21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

ARMA-HMM: A New Approach for Early Recognition of Human Activity

Kang Li and Yun Fu
Department of ECE and College of CIS, Northeastern University, Boston, MA
Department of CSE, State University of New York, Buffalo, NY
kangli@buffalo.edu

Abstract

Early Recognition of human activities is a highly desirable functionality for many visual intelligent systems. However, in computer vision, very few work have been devoted to this challenging and interesting task. In this paper, we address human activity early recognition as a pattern recognition problem of time series data. A new model called ARMA-HMM is introduced to integrate both the predictive power of sequential model HMM and time series model ARMA. We also present a novel feature called Histogram of Oriented Velocity (HOV) to encode activity video as a sequential observation of motion signals. Experiments on a daily activity dataset and a realistic YouTube sports dataset show promising results of the proposed method.

1. Introduction

In computer vision community, many studies have been dedicated to human behavior analysis, especially action recognition and event detection. In contrast, there is less attention paid to time-critical applications, such as early recognition of human activities, where early prediction of ongoing activity is extremely valuable. Variety of intelligent applications can benefit from early recognition of human activity, such as assistive systems, surveillance system, human-computer interaction systems, etc. For instance, in the battle field, intelligent unmanned ground vehicles or robots can provide real-time surveillance, detect suspicious activities, and raise alarms for emergencies or attacks before they happen. However, activity early recognition is harder than conventional activity recognition, as we need to attempt a reliable judgment with only observation of the initial phase of the action.

Our work is partly motivated by [8], which explicitly raised this problem to the computer vision community for the first time. They proposed an extension of bag-of-words paradigm, called dynamic histogram of spatio-temporal features (Dynamic BoW), to model how histogram distribution changes over time. However, in their formulation, the progress level of the activity A_p is indicated by the frame index d, which in fact assumes that activities within the same class always have the same speed and duration. Unfortunately, it is not the case in most cases for human activity.

In this paper, we propose a new approach for activity early recognition from the perspective of time series analysis, which can handle activities with variable time duration and speed. The major contributions of our work include: (1) we propose a novel frame-based global feature called Histogram of Oriented Velocity (HOV) to provide a continuous representation of human activity; (2) we present a early recognition model ARMA-HMM to achieve early recognition of human activity by integrating the predictive power of hidden Markov model (HMM) and autoregressive-moving-average model (ARMA).

1.1 Related work

To the best of our knowledge, [8, 3] are the only two attempts of activity early recognition in the computer vision literature. Ryoo [8] argues that the goal of activity early recognition is to recognize unfinished activity from observation of its early stage. Extensions of bag-of-words models and 3-D space-time local feature are used to formulate the problem. Hoai and Torre [3] proposes a maximum-margin approach for training temporal event detectors to recognize partial observed actions.

Though very few work have been done on this early recognition task directly in the computer vision field, plenty of research works from other fields of artificial intelligence motivated our research, particularly from the field of *intent/plan recognition*. Intent recognition refers to the task of inferring the plan or intentions of an intelligent agent from the agent's actions or the effects of those actions. By incorporating machine learning method, Bui et al [1] proposed a model of plan recognition based on a variant of HMMs which can automatically acquire plan models from sample data. Liao et

978-4-9906441-1-6

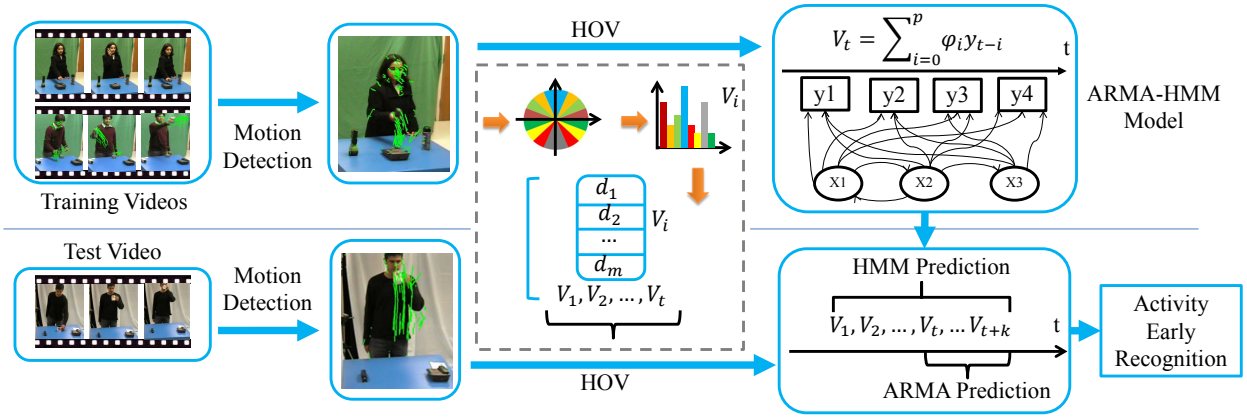

Figure 1. Overview of our proposed method.

al [4] used HMM as a plan recognition model in a cognitive assistive system, they track the location changes of subject based on GPS sensor data. The system can recognize different patterns to reveal the different intentions of activity, such as "*Shopping*" and "*Dining Out*".

2 Proposed approach

2.1 Activity representation

In this paper, we argue that even though bag-of-words model shown its effectiveness and robustness in many real world applications of human activity recognition, it is not an appropriate representation for the task of activity early recognition. The reason is that it ignores all the temporal or sequential information of activity. And in many cases, the initial observation of activity may not be sufficient to build a discriminative bag-of-words representation. We demonstrate in this paper that global motion features of activity (one feature vector per frame) and corresponding time series representation are good fit for early recognition problem. It provides us a convenient representation for modeling dynamics of human action and exploring predictability of time series signal.

In this section, we propose a novel feature called histogram of oriented velocity (HOV), which is inspired by the popular local motion feature histogram of oriented gradient (HOG) [7] descriptor and the successful use of velocity information in activity recognition [5]. As shown in the dashed box of Figure 1, there are following several steps to encode activity as a sequential observations of motion signals:

1. Detect salient key points by using Harris corner method;
2. Compute optical flow to obtain trajectories at each key point;
3. For each frame, accumulate the displacements of trajectories which are lying in the same orientation bins respectively.

Specifically, the HOV feature can be computed as follows:

$$V_t^{d_j} = \sum_{p(x_{i,t}, y_{i,t}) \in b_j} \sqrt{(x_{i,t} - x_{i,t-1})^2 + (y_{i,t} - y_{i,t-1})^2},$$

where V_t represents the global motion feature at frame f_t; $p_{i,t}$ is the ith interest point found in frame f_t; $\overrightarrow{p_{i,t-1}p_{i,t}}$ is its corresponding trajectory history; and b_j is the jth orientation bin.

2.2 Early recognition model

HMM and ARMA are two important models in time series analysis. And each of them has a unique power in terms of early recognition. HMM is good at predicting the *global* pattern of sequential observations of data, while ARMA is good at predicting future values in a *local* range of time series. Hence, we propose a hybrid model, called ARMA-HMM to draw predictive power from both of these two models.

An **HMM** of N hidden states and Gaussian emission distributions can be specified by the 4-tuple $M = \{\pi, A, \mu(X), \sigma(X)\}, 1 \leq X \leq N$, where X is the hidden state variable, π is the initial state probabilities, A is transition matrix of hidden states, and $\mu(\cdot), \sigma(\cdot)$ parameterize emission distributions for each state. The observation V_t at time t can be generated by

$$V_t = \mu(X_t) + \sigma(X_t)\varepsilon_t, \text{ where } \varepsilon_t \sim \mathcal{N}(0, 1). \quad (1)$$

And the likelihood of a multivariate time-series V is

$$p(V|M) = \sum_{S \in S^T} [P_{X_1} \mathcal{N}(V_1; \mu(X_1), \sigma(X_1)) \prod_{t=2}^{T} P_{X_t|X_{(t-1)}} \mathcal{N}(V_t; \mu(X_t), \sigma(X_t))], \quad (2)$$

where $S = \{X_1, X_2, ..., X_T\}$ is a hidden state sequence, and D_S is the set of all possible sequences.

An **ARMA** model is a tool for understanding and, perhaps, predicting future observations of time series V. The model consists of two parts, an autoregressive (AR) part and a moving average (MA) part:

$$V_t = \sum_{i=1}^{p} \varphi_i V_{t-i} + \sum_{i=1}^{q} \theta_i \eta_{t-i} + c + \eta_t, \qquad (3)$$

where $\varphi_1, ..., \varphi_p$ are the parameters for autoregressive part, $\theta_1, ..., \theta_q$ are the parameters for moving average part, c is a constant and η_t is white noise.

ARMA-HMM, as shown in Figure 1, is a three level generative model. It is essentially an ARMA-filtered HMM, which can be defined in the following form:

$$M^{(p,b,q)} = \{\pi, A, \mu(X), \sigma(X), \boldsymbol{\varphi}, \boldsymbol{\omega}, \boldsymbol{\theta}\}, \qquad (4)$$

with coefficients $\varphi = (\varphi_1, ..., \varphi_p)$, $\omega = (\omega_0, ..., \omega_b)$, $\theta = (\theta_0, ..., \theta_q)$, $\omega_0, \theta_0 \neq 0$. And the observation V_t at time t can be generated by:

$$V_t = \sum_{i=0}^{p} \varphi_i \mu(X_{t-i}) + \sum_{i=0}^{b} \omega_i \sigma(X_{t-i}) \varepsilon_{(t-i)} + \sum_{i=0}^{q} \theta_i \eta_{(t-i)}. \qquad (5)$$

To estimate the likelihood of observation sequence V, we adopt the approximation method presented in [6].

Above discussion of ARMA-HMM is for the training phase. Actually, in the testing phase, ARMA can also be used to enrich testing input sequences right before we solve the early recognition problem as an optimal-state sequence problem for the trained model (shown in lower right box of Figure 1). Specifically, we treat each dimension of HOV as an univariate time series, so ARMA models can be quickly trained respectively for each dimension based on the observed sequence $V_1, ..., V_t$. Then we use all m ARMAs to do k steps prediction which enrich the testing input observation sequence V by incorporating k predicted HOV $V_{t+1}, ..., V_{t+k}$. Finally, we calculate $Pr(V_{1:t}V_{t+1:t+k}|M_i)$ for each trained ARMA-HMM M_i and select M_c^*, where $c^* = \arg\max_i(Pr(V_{1:t}V_{t+1:t+k}|M_i))$.

3 Experiments and result analysis

To measure the usefulness of the proposed approach, we have applied our activity early recognition framework on two challenging datasets including Maryland Human-Object Interactions (MHOI) dataset [2] and UCF50 dataset [9].

3.1 Results on MHOI dataset

MHOI dataset consists of 5 annotated activities: *answering a phone call, making phone call, drinking water, lighting a flash, pouring water into container*. For each class of activity, we have 9 or 10 video samples performed by different subjects. And there are 44 video clips in total. Examples in this dataset are shown in Figure 1. We train a 3-state ARMA-HMM with 18 mixture components, and $p = 6, q = 3$ for autoregressive order and moving average order respectively. We use standard leave-one-out cross-validation method to evaluate model performance. In addition, we implemented several previous human activity early recognition approaches to compare them with our method. Three types of previous early recognition model using the same features (*i.e.* STIP features) were implemented: (1) Dynamic Bag-of-Words model [8], (2) Integral Bag-of-Words model [8], and (3) a basic SVM-based approach. Figure 2 and Table.1 shows early recognition performance on MHOI dataset.

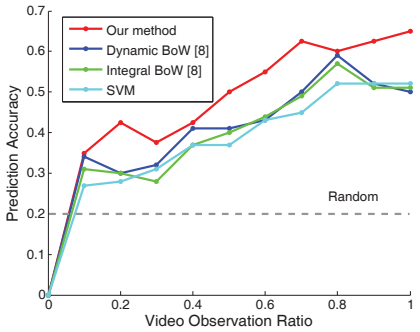

Figure 2. Activity recognition performances with respect to the observed ratio, tested on the MHOI dataset.

Table 1. Recognition performances measured on the MHOI dataset.

Methods	MHOI dataset			
	30% observed	50% observed	70% observed	100% observed
I-BoW [8]	0.28	0.40	0.49	0.51
D-BoW [8]	0.32	0.41	0.50	0.50
SVM	0.31	0.37	0.45	0.52
Our Model	**0.38**	**0.50**	**0.63**	**0.65**

3.2 Results on UCF50 dataset

UCF50 (Figure 3) is an action recognition dataset with 50 action categories, consisting of realistic videos taken from YouTube. We design 3 sets of experiments for different models to evaluate the effectiveness of ARMA-HMM model for difficult real-world human activity early recognition tasks. We compare the learned ARMA-HMM to several alternative early recognition models including Discrete-HMM, Gaussian-HMM. In these experiments we demonstrate that the ARMA-HMM achieves best performance (Figure 4 and Table

(a) Fencing (b) Golf Swing (c) High Jump

Figure 3. UCF50 dataset, YouTube Videos.

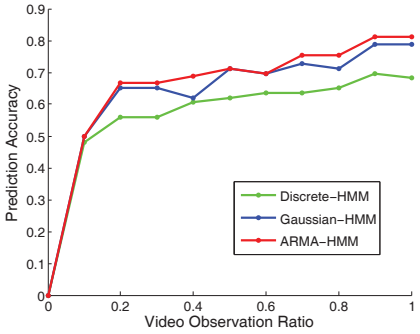

Figure 4. Activity recognition performances with respect to the observed ratio, tested on the UCF50 dataset.

2). In Figure 5 we present the detailed early recognition results for 10 most difficult categories of activities. Specifically, in these experiments, we choose 100 clips for each activity category for training and testing. We use standard leave-one-out cross-validation method to evaluate model performance. We train a 3-state ARMA-HMM with 22 mixture components, and $p = 2, q = 2$ for autoregressive order and moving average order. For comparison, a 3-state Discrete-HMM with 14 observation states and 3-state Gaussian-HMM with 20 mixture components are learned using EM algorithm run until convergence.

4 Conclusion and Future Work

In this paper, we have proposed a novel approach to model human activity as time series signals for activity early recognition. The major contributions include a global motion feature HOV and a early recognition model, ARMA-HMM. We have empirically shown that the proposed method is effective and robust for the activity early recognition task. Since our approach relies on a good global representation of human motion, activities involving multiple persons are not suitable for this model. For the future work, linear dynamic systems and other variants of HMM can be explored further.

Acknowledgement

This research is supported in part by the NSF CNS 1135660, Office of Naval Research award N00014-12-1-0125, and U.S. Army Research Office grant W911NF-11-1-0365.

Table 2. Recognition performances measured on the UCF50 dataset.

Methods	UCF50 dataset			
	30% observed	50% observed	70% observed	100% observed
D-HMM	0.56	0.62	0.64	0.68
G-HMM	0.65	0.71	0.73	0.79
ARMA-HMM	**0.67**	**0.71**	**0.75**	**0.81**

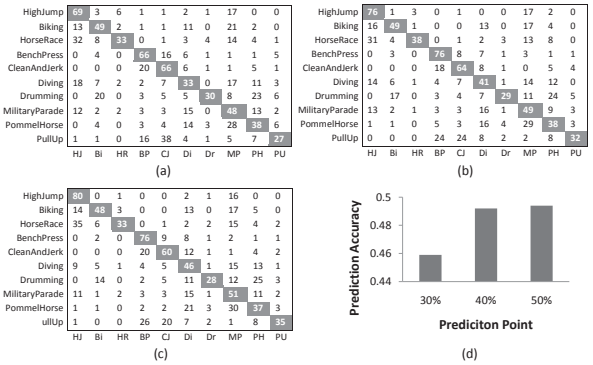

Figure 5. Early recognition results on 10 most difficult categories of activities. Observation ratio: (a) 30%, (b) 40%, (c) 50%, (d) average.

References

[1] H. H. Bui, S. Venkatesh, and G. West. Policy recognition in the abstract hidden markov model. *Journal Of Artificial Intelligence Research*, 17:451–499, 2002.

[2] A. Gupta, A. Kembhavi, and L. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE PAMI*, 31(10):1775–1789, 2009.

[3] M. Hoai and F. De la Torre. Max-margin early event detectors. In *Under review for the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[4] L. Liao, D. Fox, and H. Kautz. Location-based activity recognition using relational markov networks. In *IJCAI*, pages 773–778, 2005.

[5] R. Messing, C. Pal, and H. A. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *IEEE ICCV*, 2009.

[6] S. Michalek, M. Wagner, and J. Timmer. A new approximate likelihood estimator for arma-filtered hidden markov models. *IEEE Transactions on Signal Processing*, 48(6):1537–1547, 2000.

[7] B. T. Navneet Dalal. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[8] M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *IEEE ICCV*, 2011.

[9] M. Sullivan and M. Shah. Action mach: Maximum average correlation height filter for action recognition. In *CVPR*, 2008.