Graz, Austria, pp. 110–123 (2006)
9. Junejo, I.N.: Using dynamic Bayesian network for scene modeling and anomaly detection. *Signal Image Video Process.* **4**(1), 1–10 (2010)
10. Zhou, B., Wang, X., Tang, X.: Random field topic model for semantic region analysis in crowded scenes from tracklets. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado, pp. 3441–3448 (2011)
11. Zou, J., Chen, X., Wei, P., Han, Z., Jiao, J.: A belief based correlated topic model for semantic region analysis in far-field video surveillance systems. In: *Pacific-Rim Conference on Multimedia (PCM)*, Nanjing, China, pp. 779–790 (2013)
12. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. *Pattern Anal. Mach. Intell. IEEE Trans.* **35**(8), 1798–1828 (2013)
13. Salakhutdinov, R., Hinton, G.E.: Deep Boltzmann machines. In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*, Florida, pp. 448–455 (2009)
14. Hinton, G.E., Salakhutdinov, R.: Replicated softmax: an undirected topic model. In: *Proceedings of the Advances in Neural Information Processing Systems*, Nevada, pp. 1607–1614 (2009)
15. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003)
16. Srivastava, N., Salakhutdinov, R., Hinton, G.E.: Modeling documents with deep Boltzmann machines. In: *Presented in Uncertainty in Artificial Intelligence*, Seattle, USA (2013)
17. Rumelhart, D.E., McClelland, J.L.: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, vol. 1. MIT Press, Cambridge, MA (1986)
18. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. *Neural Comput.* **14**(8), 1771–1800 (2002)
19. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep Boltzmann machines. In: *Proceedings of the Advances in Neural Information Processing Systems*, Nevada, pp. 2231–2239 (2012)
20. Larochelle, H., Lauly, S.: A neural autoregressive topic model. In: *Advances in Neural Information Processing Systems*, Nevada, pp. 2717–2725 (2012)
21. Hinton, G.E.: A practical guide to training restricted boltzmann machines. In: *Neural Networks: Tricks of the Trade*, vol. 7700, pp. 599–619. Springer, Berlin (2012)