A Survey on Bayesian Learning Model for Human Action Recognition

Qi Xuan WANG
College of Information Science and Technology / College of Cyber Security, Jinan University, Guangzhou 511400, China
2415675318@qq.com

Abstract. Human action recognition is to judge human action by analyzing action characteristics in the fields of computer vision and video surveillance. As the development of machine learning technique, the application of Bayesian Learning Model is increasing in related fields. In order to analyze the characteristics of human action and then recognize human action, this paper introduce a survey on Bayesian Learning Model for Human Action Recognition. The paper focuses on Bayesian handcrafted and deep learning models, and evaluate the state-of-the-art benchmark datasets, e.g., Weizmann, KTH, MSR-3D, HOHA, and UCF101. In this paper, all papers are published ranging from 2007 to 2016, which provides an overview of the progress in this area.

1. Introduction
As computer vision and video surveillance develops, the information in images and videos is so abundant and detailed that it is qualified to be analyzed for human action recognition [2]. To improve the artificial intelligence of computer device in the surveillant system, many researchers struggles to develop techniques to observe and understand the image content and meaning automatically. So how to express the image information through the features accurately and utilize them to recognize the human action has become the key to human action recognition.

As far, there are many well-known human action datasets. In this paper, we selected five human action datasets as research data. The Weizmann dataset is one of the widely-used action recognition dataset and because of its widespread use, the recognition rate of the experiment using this data set is as high as 90% or more. The KTH dataset [3] has lots of samples in different scenes, it is also widely used by researchers in the behavior recognition algorithm test. The MSR-Action3D dataset [4] is an action dataset of depth sequences captured by a depth camera and is widely used in human action recognition based on depth image. HOHA (Hollywood Human Actions)-2 dataset [5] intends to provide a comprehensive benchmark for human action recognition in realistic and challenging settings. The dataset is composed of video clips from 69 movies. UCF101 dataset [6] is an action recognition dataset of realistic action videos. It has 13320 videos from 101 action categories and gives the largest diversity in terms of actions and with the presence of large variations in camera motion.

In the field of human action recognition, many researchers have put forward a lot of creative ideas in many different ways, such as intelligent monitoring system, human-computer interaction, movement analysis and so on. With the increase in the content of video surveillance, the intelligent monitoring system becomes more and more development potential, such as VSAM [7] and W4 [8]. Then, a team in Reading University has also conducted a study of the tracking and interaction of vehicles and pedestrians. The application of human action recognition in human-computer interaction is mainly reflected in the
capture of human action, according to the result of the capture, people can control the machine by a series of actions. A well-known example is the Kinect provided by Microsoft. This device is used in the task of movement analysis, it is mainly used in sport movement analysis and the medical analysis. It can see that human action recognition still has great potential in various application fields of human action recognition.

Because the Bayesian model has the advantages of simple algorithm, easy training, insensitive to missing data, etc., it has been used by many researchers to model the classification problem for a long time, such as Bayesian Deep Learning (BDL), Bayesian Classification Trees (BCT), Bayesian Belief Networks (BBN). Next, I will briefly introduce these three Bayesian learning models. First, Bayesian Deep Learning. Basic deep learning is used in many areas, but there may be some problems in deep learning such as not being robust enough and difficult to explain the learning process of deep learning. When the deep learning is combined with the Bayesian theorem, these problems can be solved by the new models with strong fitting ability of deep learning. For BCT, the introduction of Bayesian theorem in random forests provides a solution to the problem that the random forests have a tendency to produce oversized trees and the problem cause by overfitting in order to enhance the performance. About the Bayesian Belief Network, BBN is very suitable for dealing with incomplete data, even if the information is missing, is can still classify the data. Because the data and the prior knowledge are combined in a probabilistic way, BBN has a good robustness to the over-fitting of the model.

This paper presents the state-of-the-art modeling and applications of Bayesian learning model for human action recognition in recent years. In Section 2, the datasets employed in experiments are reviewed. Section 3 and Section 4 respectively describe the methods on modeling and applications, and Section 5 presents the conclusion.

2. Dataset
In this section, we introduce the five datasets mentioned in the previous section. We will also provide some examples of each dataset. The evaluations and descriptions of all datasets are shown in Table 1. Among these five data sets, there are examples of human actions for specific situations.

| Table 1 Comparison of five datasets |
|----------------------------------|
| **Action Type** | **Resolution** | **Background** | **Number of Videos** |
| WEIZMANN[1] | 10 | $180 \times 144$ | static | 90 |
| KTH [3] | 6 | $160 \times 120$ | static | 600 |
| MSR-3D [4] | 12 | N/A | cluttered | massive |
| HOHA [5] | 12 | N/A | cluttered | 3669 |
| UCF101[6] | 101 | $320 \times 240$ | cluttered | 13320 |

2.1. The Weizmann dataset
This dataset is composed of 90 videos, a total of 10 of human behavior. The ten kinds of human action are walking, running, jumping, galloping sideways, bending, one-hand waving, two-hands waving, jumping in place, jumping jack, skipping. Each action is performed by nine persons respectively. This data set is captured by a static camera, so the background of each video is not complicated. And because of its wide use, the dataset has performed well in the experimental results of human behavior recognition. Examples of this dataset are given in Fig. 1.

![Fig. 1 Three types of human action in Weizmann dataset](image1)

![Fig. 2 Examples of the KTH dataset](image2)
2.2. The KTH dataset
This data set consists of 600 video components. In the KTH dataset, all video is divided into six human behaviors, including walking, jogging, running, boxing, hand waving, and hand clapping. Each group was composed of 25 people, divided into training groups (8 persons), the verification group (8 persons) and the test group (9 person). In this data set, the change in focal length is small, and the background of the frame is static. Therefore, the data set is used as the basic data in the experiment. Examples of this dataset are shown in Fig. 2.

2.3. The MSR-Action3D
The MSR-Action 3D dataset consists of a depth sequence of human actions taken by a depth camera. This dataset includes 12 types of human action such as high arm waving, horizontal arm waving, hammering, hand catching, forward punching, high throwing, drawing x, drawing tick, drawing circle, hand clapping, two hand waving, side-boxing, bending, forward kicking, side kicking, jogging, tennis swing, tennis serving, golf swing, pick up & throw. Since the data set provides data on human behavior in three-dimensional space, it is commonly used in experiments focused on 3D human motion recognition. Examples of this dataset are shown in Fig. 3.

2.4. HOHA (Hollywood Human Actions)-2 dataset
This dataset consists of 430 videos which are divided into a training set (219 videos) and a testing set (211 videos). This data set has eight categories of human action, answering phone, getting out car, hand shaking, hugging person, kissing, sitting down, sitting up and standing up. Since the dataset was filmed with a non-static camera and all are realistic background, all of the scenarios were complex, so it was a challenging dataset for the experimenter. Examples of this dataset is given in Fig. 4.

2.5. UCF101 dataset
The UCF101 is an action recognition dataset, which is a lifelike action from the 101 action categories collected on YouTube. The dataset contains 13,320 videos from 101 action categories. The data set can be divided into 25 groups, each consisting of 4 to 7 actions video. This data set is probably the most challenging data set, because it may have the greatest operational diversity. Examples of this dataset are shown in Fig. 5.
3. Bayesian learning models for human action recognition

Khan et al. propose a condition Bayesian network for human action recognition [9]. Actually, their work is a complement of the same work in [10] where the Bayesian Network is used in the lower level inferences while the higher-level inferences is made by an HMM. In this work, an action recognition approach is proposed, which recognizes an action by detecting its components and reasoning for their consistency with the action structure. The action decomposition takes the actor as centered. An action component is a group of primitive actions and actor object relationships, also including their temporal order. They use HMM to recognize sub-components which is the primitive actions and relationships like walking and crouching, from noisy human and object tracks in linear time. Then, they use the condition Bayesian Network to combine the evidence, collected by each action, over the interval of its components. Their model is more robust to missing data.

Yang et al. propose a Multi-Feature Max-Margin Hierarchical Bayesian model (M3HBM) for action recognition [11]. Their model jointly learns a multi-feature hierarchical generative model as the representation part together with max-margin classifiers in a unified Bayesian framework for action recognition. There are three layers in the model: point-level visual observations, region-level local STPs distributed in many different small neighborhoods, and top-level global STPs shared by all different classes without position limitation. Furthermore, the Gibbs classifiers is introduced to minimize the expected loss based on the max-margin principle, and Gaussian priors are used to perform Bayesian estimation for classifier parameters. Eventually, the experiment shows that their methods for representation has many advantages and the M3HBM is also very effective in human action recognition.

As for Bayesian Classification trees, Bulo et al. provide an online learning model in [12]. Their work tries to overcome some limitations in traditional random forest. Their goal is to learn tree classifiers with a shallow structure and high generalization capability. To do this, they jointly take into account multiple feature dimensions and their correlations, and use Bayesian method in their online learning procedure with a simpler, parametric distribution. The main features of their algorithm is the update rules for the tree hyper-parameters that are free from cumbersome learning rate selection and allow them to naturally absorb the information from each new sample. Their model allows them to update over time a posterior distribution over the space of decision trees because of the online trained decision tree. In the end, it is validated that their approach is able to perform a better online forest algorithm on a variety of classification tasks, while using smaller models.

In the facial action recognition, Campos et al. propose a constrained maximum likelihood learning of Bayesian network [13]. Bayesian network is a very compact graph structure that encode a joint probability distribution for its variables. Similar to the learning problem of other network, the parameter learning of Bayesian network depend on the training data. In order to overcome problem of the incomplete and scarce data, this paper improves the parameter learning and introduces a framework based on non-linear convex optimization and combine quantitative data and domain knowledge in the form of qualitative constraints. For complete data, they directly apply convex optimization to obtain a global optimum of the constrained maximum likelihood estimation. On the other hand, for incomplete data, they extend the EM (Expectation-Maximization) method by introducing a constrained maximization in the M-step. Their experiments with facial expression recognition in real images show the benefits of qualitative constraints in parameter learning and classification accuracy in [14].

Singh et al. introduce a method that Human Action Recognition using a Dynamic Bayesian Action Network (DBAN) with 2D Part Models. They suggest that use an intermediate 2D body part representation of the human model to accurately match the human model and image observations across shape variations and observation noise [15]. They refer to the extended DBAN model which can be used in the action representation part, and suggest that the 3D pose of a person’s approximate viewpoint is orthographically projected to 2D to determine the visible part. Then, they use 2D part based model over the visible part to accurately align the 2D pose by belief propagation. Next, they compute the likelihood of the pose to recognize human actions over the aligned parts. This method computes the likelihood of a sampled pose by matching the foreground feature vectors computed over the projected human model with that obtained from the observed image. However, the matching is not straightforward because they
have to account the person scale and shape variations across different actors.

4. Conclusion

Bayesian learning model as a basic model of machine learning, is widely used in many aspects. There are a lot of very representative Bayesian learning model that we mentioned above, such as Condition Bayesian network, Bayesian Classification trees, and so on. Many researches have achieved good experimental results. Bayesian learning model for human action recognition is a very promising research direction. Benefit from the principle of priori probability in Bayesian model, the event information is obtained for further correction model from lots of training data. Moreover, some above-mentioned methods only use the Bayesian learning model in some special part, but their outputs can also lead to more accurate results using the same principles.

References

[1] Lena G, Moshe B, Eli S, et al. Actions as space-time shapes. [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29(12):2247-2253.
[2] Poppe R. A survey on vision-based human action recognition [J]. Image & Vision Computing, 2010, 28(6):976-990.
[3] Schuldt C, Laptev I, Caputo B. Recognizing Human Actions: A Local SVM Approach[C]// Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03. IEEE Computer Society, 2004:32--36.
[4] Wanqing Li, Zhengyou Zhang, Zicheng Liu. Action recognition based on a bag of 3d points[C]. Computer Vision and Pattern Recognition-Workshops, 2010 IEEE Computer Society Conference on. IEEE, 2010:9-14.
[5] Laptev I, Marszalek M, Schmid C, et al. Learning realistic human actions from movies[C]// Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2008:1-8.
[6] Khurram Soomro, Amir Roshan Zamir and Mubarak Shah, UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild, CRCV-TR-12-01, November, 2012.
[7] R Collins, A Lipton. A System for Video Surveillance and Monitoring: VSAM Final Report[R]. Pittsburgh: Robotic Institute Carnegie Mellon University, 2000: 51-68.
[8] Haritaoglu I, Harwood. D, Davis. L. VV: real-time surveillance of people and their activities[J]. IEEE Trans Pattern Analysis and Machine Intelligence,2000, 22（8）: 809-830.
[9] Furqan M Khan, Sung Chun Lee and Ram Nevatia. Conditional Bayesian Networks for Action Detection[J]. IEEE International Conference on Advanced Video and Signal Based Surveillance,2013,ICAVSBS2013
[10] S. S. Intille and A. F. Bobick. A framework for recognizing multi-agent action from visual evidence. In AAAI, 1999.
[11] Shuang Yang, Chunfeng Yuan, Baoxin Wu, Weiming Hu. Multi-Feature Max-Margin Hierarchical Bayesian Model for Action Recognition. NLPR, Institution of Automation, CAS. In CVPR, 2015.
[12] Samuel Rota Bul’o, Peter Kontschieder. Online Learning with Bayesian Classification Trees. In CVPR, 2016.
[13] Cassio P. de Campos, Yan Tong, Qiang Ji. Constrained Maximum Likelihood Learning of Bayesian Networks for Facial Action Recognition. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part III, LNCS 5304, pp. 168–181, 2008.
[14] Pantic, M., Bartlett, M.: Machine analysis of facial expressions, pp. 377–416. I-Tech Education and Publishing, Vienna, Austria (2007)
[15] Vivek Kumar Singh, Ram Nevatia. Human Action Recognition using a Dynamic Bayesian Action Network with 2D Part Models. ICVGIP ’10, December 12-15, 2010, Chennai, India. 2010 ACM 978-1-4503-0060-5/10/12.