International Review on Computers and Software (I.RE.CO.S.), Vol. 11, N. 11
ISSN 1828-6003 November 2016

Advanced Model for Human Action Annotation Based on Background Subtraction Using Learning Vector Quantitation with Co-occurrence Matrix Features

M. A. Soeleman[1,2], Eko M. Yuniarno[2], Mochamad Hariadi[2], Mauridhy H. Purnomo[2], Masanori Kakimoto[3]

Abstract – *This paper presents an advanced model for human action annotation. The proposed technique is splitting human objects in two parts, upper and under part of human beings. From these two parts, a model to extract the feature vectors by GLCM was proposed as feature for classification. The first method is to extract the Haralick features out of GLCM and the next step is the normalization process for converting co-occurrence feature matrix into various vectors as feature for classification. The research employs learning vector quantification to classify all feature vectors. Finally, the experiment conducted by utilizing Weizmann dataset shows that this approach method achieves an accuracy of 84.7%.* **Copyright © 2016 Praise Worthy Prize S.r.l. - All rights reserved.**

Keywords*: Annotation, GLCM, Classification, Learning Vector Quantification*

Nomenclature

f_1	Angular Second Moment
f_2	Contrast
f_3	Entropy
f_4	Correlation
f_5	Angular Second Moment
f_6	Clashed
f_7	Cluspro
f_8	Max Pro
f_9	Dissimilarity
f_{10}	Autocorr
f_{11}	Inertia
f_{12}	Different Entropy
f_{13}	Sum Entropy
LVQ	Learning Vector Quantization
GLCM	Gray Level Co-occurrence Matrix

I. Introduction

With the rapid growth in multimedia-based equipment, especially with the use of cameras and video, hundreds to thousands of videos produced by users have been uploaded to the internet. But the number of videos uploaded is not proportional to the number of images considered as a dataset. These circumstances lead to a fairly high disparity between the two, particularly in the images and videos related to human movement. The growing demand of video applications for annotations was encouraged to generate datasets for annotations in particular for human movement. But the process of video annotations on a complex action tends to be subjective, with one another having different opinions on human movements. This thus has an impact on the result dataset video annotations. To bridge these problems then the movement annotations are more likely to use the dataset to atomic action recognition.

Several approaches for human action recognition can be categorized as: spatiotemporal analysis shape template based approach [1] [2] [3], tracking based approach[4], flow based analysis [5] and interest points based approach [6]. The first approaches for the spatiotemporal analysis have been proposed by Bobick et al. in [1]. The authors used the motion energy and the process of moving history from images sequences as information temporal to recognize human aerobics movements. Weinland et al. in [7] matching with previous research, they employed multiple cameras to construct the process of moving history volumes and to execute the state of acting to classification with Fourier analysis for cylindrical coordinates. Blank et al. in [2] and Yilmaz et al. in [8]presented a related work for 3D approach in which the flow based approaches to optical flow computation used to describe motion are sensitive to noise and cannot reveal the true motions. Spatio-temporal appearance based techniques consider the action identification problem as a 3D object identification problem and extract characteristic from the 3D volume. The extracted distinctive traits were very huge, so the computational cost is incapable for real-time application. Tracking based techniques suffer from the same problem.

There are several variations of techniques that can be used to make annotations on the video.

To perform the annotations, humans can specify the annotations in several criteria such as time, location, and activity.

In the application of annotation the techniques can be performed automatically or semi-automatically. The most basic model is a free text description that can be added to the video. This method does not use the definition of the initial structure to perform the annotation [9]. For a pattern model, when publishing a video on internet in a social media, the user can type a description of the video that will be uploaded. Some combinations of words and sentences can be carried out to establish a free description. Some types of annotations helps in accessing the video, by not using annotations structure is easy to do but the efficiency in retrieval techniques should be performed.

In many video annotation mode based on text, the textual information can remain on a video, which the image sequences are named as additional text. This technique is utilized as a key that possibly makes the designation by some term obtained from an isolated text. According to the text, data is a source of high different meaning from knowledge if it is readily obtainable, then it will be a key filter and it can be searched from video data by the user involving intuition and natural methods. Text that is embedded in the image and video in particular to provide more detail and important to the content of knowledge communicated for instance the name of the player, speaker, title, place, date of an event occurrence [10]. The meaning can be derived from a video text. SunithaAburru in [9] using an extreme imposed text to derived the semantics of a video in which there have been growth being efficient of retrieval system. A semi-automatic approach is used to generate annotations to videos, to process semantic derived by analyzing the content of the rule-based techniques.

Exhausted equable features have been derived from the video or image. There are a few variations such as machine learning the Support Vector Machine, Clustering and the Bayesian networks resemblances and learning to do. In [11] an essential supporting structure for high meaning annotations video event can be traced by using the operations of global features, local features and motion features. By using these features, a digital image of sequences clip can be converted into form code as a set of distinctive attribute vectors. With a variety of different features that exist SVM can classify and perform learning and establish a code of chromosome-based genetic algorithm optimization to obtain relevant classification and weighting based learning.

Yang C et al. in [12] annotations of an instructional supervised learning called as Multiple-Instance Learning (MIL)can be done by expanding the prevalent method of Support Vector Machine-based MIL algorithms. By maximizing the boundary into the pattern of MIL constraints, the obstacle occurred in the MIL can be converted by the former method. Barrat et al. in [13] performed on the image of annotation classification weakness that there was only a small part of a key-code set of database annotation.

In the video annotation ontology [14] has described as a clearly developed a detailed description of a conceptualization. In the classification system large-sized system that the classifier differentiate on aspects into the category hierarchy, for example, that chapter is part of the book. In the same study [15] proposed a framework for ontology useful to enrich the semantic annotation on CCTV camera video, a semantic text and visuals could associate with the key granted by domain experts. In the video segmentation have done to discover a moving object in the video and the classification as an agent, action and receiver. Video ontology-based annotation can also be included rule or machine learning.

In the study of pattern recognition, feature extraction is very important to get the maximum results. There are several approaches to extract a key feature in many pattern recognitions based on Gabor features. Daugman in [16] proposed characterizing features at different frequencies and orientations. The function of this approach is to equalize between the 2D Gabor function and maintenance and characteristics of the visual system of mammals.

Leen-KiatSoh et al in [17] conducted a research on the analysis of image texture of ice on the sea surface for mapping. 10 GLCM with features such as entropy, contrast, correlation, homogeneity, energy, autocorrelation, dissimilarity, shade cluster, cluster prominence, the maximum probability level 64, the distance *D*is 1 and *D*is 2 proved more effective for feature extraction image texture of ice. For orientation has no effect on the classification of sea ice images. This study uses a Bayesian Classifier for classification methods

Daniel S. et al. in [18] conducted a research about the classification of the image of the battle scene on the landscape. This study used invariant moment and GLCM from 8 features such as energy, inertia, entropy, homogeneity, max probability, contrast, inverse, correlation to its extraction. The results showed SVM with radial kernel has a higher accuracy than the feed forward back propagation algorithm of ANN. Therefore in this work a LVQ for human action annotation was proposed using GLCM feature.

II. Related Research

In video frame, visual features can be done directly from a lower-level of video frame. By using the low level, the features can be used to make annotations, but they are still a crevice which appear in the midst of the information that could be derived automatically out of the visual data and interpreted. The identical data can be owned by a user to set of high-level concept of the low level descriptions.

Jardon et al. [19] conducted a research based on rule-based approach using fuzzy logic to represent from the initial definition and limitations in adaptation to different contexts. Dorado et al. in [20] proposed a rule-based approach to video annotations, where the proposed method automatically uses the knowledge of an initial annotation dataset. This can create a representation of a set of lower level from video characteristic and associated rules between lower level and upper level. In [21] conducted a study of the average on fuzzy decision tree (FDT), where rule-based automated sample is based on a limitation of exploitation in which there are measures to reduce the desire of human beings in the use of the index process.

The similarity of the features of audiovisual demonstrated to detect semantic event has been proposed in [22], while this approach detects annotation of broadcasting a soccer player in video. In proposals made by a fuzzy rule-based, the design to classify result from the adoption of statistical output through a collection of audio-visual attribute as usage and delivery a crisp set of semantic concepts corresponds to the events that occurred. Doing the extraction of hidden knowledge between tuple and the mutual relationship between the distinctive attribute related to the event can be done by the construction of a decision tree. Other approach has been successfully applied to some content such as video-based analysis [23], [24].

The authors proposed a graph-based learning that a method based approach has function to semi-supervise. In the graph method approach, process of labeling and non-labeling perform on the vertices. This is a sample in which the vertices on the edges exist can cause a reflection of the similarities among the sample regarded as a unit. A purpose intended by this method that measured on graph is based on labeling more subtle assumptions. In study [25], authors argued a method to study the concept in relation using graphical models to improve the results of annotation.

The authors argued that video annotations primarily assist in marking on some single or multi concept in the process of labeling the target dataset, where the target is carried out often without regarding to the concept of internal independent in relationships. Then, research-based multi graph semi-supervised has been proposed by [24]. The method proposes merge graph and semi-supervised learning on merging function graph.

III. Methodology

This paper presents a new approach for human action annotation. The goal of the proposal is to represent actions by extracting feature texture using gray level co-occurrence matrix to describe the object feature. The proposed method is described in Fig. 1.

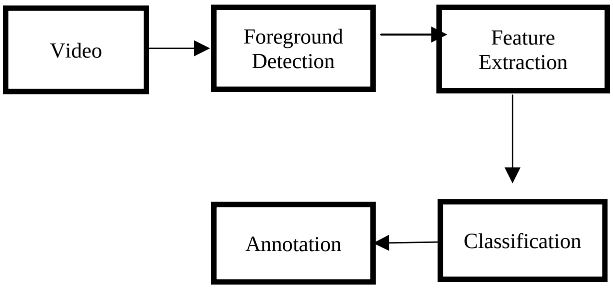


Fig. 1. Proposed Method for the Annotation Human Action

III.1 Human Detection

In the first stage for the identification of human movement, we detect the moving object, which is human, was detected.

The approach to detect objects used a background subtraction method [26]. Approach to background subtraction is done by reducing the current fr_i frame with the previous fr_{i-1} frame in the video sequence.

From the results of background subtraction technique, it can be obtained a foreground object moving form of human movement that can be identified from the background:

$$fr_{i-1} = | fr_i - fr_{i-1} | \quad (1)$$

$$fr_{i+1} = | fr_{i+1} - fr_i | \quad (2)$$

To perform better results of detecting moving objects, the background of subtraction morphological process has been applied to the frame that is being done at background subtraction. This step is performed in order to reduce noise that appears so as to obtain better results. The process morphology [27] applied includes the applied process of erosion, and dilation in order to obtain the optimum results against the background subtraction. The results of this process will be the next object to be extracted as data for segmentation are done.

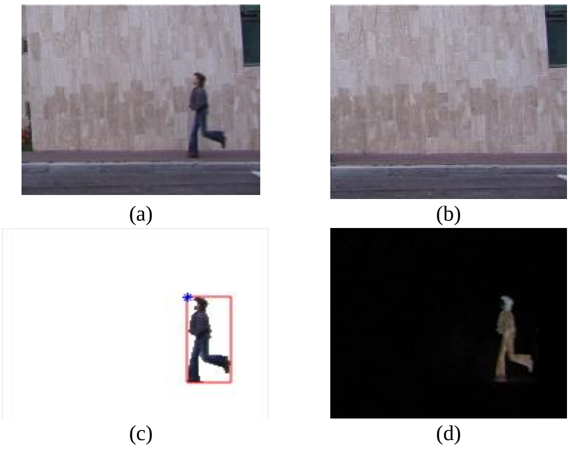


Fig. 2. Moving objects Detection using Background Subtraction (a) Current frame (b) previous frame (c) detection Object with bounding box (d) subtraction frame

After successfully detecting human as a moving object, the research continued by segmenting human action. Human body was divided in two parts, upper and lower by splitting the segmented human, as shown in Fig. 3.

III.2. Feature Extraction

This stage uses the Gray Level Co-occurrence Matrix (GLCM) texture features to take out the attribute from the human segmented. Gray Level Co-occurrence Matrix invented by Harralick in [28] proposed method for extracting texture feature from objects for feature extraction.

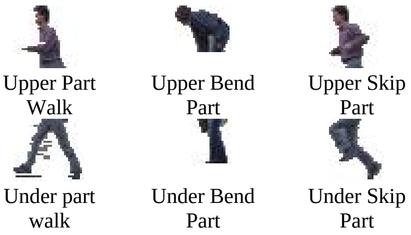
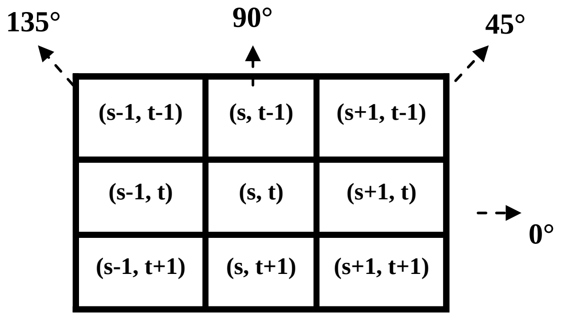

Upper Part Walk
Upper Bend Part
Upper Skip Part
Under part walk
Under Bend Part
Under Skip Part

Fig. 3. Human body split from segmented

Texture is an important aspect of an image that has been widely used in the classification such as visual inspection and pattern recognition. GLCM which is a tabulation of the frequency, or how often to different combinations of sharpness pixels occurs in a frame. The computation of math-based frame of image is shown in the following equation where $p(s,t)$ is the value $(s,t)-th$ is an input in co-occurrence normalized matrix, N denotes co-occurrence dimension of the matrix (the number of gray levels) $p_x(s)$ and $p_y(s)$ is in the margin where the probability is:

$$p_x(s)=\sum_{t=1}^{N}p(s,t),p_y(t)=\sum_{s=1}^{N}p(s,t) \tag{3}$$

Texture is generally a complex visual model that has attributes such as clarity, coloring, and gradient. A consistency of a surface attribute can be extracted in various approaches specifically, statistics, constructional, and modifications of information-based models. There is a 4-way computing in GLCM, i.e. grade= 0 °, grade= 45 °, grade = 90 °, grade = 135 °



(s-1, t-1)	(s, t-1)	(s+1, t-1)
(s-1, t)	(s, t)	(s+1, t)
(s-1, t+1)	(s, t+1)	(s+1, t+1)

Fig. 4. Matrix of GLCM

Each method has a different technique. Jian in [29] proposed Gabor and co-occurrence of this method for object recognition. An established algorithms to extract texture features called Gray Level Co-occurrence Matrix (GLCMs) is proposed by Roberts et al. in [28], which has statistical methods. The features of GLCM containing secondary-order relate to the use of relevant statistics information of space intercourse one of many features from an image. GLCM is resulted from a gray scale image. GLCM contains information about how often pixels with gray-level value (scale of intensity or gray tones) *s* occurs either horizontally, vertically, or diagonally to adjacent pixels with the value *t*. The parameters *s* and *t* indicate the value of gray level (tone) in an image:

$$f_1=\sum\nolimits_s\sum\nolimits_t P(s,t)^2 \tag{4}$$

$$f_2=\sum_{s=0}\sum_{t=0}(s-t)^2P(s,t) \tag{5}$$

$$f_3=-\sum_{s=0}^{L-1}\sum_{t=0}^{L-1}P(s,t)log\,P(s,t) \tag{6}$$

$$f_4=-\sum_{s=0}^{N-1}\sum_{t=0}^{N-1}\frac{(s-\mu_x)(t-\mu_y)}{\sqrt{\sigma_x\sigma_y}}ps_{st} \tag{7}$$

$$f_5=\sum_{s=0}^{N-1}\sum_{t=0}^{N-1}p_{st}^2 \tag{8}$$

$$f_6=\sum_{s=0}^{N-1}\sum_{t=0}^{N-1}\left(s+t-\mu_x-\mu_y\right)^3p_{st} \tag{9}$$

$$f_7=\sum_{s=0}^{N-1}\sum_{t=0}^{N-1}\left(s+t-\mu_x-\mu_y\right)^4p_{st} \tag{10}$$

$$f_8=max\left(p_{st}\right) \tag{11}$$

$$f_9=\sum_{s=0}^{N-1}\sum_{t=0}^{N-1}|s-t|p_{st} \tag{12}$$

$$f_{10}=\sum_{s=0}^{N-1}\sum_{t=0}^{N-1}(st)\,p_{st} \tag{13}$$

$$f_{11}=\sum_{s=0}^{N-1}\sum_{t=0}^{N-1}(s-t)\,p_{st} \tag{14}$$

$$f_{12}=\sum_{s=0}^{N-1}p_{x+y}(s)log\left(p_{x+y}(s)\right) \tag{15}$$

$$f_{13}=\sum_{s=0}^{2N-2}p_{x+y}(s)log\left(p_{x+y}(s)\right) \tag{16}$$

For all of the GLCM features of human object segmented we employed the normalization so all parameters could be classified for human action annotations.

III.3. Classification

Learning Vector quantization in [30]is used at this stage to classify the texture GLCM features that have been extracted. In LVQ existing data on the features of GLCM are regarded as an input vector. Data can be denoted as $V_s \in R^d$ with value of s = 1, 2, n, whereas the denomination of n is a number from the inside of data. From the contained data, any training was conducted in accordance with appropriate human movement patterns. To facilitate the giving of labeling the movement patterns, each data is denoted with the following models as $X_s \in \{1,2,3..z\}$ with value of s =1, 2,3,*n*, wherein *n* represents the quantity of data values and z is the amount of patterns that doing the training. At the stage of identification of the patterns of movement, LVQ algorithm will group them into a pattern with a Euclidean distance of the closest. The LVQ models are used as follows:

The algorithm of LVQ can be explained as follows:

1. Determine the value of the weight (w), the maximum epoch (Max Epoch), the estimated value of the minimum error (Eps) and Learning rate (α)
2. Input value :
   a. Input: v(z,n);
   b. Target: T(1,n)
3. Specify the initial value:
   a. Epoch = 0;
   b. Err = 1;
4. Make the process if (Epoch < MaxEpoch) or (Learning rate > Eps)
   a. Epoch = epoch + 1;
   b. Do for value of *s* = 1 until n
      i. Determine the value of *t* so $\|v - w_t\|$ has a minimum value stated C_t
      ii. make improvements on the value of w_j the provisions:
   - If the value of T = C_t therefore $w_t(new) + \alpha[v - w_t(old)]$
   - If value of $T \neq C_t$ then $w_t(new) = w_t(old) - \alpha[v - w_t(old)]$
   - Subtract the value of α

where:

v is vector training $(v_1, v_2, ...v_n)$

T is a category that is true for training vector

w_t is weight vector output unit *t* $(w_{st}, ..., w_{nt})$

C_t value are the categories represented by the output unit *t*

$\|v - w_t\|$ the value of the Euclidean space is not far from the input vector and the weight vector to unit output *t*

IV. Experimental Results

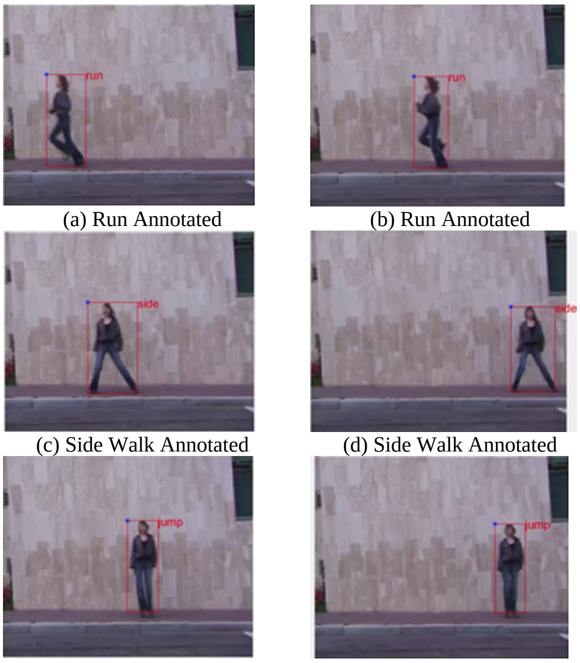
This experiment applies a dataset of video Weizmann. Dataset consists of 8 actions, and every movement consists of 8 different people. Each dataset consists of 50 frames with a resolution (180×144) and with has a frequency of 25 fps. The evaluation for the classification is done by applying confusion matrix. The confusion matrix will show the value of right and wrong in recognizing the movement action in humans. The result of human segmented and extraction feature shown in Fig. 6 and Fig. 8. In this stage to test the success of the performance results of the classification LVQ on the proposed model, the confusion matrix was used as a representation, as shown in Table I. By using the confusion matrix the accuracy of the classification results obtained can be calculated by the following equation:

$$Accuration = \frac{TP + TN}{TP + TN + FP + FN} \quad (17)$$

The four conditions in the confusion matrix can be explained as follows:

- *True Positive* (TP) is a positive embodied value in which the classification results indicate a positive truth value.
- *False Negative* (FN) defines the value which predicts opposite results shown by the value of one. - *True Negative* (TN) is a negative value which is appropriately classified as a negative value
- *False Positive* (FP) is the value which predicts positive results but it shows negative and is calculated as False Positive.

Figs. 5 show the result of human action annotation using LVQ.

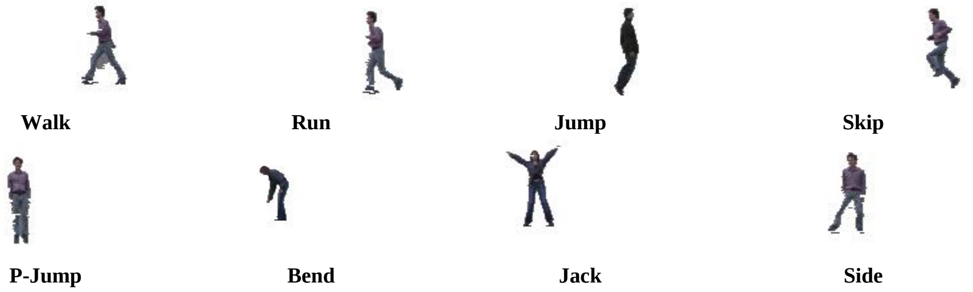


(a) Run Annotated (b) Run Annotated

(c) Side Walk Annotated (d) Side Walk Annotated

(e) Jump Annotated (f) Jump Annotated

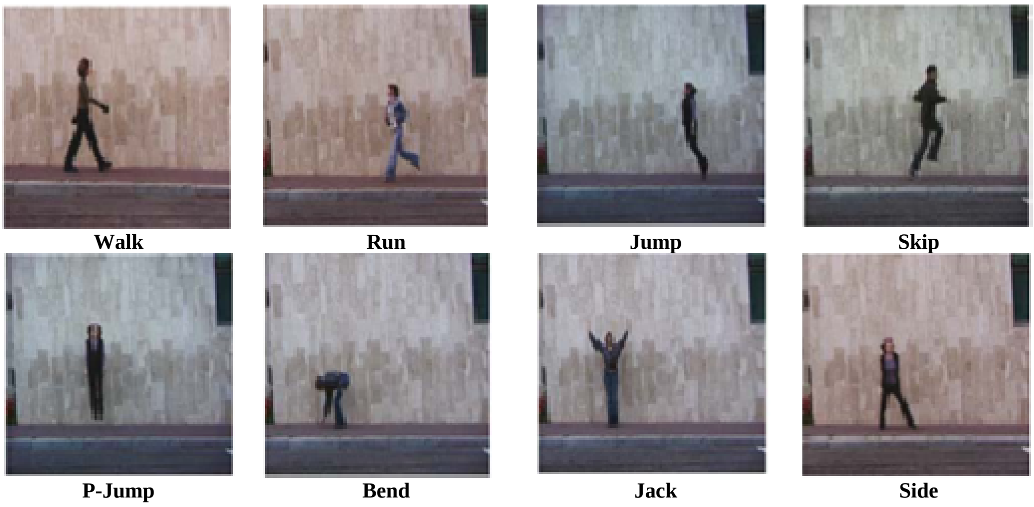
Figs. 5. Result of Annotation Human Action using LVQ



Walk **Run** **Jump** **Skip**

P-Jump **Bend** **Jack** **Side**

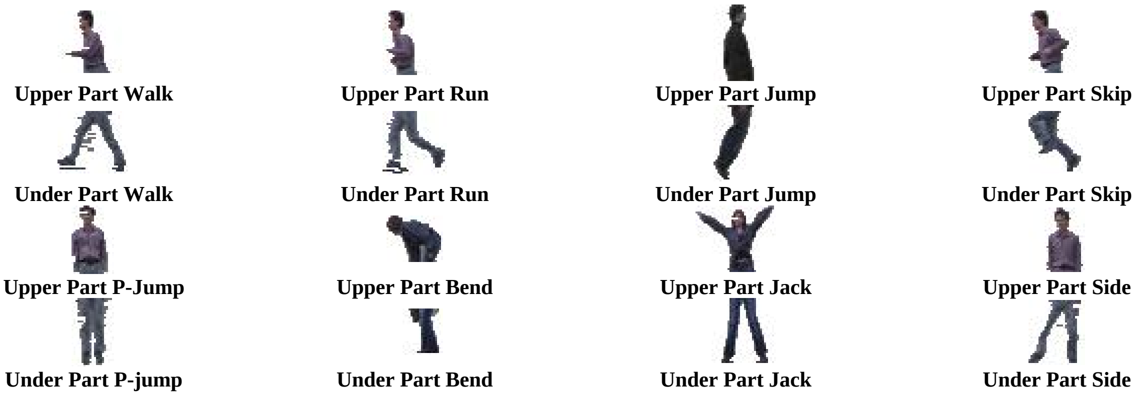
Fig. 6. Human Segmented for Feature Extraction



Walk **Run** **Jump** **Skip**

P-Jump **Bend** **Jack** **Side**

Fig. 7. Human Action from Weizmann Dataset[1]



Upper Part Walk **Upper Part Run** **Upper Part Jump** **Upper Part Skip**

Under Part Walk **Under Part Run** **Under Part Jump** **Under Part Skip**

Upper Part P-Jump **Upper Part Bend** **Upper Part Jack** **Upper Part Side**

Under Part P-jump **Under Part Bend** **Under Part Jack** **Under Part Side**

Fig. 8. Feature Extraction from Human Segmented

TABLE I
CONFUSION MATRIX OF HUMAN ACTION CLASSIFICATION RESULT

		Actual								
		Bend	Jack	Jump	P Jump	Run	Side	Skip	Walk	Class Precision
Prediction	Bend	346	1	7	4	2	4	3	2	93.77%
	Jack	1	603	2	14	2	5	5	16	93.06%
	Jump	7	0	152	0	3	1	2	10	86.86%
	P Jump	1	15	0	232	0	23	0	5	84.06%
	Run	1	1	1	0	256	5	56	37	71.71%
	Side	0	10	0	15	5	322	4	7	88.71%
	Skip	0	2	4	1	92	2	302	46	67.26%
	Walk	1	6	1	2	33	8	30	494	85.91%
	Class Recall	96.92%	94.51%	91.02%	86.57%	65.14%	87.03%	75.12%	80.06%	84.7 %

[1] http://www.ecse.rpi.edu/~cvrl/database/activity_dataset.htm

The action of human can be annotated in various actions. In the same case during the experiment, human walk sideway can be annotated by proposed jumping model. It is shown that the identifiable human action is moving sideways but the annotation is detected as a jumping movement, because the human action clenched legs and jumped sideways in the feature.

In the experiments, the data set are divided in training and testing with a split proportion of 70:30. 70% of the data will be used as training and 30% is used as data testing. The Table I reveals the classification results of experiments conducted using LVQ, resulting in an accuracy of 84.27%. The figure was derived from the total number of TP and TN divided by the amount of data being tested. Value TP and TN are arranged diagonally according to their respective category.

Table I shows that the TP number to the category of Bend is 346 and the number of TN to categories other than Bend action is 2361. The value of TP Bend is the amount of data that is classified correctly as Bend, and the value TN is the amount of data in addition to Bend classified by right in accordance with their respective category.

The number of FP in the category of Bend is 11 and the figure derived from the sum of the values in the column is reduced by the value of TP Bend, the Bend category (357-346 = 11). The FN derived from the number of columns in each category other than the category is reduced by TN Bend in each category so that the FN is 2855-2361 = 494.

V. Conclusion

In this study, an annotation of human action has been detected by applying LVQ and Co-occurrence matrix as feature extraction. It can be noted that co-occurrence matrix feature is able to derive as feature for human action annotation and be learning as classification as vector feature by Learning Vector Quantification. The result has given potential accuracy of 84.27% for recognition human action annotations. Our proposed method is different from other human annotations especially in using features of human objects for classification. In many approach for feature extraction, GLCM is usually used in object except for human feature. So, in my proposed methods, the GLCM features can be the solution for feature extraction in human annotation technique.

References

[1] Bobick A and Davis J, "The recognition of human movement using temporal template," *Pattern Anlysis and Machine Intelligent, IEEE* , vol. 23, no. 3, pp. 257-267, 2001.

[2] G. L. Blank M, Shectman E, Irani M and Basri R, "Actions as space-times shapes," in *Interational Conference Computer Vision, ICCV IEEE*, 2005.

[3] Ke Y, Sukthanka R and Herbert M, "Efficient visual event detection using volumetric features," in *Int Conference Computer Vision, ICCV IEEE*, 2005.

[4] Sheikh Y, Sheikh M and Shah M, "Exploring the space of a human actions," in *Int. Conference Computer Vision, ICCv IEEE*, 2005.

[5] Fathi A and Mori G, "Action recognition by learning mid-level motion features," in *Computer Vision Pattern Recognition CVPR IEEE*, 2008.

[6] Schuldt C, Laptev I and Caputo B, "Recognizing human actions : a local SVM approach," in *Int. Conf Pattern Recognition, ICPR IEEE*, 2004.

[7] Weinland D, Ronfard R and Boyer E, "Free viewpoint action recognition using motion history volumes," *Computer Vision IU*, vol. 104, no. 2-3, pp. 249-257, 2006.

[8] Yilmaz A and Shah M, "Action scetch : a novel action representation," in *Proc. CVPR* 1, 2005.

[9] S. Abburu, "Multi level Semantic Extraction for Cricket Video by Text processing," *International Journal of Engineering Science and Technology*, vol. 2, no. 10, pp. 5377-5384, 2010.

[10] D. Palma, J. Ascenso and F. Pereira, "Automatic text extraction in digital video based on Motion analysis," in *Int. Conf. on Image Analysis and Recognition (ICIAR)*, Porto, 2004.

[11] J. Lu, Y. Tian, Y. Li, Y. Zhang and Z. Lu, "A framework for video event detection using weigthed SVM classifiers," in *Artificial Intelligence and Computational Intelligence, AICI, International Conference*, 2009.

[12] C. Yang and M. Dong , "Region-based image annotation using Asymmetrical Support vector Machine-based Multiple-instance Learning," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* , 2006.

[13] S. Barrat and S. Tabbone, "Classification and Automatic annotation extension of images using Bayesian network," in *SSPR International Conference*, 2008.

[14] T. Gruber, "A translation approach to portable ontology specifications," *Knowledge Acquisition* , vol. 5, no. 2, pp. 199-220, 1993.

[15] B. Vrusias, D. Makris and J. Renno, "A framework for ontology enriched semantic annotation of CCTV Video," in *Eight International workshop on image analysis for multimedia interactive services,IEEE*, 2007.

[16] J. Daugman, "Complete discrete 2-D Gabor Transform by Neural Network for image analysis and compression," *IEEE Trans Accoust Speech SIgnal for Image analysis and Compression*, vol. 36, pp. 1169-1179, 1988.

[17] L. Soh and Tsatsoulis, "Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices," *Geoscience and Remote Sensing, IEEE Transactions* , vol. 37, no. 2, pp. 780-795, 1999.

[18] S. Raja and Shanmugam, "ANN and SVM Based War Scene Classification Using Invariant Moments and GLCM Features," *Machine Learning*, vol. 2, no. 6, pp. 869-873, 2012.

[19] R. Jardon , S. Chaudhurry and K. Biswas, "Generic video classification : An Evolutionary learning based fuzzy theoretic approach," in *Int. Conf. Indian Computer Vision Graphics and Image Processing*, 2002.

[20] A. Dorado, J. Calic and E. Izquierdo, "A Rule-based video annotation System," *IEEE Transactions on Circuits and System for Video Technology*, vol. 14, no. 5, 2004.

[21] M. Detyniecki and C. Marsala, "Automatic Video annotation with forests of Fuzzy Decision Trees," in *Mathware and Soft Computing*, 2000.

[22] M. Hosseini and M. Moghadam, "Fuzzy rule-based reasoning approach for event detection and annotation of broadcast soccer video," *Appl. Soft. Computer* , 2012.

[23] H. Tong, J. He, J. Li, S. Zhang and W. Ma, "Graph Based multi modality learning," in *ACM Multimedia*, Singapore, 2005.

[24] M. Wang , X. Hua, R. Hong , J. Tang and Y. Song, "Unified Video annotation via Multigraph Learning," *IEEE Trans. Circuits Syst. Video Tech.* , vol. 19, no. 5, 2009.

[25] M. Weng and Y. Chuang, "Multi-cue fusion for semantic video indexing," in *ACM Multimedia*, 2008.

[26] M. A. Soeleman, M. Hariadi and M. H. Purnomo, "Adaptive threshold for background subtraction in moving object detection using fuzzy c-means," in *IEEE Tencon Philippine Section*, Philippine, 2012.

[27] P. Spagnolo, T. Orazio, Distante and M. L. A, "Robust foreground segmentation from color video sequence using

background subtraction with multiple threshold," *Journal Image and Vision*, vol. 24, pp. 441-423, 2006.

[28] H. M. Robert, S. K and D. Its'Hak, "Texture Features for Image Classification," *IEEE Transactions On Systems, Man and Cybernetics*, vol. 6, no. 3, pp. 610-621, 1973.

[29] Z. Jian, L. Chuan-Cai, Z. Yue and L. Gui-Fu, "Object recognition using Gabor co-occurrence similarity," *Pattern Recognition, Elsevier*, vol. 46, pp. 434 - 448, 2013.

[30] B. Marcin and D. Włodzisław , "LVQ algorithm with instance weighting for generation of prototype-based rules," *Elsevier, Neural Network*, vol. 24, p. 824–830, 2011.

[31] O. M. Jafar and R. Sivakumar, "Distance Based Hybrid Approach for Cluster Analysis Using Variants of K-means and Evolutionary Algorithm," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 8, no. 11, pp. 1355-1362, 2014.

[32] G. Wenzong and C. Guolong, "Human action recognition via multi-task learning base on spatial-temporal feature," *Information Sciecne, Elsevier*, vol. 320, pp. 418-428, 2015.

[33] A.-A. A. Haiam and H. Elsayed E, "Human action recognition using trajectory-based representation," *Egyptian Informatics Journal, Elsevier*, vol. 16, pp. 187-198, 2015.

[34] S. Manel, M. Mahmoud and A. B. Chokri, "Human action recognition based on multi-layer Fisher vector encoding method," *Pattern Recognition Letters, Elsevier*, vol. 65, pp. 37-43, 2015.

[35] N. Jalal A, "Energy-based model of least squares twin Support Vector Machines for human action recognition," *Signal Processing, Elsevier*, vol. 104, pp. 248-257, 2014.

[36] K. Villi, Z. Guoying and P. Matti, "Recognition of human actions using texture descriptors," *Machine Vision and Application, Springer*, pp. 26-39, 2009.

[37] M. Mona M, H. Elsayed, F. Magda B and E. N. Heba A, "An enhanced method for human action recognition," *Journal of Advanced Research, Elsevier*, vol. 6, pp. 163-169, 2015.

[38] Csurka G, Dance C, Fan L, Willamowski J and Bray C, "Visual categorization with bags of key points," *in ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.

Authors' information

[1]Computer Science Department, Dian Nuswantoro University,
Semarang, Indonesia.

[2]Electrical Engineering Department,
InstitutTeknologiSepuluhNopember, Surabaya, Indonesia.

[3]School of Media Science, Tokyo University of Technology, Japan.

M. A. Soeleman received bachelor and master degree in 1999 and 2004 respectively from Computer Science Department Dian Nuswantoro University, Semarang, Indonesia. Since 2010, he has studying at the Electrical Engineering of Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia as a doctoral student. He works as a lecturer of Computer Science Faculty, Dian Nuswantoro University, Semarang, Indonesia. His research interest includes image processing, computer vision and video processing.

Mochamad Hariadi graduated from Electrical Engineering Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia for his bachelor in 1995 and received MSc and PhD degree in 2003 and 2006 respectively from Graduate School of Information Science (GSIS) Tohuku University, Japan. He is currently lecturer of Electrical Engineering Department ITS. His research interests include multimedia processing and artificial intelligence. He is IEEE member and IEICE member.

Eko Mulyanto Yuniarno graduated from Electrical Engineering Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia for his bachelor in 1995 and received his Master and Doctor degree in 2005 and 2013 respectively from Graduate School of Electrical Engineering ITS Surabaya. He is currently lecturer of Electrical Engineering Department ITS. His research interests include computer vision, image processing and multimedia processing.

Mauridhi Hery Purnomo received bachelor degree from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia in 1985. He received his M.Eng and Ph.D degree from Osaka City University in 1995 and 1997 respectively. He joined ITS in 1985 and has been a Professor since 2004. His research interests include intelligent system application, electric power system operation, control and management. He is an IEEE member

Masanori Kakimoto earned his bachelor and Ph.D. degrees from The University of Tokyo, Tokyo, Japan in 1982 and 2005 respectively. He joined Schoolof Media Science,Tokyo University of Technology, Hachioji, and Tokyo, Japan and has been a Professor since 2012. His current interests include Computer Graphics, Visual Simulation, and Visualization. He is a member an ACM SIGGRAPH