Person Recognition from Activity using Bag of Words

Vidhya.V.S.Nair, Subha V

Abstract— ***In this paper the discriminant pattern hidden in the way of doing an activity for every person is explored. This pattern can be utilized for person recognition purpose in uncontrolled scenarios unlike finger print, iris, retina etc. (based on physical biometrics). This method is based on single video camera based data. From the video of various activities, background subtraction is done to remove insignificant data. From the binary video obtained after background subtraction structural tensor based features are detected and extracted. The extracted features defines the variation from the mean position are then clustered by means of k-means clustering. Histogram of cluster centroids is calculated using Bag Of Words (BOW) and classified by category classifier. Histogram of input video action sequence is compared with each of dataset and predicts the category, which corresponds to the label of person.***

Index Terms**:** ***Activity based identification, Background subtraction, Silhouette, Structural Tensor, Bag Of Words, Category classifier, Structured Support Vector Machine.***

I. INTRODUCTION

Traditional person recognition system is based on identity card or smart card or password, but the rate of missing or fraudulence is high. Then further researchers explores the unique pattern exists in biologically, that is biometric features. Biometric features are classified into physical biometric features like fingerprint, iris, palm prints or DNA [1] and behavioural biometric features like gait, keyboard typing or voice [2]. Behavioural biometrics are used in some circumstances where person is supposed to do some particular action like typing, talking and walking. Use of biometric data reduce the risk factor with respect to the traditional credentials, because its based on the features each person posses in a unique way. In gait based person recognition system [3], it works by identifying the unique pattern or style hidden while one walks. Different from other biometric patterns, gait based features can be easily acquired even from an uncooperative subject. Features like trajectories of lower body joint angles translational and angular velocities are utilized in gait based recognition. Main disadvantage of gait based person recognition is the case if the subject is supposed to do some other activity other than walking like jump, bend or jogging. In such cases gait features must be extended to suit any activity. Like walking every activity possess a discriminant pattern, which will be different among persons. Activity based person recognition is an emerging field of research based on which only a few methods has been proposed yet.

Manuscript published on 30 June 2016.

***** Correspondence Author (s)
Vidhya.V.S.Nair, Department of Electronics and Communication, SCT College of Engineering, Pappanamcode, India.
Subha V, Department of Electronics and Communication, SCT College of Engineering, Pappanamcode, India.

Activity based person recognition can be considered as an extension of gait based person recognition, by extending leg features to whole body features.

In activity based person recognition full posture of subject in each frame is utilized for recognition. Two approaches used for activity based recognition are: Video sequence based [4] and Motion capture data based [5] methods. In video sequence based person recognition system a multi camera setup is used to capture features from different angles. Features from N cameras are concatenated to produce posture vectors and then by means of clustering D representative human body pose prototypes (Dynemes) are formed. To map the features to a discriminant space, fuzzy similarities between posture vector and dynemes are calculated. In identification phase test video is mapped to same discriminant space and calculates similarity for recognition. In some paper multi camera based data are used to construct 3D structures.

Motion capture system can be classified into two: Marker based and marker-less system. Marker based system is suitable for conscious case (cooperative subject), where motion features are extracted based on the optical flow of reflective markers attached to skin to identify bone landmarks. Markers can be active or passive. Passive optical system use markers coated with a retro-reflective material to reflect light that is generated near the cameras lens. The camera's threshold can be adjusted so only the bright reflective markers will be sampled, ignoring skin and fabric. Active optical systems triangulate positions by illuminating one LED at a time very quickly or multiple LEDs with software to identify them by their relative positions, somewhat akin to celestial navigation. One way of the marker-less method of extracting features is the use of kinect. Kinect is a line of motion sensing input device with depth sensor consists of infra-red laser. Both

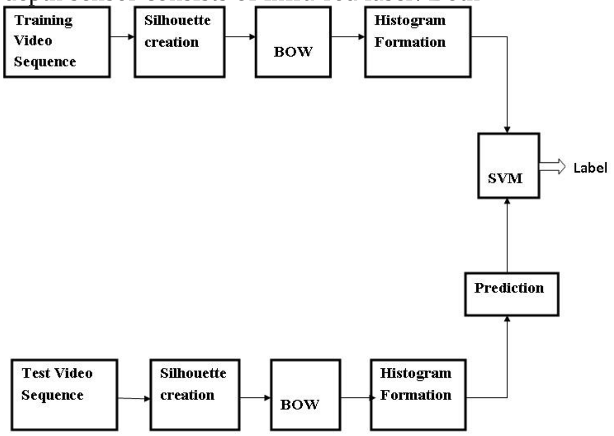

Figure 1: Overview Of Activity Based Person Recognition

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

system works by placing a skeleton model above the corresponding measurements. Each joint angles are represented as quaternions. In one method a correspondence matrix is constructed by comparing the input quaternion set with each trained quaternion set in database. Based on the nature of the correspondence matrix (position of unity) label of input feature is given. In the second method, Locality Preserving Projections (LPP) is applied before dyneme calculation as pre-processing. Bag Of Words (BOW) is applied to the codebook of pose vector and then classified using Structured SVM.

Marker-less motion capture techniques have gained great interest in the field of feature extraction, but due to limited number of methods to explore repeatability of joint kinematics it is not much used. Main disadvantage of these existing methods are system complexity.

Specific hardware and special software programs are required to obtain the motion data. The proposed method is based on single camera setup and a data processing software which is much simpler than other existing methods. In the training phase posture vectors of each activity is extracted by applying background subtraction, assuming a constant background. Frame level processing is done to detect and extract structural tensors among frames. From this bulk feature space, using k-means clustering features are clustered to k centroids and then Bag Of Words(BOW) is applied to obtain the histogram of each feature centroid. Creation of histogram is the discriminant space and further classified using category classifier. In the identification phase, histogram of k visual words are constructed from the input video and then predict the nearest classifier. Label of person is given based on the prediction.

II. PROPOSED METHOD

The proposed method of activity based person recognition can be subdivided into 3 parts: Feature detection and extraction, Bag Of Words creation and Classification. Person under analysis is free to perform a number of activities included in the action data set. Fig.1 depicts the overall flow of proposed method.

A. *Feature detection and extraction*

The dataset and input to the system is colour video sequence. Let $X_{is} = [x_1, x_2, \ldots \ldots \ldots, x_n]$ be ith training video sequence with n frames. Video sequence contains unnecessary things and background details. To remove all these unnecessary data an image segmentation technique, background subtraction [6] is applied to every frames. Background subtraction can only be applies where the background is known. Thus after subtracting the known background from each frames, a thresholding is done to obtain a binary sequence. Thereby the resulting silhouette possesses a dark background with white moving subject. To reduce the memory usage the region of interest based on moving body is cropped with respect to centre of mass. Fig. 2 depicts the silhouette of walking action. Silhouette vector of each frame is considered as instant posture vector I_{ij} , where it represents the binary posture vector of i^{th} person in j^{th}frame. Next step is the detection and extraction of salient features using structure tensor. Structure tensor is second moment matrix, to represent the gradient of a function. It also has a more powerful description of local patterns as opposed to the directional derivative through its coherence measure.

The term ”tensor” refers to a representation of an array of data. For a function I of two variables (x,y), the structure tensor is the 2 * 2 matrix :

$$S_w[p] = \begin{bmatrix} \sum_r w[r](I_x[p-r])^2 & \sum_r w[r]I_x[p-r]I_y[p-r] \\ \sum_r w[r]I_x[p-r]I_y[p-r] & \sum_r w[r](I_y[p-r])^2 \end{bmatrix} \quad (1)$$

Where I_x and I_y are the partial derivatives of I with respect to x and y; the integrals range over the plane R^2; and w is some fixed ”window function”, a distribution on two variables.

Note that the matrix S_w is itself a function of (x,y). Eigen-decomposition is then applied to the structure tensor matrix ’S’ to form the eigenvalues and eigenvectors (L_1, L_2) and (e_1, e_2) respectively, IT summarize the distribution of the gradient $\nabla I = (I_x, I_y)$ of I within the window defined by w .These new gradient features allow a more precise description of the local gradient characteristics. To map this features to a discriminant space Bag Of Features model is applied.

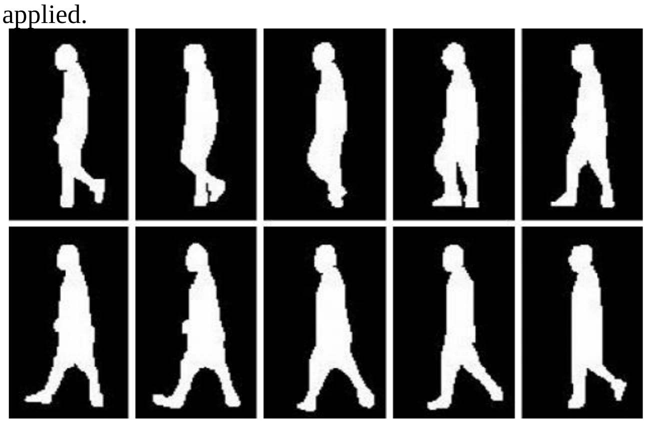

Figure 2: Near-field video: Example of walking action

B. *Bag Of Words*

The concept of Bag of Visual Words is used to describe the visual content or visual features of frames by visual description of all regions of interest [12] and represent it in a histogram model. Simply Bag Of Words model helps to describe each features and simply name it using of alphabets or numbers. Resulting feature space is a bulk data which reduce the complexity of system. Thus k-means clustering is applied to this bulk data as a part of quantization, results k number of centroids. The set of visual words are called visual vocabulary. The main purpose of clustering is quantization of feature space which is helpful in creating histogram. Additionally, the bag Of Words object provides an encode method for counting the visual word occurrences in an image. It produced a histogram that becomes a new and reduced representation of an image. and is represented as $H_i = [h_{1i}, h_{2i}, \ldots \ldots h_{ki}]$.

The histogram length corresponds to the number of visual words that the Bag Of Words object constructed. The histogram becomes a feature vector for the image. The main advantages of the BOW representation are: (1) its compactness, i.e. reduced storage requirements, and (2) the rapidity of search due to an inverted file system.

Encoded training images from each category are fed into a classifier training process invoked by the Train Image Category Classifier function and this function relies on the Structured SVM classifier. Structured Support Vector Machines are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis.

For a set of ` training instances, $(x_n, y_n) \in H \times Y$, n = 1,2,....l from a sample space H and label space Y , the structured SVM minimizes the following regularized risk function.

$$\min_{w} \quad \|w\|^2 + C \sum_{n=1}^{\ell} \max_{y \in \mathcal{Y}} \left(\Delta(y_n, y) + w'\Psi(x_n, y) - w'\Psi(x_n, y_n) \right)$$

At test time, only a sample $x \in H_t$ is known, and a prediction function $f: H_t \to Y$ maps it to a predicted label from the label space Y . For structured SVMs, given the vector w obtained from

training, the prediction function is the following.

$$f(x) = \operatorname{argmax}_{y \in Y} \; w'\varphi(x, y) \qquad (2)$$

Therefore, the maximizer over the label space is the predicted label.

III. EXPERIMENTAL RESULTS

We evaluate the proposed activity based person recognition method using KTH video dataset. In our experiments we have used the videos of 6 persons depicting performing four actions, i.e., walk, jog , jump forward and wave one hand. In order to remove the background effect background subtraction is applied to the colour video frames by assigning first frame as known background. Thereby silhouette video (binary frames) is obtained. Then subdivide each video sequences to 20 instances. The algorithm has been trained by using 10 instances of each action class and tested by using the remaining action instances. This procedure has been repeated multiple times (folds), one for each set of test action instances, in order to complete an experiment.

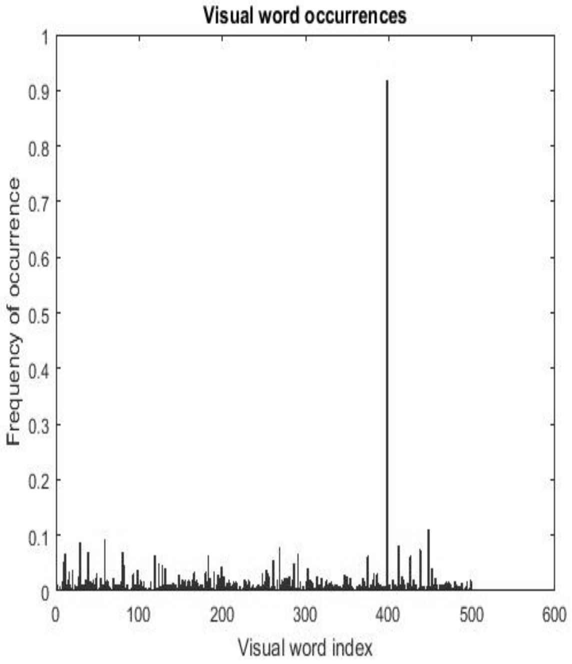

Figure 3: Histogram of Person 1 While walking

Extracted structural tensors using the Grid selection method. The Grid Step is [8 8] and the Block Width is [32 64 96 128]. 47694 structural tensor based features are extracted . Regarding the value of k, optimal recognition is resulted for k = 500. Table I shows the recognition rate with the value of k. Histogram plot of two persons performing the same.

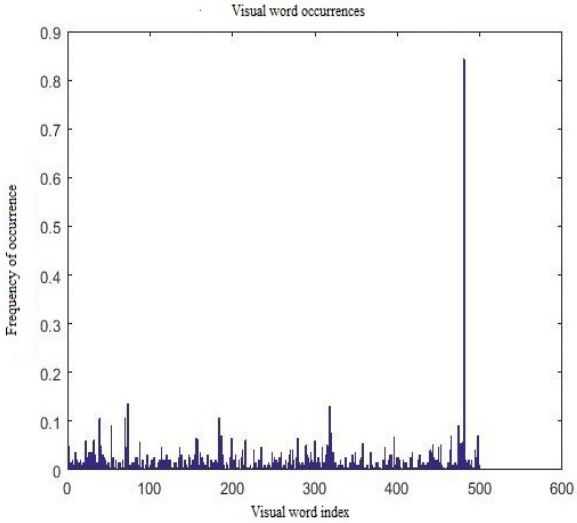

Figure 4: Histogram of Person 2 While walking

Table I

k	Recognition rate
200	0.86
300	0.88
400	0.91
500	0.93

action, walking, is shown in figure 3 and figure 4. The plot clearly depicts the discriminant nature of histogram feature space.

IV. CONCLUSIONS

Key feature of the proposed method is its simplicity based on a single video camera. Existing methods of activity based person recognition needs extra system requirements and additional hardwares like kinect sensor and multiple cameras. In the proposed system recognition is done using the structural tensor extracted through a single camera video. Thus human movements other than walking (gait) can be successfully used for person recognition using a simple system image processing sysem. Bag Of words creation along with k- means clustering makes the system scale invariant, thus recognition rate is independent of camera calibration to a medium extend.

ACKNOWLEDGEMENT

We would like to thank reviewers for their valuable and helful suggestions. We would also like to thank the faculties, the resource providers and the system administrator for the useful feedback and constant support.

REFERENCES

1. Jain, A. Ross, S. Prabhakar,"An introduction to biometric recognition", IEEE Trans. Circuits Syst. Video Tech., vol. 14,pp. 420,2004.
2. R.V. Yampolskiy, V. Govindaraju,"Behavioural biometrics: a survey and classification", Int. J. Biom.,pp. 81113, 2008
3. R. Tanawongsuwan, A. Bobick,"Gait recognition from time-normalized jointangle trajectories in the walking plane," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR,pp. II- 726II-731.,2001.
4. Iosifidis, Anastasios Tefas and Ioannis Pitas, " Person Identification From Actions Based On Dynemes And Discriminant Learning",IEEE vol.978, No.1, pp. 4673- 4989,2013
5. Eftychia Fotiadou and Nikos Nikolaidis , "Activity-based methods for person recognition in motion capture sequences," Pattern Recognition Letters , vol.49 ,pp.4854, 2014
6. Y. Benezeth, P. Jodoin, B. Emile, H. Laurent, and C. Rosenberger, "Review and evaluation of commonly implemented background subtraction algorithms", IEEE International Conference on Pattern Recognition, pp. 14,2008.
7. H. Bay, A. Ess, T. Tuytelaars and L.Van Gool, "SURF: Speeded Up Robust Features.", Computer Vision and Image Understanding,vol.110, no.3, pp.346-359, 2008.
8. J. Wang, M. She, S. Nahavandi, and A. Kouzani,"A review of vision-based gait recognition methods for human identification," International Conference on Digital Image Computing: Techniques and Applications, pp.320327, 2010.
9. D.A.R. Vigo, F.S. Khan, J. van de Weijer, T. Gevers,"The Impact of Color on Bag-of-Words Based Object Recognition,", International Conference on Pattern Recognition (ICPR), pp. 1549 - 1553 ,2010.
10. CJC. Burges,"A tutorial on support vector machines for pattern recognition," Data Mining Knowledge Discovery, vol. 2, no. 2, pp. 121167, 1998.
11. S. Das, R. Wilson, M. Lazarewicz, L. Finkel,"Gait recognition by two-stage principal component analysis," 7th International Conference on Automatic Face and Gesture Recognition, pp. 579584, 2006.
12. D. Gokalp and S. Aksoy," Scene classification using bag-of-regions representations", IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07), pp. 18, June 2007.