

Pose-based Human Action Recognition with Extreme Gradient Boosting

Vina Ayumi

Machine Learning and Computer Vision Laboratory
Faculty of Computer Science
Universitas Indonesia
Depok West Java, Indonesia
vina.ayumi@ui.ac.id

Abstract—This Paper investigate action recognition by using Extreme Gradient Boosting (XGBoost). XGBoost is a supervised classification technique using an ensemble of decision trees. In this study, we also compare the performance of Xboost using another machine learning techniques Support Vector Machine (SVM) and Naive Bayes (NB). The experimental study on the human action dataset shows that XGBoost better as compared to SVM and NB in classification accuracy. Although takes more computational time the XGBoost performs good classification on action recognition.

Keywords—*pose-based human action recognition; Extreme Gradient Boosting; Machine Learning*

I. INTRODUCTION

Action, gesture, or motion is one of the most important communication tools used by humans. Often people communicate using their body parts movement like hands and head rather than speaking [1]. Human action recognition is the most active topic in computer vision. Many important applications in human action recognition such as surveillance system, video analysis, video retrieval, robotics, and human-computer interaction. There are extensive literatures in action recognition in a number of fields, including computer vision, machine learning, pattern recognition, etc [2][3]. Aims to analyze the various kinds of gesture in human activity automatically, human action recognition method has been developed using video data, motion capture, depth data or some the combination of these modalities [4].

The two stages that usually carried out in the human action recognition procedure: a). Human detection and action feature extraction and b). Action classification [5]. The selection of methodologies used in both stages influenced the final recognition performance [3].

In human action recognition, selection of classification methods plays an important role. There are various techniques of pattern recognition, such as k-Nearest Neighbor (k-NN), Hidden Markov Model (HMM), Conditional Random Field (CRF), Extreme Learning Machine (ELM), Random Decision Forest (RDF), Support Vector Machine (SVM), and Relevance

Vector Machine (RVM). These techniques have been developed and applied in human action recognition.

One of the most effective and widely used method in machine learning is tree boosting. Tree boosting algorithm has been used to give state-of-the-art results in many standard classification benchmarks [6]. Recently, Chen and guesstrin proposed a scalable end-to-end tree boosting system called Extreme Gradient Boosting (XGBoost) for learning an ensemble of decision trees. The boosting technique was adjusted to enhance a Taylor expansion of the loss functions. The model is also insensitive to imbalanced data phenomenon because it enables to select AUC measure for evaluation and forces proper ordering of the imbalanced data. This method is widely used in many machine learning and data mining challenges and successfully applied to many classification problems to achieve state-of-the-art of the result [7]. In this paper, we proposed Extreme Gradient Boosting for posed-based human action recognition.

II. RELATED WORK

Several previous studies in human action recognition have been done, Starner et al. perform recognition of American Sign Language using the Hidden Markov Model (HMM) [8]. Chung and Yang proposed threshold models with Conditional Random Field (CRF) to gesture recognition by using depth information of Microsoft Kinect sensor. The threshold models with Conditional Random Field (CRF) used to differentiate vocabulary gestures and non-vocabulary gestures [9]. Chen and Koskela in [10], was proposed motion capture recognition using RGB depth camera and Extreme Learning Machine (ELM) as classifier. The gesture classification using K- nearest neighborhood (k-NN) classifier was considered in [11]. Schuldt et al. Proposed action recognition using local measurements features combined with SVM as a robust classifier approach [12]. Wiehwa et al. addressed the issues of human action recognition by introducing multiclass relevance vector machine (mRVM), the experimental result using Weizmann dataset shows the proposed method offers superior performance [5]. Ayumi and Fanany demonstrate a comparison of RVM and SVM for human action recognition. The experimental result shows although takes more training time RVM is better as

compared to SVM in human action recognition [13]. The studies in [14] proposed distribution sensitive learning on RVM to deal with the unbalanced data problem in human gesture dataset. Experimental result on the gesture dataset shows distribution sensitive learning on RVM can improve the performance of recognition.

A. Extreme Gradient Boosting

XGBoost (eXtreme Gradient Boosting) first developed by Chen and Guestrin is an open source project to implement an efficiently, fast, and scalable machine learning system called Gradient Tree Boosting in a wide variety of machine learning problems [7]. Xgboost is an ensemble of Regression Trees (CART) $\{R_1(x_i, y_i), \dots, R_k(x_i, y_i)\}$ and C Classification where x_i and y_i are the training and the class label respectively. The prediction scores is summed up to get the final score. Then the final score evaluated through C additive functions, as shown in Equation (1):

$$\hat{y}_i = \sum_{k=1}^C f_k(x_i), f_k \in F \quad (1)$$

Where f_k and F represent the independent tree structure with leaf scores and the space of all CART respectively. The regularized objective to optimize is given by Equation (2):

$$Obj(\Theta) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (2)$$

The first term represents the differentiable loss function (l), that measures the difference between the predicted \hat{y} and the target y_i . The second term is a regularization Ω to avoid overfitting which penalizes the complexity of the model. It is given by $\Omega(f) = \gamma R + \frac{1}{2} \lambda \sum_{j=1}^R w_j^2$. Where R is the number of leaves and w is the weight of each leaf. The constants γ and λ are to control the regularization degree. Irrespectively from the use of regularization, two additional techniques used to prevent overfitting are shrinkage and descriptor subsampling [7].

For a training dataset of action with the vectors of descriptors and the corresponding class labels, the Xgboost training procedure is summarized as follows;

- i. For every descriptor,
 - Sort the numbers
 - Scan the best splitting point
- ii. Choose the best splitting point descriptor that optimizes the training objective,
- iii. Continue splitting (as in (i) and (ii)) until the predetermined most extreme tree profundity is came to,
- iv. Assign score of prediction to the leaves and prune any negative nodes (nodes with negative additions) in a bottom-up order,
- v. Repeat the above steps in an added substance way until the predetermined number of rounds (trees K) is achieved.

Since added substance preparing is utilized, the prediction \hat{y} at step t expressed as

$$\hat{y}_i^{(t)} = \sum_{k=1}^K f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (3)$$

Equation (2) can be written as

$$Obj(\Theta)^{(t)} = \sum_i l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) \quad (4)$$

Furthermore, more generally by taking the Taylors expansion of the loss function to the second order

$$Obj(\Theta)^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) + \Omega(f_t) \quad (5)$$

Where $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ and $h_i = \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)})$ are the first and second order statistics on the loss function respectively. A simplified objective function without constants at step t is as the following

$$Obj(\Theta)^{(t)} = \sum_{i=1}^n [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \quad (6)$$

The objective function can be composed by expanding the regularization term as

$$\begin{aligned} Obj(\Theta)^t &= \sum_{i=1}^n [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \\ &= \sum_{j=1}^T [\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 + \gamma T] \quad (7) \end{aligned}$$

the $I_j = \{i | q(x_i) = j\}$ is the instance set of leaf j , for a given structure $q(x)$. w_j^* and Obj^* are the optimal leaf weight and objective function which measure how good the structure is are respectively given by Equations (8) and (9)

$$w_j^* = -\frac{G_j}{H_j + \lambda} \quad (8)$$

$$Obj^* = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (9)$$

where $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$. Equation (10) is utilized to score a leaf node during splitting. The first term of the equation remains for the score on the left, the second term for the score on the right and the third term are the original leaf. In addition, the last term, γ , is regularization on the additional leaf.

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \quad (10)$$

III. EXPERIMENTAL SETUP

This experimental study is carried out by implementing the action recognition task with XGBoost. The specification of hardware and software that used in this experiment are: Processor Intel Core i7-3770 CPU @ 3.40GHz, Memory DDR2 RAM 64.00 GB, Ubuntu 14.04 LTS 64 bit. The XGBoost algorithms are implemented in Python language <https://github.com/dmlc/xgboost> [7].

To show the performance of proposed method, we used action dataset namely the Microsoft Research Cambridge-12 (MSRC-12) Kinect gesture database [19], UTKinect-Action3D dataset [15], Badminton sports action dataset [16], and Bali dance motion dataset [17].

A. MSRC-12 Kinect Gesture Database

The Microsoft Research Cambridge-12 (MSRC- 12) Kinect database is the dataset that consists of sequences of human skeletal body movements with its meaning which will be

recognized by the system. Consists of 594 sequences collected from 30 people performing 12 gestures, the dataset was captured using a Kinect depth sensor in Microsoft platform. In every sequence, one subject performs an action in several times. The 12 gestures shown on figure 1 they are: wind up, beat both lift arms, goggles, shoot, bow, throw, duck, push right, had enough, change weapon, and kick. Different types of introduction are giving to them to show the effect on movements of subjects. Hence, the dataset is built not only to measure the performance of recognition system but also to evaluate all the instruction such as by text, picture and video. Kinect Pose Estimation pipeline used to estimate 20 3D joints in every frame [18].













Metaphoric gestures	Main frames	Iconic gestures	Main frames
Start music\ raise volume (G1)		Crouch or hide (G2)	
Navigate to next menu (G3)		Put on night vision goggles (G4)	
Wind up the music (G5)		Shoot with a pistol (G6)	
Take a bow to end the session (G7)		Throw an object such as a grenade (G8)	
Protest the music (G9)		Change weapon (G10)	
Lay down the tempo of a song (G11)		Kick to attack an enemy (G12)	

Fig. 1 Example motion of MSRC-12 Kinect gesture data set [10]

1) UTKinect-Action3D Dataset

The UTKinect-Action3D dataset containing 10 types of human actions in indoor settings. Using a single stationary Kinect, the video was captured. The are 10 actions include: walk, sit down, stand up, pick up, carry, throw, push, pull, wave and clap hands. Figure 2 shows samples of 10 actions. Each action was collected from 10 different persons for 2 times: 9 males and 1 female. One of the persons is left-handed. The dataset contains 6220 frames of 200 action samples. Three channels were recorded: RGB, depth and skeleton joint locations. The skeletal joint locations, each row contains the data of one frame, the first coloumn is frame number, the following coloumn are the locations of joint 1-20 in x, y, and z. The x, y, and z in meters are the coordinates relative to the sensor array[15].



Fig. 2 Sample images of 10 actions [15]

B. Badminton Sports Action Dataset

Consist of sequences of human skeletal body movements and its meaning, the Badminton Sports Action dataset was captured by using Kinect depth sensor consists of eight motions they are Short Service, Long Service, Forehand Stroke, Backhand Stroke, Overhead Loop, Drop Shot, Underhand lob, and Smash [20] .

C. Bali Dance Motion Dataset

The Bali dance motions dataset are captured using a static mounted depth sensor camera positioned in front of the performer that produces skeleton coordinates at the rate of 30 fps. Data recording are started with depth sensor calibration to track initial position of the dancer skeleton joints. The pattern classes used in this study are some basic gestures from Balinese Pendet traditional dance. The data consist of 75,197 poses which is segmented into 705 dance motion segments that are annotated into 10 classes of dance motions. The motion primitives for dance pattern are basic gesture of Bali traditional dance namely: *agem kanan*, *agem kiri*, *piles*, *ngeseh*, *luk nerudut*, *ngegol*, *ulap-ulap*, *tanjek kanan*, *tanjek kiri*, and *tabur bunga*. Some of the motions primitives are shown on figure 3. Those motion primitives serve as common basic gestures of Bali traditional dance [17].

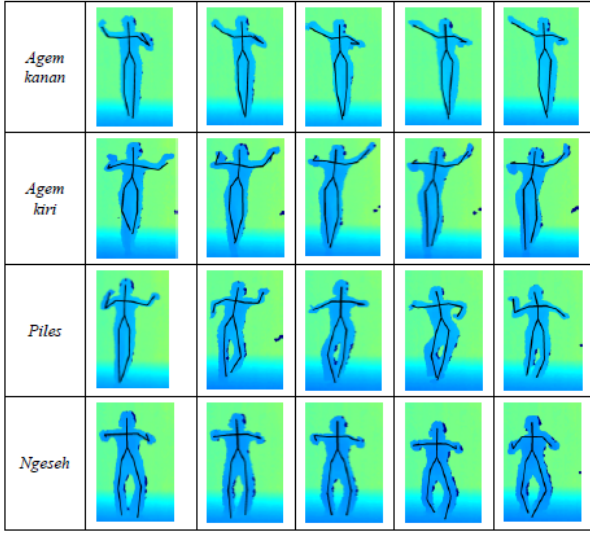


Fig. 3 Key poses of some motion primitives [17]

IV. RESULT

This study successfully implemented XGBoost for the action recognition dataset. In this study, the data are separated into training data and testing data using 10 k-cross validation. Tabel 1 shows the number of pose in each dataset

TABLE I. NUMBER OF POSE/ FRAME/ ROW IN EACH DATASET

Dataset	Number of pose
Microsoft	37927
UTKinect-Action3D	2388
Badminton Sports Action	6425
Bali Dance Motion	75197

In this study we also compared the performance of XGBoost with two machine learning algorithms that have been used in the previous studies for activity prediction they are Support Vector Machine (SVM) and Naïve Bayes (NB).

TABLE II. ACCURACY AND COMPUTATIONAL TIME FOR XGBOOST AND SVM

Data/Methods		MSRC-12	UTK3D	Badminton Sports Action	Bali Dance Motion
XGBoost	Acc (%)	99.82	96.64	98.91	99.88
	Time (s)	2578.95	127.77	142.84	1173.31
SVM	Acc (%)	53.40	64.81	97.66	98.87
	Time (s)	1089.78	0.8423	4.50	132.09
NB	Acc (%)	37.92	68.52	82.55	82.82
	Time(s)	0.41	0.0246	0.0488	0.43

TABLE III. F1-SCORE, PRECISION AND RECALL FOR XGBOOST AND SVM

Data/Methods		MSRC-12	UTK3D	Badminton Sports Action	Bali Dance Motion
XGBoost	F1-score	0.9980	0.9641	0.9894	0.9979
	Precision	0.9980	0.9661	0.9882	0.9995
	Recall	0.9981	0.9648	0.9908	0.9962
SVM	F1-score	0.5342	0.4288	0.9759	0.9831
	Precision	0.5935	0.4251	0.9760	0.9941
	Recall	0.5229	0.4813	0.9759	0.9744
NB	F1-score	0.3718	0.6266	0.8380	0.8432
	Precision	0.4359	0.6994	0.8450	0.8483
	Recall	0.3752	0.6341	0.8460	0.8787

Table II gives comparison accuracy and computational time of XGBoost, SVM, and NB. The result shows that XGBoost is better than SVM and NB in accuracy rate. The comparison of computational time for the machine learning method is one of good aspect for classification task in supervised technique. The experiment result shows that XGBoost takes more computational time than SVM and NB.

Figure 4, 5, 6, 7 gives the confusion matrix comparison of XGBoost, and SVM with the action datasets. The confusion matrix show that XGBoost learning is more balanced and also more robust for predicting classes than SVM in action datasets.

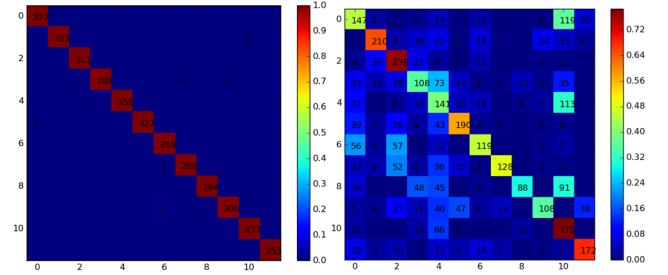


Fig. 4 Confusion Matrix MSRC-12 with XGBoost and SVM.

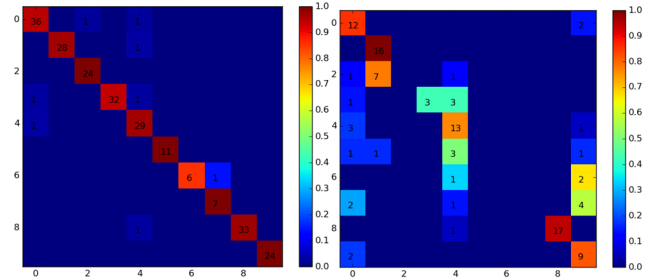


Fig. 5 Confusion Matrix UTKinect-Action3D with XGBoost and SVM.

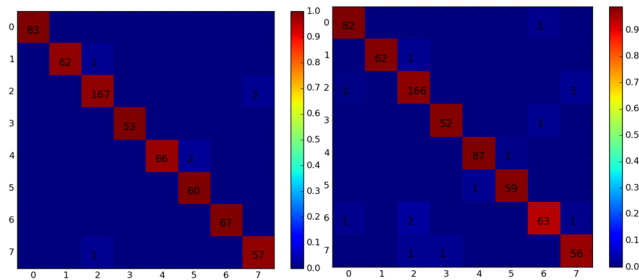


Fig. 6 Confusion Matrix Badminton Sport Action with XGBoost and SVM.

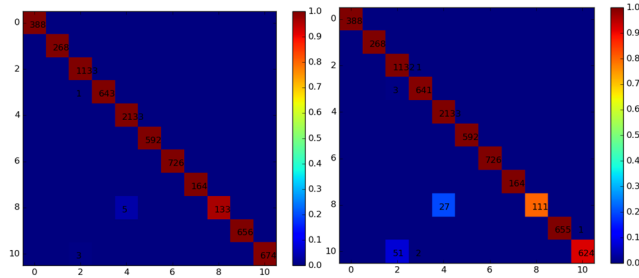


Fig. 7 Confusion Matrix, (a) Bali Dance Motion with XGBoost and SVM.

V. CONCLUSION

This study investigated the performance of XGBoost in action recognition task. The Experiment study conducted using action dataset the Microsoft Research Cambridge-12 (MSRC-12) Kinect gesture database, UTKinect-Action3D dataset, Badminton sports action dataset and Bali dance motion dataset. Our experimental results showed that although takes more computational time than SVM and NB, the XGBoost performs good classification on action recognition dataset.

ACKNOWLEDGMENT

This work is supported by Higher Education Center of Excellence Research Grant funded by Indonesian Ministry of Research, Technology and Higher Education (Contract no. 1068/UN2.R12/HKP.05.00/201

REFERENCES

- [1] K. K. Biswas, "Gesture Recognition using Microsoft Kinect ®," vol. 2, pp. 100–103, 2011.
- [2] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine Recognition of Human Activities: A Survey BT - Circuits and Systems for Video Technology, IEEE Transactions on," vol. 18, no. 11, pp. 1473–1488, 2008.
- [3] J. K. Aggarwal and M. S. Ryoo, "Human Activity Analysis : A

Review," 2007.

- [4] Q. Chen, N. D. Georganas, E. M. Petriu, K. Edward, A. Ottawa, and C. Kin, "Real-time Vision-based Hand Gesture Recognition Using Haar-like Features," 2007.
- [5] W. He, "Recognition of human activities using a multiclass relevance vector machine," *Opt. Eng.*, vol. 51, no. 1, p. 017202, Feb. 2012.
- [6] P. Li, "Robust logitboost and adaptive base class (abc) logitboost," *arXiv Prepr. arXiv1203.3491*, no. 2, pp. 1–30, 2012.
- [7] T. Chen and C. Guestrin, "XGBoost," *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '16*, pp. 785–794, 2016.
- [8] T. Starner, S. Member, J. Weaver, A. Pentland, and I. C. Society, "Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video," vol. 20, no. 12, pp. 1371–1375, 1998.
- [9] H. Chung and H.-D. Yang, "Conditional random field-based gesture recognition with depth information," *Opt. Eng.*, vol. 52, no. 1, p. 017201, Jan. 2013.
- [10] X. Chen and M. Koskela, "Skeleton-Based Action Recognition with Extreme Learning Machines," no. October, 2013.
- [11] X. Jiang and F. Zhong, "Robust Action Recognition Based on a Hierarchical Model," pp. 191–198, 2013.
- [12] C. Sch and L. Barbara, "Recognizing Human Actions : A Local SVM Approach *," pp. 3–7.
- [13] V. Ayumi and M. I. Fanany, "A comparison of SVM and RVM for human action recognition," *Internetworking Indones. J.*, vol. 8, no. 1, pp. 29–33, 2016.
- [14] V. Ayumi and M. I. Fanany, "Distribution-Sensitive Learning on Relevance Vector Machine for Pose-Based Human Gesture Recognition," *Procedia Comput. Sci.*, vol. 72, pp. 527–534, 2015.
- [15] L. Xia, C. C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 20–27, 2012.
- [16] A. Budiman and M. I. Fanany, "Pose-based 3D human motion analysis using Extreme Learning Machine," *2013 IEEE 2nd Glob. Conf. Consum. Electron.*, no. L, pp. 3–7, 2013.
- [17] Y. Heryadi, M. I. Fanany, and A. M. Arymurthy, "A method for dance motion recognition and scoring using two-layer classifier based on conditional random field and stochastic error-correcting context-free grammar," *2014 IEEE 3rd Glob. Conf. Consum. Electron. GCCE 2014*, no. July 2016, pp. 771–775, 2015.
- [18] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," *Proc. 2012 ACM Annu. Conf. Hum. Factors Comput. Syst. - CHI '12*, p. 1737, 2012.