A Comparison of SVM and RVM for Human Action Recognition

Vina Ayumi, Mohamad I. Fanany

***Abstract*—Human action recognition is a task of analyzing human action that occurs in a video. This paper investigates action recognition by using two classification techniques, namely Relevance Vector Machine (RVM) and Support Vector Machine (SVM). SVM is a technique for supervised classification that used in statistics and machine learning. By separating the distinct class with a maximum possible wide gap, SVM tries to predict the respective class given a set of input data. On the other hand, RVM is a Bayesian model of Generalized Linear Model (GLM) that has an identical function with SVM. RVM uses significantly fewer basis functions as it uses Bayesian inference with a prior distribution on weight thus makes solution sparse. Experimental studies on a human action dataset show that RVM is better as compared to SVM on action recognition. Although RVM takes more training time, however, it requires fewer testing time than SVM. RVM model is more general because it contains minimum basis function. Therefore, it is more robust compared to SVM. RVM performs good classification on action recognition that contains large dataset.**

***Index Terms*—Action Recognition, Relevance Vector Machine (RVM), Support Vector Machine (SVM).**

I. INTRODUCTION

HUMAN actions and gestures are important communication tools used by humans. People sometimes communicate by using body movements such as hands or head rather than speaking [1]. Human action and gesture recognition are one of the most active research areas in Machine Learning, Computer Vision, and Human-Computer Interaction. Human action recognition aims to provide an automated analysis of various kinds of human activities [2]. Action recognition is applied in various areas such as video surveillance, human-machine interaction, video retrieval, sports analysis, gaming, biometric, analysis of sign language, and robotic. Action and gesture recognition methods have been developed starting from either video, motion capture, depth data or some combination of these modalities [2]. There are two stages in the human action recognition procedure: a). Human detection and action feature extraction and b). Action classification. Final recognition performance is influenced by the selection of methodologies used in both stages [3].

Manuscript received July 6, 2015.

Vina Ayumi, Mohamad Ivan Fanany is with the Computer Science Departement, University of Indonesia, Depok, West Java, Indonesia (e-mail: vinaayumi@gmail.com, ivan@cs.ui.ac.id).

Several previous studies have been done in human action recognition, Starner et al. using the Hidden Markov Model (HMM) to perform recognition of American Sign Language [4]. The studies in [5] proposed gesture recognition by using depth information of Microsoft Kinect sensor and threshold models with Conditional Random Field (CRF) to differentiate vocabulary gestures and non-vocabulary gestures. In [2], it was proposed a recognition of motion capture with RGB depth camera using Extreme Learning Machine (ELM). In [6], Random Decision Forest (RDF) was used to recognize gestures from RGB depth camera. The gesture classification using K-nearest neighborhood (*k*-NN) classifier was considered in [7].

The selection of classification methods plays an important role in human action recognition. Various techniques of pattern recognition, such as k-Nearest Neighbor (*k*-NN), Extreme Learning Machine (ELM), and Support Vector Machine (SVM) have been developed and applied in action recognition. The research community strongly agrees that the performance of support vector machine is good on the task of action recognition. A new approach in classification technique is a Bayesian model of Generalized Linear Model (GLM) that has an identical function with SVM and it is termed as relevance vector machine (RVM). On the same problem with SVM, RVM tries to define a probabilistic vector. RVM offers several advantages over the SVM. First, RVM is a non-linear probabilistic model with a prior distribution on weight that makes solution sparse. RVM can produce a decision function fewer than SVM, and can maintain accuracy. RVM does not require any regularization parameter tuning during the training, nor does it require kernel to fulfill the Mercer's condition like SVM. This paper investigates performances of SVM and RVM learning methods on the action recognition task. This study has been carried out on action dataset namely the Microsoft Research Cambridge-12 (MSRC-12) Kinect gesture database and the Badminton sports action dataset.

II. STATE OF THE ART

Support vector machine (SVM) was first proposed for numerical data [8]. The main idea is to determine separators that best separate the different possible classes. Several studies for action recognition have been proposed recently using SVM [1]. Gesture recognition using Kinect depth camera is carried out by performing three stages. First, pre-processing to eliminate background on the depth image. Second, feature extraction to the extracting region of interest obtained by

placing 14x14 grid in the foreground. Then, training and testing are performed by using SVM. A multiclass SVM is used to train the system for the classification of the eight gestures. In [5], it was proposed sign language recognition by detecting and recognizing manual signals (MS) and non-manual signals (NMS). After that, hierarchical CRF (HCRF) and boost map embedding are used to detect MS and NMS. It is then followed by using SVM and appearance model for recognizing MS and NMS. In [9], it was proposed a detecting interest points techniques by using SIFT (scale invariant feature transform) from each frame of video and Bag of Video Words (BoVW) approaches and multiclass SVM for action classification. In [10], it has been performed feature extraction such as joint positions, joint velocities, joint angels, and joint angular velocities and used some of the machine learning methods such as Naive Bayes, SVM, and Random Forest. In [6], RDF has been used to model the data and identify the most effective features and put the features based on position and time. SVM classifier has been used to training of the features that have been selected.

On the other hand, several studies for action recognition have been proposed recently by using RVM. RVM tries to learn data given the dataset based on the Bayesian inference of a linear model with an appropriate prior [14]. In [11], it has been proposed human action recognition using the sparse representation (RVM) on image sequences as the collection of spatiotemporal events. In [12], it was introduced Multiclass Relevance Vector Machine (mRVM) algorithm to the human pose classification from single stationary camera to video surveilance application, i.e. first, performed foreground blobs extraction from the edge that obtained, and then performed normal and abnormal behavior classification using RVM. In [3], it was proposed a recognition framework that consisted of three modules: detecting the human silhouette blobs from image sequence by background subtraction; performed shape and motion feature extraction from Variation Energy Image (VEI); performed human action recognition by using mRVM. RV was proposed as regression to recovering 3D human body pose from a single image and monocular image sequences in [13].

A. Support Vector Machine

Support Vector Machine (SVM) is a technique for supervised classification that used in statistics and machine learning. SVM traditionally is a two-class classifier, that tries to predict the respective class given a set of input data (features). SVM is a non-probabilistic binary linear classifier [14]. Given input set of training instance from the two classes, SVM tries to build a model. This model can be used to classify and to predict new and unseen instance based on learning from the training data. Given a set of data points that belong to either of two classes, it represents the data instance into space. It finds a maximum possible wide gap that maximizing the distance of either class, and find the optimal separating hyperplane that minimizes the risk of misclassifying the training samples and unseen test samples.

Given set of training data (x_1,y_1), (x_2,y_y), ...(x_n,y_n), the problem is to separate the set of training data into two classes where $x_i \in R^N$ is a feature vector and $y_i \in \{-1,+1\}$ is a class label. Two classes can be separated by a hyperplane $w.x + b = 0$ in some space H. We have no prior knowledge about the distribution, then the optimal hyperplane is the one which maximizes the margin [15]. The optimal values for w and b can be found by solving a constrained minimization problem, using Langrange multipliers $\alpha_i (i = 1, \dots n)$.

$$f(x) = sgn(\sum_{i=1}^{n} \alpha_i, y_i K(x_i, x) + b) \qquad (1)$$

where α_i and b are found by using SVC learning algorithm [15]. The support vectors are x_i with nonzero α_i. $K(x_i, y) = x.y$, corresponds to kernel that constructing an optimal separating hyperplane in the input space R^N [16].

B. Relevance Vector Machine

RVM is a regression and classification technique introduced by Tipping [17]. RVM is a Bayesian model of Generalized Linear Model (GLM) that has an identical function with SVM. RVM offers some advantages over the SVM. First, RVM is a non-linear probabilistic model with a prior distribution on weight that makes a sparse solution. RVM can produce a decision function fewer than SVM, and can maintain accuracy. It does not require any regularization tuning parameters during the training and kernel in order to fulfill the Mercer's condition like SVM.

Given a dataset of input vectors $\{x_n\}_{n=1}^{N}$, $x_n \in R^d$, along with corresponding targets $\{t_n\}_{n=1}^{N} t_n \in \{0,1,2, \dots\}$, where R is a set of real numbers and d is the dimension. A sigmoid function based on the kernel principle [17] can be used as a classifier

$$y_n = \frac{1}{1+exp(-\varphi_n.w)} \qquad (2)$$

where $\varphi_n = (\Phi(x_n, x_1), \Phi(x_n, x_2), \dots., \Phi(x_n, x_N))^T$. In the above equation, $\Phi(x_n, x_m)$ is commonly implemented using a radial basis function in many vector machines where the inputs are numerical. The likelihood function of the classification model using the cross-entropy function is as follows.

$$p(t|w) = \prod_{n=1}^{N} y_n^{t_n} (1 - {y_n}^{1-t_n}) \qquad (3)$$

An automatic relevance vector determination (ARD) prior [18] is used to prevent over-fitting over the coefficients [17]

$$p(w|a) = \prod_{n=1}^{N} G(0, \alpha_n^{-1}) \qquad (4)$$

where $\alpha = (\alpha_1, \alpha_2, , \alpha_N)^T$. The posterior of the coefficients is as follows

$$p(w|t,\alpha) \propto |\sum|^{-1/2} exp\{-\frac{1}{2}(w-u)^T \sum^{-1} (w-u)\} (5)$$

The mean vector and the covariance matrix are

$$u = \sum \Phi^T B t \quad (6)$$

and

$$\sum = (\Phi^T B \Phi + A)^{-1} = H^{-1} \quad (7)$$

respectively, where $t = (t_1, t_2, \ldots t_l)^T$, $B = diag\{y_n(1-y_n)\}$, $A = diag\{\alpha_1, \alpha_2, , \alpha_N\}$, and Φ is a squared input kernel matrix. The marginal likelihood can be obtained through integrating out the coefficients.

$$p(t|\alpha) \propto |B^{-1} + \Phi A^{-1} \Phi^T|^{-1/2} exp\{-\frac{1}{2} t^T (B^{-1} + \Phi A^{-1} \Phi^T)^{-1} t\} (8)$$

In learning, α can be estimated

$$\alpha_n(\tau+1) = (1 - \alpha_n(\tau) \sum_{nn} / u_n^2(\tau)) \quad (9)$$

where τ is the iterative time. The weight update can follow

$$\Delta w = -H^{-1} \nabla L \quad (10)$$

where

$$L = -log\{|\sum|^{-\frac{1}{2}} exp[-\frac{1}{2}(w-u)^T \sum^{-1} (w-u)]\} (11)$$

H is the Hessian matrix. Equation (9) is a closed form where we have to use an iterative algorithm to update weights.

III. Experimental Setup

This experimental study is carried out by implementing the action recognition task with SVM and RVM. The algorithms are implemented in Matlab programming language. RVM and SVM are evaluated by performing experiments on human action dataset.

A. Dataset

To show the performance of the proposed method, we used action dataset namely the Microsoft Research Cambridge-12 (MSRC-12) Kinect gesture database [19] and Badminton sports action dataset. The Microsoft Research Cambridge-12 (MSRC-12) Kinect database consists of sequences of human skeletal body movements and its meaning which will be recognized by the system. The dataset that was captured using a Kinect depth sensor in Microsoft platform consists of 594 sequences collected from 30 people performing 12 gestures. One subject performs an action several times in every sequence. The 12 gestures are: lift arms, duck, push right, goggles, wind up, shoot, bow, throw, had enough, change weapon, beat both and kick. Different types of introduction are giving to them to show the effect on movements of subjects. Therefore, the dataset is constructed to not only 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.

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

IV. Result

A. MSRC-12 Kinect Gesture Database

This study successfully implemented SVM and RVM for the MSRC-12 Dataset. In this study, the data are separated into training data and testing data using cross-validation. Traning data took 2/3 of the dataset of each class and the rest as a testing data. The dataset consists of 12 actions performed by six persons. The features consist of 20 3D joint human body skeleton so that we have 60 features. To make comparison are equally represented, we used the same kernel function which is the Radial Basis Kernel (RBF) with scale 0.3. In this experiment we also compared the result of classification using the data of 1, 2 3, until six persons.

TABLE I
TABLE NUMBER OF FRAMES

Total Person	Number
1	8180
2	16235
3	22862
4	30182
5	37952
6	44365

The accuracy of the classification results is shown in Table II. The result shows that RVM is better than SVM in accuracy rate. It is an indication that RVM can be used for generalized action recognition. When the amount of data increases, RVM accuracy more stable and SVM accuracy tend to be decreased. It is mainly because the dataset contains large of data points for building better prior probabilistic information. The comparison of training time for the model is one of good aspect for classification task in supervised technique.

TABLE II
TABLE ACCURACY AND SUPPORT/ RELEVANCE VECTOR OF 60 FEATURES DATA

Total Person	SVM		RVM	
	Acc	SV	Acc	RV
1	100	13	100	17
2	99.37	33	99.72	35
3	95.38	81	98.73	67
4	92.42	176	98.58	85
5	91.28	447	97.67	80
6	89.18	688	97.83	93

Table III gives comparison of SVM and RVM time in second for building model for classification. The RVM training time is higher than SVM because it computes prior information for prediction of the class relationship that is not the case with SVM. RVM testing time is fewer compared to SVM, as it produces fewer relevance vectors than support vectors as shown in Table III.

TABLE III
TABLE TRAINING AND TESTING TIME IN SECOND

Total Person	SVM		RVM	
	Training Time	Testing Time	Training Time	Testing Time
1	9.68	0.97	3510	0.02
2	9.10	1.25	9366	0.01
3	65.71	3.28	53954	0.0001
4	194.58	6.19	63671	0.15
5	202.703	10.63	26258	0.02
6	352.806	20.17	79160	0.15

In this study, we also tried to reduce the number of features in the dataset by using Principal Component Analysis (PCA). The number of the previous features are 60 then we reduced the features become 10. The result of the classification using ten features is given in the Table VI. Although the number of features is reduced in RVM, RVM has accuracy which is higher than SVM.

TABLE IV
TABLE ACCURACY OF 10 FEATURES DATA

Total Person	Accuracy	
	SVM(%)	RVM(%)
1	66.98	97.44
2	66.98	93.01
3	53.88	98.84
4	68.16	97.72
5	52.14	86.00
6	44.78	94.28

Table V shows the accuracy of each class using five-person data. These experimental results show that RVM has an accuracy rate compared with SVM for each class in the dataset.

TABLE V
TABLE ACCURACY OF EACH CLASS SVM AND RVM

Class	Accuracy	
	SVM(%)	RVM(%)
Start Music	91.33	97.46
Crouch and hide	87.49	98.37
Navigate to next menu	97.41	99.11
Put on night vision goggles	91.81	94.81
Wind up the music	94.05	98.97
Shoot a pistol	87.10	94.28
Take a bow end music session	90.30	99.37
Throw an object	92.05	98.40
Protest the music	88.99	90.89
Change weapon	92.29	99.15
Move up the tempo of the song	94.16	95.39
Kick	87.63	99.90

B. Badminton Sports Action Dataset

This study also implemented both SVM and RVM for the Badminton sports action dataset. In this study, the data are separated into training data and testing data using cross-validation. Traning data took 2/3 of the dataset of each class and the rest were the testing data. The dataset consists of 8 actions that performed approximately 10 times by 5 different person. The features consist of 14 joint human body skeleton in ρ and θ so that we have 28 features. The experimental result is shown in Table IV. Experimental result on badminton dataset show that RVM is better compared with SVM. Altough it takes more training time, RVM is more robust because the models contains fewer relevance vector than support vector on SVM.

TABLE VI
TABLE OF ACCURACY, SV/RV, TRAINING AND TESTING TIME IN SECOND

	Badminton	
	SVM	RVM
Accuracy (%)	95.89	97.00
SV/RV	28	6
Train (s)	0.11	5.43
Test (s)	0.01	0.01

V. CONCLUSION

This paper investigated action recognition by using SVM and RVM. Our experimental results showed that RVM had an accuracy rate compared with SVM. When the amount of data increased, RVM accuracy was more stable than SVM that inclined. When the number of features was reduced, RVM has superior accuracy. Hence, we conclude that RVM took more

training time. However, it required fewer testing time than SVM. RVM model more was more robust because it had minimum basis function and more general. RVM performed good classification on action recognition in large dataset.

REFERENCES

[1] K. K. Biswas, "Gesture Recognition using Microsoft Kinect," in *Proc. 5th International Conference on Automation, Robotics and Applications*, New Zealand, vol. 2, pp. 100–103, Dec 2011.

[2] X. Chen and M. Koskela, "Skeleton-Based Action Recognition with Extreme Learning Machines," no. October, 2013.

[3] W. He, "Recognition of human activities using a multiclass relevance vector machine," in *Optical Engineering* SPIE., vol. 51, no. 1, p. 017202, Feb. 2012.

[4] 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,"*IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1371–1375, 1998.

[5] H. Chung and H.-D. Yang, "Conditional random field-based gesture recognition with depth information," *Optical Engineering* SPIE, vol. 52, no. 1, p. 017201, Jan. 2013.

[6] Negin, F. Özdemir, C.B. Akgül, K.A. Yüksel, and A. Erçil, "A decision forest based feature selection framework for action recognition from rgb-depth cameras," *In Image Analysis and Recognition*, pp. 648-657. Springer Berlin Heidelberg, 2013.

[7] X. Jiang and F. Zhong, "Robust Action Recognition Based on a Hierarchical Model," in *International Conference on Cyberworlds*, pp. 191–198, 2013.

[8] C. and V. V. Cortes, "Support-Vector Networks," vol. 7.

[9] M. M. Moussa, E. Hamayed, M. B. Fayek, and H. A. El Nemr, "An enhanced method for human action recognition," *J. Adv. Res.*, 2013.

[10] H. Zhang, W. Du, and H. Li, "Kinect Gesture Recognition for Interactive System," pp. 1–5.

[11] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal salient points for visual recognition of human actions.," *IEEE Trans. Syst. Man. Cybern. B. Cybern.*, vol. 36, no. 3, pp. 710–719, Jun. 2006.

[12] B. Yogameena, S. V. Lakshmi, M. Archana, and S. R. Abhaikumar, "Human Behavior Classification Using Multi-Class Relevance Vector Machine," vol. 6, no. 9, pp. 1021–1026, 2010.

[13] A. Agarwal, B. Triggs, and D. Europe, "3D Human Pose from Silhouettes by Relevance Vector Regression," in *Proc. of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[14] M. Rafi and M. S. Shaikh, "A comparison of SVM and RVM for Document Classification," *Procedia Comput. Sci.*, vol. 00, pp. 3–8, 2013.

[15] V. N. Vapnik, "An Overview of Statistical Learning Theory," in *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[16] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," I*n Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference*, vol. 3, pp. 32-36, IEEE, August, 2004.

[17] T. ME., "Sparse Bayesian learning and the relevance vector machine," *J Mach Learn Res*, vol. 1, 2001.

[18] D. J. C. MacKay, "A Practical Bayesian Framework for Backpropagation Networks," *Neural Comput.*, vol. 4, no. 3, pp. 448–472, 1992.

[19] 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*, pp. 1737, 2012.

[20] A. Budiman and M. Fanany, "Pose-based 3d human motion analysis using extreme learning machine,"in *Consumer Electronics (GCCE), 2013 IEEE 2nd Global Conference* on, Oct 2013, pp. 3-7, 2013.

Vina Ayumi received her B.Sc. in Informatics Techniques (State Islamic University Jakarta, Indonesia, 2012). She is now taking her Master degree in Computer Science (University of Indonesia). Her research interest includes machine learning, data-mining, and image processing.

Mohamad I. Fanany received his B.Sc in Physics Departement, Faculty of Science and Mathematics (University of Indonesia), M.Sc in Computer Science from Faculty of Computer Science, University of Indonesia, and Ph.D. from Departement of Computer Science, Tokyo Institute of Technology. Now he works as a researcher and lecturer at Faculty of Computer Science and Graduate School of Biomedical Engineering Univerity of Indonesia). His research interest includes imaging science and engineering, 3D perception, reconstruction, recognition, and data-mining for autonomous and assisted driving, multi-modal sensor fusion for real-time decision and planning, combining vision and graphics for the application of advanced machine learning especially in remote sensing, climate modeling, biomedical, automobile, broadcasting, and robotics industry.

.