Video Person Reidentification based on Neural Ordinary Differential Equations and Graph Convolution Network

Li-qiang ZHANG (dxl@zznu.edu.cn)

Electrical engineering department,Zhengzhou Technical College https://orcid.org/0000-0003-0868-4083

Long-yang HUANG

CAFUC University: Civil Aviation Flight University of China

Xiao-li DUAN

Zhengzhou Teachers College: Zhengzhou Normal University

Research

Keywords: person reidentification, graph convolutional network, neural ordinary equations

Posted Date: May 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-466426/v1

Version of Record: A version of this preprint was published at EURASIP Journal on Advances in Signal Processing on August 5th, 2021. See the published version at https://doi.org/10.1186/s13634-021-00747-1.

Video person reidentification based on neural ordinary differential equations and graph convolution network

[1] * Li-qiang ZHANG, [2] Long-yang HUANG, [3] Xiao-li DUAN

1 .Electrical engineering department , Zhengzhou Technical College, Zhengzhou,450121, China，dxl@zznu.edu.cn

2. College of Air Traffic Management, Civil Aviation Air Flight University of China， Guanghan，618307, China，longyanghuang@cafuc.edu.cn

3.School of Economics & Management，Zhengzhou Normal University， Zhengzhou，450044, China, dxly2005@126.com

Abstract Person reidentification rate has become a challenging research topic in the field of computer vision due to the fact that person appearance is easily affected by lighting, posture and perspective. In order to make full use of the continuity of video data on the time line and the unstructured relationship of features, a video person reidentification algorithm combining the neural ordinary differential equation with the graph convolution network is proposed in this paper. First, a continuous time model is constructed by using the ordinary differential equation (ODE) network so as to capture hidden information between video frames. By simulating the hidden space of the hidden variables with the hidden time series model, the hidden information between frames that may be ignored in the discrete model can be obtained. Then, the features of the generated video frames are given to the graph convolution network to reconstruct them. Finally, weak supervision are used to classify the features. Experiments on PRID2011 data sets show that the proposed algorithm can significantly improve person reidentification performance.

Keywords: person reidentification, graph convolutional network, neural ordinary equations

1. Introduction

In recent years, with the country's increasing attention to public security issues and the development of video surveillance technology, more and more cameras are deployed in crowded places [1][2]. However, the operation of large-scale video surveillance system generates a huge amount of surveillance data, which is difficult to analyze and process quickly by means of solely relying on human resources. Therefore, the intelligent surveillance system that automatically completes the monitoring task by computer vision technology emerges at the right moment [3].

* *Corresponding author.*
E-mail address: dxl@zznu.edu.cn.

Although the current face recognition technology has been relatively mature, it is often impossible to obtain effective face images in the actual monitoring environment. Thus, it becomes very important to use the whole body information to lock and search persons. This also makes person reidentification technology gradually become a research hotspot in the field of computer vision, which attracts extensive attention[5].

Person reidentification aims to accurately identify a person who appears in one camera when he appears again in other cameras [5][6]. Due to the influence of camera point of view, dramatic changes in moving human body posture, lighting, shielding and chaotic background, etc. [7][8], person reidentification algorithm still faces great challenges. At present, the research methods of person reidentification are mainly divided into single-frame image-based and video-based person reidentification [9].

Early video person detection methods are usually based on image detection, and the static features of the image are extracted to determine whether there is a person in each frame. However, with the extensive application of depth model in the field of video detection, recent researches have paid more attention to the characteristics of video information, such as the temporal nature and dynamic characteristics. Graphic convolutional network (GCN) and ordinary differential equation (ODE) are among the latest achievements in machine learning, which apply unstructured and continuous models to a variety of learning tasks. In this paper, a video continuum model is established through the ordinary differential equation, and a continuous time airspace person detection model based on video stream is proposed in combination with the graphic convolution network.

2. Methods

2.1. Graph convolutional network

Most of the graph neural network models use graph convolution, whose core is the convolution kernel parameter sharing in local areas. The same convolution kernel is used for the convolution operation of each graph node, which can greatly reduce the number of model parameters. Parameter update of the convolution kernel can be seen as learning a graph function $G=(\upsilon,\varepsilon)$, which respectively represent the connecting edge between vertices in the graph. The input is eigenmatrix $X \in R^{N\times D}$, N isthe number of vertices, D is the characteristic dimension, and the matrix expression of the graph structure (usually expressed as adjacency matrix A). The output is $Z \in R^{N\times F}$ and F is the output dimension of the convolution layer of the graph. Each graph convolution layer can be represented as the following nonlinear function

$$H^{(l+1)} = f\left(H^{(l)}, A\right) \tag{1}$$

Where, $H^{(0)} = X, H^{(l)} = Z$, and l is the number of convolution layers. For different models of the task, the appropriate convolutional function $f(\cdot,\cdot)$ will be selected and parameterized. This paper uses the same convolution function as Kipf et al. [9], whose basic form is

$$f\left(H^{(l)}, A\right) = \sigma\left(AH^{(l)}W^{(l)}\right) \tag{2}$$

Where $W^{(l)}$ is the weight parameter of the l-level neural network, and $\sigma(\cdot)$ is the nonlinear activation function, usually ReLU (rectified linear unit). After the above improvement, the graph convolution function can be calculated as

$$f\left(H^{(l)}, A\right) = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{3}$$

Where $\hat{A} = A + I$, I is the identity matrix, $\hat{D}$ is the diagonal vertex degree matrix of $\hat{A}$.

2.2. Extraction of continuous hidden state characteristics based on The Constant differential equation

The constant differential equation network is a new branch in the field of neural network. It makes the neural network continuous and uses ordinary differential equation solvers to fit the neural network itself. Its basic problem domain equation is as follows

$$h_{t+1} = h_t + f\left(h_t, \theta_t\right) \tag{4}$$

$$\frac{dh(t)}{dt} = f\left(h(t), t, \theta\right) \tag{5}$$

$$h(T) = h(t_0) + \int_{T_0}^{T} f\left(h(t), t, \theta\right)dt \tag{6}$$

Where h_t stands for the hidden state, and $f(\cdot)$ represents the nonlinear transformation of a monolayer neural network. Equation (5) represents the forward propagation process between the residual blocks in the standard residual network. The neural network of each layer fits the residual term, while in Equation (6), the output of the neural network is regarded as the gradient of the hidden state. Then the hidden state value of t can be obtained at any time by solving the equation integrally. The

number of evaluation points can be considered as equivalent to the number of model layers of the discrete neural network. In this paper, basic applications of ODE network on various mainstream model structures are proposed. The convolutional neural network model and the implicit state model based on time span are referred.

The feature extraction of video pedestrian mainly includes two aspects. On the one hand, it is the static feature extraction of video frame images in regular space, including pedestrian edge, color and other features. In this respect, the mainstream neural network has been able to obtain a high recognition rate. Experiments show that the static feature extraction of pedestrians does not need too deep network scale. On the other hand, it is also one of the difficulties of video pedestrian detection, which is the spatiotemporal dynamic characteristics of pedestrian in time span. Many scholars have proposed different methods to extract the dynamic characteristics of pedestrians. However, none of the current methods take into account the continuous information lost between discrete video frames. From the perspective of continuous events, this paper attempts to fit the probability distribution of the hidden dynamic characteristics of person $Z(t)$ through the hidden state model of ODE network, as shown in Figure 1.

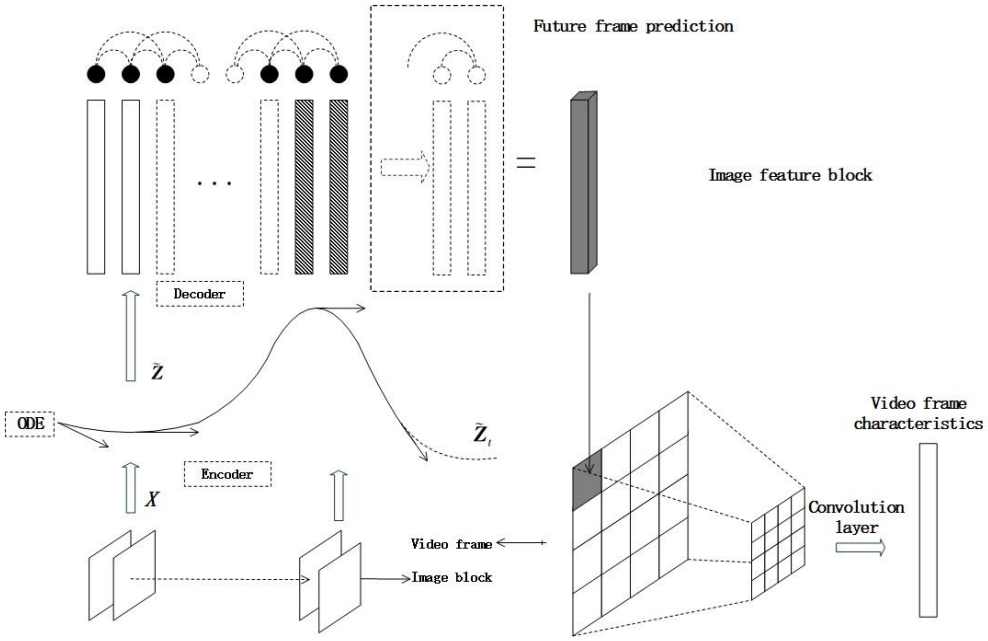

Figure 1 Feature extraction of continuous hidden state model

Firstly, the static feature vector X of a single frame is extracted by common convolutional network, such as residual network. If the feature of the image block is extracted, an additional layer of convolutional network is added to predict the

category of the complete image of the frame. The feature vector of the complete image can be obtained from the sequentially arranged feature of the image block, and the pooling layer can also be used. Secondly, the static feature X is sampled in reverse chronological order, and the predicted initial hidden state $\hat{Z}^{\prime}$ is obtained through the timing network (cyclic neural network is used in this paper). The hidden state probability distribution P($\hat{Z}^{\prime}$) is obtained from the ODE network, and the hidden state value $\hat{Z}_t^{\prime}$ can be predicted at any time. Finally, the implicit state value is converted to target feature vector X^* by the decoder.

2.3. The video person reidentification based on ODE and GCN

The video frames firstly are considered on the time span of contact with $G_k = (\upsilon_k, \varepsilon_k)$ graph model, the window size is $2k+1$, at the current moment as the center. With the current moment as the center, each video frame has k entry and exit edges and a self-ring edge, a total of $2k+1$. And the undirected graph is used for the reason of considering the relevance of before and after the event simutaneously. Each layer of graph convolutional network can contain ***n*** such windows, depending on the size of the network and usually determined by the length of the video block. Thus, the state update equation of each node in the middle layer of the graph convolution network can be expressed as

$$X_t^{l+1} = \sigma\Big(\sum_{t_i=t-k}^{t+k} \frac{1}{\tilde{Z}} X_{t_i}^{l} W_{\mathrm{t}i}^{l} \Big)\ k = 1,2,\Lambda \tag{7}$$

Where, $\tilde{Z}$ is the normalization factor, the same as Equation (7). l represents the number of layers in the graph convolution network, and σ is the activation function.

Video detection still boils down to classification. In the classification task, there are mature full supervision algorithms. However, a large amount of high-quality data is required, while the acquisition of high-quality video in real scenes is a difficulty, and relevant real data is lacking at this stage. The transfer learning for small data sets and the weak supervision algorithm with low requirements for the tag quality of data samples show outstanding advantages.

Input: Graph model function $G(\upsilon, \varepsilon)$, graph function parameter W, video block window size k, initialize the graph model adjacency matrix A_0, the input feature of the node in the figure X_0, the margins of positive and negative samples in triplet

losses were margin.

Output: X^* can distinguish target characteristics, characteristics of person sample space center $\boldsymbol{C_p}$, the non-person sample $\boldsymbol{C_n}$ feature space center;

1) initialize, $A = A_0, C_P = 0, C_n = 0$

2) randomly sampling Triplet $\hat{X}_i = Triplet(X_a, X_p, X_n)$ from input feature sample X_0, where X_a is the anchor point, X_p is the same random sample point as the anchor point category, and X_n is the opposite of the anchor point category;

3) repeat
4) forward transmission:
5) for X in Triplet $Triplet(X_a, X_p, X_n)$ do:
6) for Gall layers do:
7) Generate its diagonal node degree matrix $\boldsymbol{D}$ from $\boldsymbol{A}$;
8) to calculate the normalized coefficient $\frac{1}{Z} = \tilde{D}^{-\frac{1}{2}}\hat{A}\tilde{D}^{-\frac{1}{2}}$;
9) for X_t in X do:
10) node status update $X_t = \sigma\left(\sum_{X_{ti} \in neighbor^k X_t} \frac{1}{\tilde{Z}} X_{ti} W_{ti}\right)$
11) end for
12) update the adjacency matrix A from the new graph node state;
13) end for
14) return X, update triplet set $\hat{X}$
15) end for
16) return $\hat{X}$ as X^*
17) back propagation:
18) for $Triplet(X_a, X_p, X_n)$ in X^* do
19) computing triple loss

$$L = \max\left(\left\|X_a - X_p\right\|_2^2 - \left\|X_a - X_n\right\|_2^2 + m\arg in, 0\right)$$

20) end for

21) calculate the average loss $\overline{L}$;

22) back propagation $\overline{L}$ gradient using Adam algorithm;

23) until convergence or reach the maximum number of training

3. Experimental

The algorithm experiment was conducted in two video person data sets PRID2011 and The iLIDS-VID dataset [10]. PRID2011 data set contains two static camera collection of video. Camera A recorded video information of 385 persons, and camera B recorded video information of 749 persons, among which 20 were collected by camera A and B at the same time. The video sub-set of each person contained video frames ranging from 5 to 675. In order to ensure the validity of the spatiotemporal features, 178 video frames of person video subsets were selected. All the videos in this data set were shot in an outdoor environment with less occlusion and no crowding. Each person had abundant walking posture images. Some examples of person video frames are shown in Figure 2.

Figure 2 Example of PRID2011

The iLIDS-VID dataset consists of 300 different pedestrians viewed through two disconnected cameras in a public open space. The data set was created from two non-overlapping camera views of pedestrians observed from the i-LIDS Multi-camera Tracking Scene (MCTS) data set, captured under a multi-camera closed-circuit television network in an airport arrival hall. It consists of 600 sequences of images from 300 different individuals, each with a pair of sequences from two camera views.

Each image sequence has a variable length, from 23 to 192 frames, with an average of 73 frames. Some examples of person video frames are shown in Figure 3.

Figure 3 Example of iLIDS-VID

The experiment is based on pytorch deep learning framework. The hardware configuration is 32GB memory, Intel(R) Core(TM) I7-4790K processor and NVIDIA GTX1080 8GB graphics card. For each test, the training set and test set were randomly generated for each experiment, and 10 experiments were repeated under the same conditions. The average value of the results of the 10 experiments was taken as the final result of this test. Experimental results evaluate the performance of the algorithm by recognition rate.

On the video data set, this paper trains every 5 frames, and tests all the data. Meanwhile, the detection rate and false alarm rate of the hidden state model on the untrained frames are also tested. The total images were taken for training, and all frames were tested. A small batch of 128 frames was used for training, the number of training iterations was 2000, the initial learning rate was 0.000 1, and the gradient descent method was Adam. The hidden dimension of the ODE network convolution model is 64, and the specific structure is shown in Table 1. In this paper, group normalization is used for all normalized layers, and the maximum number of groups is 32. In the classification training, cross entropy loss and triplet loss were used for full and weak supervision training, and the resulting models were CGN_XE and CGN_WK respectively. The encoder of the implicit state model uses long and short

term memory network (LSTM). The decoder is the full connection layer, the model's hidden layer depth is 128, and the window value K is 2. The implicit state sampling uses monte Carlo method, and the sampling points of each video segment is 100+50, which are respectively the detection of the current period and the prediction of the future period.

Table 1 Gradient estimation model architecture

The module	Gradient estimation model
Subsampling module (optional)	[Normalized layer, 3×3 convolutional layer, ReLU layer]×1 [Normalized layer, 4×4 convolutional layer (step size 2), ReLU layer]×2
The ODE module	[Normalized layer, ReLU layer, 3×3 convolutional layer]×2 [Normalized layer]×1
Convolution model Tacit module	[Linear transformation layer, ReLU layer]×3
Fully connected module	[Normalized layer, ReLU layer, global maximum pooling layer, linear transformation layer]

Note: The unmarked convolutional layer step size defaults to 1. The number after the multiplication sign in the brackets represents the number of repetitions of the submodule in the brackets.

This article uses cross-validation on image data sets. The batch number is 32, the hidden dimension of the model is 32, the number of cycles is 100, and the initial learning rate is 0.001. Adam method is used for back propagation gradient. Meanwhile, cross entropy and triples loss are also used for the classification training for full supervision and weak supervision respectively, and the best performance of each comparison index is obtained.

4. Result and discussion

In the experiment, 300 persons sequencerandomly selected constitutes a training set, while the remaining 300 persons test set. The obtained experimental results are compared with other typical algorithms, the Dynamic RNN-CNN network [11], the accumulative motion context (AMOC) network [12], the algorithm using shared

attention of the matrix [13], and the ReRank application method based on shared attention of the matrix [9] . The results of person reidentification rates are shown in table 2.

Table 2 Person reidentification rates of different methods of PRID2011

Algorithm	Rank1	Rank5	Rank10	Rank20
Paper [11]	58.0	84.0	91.0	96.0
Paper [12]	68.7	94.3	98.3	99.3
Paper [13]	62.0	86.0	94.0	98.0
Paper[9]	76.0	94.0	97.0	98.0
Proposed	**80.6**	**95.5**	**98.4**	**99.4**

According to the data in Table 2, the recognition rate of proposed algorithm is significantly improved compared with the existing algorithms. Among them, the Rank1 reaches 80.6%, and has been improved 4.4% compared with the method proposed by paper [9]. In Rank5 and Rank20, there were some improvements compared with other algorithms.

Table 3 Person reidentification rates of different methods of iLIDS-VID

Algorithm	Rank1	Rank5	Rank10	Rank20
Paper [11]	70.1	89.6	96.2	96.3
Paper [12]	83.7	95.7	98.4	97.3
Paper [13]	78.1	95.0	98.0	99.1
Paper[9]	83.0	96.2	98.2	98.0
Proposed	**87.5**	**97.6**	**99.2**	**99.8**

From the experimental data of ilDS-VID data set in Table 3, it can be seen that the proposed method has a higher recognition rate compared with the existing mainstream methods. Experiments further verify the effectiveness of the proposed algorithm.

5. Conclusion

Person reidentification is a topic of great research value in the field of computer vision. In order to improve the performance of video person reidentification, a video person reidentification algorithm combining the ordinary differential equation (ODE) and graph convolution network is proposed in this paper. First, the ODE tacit model fitting person hidden in video distribution is used to replenish lost information between frames. Then, learning by figure convolution network connection between video frames for continuous and interval, the unstructured relationship between the characteristics is built, so that the positive and negative samples can be divided. Finally, classification results are obtained by choosing to use full connection layer or direct calculation characteristics with positive and negative samples center distance. The experimental results show that the proposed method can significantly improve the performance of video person reidentification, which is of great significance for the research of video person reidentification.

ABBREVIATIONS

ODE: ordinary differential equation
GCN: graphic convolutional network
MCTS: multi-camera tracking scene
LSTM: long and short term memory network
ReLU: rectified linear unit
AMOC: accumulative motion context

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable

CONSENT FOR PUBLICATION

Not applicable

AVAILABILITY OF DATA AND MATERIAL

The labeled dataset used to support the findings of this study are available from the corresponding author upon request.

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

Scientific research projects of Civil Aviation Air Flight University of China , Research on accurate track prediction and real-time collision detection technology （J2010-33 ）;Special Research Project on University Integrity in Henan Province in 2020 （ 2020-LZZD03 ） ： A Study on the Evaluation of Financial Disclosure Transparency in Henan Province —— Based on the Analysis of Financial Disclosure Data of 38 Provincial Administrative Colleges.

AUTHORS' CONTRIBUTIONS

Li-qiang ZHANG, as the primary contributor, completed the analysis, experiments and paper writing.

Long-yang HUANG and Xiao-li DUAN helped perform the analysis with constructive discussions.

ACKNOWLEDGEMENTS

Not applicable

Reference

[1] Wells J. " This Video Call May Be Monitored and Recorded": Video Visitation as a Form of Surveillance Technology and Its Effect on Incarcerated Motherhood[J]. Screen Bodies, 2019, 4(2): 76-92.

[2] Cai Z, Li D, Deng L, et al. Smart city framework based on intelligent sensor network and visual surveillance[J]. Concurrency and Computation: Practice and Experience, 2019: e5301.

[3] Karanam C R, Korany B, Mostofi Y. Tracking from one side: multi-person passive tracking with WiFi magnitude measurements[C]//Proceedings of the 18th International Conference on Information Processing in Sensor Networks. 2019: 181-192.

[4] Yu H X, Zheng W S, Wu A, et al. Unsupervised person re-identification by soft multilabel learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 2148-2157.

[5]Chen Y C , Zhu X , Zheng W S , et al. Person Re-Identification by Camera Correlation Aware Feature Augmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018, 40(2):392-408.

[6] Fu Y, Wei Y, Zhou Y, et al. Horizontal pyramid matching for person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 8295-8302.

[7] Lin Y, Zheng L, Zheng Z, et al. Improving person re-identification by attribute and identity learning[J]. Pattern Recognition, 2019, 95: 151-161.

[8] Kang J K, Hoang T M, Park K R. Person re-identification between visible and thermal camera images based on deep residual CNN using single input[J]. IEEE Access, 2019, 7: 57972-57984.
[9] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[EB/OL].[2019-05-07]. https://arxiv.ovg/pdf/1609.02907.pdf.
[10] Wang Z, He L, Gao X, et al. Multi-scale spatial-temporal network for person re-identification[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 2052-2056.
[11] Cho Y. Dynamic RNN-CNN based Malware Classifier for Deep Learning Algorithm[C]//2019 29th International Telecommunication Networks and Applications Conference (ITNAC). IEEE, 2019: 1-6.
[12] Liu H, Jie Z, Jayashree K, et al. Video-based person re-identification with accumulative motion context[J]. IEEE transactions on circuits and systems for video technology, 2017, 28(10): 2788-2802.
[13] Khatun A , Denman S , Sridharan S , et al. A Deep Four-Stream Siamese Convolutional Neural Network with Joint Verification and Identification Loss for Person Re-Detection[J]. 2018.

About the Author

ZHANG Li-qiang received the B.S. degrees at the School of Computer Science and Technology, Nanjing University of Technology,Nanjing, China, in 1999, His research interests include image/video processing, computer vision, and super-resolution.

Huang Longyang received the B.S. degree at the School of Electronic Engineering, Changchun University Of Science and Technology,Changchun, China, in 1996, and the M.S. degree in Telecommunications Engineering from University of Electronic Science and Technology of China, Chengdu, China, in 2004, and the Ph.D. degree in Circuits & Systems from Beijing University of Posts and Telecommunications, Beijing, China,in 2009. he's an associate professor at the Civil Aviation Flight University of China. His research interests include signal processing, aeronautical telecommunications, and air traffic management.

Duan Xiaoli received the B.S. degree at the School of Electronic Engineering, Zhongyuan University Of Technology,Zhengzhou, China, in 2002, and the M.S. degree in Electronic Engineering from Xiamen University, Xiamen, China, in 2006, and the Ph.D. degree in Electronic Engineering from Sichuan University, Chengdu, China,in 2019. She's an professor at the Zhengzhou Normal University. Her research interests include signal processing.

Correspondence
ZHANG Li-qiang
Electrical engineering department
Zhengzhou Technical College,
Zhengzhou City, HeNan Province, China

Postcode：450121
Phone: 15038310615

Figures

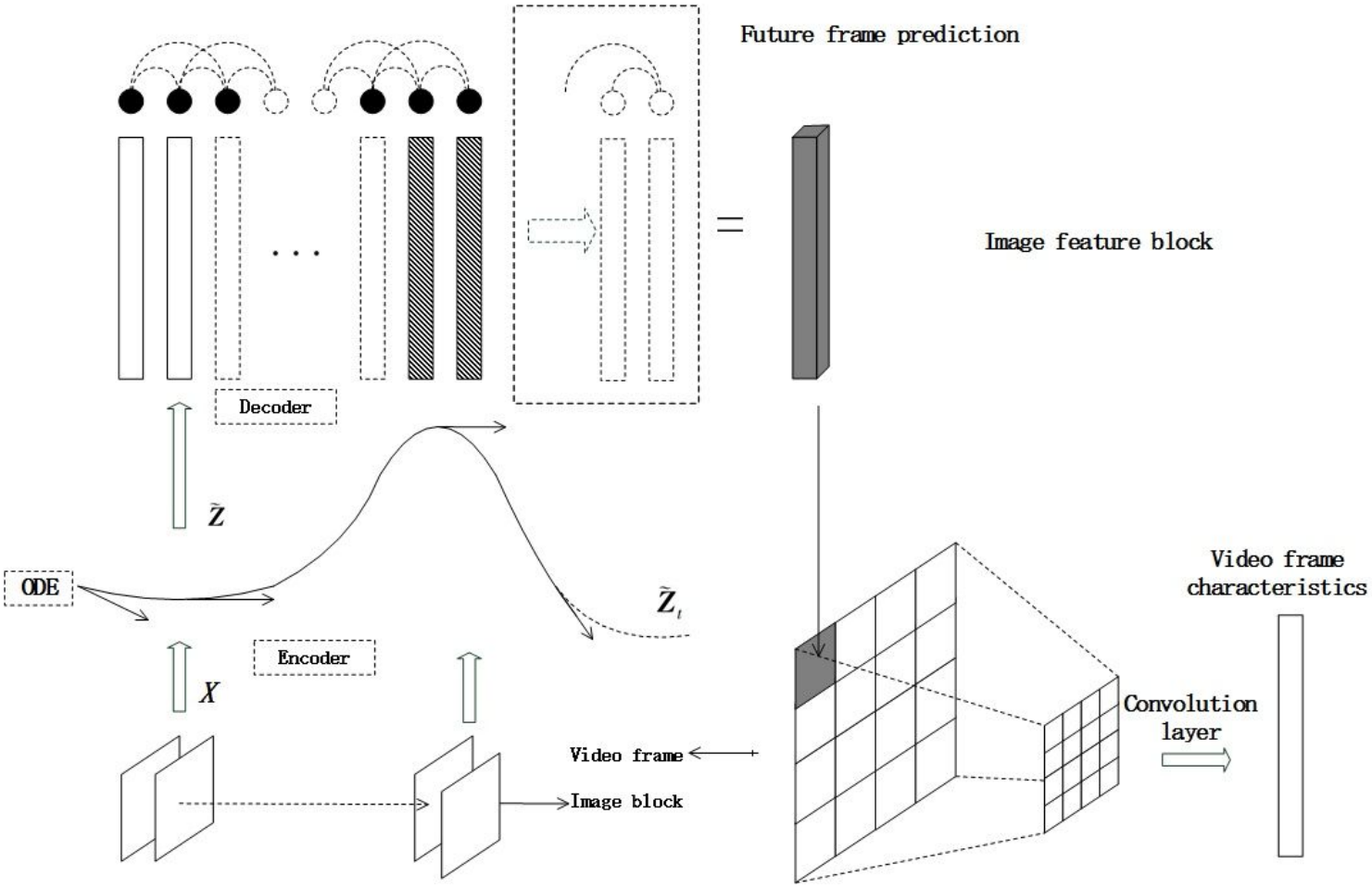

Figure 1

Feature extraction of continuous hidden state model

Figure 2

Example of PRID2011

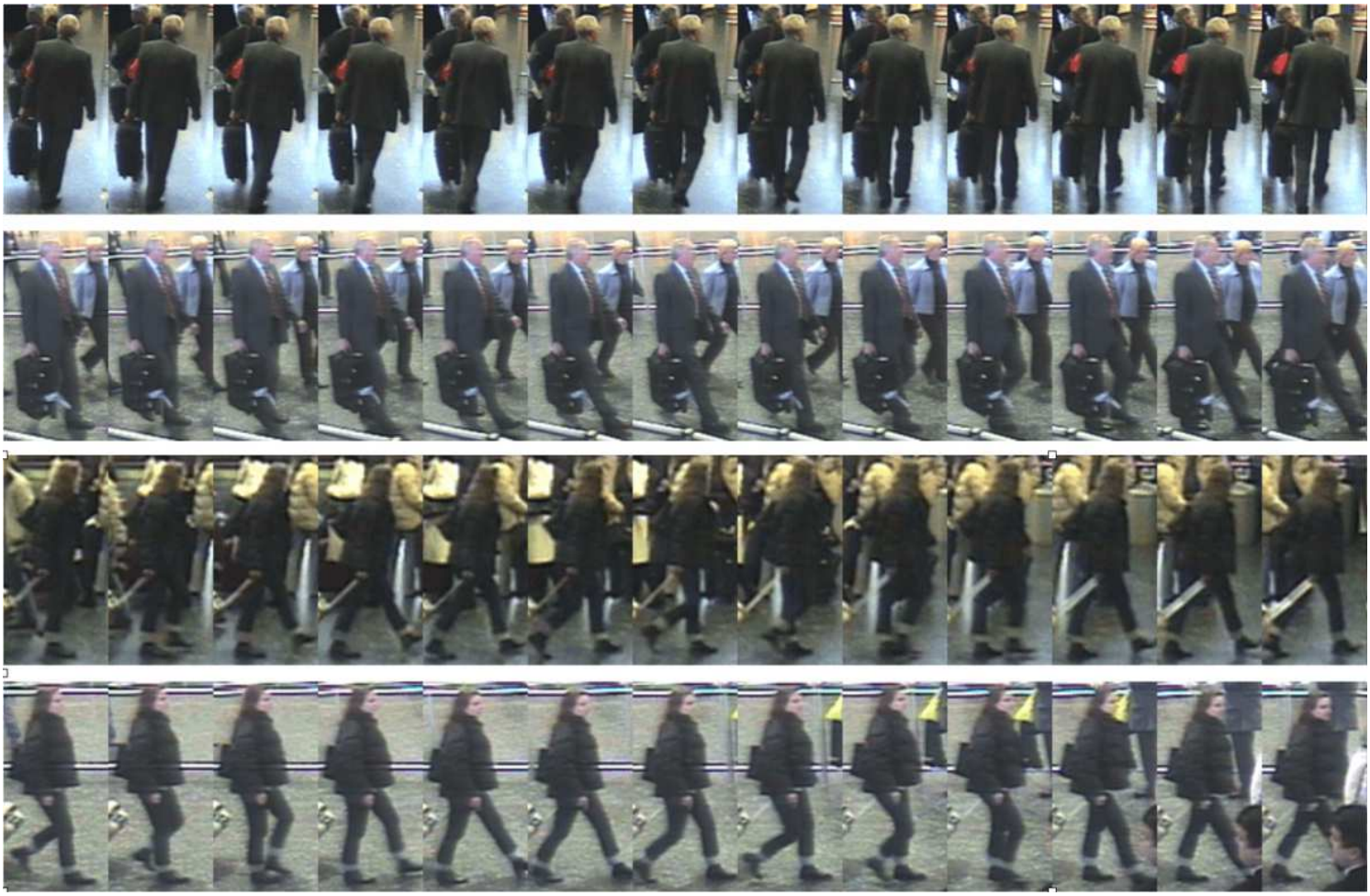

Figure 3

Example of iLIDS-VID