Recognizing Human Actions by Local Space Time and LS-TSVM over CUDA

Mohsin Raza Siyal, Muhammad Saeed, Jibran R. Khan and Farhan A. Siddiqui
Department of Computer Science,
University of Karachi, Pakistan

Kamran Ahsan
Department of Computer Science,
Federal Urdu University of Arts, Science and Technology,
Pakistan

***Abstract*—Local space-time features can be used to make the events adapted to the velocity of moving patterns, size of the object and the frequency in captured video. This paper purposed the new implementation approach of Human Action Reorganization (HAR) using Compute Unified Device Architecture (CUDA). Initially, local space-time features extracted from the customized dataset of videos. The video features are extracted by utilizing the Histogram of Optical Flow (HOF) and Harris detector algorithm descriptor. A new extended version of SVM classifier which is four time faster and has better precision than classical SVM known as the Least Square Twin SVM (LS-TSVM); a binary classifier which use two non-parallel hyperplanes, is applied on extracted video features. Paper evaluates the LS-TSVM performance on the customized data and experimental result showed the significant improvements.**

Keywords*—Motion detection; human action recognition; LS-TSVM; GPU Programming; Compute Unified Device Architecture (CUDA)

I. Introduction

Action recognition from video objective is to recognize the action and goal by analyzing the series of frames and their relationship that define the classification of the action. The pose estimation is an element of computer vision to transform 3D looking objects into 2D images from the video feeds to detect the corners and edges by using free-form contours [1]. Mostly it uses multiple methods and combine them consecutively to prevent the limitations of each. The pose assessment and action reorganization, both are vital elements of vision based human motion understanding. They are used in many applications like intelligent surveillance system, learning humanly moves in games and human interaction with computer systems [2], [3].

Another very important use of action recognition and pose estimation could be the storage of video as an abstract data like the human brain does. Harris detector algorithm is very efficient approach for corner detection in image processing. Motivation for the Harris detection is matching problem during the motion of pictures, pitch problem to find the best patch from first image to second. Histogram of Optical Flow (HOF) is used to detect edges from the images, it also supports the gradient structure which has property of photometric transformation, human detection, local shape, relatively invariant to local geometric transformation coarse spatial sampling and fine orientation sampling works best [4], [5].

Support Vector Machines (SVM) is used to perform classification in a nonlinear manner. It can also be worded as function estimation, with the optimization of convex accompanied by the primal-dual interpretation and distinctive solution [6]. Whereas, the LS-TSVM can be used as both linear and nonlinear classification including function estimation, solving linear systems, regularization networks, link with Gaussian processes and valid in primal-dual optimization formulations and high dimensional input spaces [7], [8], kernel versions of Fisher Discriminant Analysis (FDA) and Sparse approximation and robust regression [9].

Finally, LS-TSVM uses two non-parallel hyperplanes in such a manner that each of the hyperplane is close to the one of the other classes and leaves the existing concurrently. TWSVM proves itself four times faster than a normal SVM by solving two smaller size two smaller-sized Quadratic Permutation Polynomial (QPPs). SVM and TWSVM are initially developed for solving the binary classification problems. Yet, classification of the multi-class problems is usually come across in real-world situations. That's why extension to multi-class classification problems from classical SVM and TWSVM are still ongoing research. Nevertheless, there are two serious problems in SVM for multi-class classification problems. One is the how fast a machine can learn a model and other is methods for handling potential unbalance of samples in dissimilar classes. For two different classes, the purposed LS-TSVM method, solves the unbalance problem by using different variable. Henceforth, solving linear equation system, enhanced the model learning speed and turn out to be faster. On the basis of this analysis, the paper aims to expand from SVM to LS-TSM in HAR. Linux system with GPU installed and programming was done over CUDA to increase the algorithm performance.

II. Representation

To represent the corner detection paper uses Harris detection method which uses a gradient formulation to detect response at any shift (x,y) [5].

$$E(u,v) = \sum_{x,y} w(x,y)[I(x+u,y+v) - I(x,y)]^2 \quad (1)$$

If E(u,v) is close constant patches, it will be near 0. E(u,v) will be higher provided unique patches. It is clear that E(u,v) should be higher. In this work, bilinear approximation for small shifts [u, v] is used and is shown below.

$$E(u,v) \cong [u,v]M\begin{bmatrix}u\\v\end{bmatrix} \quad (2)$$

In the above equation, M is 2×2 matrix that is calculated by following image derivations equation:

$$M = \sum_{x,y} w(x,y)\begin{bmatrix}I_x^2 & I_xI_y\\ I_xI_y & I_y^2\end{bmatrix} \quad (3)$$

Calculating a weighted sum (simple case, w=1) which is windowing function where Ix, Iy are the product of components of gradient and the calculating the corner response by:

Measure or corner response:

$$R = detM - k(traceM)^2$$

$$detM = \lambda_1\lambda_2$$

$$traceM = \lambda_1+\lambda_2 \quad (4)$$

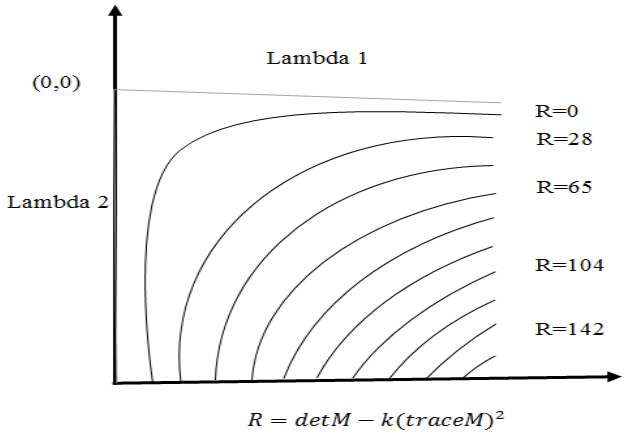

Fig. 1. Corner response map.

The 'k' is empirically defined constant, whose value is k = 0.04-0.06

'R' only depends on eigenvalues of M, for a corner R is higher, with higher edge magnitude R is negative and for a flat region |R| is small [3], [5] as shown in Fig. 1.

Joint angles are parsed out from the captured video feeds and all the theoretical methods of action recognition are applied to them. There are so many challenges are involved while applying such approaches to the video feeds including accurately and precisely detecting and extracting joints, tracking the joints with limitation of visual, variations in size, scale, pose etc. In the field of object reorganization, the paper suggests the idea of using optical flow in motion sequence is very much efficient, based on the research and successful feature histogram results extraction. Yet, as it is known that the size of the descriptor or the number of pixels in person varies eventually. Also, there are some issues involve in using optical flow to minimize the background noise of the image and in computation as well, abnormality in scale changes, problem with direction of motion. To prevent these problems, the optical flow distribution is used. It is obvious that when the object moves with a fixed background in scene, it creates a very specific profile of optical flow. For example, a sample for waving hand sequence depicts optical flow patterns; the optical flow profile will be different at different scale of same motion or activity such as zoom-in and zoom-out. In case of zoomed out the magnitude of OF vector would be smaller and vice versa. Likewise, if the waving person direction changes, the OF examined would be an image in the vertical axis to that examined. Therefore, based on optical flow, work has computed the feature that depicts the activity profile at every instance of time, which does not affect by the change of scale or direction of movement [10], [11].

Work uses local space time feature [12] to handle the moments of non-constant motion by primitive events belonging to progressive two dimensional images.

Authors build its scale-space representation $L(\cdot, \sigma^2, \tau^2) = f * g(\cdot, \sigma^2, \tau^2)$ to find the local features in a sequence of images f(x,y,t). this paper uses Gaussian convolution kernel $g = \exp(-(x^2 + y^2)/2\sigma^2 - t^2/2\tau^2)/\sqrt{(2\pi)^3\sigma^4\tau^2}$ for gradients of spatiotemporal image representation $\nabla L = (L_x, L_y, L_t)^T$ is used to computer second moment matrix[13].

$$\mu(\cdot; \sigma^2, \tau^2) = g(\cdot; s\sigma^2, s\tau^2) * \nabla L(\nabla L)^T \quad (5)$$

where in order to allocate position of feature using the local maxima of $H = \det(\mu) - k\ \mathrm{trace}^3(\mu)$ over (x, y, t).

In space and time, Gaussian kernel associated spatial and temporal scale parameters (σ, τ), are used to define spatiotemporal feature neighborhood. By the help of automatically selecting scales parameters (σ, τ), it is feasible to adopt the feature size to match the spatiotemporal [14], [15] level of original image structure.

Also, shape of the feature can be varied according to the speed of local patterns which makes the feature more steady and stable along with the use of dissimilar number of camera motions [16]. In order to gain scale invariance, ineffectiveness of velocity of the camera motion, paper uses both of these methods.

III. Classification Least Square Twin Support Vector Machine

LS-TSVM is four time faster and has better precision than classical SVM in binary classification. LS-TSVM uses the classical SVM and Twin SVM to prevent the limitations of each by using two non-parallel hyperplanes.

A system of linear equations can be used to solve:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\} \quad (6)$$

Where the R is the n-dimensional real space containing the x and y the i^{th} data sample and $y_i \in \{+1, -1\}$ is the class label. Likewise number of patterns are 'l'.

Decision function to classify the patterns used by SVM:

$$f(x) = sgn((w.x) + b) \quad (7)$$

SVM uses hyper-plane to separate pattern of two classes, illustrated in Fig. 2.

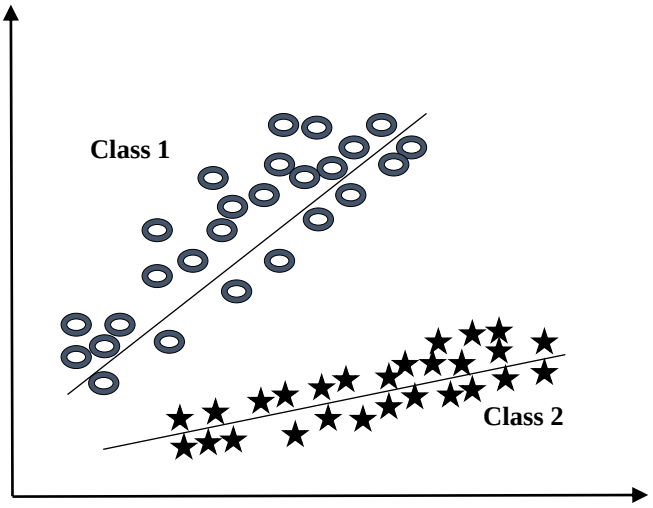

Fig. 2. Geometric representation of binary support vector machine.

The equation of hyper-plane is:

w.x=b=0

Following are the plans in with above hyper-plan lies:

$$w^T.x + b = 1 \; and \; w^T.x + b = -1 \tag{8}$$

Here, R is the normal vector in n-dimensional Real Space and $b \; \epsilon \; R$ is a bias term. To find R SVM solves QPP:

$$min_{w,b,\xi} \qquad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi i \tag{9}$$

$$y_1\big((w.x_i) + b\big) \geq 1 - \varepsilon_i \; and \; \xi_i \geq 0 \tag{10}$$

Where C>0 represents slack variables and ε_i is the penalty parameter similarly i = 1 … l. how much the data sample is misclassified is defined by the slack variable and QPP mentioned above is solved using the dual form. SVM dual formulation changes according to the amount of patterns in the dataset. Complexity for the 1 training pattern is $O(1^3)$ [6].

In order to perform classification of the patterns of two classes Twin SVM uses below mentioned decision function

$$f(x) = argmin_{i=1,2}\frac{|w_i.x+b_i|}{\|w_i\|} \tag{11}$$

By optimization of a pair of QPP TWSVM can attain two non-parallel hyper-planes in order to execute classification task. QPPs are:

$$\min(w_1, b_1, \varepsilon) \qquad \frac{1}{2}\|x_1w_1 + e_1b_1\|^2 + c_1e_2^T\xi$$

$$\text{s.t. } (x_2w_1 + e_2b_1) + \xi \geq e_2, \xi \geq 0 \tag{12}$$

$$\min(w_2, b_2, \eta) \qquad \frac{1}{2}\|x_2w_2 + e_2b_2\|^2 + c_2e_1^T\eta$$

$$\text{s.t.} \qquad (x_1w_2 + e_1b_2) + \eta \geq e_1, \eta \geq 0 \tag{13}$$

where patterns of positive c1 comes from matrices $x_1 \in R^{l_1*n}$ and negative c2 from $x_2 \in R^{l_2*n}$ and know that c1 and c2 > 0 which represents the penalty parameters for misclassification of the data sample.

Two hyperplanes defined by Twin SVM which are not parallel in n-dimensional space is as follow

$$x^Tw_1 + b_1 = 0 \text{ and } x^Tw_2 + b_2 = 0 \tag{14}$$

In order to solve smaller size QPPs, Twin SVM used the pattern of one class to provide its' constraints. Where the complexity of the Twin SVM is O(2x(1/2)3) provided that number of patterns in both classes is almost 1/2. Hence, by the above Fig. 3, it is proved that the Twin SVM is 4x speedy than simple SVM.

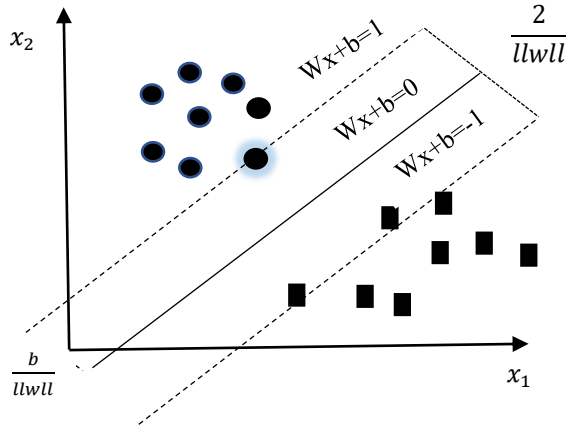

Fig. 3. Geometric representation of binary twin support vector machine.

IV. Experiments

First it detected the video feed and converted each frame into the grayscale. Then converted grayscale images into the threshold values to start analyzing image data with Harris detection and HOF to take care of the video changing effects. After that, apply two methods for motion detection and recognition one is LS-TSVM which works mutually with motion descriptor called local features (LF) and Optical Histogram local feature. Then, paper compares both methods for performance evaluation to different approaches for classification.

V. Method

This work enhanced algorithm performance by using CUDA programming, with the high-performance GPU with 8 core CPU and 12 GB of RAM.

Harris Detection method is used to detect edges using the intersection of two edges and point of intersection represents direction of change in two edges. In order to detect it, high distinction of gradient of the image play a vital role. HOF method is used to detect edges from each fame of video [9], it also supports the gradient structure which has property of photometric transformation, human detection, local shape, relatively invariant to local geometric transformation coarse spatial sampling and fine orientation sampling works best. Finally, LS-SVM is used for classification of the actions. It uses the both linear and nonlinear classification and function estimation. It improves the performance and efficiency by providing the two hyperplane and finding the least distance between each hyperplane for classification.

VI. Results

Program can detect human activities including hand waving, clapping and walking with high accuracy. CUDA programming is implemented over the GPU to enhance the rendering speed of the image processing algorithms.

Experiments results are show in the Fig. 4. In part (a) images show recognition of the hand waving, in part (b) images show the recognition of clapping and in part (c) images show the detection of human walking activity. This work has implemented different methods to perform recognition like simple SVM but LS-SVM proved to be the best.

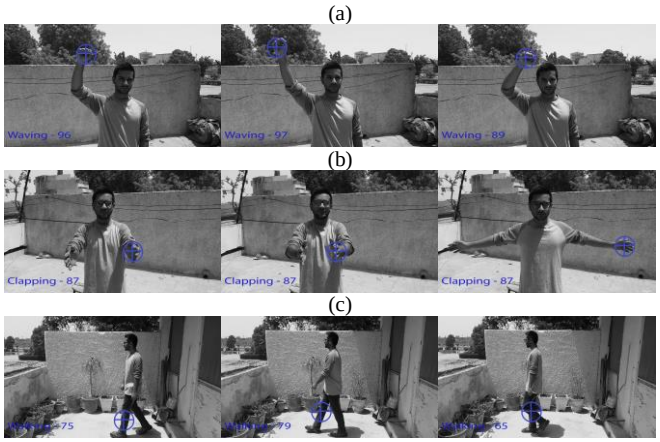

Fig. 4. Demonstration of human activity recognition.

VII. Conclusion

In the field of human action recognition and pose estimation, paper has demonstrated the LS-TSVM feature by analyzing the motion patterns over CUDA. This paper has implemented a novel method over GPU for action recognition using the both methods motion descriptor term as Local feature and Histogram Local Feature with LS-TSVM which proves much efficient and effective than the other approaches. In order to evaluate, it uses customized video dataset in human action recognition system.

References

[1] T. B. Moeslund and E. Granum, "A Survey of Computer Vision-Based Human Motion Capture," Comput. Vis. Image Underst., vol. 81, pp. 231–268, 2001.

[2] J. K. Aggarwal and Q. Cai, "Human Motion Analysis: A Review," Comput. Vis. Image Underst., vol. 73, no. 3, pp. 428–440, 1999.

[3] X. Q. H. Wang, G. Tan, "An Adaptive Corner Detector Based on Curvature Scale Space," Comput. Technol. Autom., 2007.

[4] T. J. XU Xian-feng, "An Improved Multi-scale Harris Feature Point Detection Method," Comput. Eng., vol. 38, no. 17, pp. 174–177, 2012.

[5] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," Procedings Alvey Vis. Conf. 1988, pp. 147–151, 1988.

[6] Cristianini Nello and John Shawe-Taylor, An introduction to support Vector Machines: and other kernel-based learning methods. New York, New York, USA: Cambridge University Press, 2000.

[7] J. V. Johan A K Suykens, Tony Van Gestel, Jos De Brabanter, Bart De Moor, Least Squares Support Vector Machines. K U Leuven, Belgium: World Scientific Publishing Co. Pte. Ltd, 2002.

[8] T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, and J. a. K. Suykens, Least Squares Support Vector Machines, vol. 4, no. July. 2002.

[9] M. J. Black and A. D. Jepson, "EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation," Int. J. Comput. Vis., vol. 26, no. 1, pp. 63–84, 1998.

[10] C. Gu, P. Arbeláez, Y. Lin, K. Yu, and J. Malik, "Multi-component Models for Object Detection," in Computer Vision--ECCV 2012, 2012, vol. 7575, pp. 445–458.

[11] N. Razavi, J. Gall, P. Kohli, and L. Van Gool, "Latent Hough Transform for Object Detection," in European Conference on Computer Vision, 2012, pp. 312–325.

[12] Laptev and Lindeberg, "Space-time interest points," in Proceedings Ninth IEEE International Conference on Computer Vision, 2003, pp. 432–439 vol.1.

[13] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., 2004, p. 32–36 Vol.3.

[14] I. Laptev and Institutionen för numerisk analys och datalogi (Stockholm), Local spatio-temporal image features for motion interpretation. 2004.

[15] [15] S. Belongie, C. Fowlkes, F. Chung, and J. Malik, "Spectral Partitioning with Indefinite Kernels Using the Nyström Extension."

[16] [16] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in Proceedings Ninth IEEE International Conference on Computer Vision, 2003, no. October, pp. 726–733.