Behavior Recognition Based on Complex Linear Dynamic Systems

Yun Liu[1], Haifeng Sun[2], Chuanxu Wang[3], Shujun Zhang[4]
{Lyun-1027@163.com[1], 398149427@qq.com[2], wangchuanxu_qd@163.com[3]}

School of information science and technology, Qingdao University of Science & Technology[1,2,3,4]

Abstract. Time dynamics is a very important part of human behavior recognition. The linear dynamic system can model the time dynamics, but in the traditional linear dynamic system, the transfer matrix and the output matrix are subject to permutations, rotations, and linear combinations. Therefore, each row in the output matrix can not uniquely identify the characteristics of the corresponding system. In this paper, we propose complex linear dynamic systems to extract the "invariant" features of each time series. Firstly, describing the original video using motion boundary histogram (MBH). Then, we propose to model the motion dynamics with complex linear dynamical systems (CLDS) and use the model parameters as motion descriptors. Finally, the KNN classifier is used to classify it. Experiments with the KTH and UCF sports database show that our method is more accurate than the traditional linear dynamic system.

Keywords: behavior recognition, timing modeling, linear dynamic system, complex linear dynamic system

1 Introduction

In recent years, human behavior recognition has become a hot topic in the field of computer vision. It has a wide range of applications in video surveillance, human-computer interaction and virtual reality. The surveys by Turaga *et al.* [1] and Poppe [2] provide an extensive overview of video analysis of human motion sequences. In the case of motion sequence representations, the previous work can be roughly classified into appearance-based methods and motion-based methods.

Appearanced-based method, various local [3-6] or global [7-10] visual features are usually extracted from the original video data to represent the motion sequence. For example, Niebles *et al.* [6] use a bag-of-words model by extracting and clustering local spatio-temporal interest points to represent human behavior. The main problem in these approaches is that they discard information about the time inherent to behavior and fail to capture the temporal dynamics of human activity.

Motion-based methods often model motion sequences using temporal state-space models [11-13] and consider human behavior recognition as a temporal classification problem. The linear dynamic systems (LDS) are often used to model motion sequences to capture motion dynamics. For example, Ding *et al.* [14] learning linear dynamic systems with high-order tensor data with skeleton to recognize behavior. Luo *et al.* [15] combine LDS and cuboids for human action recognition in a maximum margin distance learning framework. We find that in traditional linear dynamic systems, the transfer matrix is not unique, each row in the output

MOBIMEDIA 2018, June 21-22, Qingdao, People's Republic of China

DOI 10.4108/eai.21-6-2018.2276576

matrix can not uniquely identify the characteristics of the corresponding system. The same set of observation sequences can produce a completely different transformation matrix, so it is difficult to explain.

To solve the above problem, we employ a complex linear dynamic system to extract the "invariant" features of each time series. Firstly, we encode each frame of the motion sequence using the Motion Boundary Histogram (MBH). Secondly, we model the complex linear dynamic system on the MBH sequence to describe the global dynamics. Thirdly, calculating the the pair-wise distance between them. Then, any off-the-shelf classifier (such as KNN or SVM) can be used to classify these sequences. In this way, we expect to capture the global temporal dynamics of the motion sequence to improve the accuracy of behavior recognition. We proved that we can extract the "invariant" feature by complex LDS, it is more suitable for classification, so the recognition rate better than the traditional methods.

2 Related work

Time-dependent state-space methods such as HMM, Conditional Random Field (CRF) or dynamic systems are often used to model motion sequences to capture motion dynamics. Brand *et al.* [12] proposed a coupled HMM to represent the interaction between the subjects. Caillette *et al.* [16] used a variable length Markov model (VLMM) to describe the observations and 3D poses for each action. Hongeng and Nevatia [17] incorporate domain knowledge as a priori probability of state duration into the HMM framework, using hidden semi-HMMs for event detection. HMM is very effective for modeling time series data. However, its application is limited due to the assumptions of conditional independent observations and hidden state sequences of Markov properties. CRF, on the other hand, avoids both these assumptions and allows non-local dependencies between states and observations. Sminchisescu *et al.* [13] used CRFs for human motion recognition. They show that CRF is superior to HMM and the Maximum Entropy Markov Model (MEMM) when longer observation lengths are considered. Vail *et al.* [18] compared CRFs and HMM in detail and concluded that CRFs perform as well as HMMs or better than HMMs. However, although HMMs and CRFs model motion sequences as a time-varying sequence, they do not explicitly model motion dynamics.

The dynamic system methods captures temporal changes by decoupling action sequences into subspace pose and potential dynamics. Bregler [19] proposed a multi-level framework for learning and recognizing human dynamics. LDS is used to describe mid-level simple movements, while HMMs are learned to represent advanced and complex behaviors. Black *et al.* [20] used a mixed auto-regressive process to represent multi-class motion sequences. Model parameters are learned by combining maximum expectation (EM) and condensation algorithms. Pavlovic and Rehg [21] used switching LDS to simulate nonlinear dynamics in human motion, while model learning and inference are based on variational techniques. Turaga *et al.* [22] model the motion sequence as a concatenation of LDSs. They split the sequence simultaneously in the time dimension and learn the LDS for each segment. Wang *et al.* [23] used Gaussian process dynamics to explore the non-linearity of motion sequences. The model parameters are marginalized rather than estimated. This leads to a dynamic system that is a non-parametric model. Although dynamic system methods are very effective in describing the dynamics of motion sequences, they often require detailed statistical modeling and parametric learning. In addition, the exact reasoning is usually difficult to deal with, need to develop approximation method.

Recent work reported in system identification literature has made it easy to compare dynamic systems by directly defining the distance or kernel metrics in the model space. Martin [24] defines a metric of stable ARMA models based on comparison of their cepstral coefficients. De Cock and De Moor [25] extend this concept and propose a more stable ARMA model by using the subspace angle between the two systems. Chan and Vasconcelos [26] derive a probability kernel based on Kullback-Leibler divergence and use it for dynamic texture classification. Vishwanathan *et al.* [27] proposed a general similarity metric of dynamic scene analysis based on Binet-Cauchy theorem. Since most of the work is designed for dynamic textures, few attempts have been made to use it for human motion recognition. Bissacco *et al.* [28] extended the work in [27] and defined a new kernel-based LDS metric for human gait recognition. Chaudhry *et al.* [9] used histogram of directed optical flow (HOOF) to encode each frame of the motion sequence and used Binet-Cauchy kernel [27] to describe the HOOF sequence. In this paper, we consider that most of the previous work on learning dynamic systems neglected the invariance of features. To solve these problems, we introduced a complex LDS learning algorithm to extract the invariant dynamic features of the video sequence. We experiment to show a satisfactory recognition result on the dataset.

3 Recognition with complex LDS

LDS and its extensions [19-22] have long been the subject of human motion analysis. They show superiority to classification tasks over common HMMs, but usually require complex Bayesian modeling and reasoning. Learning LDS parameters [29], [30] in system identification theory and similarity measurements between LDS [24-27] have made LDS successful for classify high-dimensional time series data in the field of dynamic textures [31], [32]. This gives us a new way to simulate and compare the dynamics of a sequence of actions. By modeling the temporal variation with LDS, the system's theoretical methods specifically consider the global dynamics of the action sequence. The similarity between two LDSs is directly measured using the distance or kernel metric defined in LDS space. Therefore, we can use LDS to capture the dynamic information of human activities, and classify action sequences by these similarity metrics.

3.1 Linear dynamic system

Let $A \in \mathbb{R}^{n \times n}$ denote the system transition matrix, $C \in \mathbb{R}^{p \times n}$ denote the subspace mapping matrix. Here p and n are the dimensions of observation space and state space, respectively. Then, a stationary LDS can be represented by the parameter tuple $M = (A, C)$, and according to the following equation

$$\begin{cases} x_{t+1} = Ax_t + v_t \\ y_t = Cx_t + w_t \end{cases} \tag{1}$$

where $x_t \in \mathbb{R}^n$ is the state variable or a latent variable, $y_t \in \mathbb{R}^p$ is the observed random variable or feature, v_t and w_t are the system noise and the observation noise, respectively. If we assume that the noises are zero-mean i.i.d Gaussian processes, we have $v_t \sim N(0,Q)$ and $w_t \sim N(0,Q)$. Here Q and R are covariant matrices of multivariate Gaussian distributions.

In equation (1), the hidden state is modeled as a first-order Gaussian-Markov process, where x_{t+1} is determined by the previous state x_t. The output y_t depends on the current state x_t. Given a video sequence $y_{1:\tau}$, learning its intrinsic dynamics is equivalent to identifying the model parameter M. This is a typical system identification problem and is generally solved by using least squares estimation.

Let the column matrix $Y_{1:\tau} = \left[y_1, y_2, \cdots, y_\tau\right]$ and $X_{1:\tau} = \left[x_1, x_2, \cdots, x_\tau\right]$ represent the observation sequence and the state sequence, respectively. In order to obtain the closed-form estimate of the model parameter M, we first decompose the observation matrix with the singular value decomposition (SVD), $Y_{1:\tau} = U \sum V^T$. Where U ,V are orthogonal and $\sum$ is a rectangular diagonal matrix of positive non-negative real numbers on the diagonal. To get the subspace mapping matrix and underlying state sequence estimates set by

$$\hat{C} = U, \hat{X}_{1:\tau} = \sum V^T \tag{2}$$

the model dimension n is determined by preserving singular values that exceed a given threshold.

Then the least-squares estimate of A is

$$\hat{A} = \arg\min_A \left\| A\hat{X}_{1:\tau-1} - \hat{X}_{2:\tau} \right\|_F^2 = \hat{X}_{2:\tau} \hat{X}_{1:\tau-1}^+ \tag{3}$$

where $\|\cdot\|_F$ denotes the F-norm and $^+$ denotes the Moore-Penrose inverse. Given the above estimates of $\hat{A}$ and $\hat{C}$, the covariance matrix $\hat{Q}$ and $\hat{R}$ can be estimated directly from the residuals.

According to equation (2), LDS implicitly simulates the observation sequence $Y_{1:\tau}$ using the subspace mapping matrix C and its corresponding coefficient $X_{1:\tau}$. In the task of human behavior recognition, the subspace matrix C describes the action components, while the matrix A is derived from $X_{1:\tau}$ and represents the motion dynamics. Therefore, we can use $M = (A, C)$ to represent the motion sequence descriptor. Such a descriptor captures the dynamic and embedded components of a motion sequence and is very different from a local spatio-temporal gradient descriptor.

However, there is a problem with using M as a descriptor. In traditional LDS models, transformation matrix A is not unique: it is affected by permutations, rotations, and linear combinations, as is output matrix C. Therefore, each line in C can not uniquely identify the characteristics of the corresponding system. Therefore, we need to extract "invariant" features for each time series.

3.2 Complex linear dynamic system

Feature invariance is a very important attribute of LDS, and has been studied deeply in the system identification literature. Previous LDS may not generate features that correspond to the original data. Complex linear dynamic system noise variables follow the complex Gaussian distribution, and an important property of the complex Gaussian distribution is "rotation invariance." Therefore, we can be use it to obtain the "invariant" feature of the corresponding sequence. We show our model in this section and generating features that correspond to the original data is crucial for the classification.

The complex linear dynamic system (CLDS) is defined by the following equation.

$$\begin{cases} z_1 = u_0 + w_1 \\ z_{n+1} = A \cdot z_n + w_{n+1} \\ x_n = C \cdot z_n + v_n \end{cases} \tag{4}$$

the noise vector follows a complex normal distribution, $w_1 \sim CN(0, Q_0)$, $w_i \sim CN(0, Q)$, $v_j \sim CN(0, R)$. Note that unlike traditional linear dynamic systems, CLDS allows parameters to be complex values, with the constraint that Q_0, Q and R must be Hermitian positive definite matrices. Figure 1. shows the graphical model. It can be seen as a continuous linear Gaussian distribution over the hidden variable z's and the observed value x.

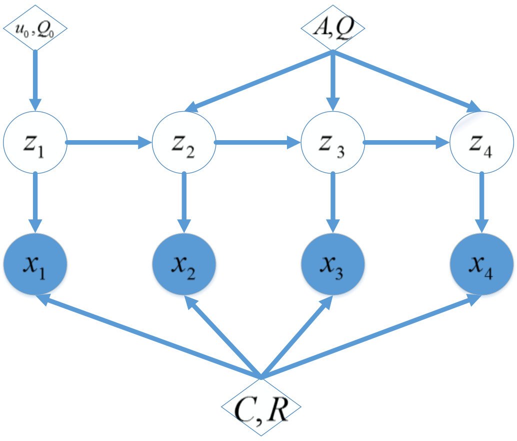

Fig. 1. Graphical Model for CLDS. x are real valued observations and z are complex hidden variables. Arrows denote linear Gaussian distributions.

The problem with learning is to estimate the best fit parameter $\theta = \{u_0, Q_0, A, Q, C, R\}$, passing a given observation sequence $x_1 \cdots x_N$. We use Complex-Fit, a novel complex valued expectation-maximization algorithm towards a maximum likelihood fitting.

The expected negative-loglikelihood of the model is

$$\begin{aligned} L(\theta) &= \mathrm{E}_{z|x}\left[-\log P(X,Z\mid\theta)\right] \\ &= \log|Q_0| + \mathrm{E}\left[(z_1-u_0)^* Q_0^{-1}(z_1-u_0)\right] \\ &\quad + \mathrm{E}\left[\textstyle\sum_{n=1}^{N-1}(z-A\cdot z_n)^*\cdot Q^{-1}\cdot(z_{n+1}-A\cdot z_n)\right] \\ &\quad + \mathrm{E}\left[\textstyle\sum_{n=1}^{N}(x_n-C\cdot z_n)\right]+(N-1)\log|Q|+N\log|R| \end{aligned} \tag{5}$$

where the expectation $\mathrm{E}[\ \ ]$ is over the posterior distribution of Z given X.

Unlike the traditional linear dynamic system, the objective here is a complex valued function that requires nonstandard optimization in complex domain. In negative-loglikelihood, there are two sets of unknowns, parameters and posterior distributions. We will briefly describe the Complex-Fit here. The M-step is derived by taking the complex derivatives of the objective function and equating them to zero. It update the parameters to optimize $L(\theta)$. During the E-step, we will compute the mean and covariance of the edge and joint posterior distribution $P(z_n \mid X)$ and $P(z_n, z_{n+1} \mid X)$. The E-step calculates the posterior distribution using a forward-backward sub steps (corresponding to Kalman filtering and smoothing in the traditional LDS). The overall idea of the Complex-Fit algorithm is to optimize the parameter set θ, just as we know the posterior distribution, and then estimate the posterior distribution with the current parameters. It then takes turns to obtain the optimal solution.

Once we have used Complex-Fit (with a diagonal transformation matrix) to best estimate these parameters, We can use the output matrix $M=(A,C)$ in the CLDS as a feature to represent the motion sequence descriptor and compute the distance for it, then classify it using any off-the-shelf classifier.

3.3 Distance Metric for complex LDS

Given an action sequence, we use the complex LDS model parameter $M=(A,C)$ as the motion sequence descriptor, with the dynamic matrix $A\in GL(n)$, where $GL(n)$ is the group of all $n\times n$ invertible matrices, and with the mapping matrix $C\in ST(p,n)$, where $ST(p,n)$ is the Stiefel manifold. Since the model space has a non-Euclidean structure and the descriptor is in non-vector form, this naturally raises the issue of how to measure the similarity between two descriptors. Martin [24] defines a metric for stable ARMA models based on a comparison of their cepstrum coefficients. De Cock and De Moor [25] improve Martin's work by using the subspace angles between two LDSs. The subspace angles are defined as the principal angles between the column spaces of infinite observability matrices $O_\infty(M_i)=\begin{bmatrix} C_i^T & (C_iA_i)^T & (C_iA_i^2)^T & \cdots \end{bmatrix}^T \in \mathbb{R}^{\infty\times n}$ for $i=1,2$.

Let $M_1=(A_1,C_1)$ and $M_2=(A_2,C_2)$ denote the two motion sequence descriptors. The computation of subspace angles is obtained by solving the Lyapunov equation

$$Q = A^TQA + C^TC \tag{6}$$

for Q, where $Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \in \mathbb{R}^{2n \times 2n}$, $A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} \in \mathbb{R}^{2n \times 2n}$, $C = \begin{pmatrix} C_1 & C_2 \end{pmatrix} \in \mathbb{R}^{p \times 2n}$.

The equation of (6) is guaranteed to exist when M_1 and M_2 are stable. The cosines of the subspace angles $\cos^2 \theta_i$ are calculated as eigenvalues of matrix $Q_{11}^{-1} Q_{12} Q_{22}^{-1} Q_{21}$, where $Q_{kl} = O_\infty (M_k)^T O_\infty (M_l)$ for $k, l = 1, 2$.

The subspace angles distance is defined as

$$d_{LDS} (M_1, M_2)^2 = -\log \prod_{i=1}^{n} \cos^2 \theta_i \tag{7}$$

When we get the distance of the LDS, we can use any off-the-shelf classifier (such as KNN) to classify. Because KNN is too simple, we propose set a certain weight, found that the effect will be more ideal.

4 Experiment

4.1 Database

In this paper, KTH and UCF sports behavioral data sets are used to test and verify the algorithm, using a leave-one-out verification method (LOOCV). The KTH dataset is the most widely recognized behavioral dataset in the field of behavioral recognition, including a total of 2391 video samples of 6 types of actions performed by 25 individuals in 4 different scenarios. UCF dataset The dataset consists of 150 video sequences and consists of 10 behaviors, as shown in Figure 2. This dataset has a wide range of perspectives and has been used extensively in many studies such as motion recognition, motion localization and saliency detection.

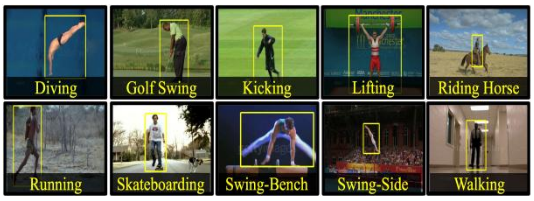

Fig. 2. UCF dataset sample.

4.2 Feature Extraction

To compute complex LDS, we extract sequential features from all the videos. Our proposed method can be used with different types of features, including raw pixels, provided that the features form a time series. Silhouettes or shape features [33] are useful, but they are difficult to obtain in unconstrained environments. In this paper, we use the motion boundary histograms [34] to characterize the action profile. The MBH encodes the relative motion between pixels by computing gradients of the x and y optical flow components separately. It suppresses most of the camera motion and background texture, and thus highlights the foreground subject. Some examples are illustrated in Figure 3.

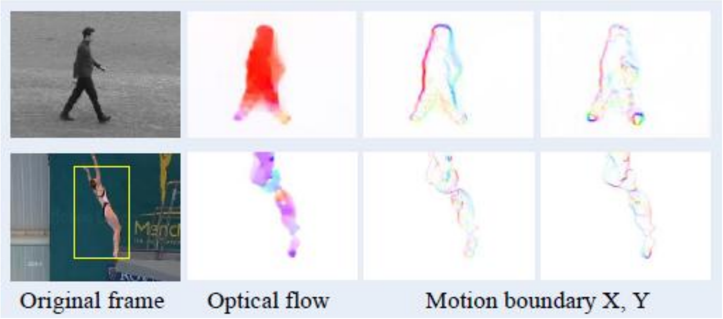

Fig. 3. Illustration of raw, optical flow and MBH (x, y) images of two action sequences from the KTH dataset (top) and the UCF sports dataset (bottom), respectively. For the optical flow and MBH images, gradient/flow orientation is indicated by color (hue) and magnitude by saturation.

As suggested in [34], we resize the sequences into 64×128 pixels. The MBH is computed by quantizing the orientations into 9 bins with 2×2 blocks of 8×8 pixel cells. To improve the performance, block overlap (0.5) is also incorporated. Thus we obtain a total of 7×15 blocks, where each block is described by a 4×9 histogram. The final histogram size is 3780 for both x and y components of MBH (i.e., $7\times15\times36$).

4.3 Complex LDS

Timing Modeling Using complex LDS When timing information is extracted, the state space dimensions of the model parameters and the number of iterations, our reference and experimental comparison, we find that the hidden state has a dimension of 6 and an iteration number of 100, will be better.

4.4 Classifier

In order to verify the validity of the algorithm, we test the recognition accuracy of the CLDS algorithm and T-LDS algorithm on KTH and UCF sports datasets. At the same time with several other timing algorithms Rb-LDS, CRF, SLDS and MEMM were compared. As can be seen from Table 1 and Table 2, the CLDS algorithm improves the recognition accuracy by 4% -5% over the traditional LDS algorithm on the KTH and UCF datasets. This shows that the extraction of "invariant" features in the processing of timing information using LDS plays an important role in video behavior analysis.

In the KTH dataset, the recognition accuracy of CLDS algorithm reaches 92.37%, reaching the highest in the same timing model. Because the jogging and running behaviors themselves have a lot of similarities, it is easy to produce confusion, the recognition accuracy is relatively low, and other types of behaviors can basically be accurately identified. For UCF sports datasets with multiplayer behavior, the highest recognition rate is 81.56%, which is also the highest in the same time series model, so the algorithm has good performance for single person behavior and multiplayer behavior.

Table 1. Comparison of six methods on the KTH dataset.

KTH	Box	Hand clap	Hand wave	Jog	Run	Walk	average
CLDS	1.0	0.90	1.0	0.84	0.87	0.89	0.9237
Rb-LDS	1.0	1.0	1.0	0.84	0.77	0.89	0.9167
T-LDS	1.0	0.86	0.97	0.82	0.73	0.87	0.8747
CRF	0.96	0.97	0.97	0.79	0.84	0.84	0.8950
SLDS	0.97	0.96	0.97	0.83	0.80	0.85	0.8955
MEMM	0.93	0.87	0.90	0.76	0.73	0.81	0.8136

Table 2. Comparison of six methods on the UCF dataset.

UCF sports	Dive	Golf	Kick	Lift	Ride	Run	Skete	Swing1	Swing2	Walk	Average
CLDS	0.94	0.88	0.85	0.83	0.68	0.63	0.75	0.84	0.86	0.82	0.8156
Rb-LDS	0.93	0.89	0.85	0.83	0.67	0.62	0.75	0.85	0.85	0.82	0.8133
T-LDS	0.88	0.83	0.82	0.80	0.65	0.60	0.73	0.81	0.83	0.80	0.7681
CRF	0.91	0.87	0.83	0.82	0.68	0.64	0.73	0.85	0.84	0.81	0.8098
SLDS	0.93	0.88	0.84	1.84	0.66	0.63	0.75	0.84	0.85	0.82	0.8130
MEMM	0.87	0.82	0.81	0.79	0.63	0.58	0.72	0.78	0.81	0.77	0.7484

5 Conclusion

In the field of human behavior recognition, capturing temporal information features in video video is a challenging issue. In this paper, we introduce a simple and efficient CLDS learning algorithm to describe the dynamics of motion sequences. CLDS noise variables follow the complex Gaussian distribution, and an important property of the complex Gaussian distribution is "rotation invariance." Therefore, we can be used to obtain the "invariant" feature of the corresponding sequence. This is crucial for our dynamic system model distance metrics. We conducted a wide range of experiments on two public data sets. We evaluated CLDS in the selection of model parameters and compared it with the traditional LDS as well as four temporal methods, namely Rb-LDS [15], MEMM [35], CRF [13], and switching LDS [21] conversion to quantify the improvement in recognition rate. We compare the current state of the art results and show the great potential of our approach. Besides, our method can also be used as a feature extraction tool in other applications such as signal compression.

Because related sequences often involve long-term changes, LDS can not describe embedded nonlinear dynamics alone. One possible approach is to develop a nonlinear dynamic system model, such as phase space reconstruction in a chaotic model. This is undoubtedly an interesting question that we will consider in our next step of work.

Acknowledgments. Our work is supported by China Nature Science Fund Project 61472196 and 61672305; also I would like to express our thanks to Shandong Nature Science Fund Project ZR2015FM012 and the Key Research & Development Plan Project of Shandong Province (No.2017GGX10127).

References

[1] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea.: Machine recognition of human activities: A survey. IEEE Trans. Circuits Syst. Video Technol., vol. 18, no. 11, pp. 1473–1488(2008)

[2] R. Poppe.: A survey on vision-based human action recognition. Image and Vision Computing, vol. 28, no. 6, pp. 976–990(2010)

[3] I. Laptev.: On space-time interest points. Int'l J. Computer Vision, vol. 64, no. 2-3, pp. 107–123, (2005)

[4] P. Doll ́ar, V. Rabaud, G. Cottrell, and S. Belongie.: Behavior recognition via sparse spatiotemporal features. in Proc. IEEE Int'l Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72(2005)

[5] P. Scovanner, S. Ali, and M. Shah.: A 3-dimensional sift descriptor and its application to action recognition. in Proc. Int'l Conf. Multimedia, pp. 357–360(2007)

[6] J. Niebles, H. Wang, and L. Fei-Fei.: Unsupervised learning of human action categories using spatial temporal words. Int'l J. Computer Vision, vol. 79, no. 3, pp. 299–318(2008)

[7] A. Efros, A. Berg, G. Mori, and J. Malik.: Recognizing action at a distance. in Proc. IEEE Int'l Conf. Computer Vision, pp. 726–733(2003)

[8] D. Tran and A. Sorokin.: Human activity recognition with metric learning. in Proc. European Conf. on Computer Vision, (2008)

[9] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal.: Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. in Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1932–1939(2009)

[10] H. Wang, A. Kl ̈aser, C. Schmid, and C. L. Liu.: Dense trajectories and motion boundary descriptors for action recognition. Int'l J. Computer Vision, vol. 103, no. 1, pp. 60–79(2013)

[11] J. Yamato, J. Ohya, and K. Ishii.: Recognizing human action in time-sequential images using hidden Markov model. in Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 379–385(1992)

[12] M. Brand, N. Oliver, and A. Pentland.: Coupled hidden Markov models for complex action recognition. in Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 994–999(1997)

[13] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas.: Conditional models for contextual human motion recognition. in Proc. IEEE Int'l Conf. Computer Vision, pp. 1808–1815(2005)

[14] W. Ding, and L. Kai.: Learning Linear Dynamical Systems with High-Order Tensor Data for Skeleton based Action Recognition. Computer Vision and Pattern Recognition, (2017)

[15] G. Luo, S. Yang, G. Tian, C. Yuan, W. Hu, and Maybank, S. J.: Learning Human Actions by Combining Global Dynamics and Local Appearance. in Proc. IEEE Trans Pattern Anal Mach Intell, pp. 2466–2482(2014)

[16] F. Caillette, A. Galata, and T. Howard.: Real-time 3-D human body tracking using learnt models of behaviour. Computer Vision and Image Understanding, vol. 109, no. 2, pp. 112–125(2008)

[17] S. Hongeng and R. Nevatia.: Large-scale event detection using semi-hidden Markov models. in Proc. IEEE Int'l Conf. Computer Vision, pp. 1455–1462(2003)

[18] D. L. Vail, M. M. Veloso, and J. D. Lafferty. Conditional random fields for activity recognition. in Proc. Int'l Conf. Autonomous Agents and Multi-agent Systems, (2007)

[19] C. Bregler.: Learning and recognizing human dynamics in video sequences. in Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 568–574(1997)

[20] A. Blake, B. North, and M. Isard.: Learning multi-class dynamics. in Proc. Ann. Conf. Neural Information Processing Systems, pp. 389–395(1999)

[21] V. Pavlovi ́c and J. M. Rehg.: Impact of dynamic model learning on classification of human motion. in Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 788–795(2000)

[22] P. K. Turaga, A. Veeraraghavan, and R. Chellappa.: From videos to verbs: Mining videos for activities using a cascade of dynamical systems. in Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1–8(2007)

[23] J. M. Wang, D. J. Fleet, and A. Hertzmann.: Gaussian process dynamical models for human motion. IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 2, pp. 283–298(2008)

[24] R. J. Martin.: A metric for ARMA processes. IEEE Trans. Signal Process., vol. 48, no. 4, pp. 1164–1170(2000)

[25] K. De Cock and B. De Moor.: Subspace angles between ARMA models. Systems and Control Letter, vol. 46, pp. 265–270(2002)

[26] A. B. Chan and N. Vasconcelos.: Probabilistic kernels for the classification of auto-regressive visual processes. in Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 846–851(2005)

[27] S. V. N. Vishwanathan, A. J. Smola, and R. Vidal.: Binet-Cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. Int'l J. Computer Vision, vol. 73, no. 1, pp. 95–119(2007)

[28] A. Bissacco, A. Chiuso, and S. Soatto.: Classification and recognition of dynamical models: The role of phase, independent components, kernels and optimal transport. IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 11, pp. 1958–1972(2007)

[29] P. Van Overschee and B. De Moor.: N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems. Automatica, vol. 30, no. 1, pp. 75–93(1994)

[30] Z. Ghahramani and G. E. Hinton.: Parameter estimation for linear dynamical systems. Dept. Computer Science, Univ. of Toronto, Technical Report CRG-TR-96-2(1996)

[31] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto.: Dynamic textures. Int'l J. Computer Vision, vol. 51, no. 2, pp. 91–109(2003)

[32] F. Woolfe and A. W. Fitzgibbon.: Shift-invariant dynamic texture recognition. in Proc. European Conf. on Computer Vision, pp. 549–562(2006)

[33] L. Wang and D. Suter.: Learning and matching of dynamic shape manifolds for human action recognition. IEEE Trans. Image Process., vol. 16, no. 6, pp. 1646–1661(2007)

[34] N. Dalal, B. Triggs, and C. Schmid.: Human detection using oriented histograms of flow and appearance. in Proc. European Conf. on Computer Vision, (2006)

[35] A. McCallum, D. Freitag, and F. Pereira.: Maximum entropy Markov models for information extraction and segmentation. in Proc. Int'l Conf. Machine Learning, pp. 591–598(2000)