International Journal of Engineering & Technology, 7 (3) (2018) 1405-1409

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i3.11645
Research paper

Local Temporal Block Difference Pattern for Action Recognition in Surveillance Videos using Tree Based Classifiers

M. Poonkodi[1*] and G. Vadivu[2]

[1]*Department of Computer Science and Engineering, SRM Institute of Science and Technology, Tamilnadu, India.*
[2]*Department of Information Technology, SRM Institute of Science and Technology, Tamilnadu, India.*
[*]*Corresponding author E-mail: poonkodi.m@vdp.srmuniv.ac.in*

Abstract

Intelligent video classification and prediction is a fundamental step towards effective retrieval system. Huge volume of video is available for navigation today and managing such video and prediction of the activity before its completion gains importance in video surveillance, human computer recognition, gesture recognition etc., An eminent Local Temporal Block Difference Pattern (LTBDP) is introduced which enable efficient feature extraction that could be given to Tree Classifiers like Random Forest and REPTree for further prediction. The proposed pattern has been evaluated on UT-interaction dataset which enable researchers to predict ongoing human actions in an efficient manner. Experimental results using LTBDP in Random Forest and REPTree classifier gives 85.6% and 66.45% accuracy respectively.

Keywords: *Action Recognition, Decision Tree, Random Forest, REPTree, Temporal Difference.*

1. Introduction

In the real world Recognition of Human activity is of prime importance which is being carried out effectively in Computer Vision through some distinguished approaches like hierarchical and non hierarchical methods. Hierarchical approaches also called as statistical approach deals with human activity recognition directly from the descriptors gained from the training set. On the other hand non hierarchical, the syntactical approach recognition is done from the sub events. Besides the above said approaches, recognition could also be done using the prior knowledge and the information contained in the activity. The latter approach could be categorized as descriptor based approach. Vital need of human action recognition and human -human interaction has emerged across the globe for diverse applications such as an aid in elderly monitoring, in video surveillance, an alert to prevent suspicious activity and so on. The prime role of human activity recognition deals with classifying videos in the labeled classes. Such action necessarily encounters many challenging issues like finding the number of persons involved in the video, spotting the person involved by using their identification, illumination changes, cluttered video and may compromise even simple movements like "handshake", "hug" considering them as low level actions.

The most primitive definition of human motion presented in Bobick, A. F. (1997) [1] categorize movement as atomic event, activity as continuous movement and action as movements in large scale normally interacting with the environment. Human action recognition (HAR) by Laptev, I. , & Pérez, P,(2007) [2] started to recognize actions (jump, walk, jog, hand wave) in restricted scenarios performed by single person. On the contrary, Turaga, P. , Chellappa, (2008) [3] proposed action as an atomic activity by an individual that would last for short span of time and activity as an intricate progression of actions executed by quite a number of humans interacting with each other. Slowly then dataset compilation were made from Hollywood movies by Laptev, I. , & Pérez, P. (2007) [2] where, more than one individual performed actions even in chaotic environment.

An advancement in human interaction challenged to recognize object and activities like hugging, handshaking which involve multiple person interaction is seen in Ryoo, & Aggarwal,(2009) [4]. Poppe, (2010) [5] espouse a pecking order that categorize basic movement as action primitive, an action to be a series of action primitives that depicts entire body movement and the activity to be numerous actions with complex explication of the movement. Aggarwal & Ryoo, (2011) [6] introduced events and sub events for human action recognition. Sub events are primordial actions which enable main event recognition like "walking", "jogging". For example "leg apart", "one leg forward and other to follow" are all sub events for main event "walking" or "jogging"'. Vishwakarma, & Agrawal, (2013) [7] extended the human action recognition towards surveillance systems that facilitate, to identify the cause for any mishap from the input video obtained from a still static camera.

Human activity recognition using hierarchical and certain reasoning methodologies is discussed by Ziaeefard, M. , & Bergevin, R. (2015) [8]. Semantic feature extraction, recognition of scenes was mostly discussed with its application. Geetha MK, Arunnehru J,(2016) [9] predicted the forthcoming action to be performed by human in its early stages before completion of the activity which may create an responsiveness if the activity is going to be suspicious. This type of human interaction recognition is considered to be a difficult task because it greatly depends on the pace at which the actors perform the action, the position of the camera placement, cluttered surroundings and most important is the involvement of the action representation by few pixels. Currently UT-interation dataset (UTID), by Ryoo and

Aggarwal(2009) [4], is a HIR dataset in which six actions performed in restricted circumstances by actors were recorded using still cameras. Another popular human interaction recognition dataset is TV Human Interactions (TVHID), by Patron-Perezet al. (2010) in which the viewpoint of the camera was unrestricted and any four possible human interaction from well-liked TV shows were hauled out.

Our work is carried out on the exigent UT-Interaction dataset for human activity recognition. Further sections of the paper is organized as follows : Section2 gives the summary of the related works and introduction to tree classifiers,Section3 describes the proposed LTBDP and the remaining two sections Section 4 and 5 elaborates the Experimental results and conclusion .

2. Related work

Action recognition and prediction are emerging in computer vision as it plays vital role in surveillance, early monitoring, health care, gesture recognition, entertainment and many more applications in real time. Such recognition initially started with the study of single person action recognition, slowly it evolved to hit upon two person involvement where actual interaction is vivid. Two person interactions may involve an object–human or human-human interaction. Early stages of human-object recognition comprises concurrent recognition of both human and object in a probabilistic manner by Gupta & Davis (2009) [10], mining interesting and idiosyncratic features by B. Yao, L. Fei-Fei (2010) [11], construction of bag of words and analyzing human parts involvement in performing that action is expressed by V. Delaitre, I. Laptev (2010) [12]. Human interaction study with two person, started with recognition based on the activity trajectory information detailed by A. Datta, M. Shah (2002) [13] for detection of violent activity. Need for study of multiple parts of the body involved in interaction is shown by Park, J.K. Aggarwal (2006) [14]. Further recognition enforced head and face orientation as important information.

Assorted approaches for action recognition in HAR have been proposed by Marín-Jiménez MJ, Yeguas (2013) [16], where Spatio-Temporal Interest points (STIP) are identified to enable the action recognition better.Spatio-Temporal Interest points were extracted from video sequences through Harris3D actions in videos where different types of low- and middle-level features have been used to predict action. Sener F, Ikizler-Cinbis, (2015) [17] elaborated the interaction between two people in action recognition. Tracing interacting persons, acquiring visual descriptors together with the distance between the actors, in the multiple learning process help to recognize the action. Action recognition from the video sequence is identified using Difference Intensity Distance Group Pattern (DIDGP) method and the extracted features are fed into SVM with polynomial and RBF kernel is shown in Geetha MK, Arunnehru J, (2015) [9, 20].

A new approach where the continuous joint parts movement is considered for prediction that represent the data in low dimensional space but with high accuracy is detailed by Victoria Bloom, Vasileios Argyriou, (2017) [15]. Identification of action is categorized in three phases like offline, online and early. Offline prediction is of minimum importance and the need to predict the action before completion is gaining importance in the real world.

3. Delineate of the Work

Human Action Recognition in Computer Vision is an active area of research which enables recognition and extracting meaningful information from complex tasks. Early prediction becomes cumbersome as it is the outcome of a series of actions, involving challenges by occlusion where one image is superimposed over the other thus blocking the identity of its exposure, intensity of lighting, position of camera etc. Our objective is to scrutinize the actions performed by two actors considering the following assumptions.

- Fixed static camera.
- Light source is stable, no shimmering.
- Background may be Contrasting but has high frame rate and resolution.

An action which comprises a number of movements in human body is a difficult task to be recognized and classified without human intervention. This work focuses on a unique identification pattern named Local Temporal Block Difference Pattern (LTBDP) and further action classification using Tree classifiers like Random Forest and REPTree classifiers. To start with ROI (Region of Interest) is extracted from the difference frame and a 5×5 matrix is constructed, this again is divided into four patches P1, P2, P3, P4 each which is 3×3 to identify the useful information in small window. Features obtained by this pattern are normalized using Gaussian convolution that contributes to the recognition of the action using Random Forest and REPTree classifiers. Our work is validated on UT-Interaction dataset comprising six classes of action like kick, punch, handshake, hug, push and point giving an accuracy of 85.6%.

3.1. Random Forest

Random Forest by L. Breiman (2001) [19] is a flexible machine learning algorithm which produces excellent and most expected results even without hyper parameter modification. Random Forest is an ensemble of Decision tree which works good both with regression and classification data. It works well in handling missing data, outlier values, dimensional reduction and few preprocessing steps. Decision tree thus constructed is totally based on random selection of data. In Computer Vision, RF was initiated by Lepetit & Fua (2006) [21] and Ozuysal, Fua & Lepetit (2007) [22]. RF constructs multiple decision trees based on the random selection of data from the training samples. For a training sample of N, the number of subsets is “m” .Decision Tree is grown for each value of m. Features are randomly chosen for every split in a RF. The correlation between the decision trees is reduced by selecting features at random to facilitate better prediction ability. The random tree implementation of WEKA (open source software) utilized for experimental purpose has the number of trees set to 100 and its depth set to 50 after analysis.

3.2. REPTree

Reduced Error Pruning (REP) Tree classifier is one of the fast decision tree algorithm that works on the principle of calculating information gain with entropy .It minimizes error that arises from variance and prunes the tree by using reduced error pruning with back fitting. Missing values are handled as in C4.5 by splitting the corresponding instances into multiple pieces. REPTree introduced by Quinlan J, (1987) applies regression logic to generate multiple trees in different iterations. REPTree algorithm is applied on UT-Interaction dataset and experimental results are tabulated.

4. Proposed Approach

The proposed approach is shown in Fig. 1. Initially the input video is converted to gray scale and further the noise is removed for fine feature generation. All frames are smoothed by Gaussian convolution method with a matrix size of 3×3 for successful feature extraction and classification. The difference image is obtained by finding the difference between two consecutive frame of an input video sequence is discussed in Section 3.1. Though the difference image got hold, exhibits the motion information further steps to spot out the depth information of the action performed is carried out using Harris Corner to identify 10 strong points from the motion information. A 80 dimensional feature vector is extracted from the selected block, which is fed to Tree classifiers for action recognition.

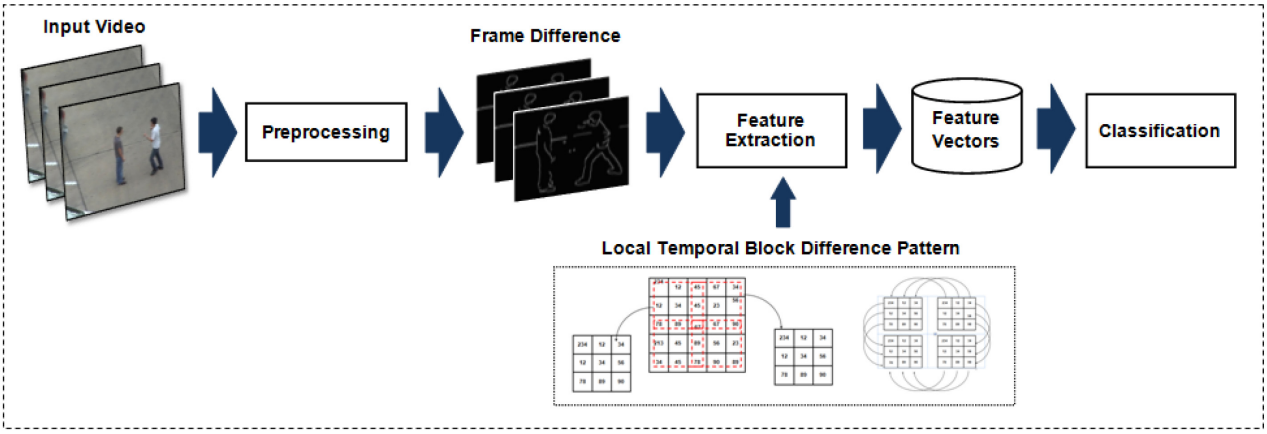

Figure 1: Block diagram of proposed LTBDP method

4.1. Frame Difference

A distinguished method to find the direction of the moving body with respect to still camera is termed as frame difference. Tangible way is to find the difference in displacement between two consecutive frames, where the moving body is considered pixel by pixel for the measurement. Let us consider F_{i-1} to be initial frame and F_n the final one, Frame Difference (FD) is determined as follows

$$FD = |F_{i-1} - F_i|, |F_i - F_{i+1}|, |F_{i+1} - F_{i+2}|, ...|F_{i+k} - F_{n-1}| \quad (1)$$

$$where \;\; 1 < i < k \quad (2)$$

The pixels retrieved in such a way is in its binary format symbolizing pixel value 1 as on the target frame and the ones with zero represent an ideal body. Frame Difference between two consecutive frames is shown in Fig. 2.

4.2. Harris Corner Detection

Computer Vision tasks where stereo and motion information is required to recognize action performed is calculated by considering the intensity values of small size window. Harris Corner (Chris Harris and Stephen Mike, 1988) [18] is a mathematical operator that enables to infer intensity values on corners. $H(x,y)$ gives the displacement from original to moved window and is robust to further changes. The motion information captured due to displacement is shown in Fig. 2.

4.3. Local Temporal Block Difference Pattern (LTBDP)

Motion information from frame difference concentrates on all strong interest points that could be obtained by Harris Corner. Large values of $H(x,y)$ ensures that it is robust to scaling, rotation and illumination variation. Ten such points are considered, and every such point is taken as centre point to form a 5×5 matrix in order to get the depth information in and around the strong interest points. This 5*5 matrix is further divided into four patches P1, P2, P3 and P4 each with a size of 3×3 as shown in Fig. 2. Clear and additional information about the region surrounded by the interest point is calculated which brings out the clear picture of the action description in that state.

4.4. Block Difference Pattern Calculation

The proposed method initially starts by finding the difference relationship between two patches out of the four patches (P1, P2, P3, P4) both in horizontal and vertical manner. As shown in Fig. 3 consider two patches P1 and P2 and compute their adjacent points

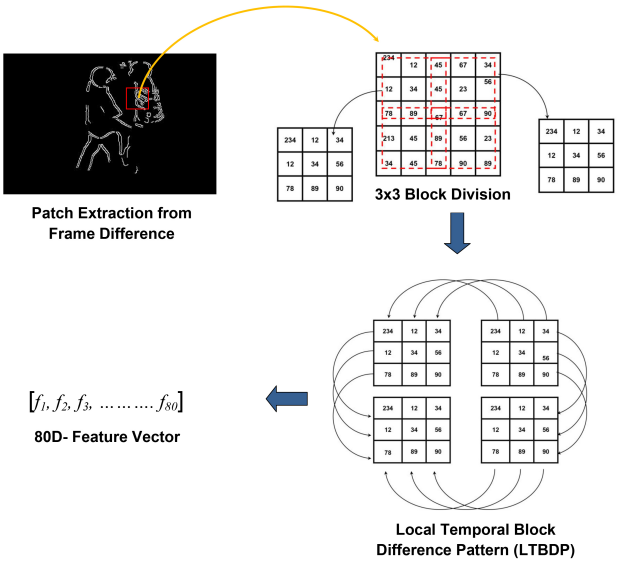

Figure 2: LTBDP feature extration from the difference frame

difference as $P1(x_i,y_j) - P2(x_i,y_j)$ which becomes the difference matrix $Dk(xi,yj)$. Vertical patch difference is calculated in the similar fashion between P1 and P3.Thus the difference calculated between patches form four 3×3 difference matrix $D_k(x,y)$.Thus in general the above 2D feature extraction may be calculated as given in Eqn. 3.

$$D_k(x_i,y_j) = \left| P_k(x_i,y_j) - P_k(x_i,y_j) \right| \quad (3)$$

$$where \; 0 < i < 2; \; 0 < j < 2; \; 1 < k < 3 \quad (4)$$

From every D_k constructed, mean (μ) and standard deviation (σ) is calculated and made as feature vector f_i. Thus for a pixel $H(x,y)$ considered as strong interest points 8 features comprising mean and standard deviation from each D_k is obtained. On an average ten interest points are taken which constitute a 80 dimension feature vector ranging from $f_1, f_2, f_3 ..... f_{80}$ as in Fig. 2.

5. Experimental Results

The experiments are carried out in Python3 with OpenCV3 in Windows10 operating system with Intel i7 Processor 3 GHz with 8 GB RAM. The extracted LTBDP features are modeled for training and testing using WEKA tool. Random forest and RepTree are used for experimental purpose.

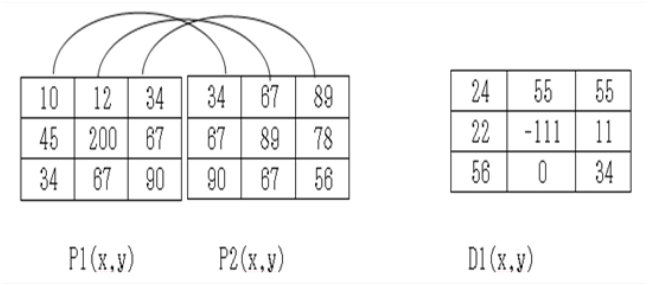

Figure 3: Difference matrix $D_k(x_i, y_j)$ from patches

5.1. UT Interaction Dataset

Ongoing human activities recognition in continuous video is the expected output of the UT Interaction [19] Dataset. This dataset has 6 classes of relentless human-human interactions namely shake-hands, hug, kick, point, punch and push. Two sets of videos set #1 ,set #2 are captured with 30 fps, each with a frame width and length of 720 $\times$ 480 and the action performing human height in the video is around 200 pixels. 10 video sequences for each set is included with an average of 8 executions of activities in each. These videos are recorded with different participants performing the same or all of the actions wearing different clothing in different background, illumination and scaling. Set #1 is recorded in an almost static background with camera jitter less when compared to Set #2. Our work concentrates on all six actions and predicts the action accurately.

5.2. 2D- Features

To start with action recognition, patches are built using proposed LTBDP where interest points are extracted using Harris Corner. Features extracted in this fashion are fed to Random Forest and RepTree classifiers in order to evaluate the performance of the obtained features. The confusion matrix using 2D features of the Tree classifiers like Random Forest and REPTree are shown in Table 1 and 2 respectively to recognize the actions handshake, hug, kick, point, push and punch.

5.3. Quantitative Evaluation

According to the proposed LTBDP feature extraction method 80 dimension features vector is fed to Random Tree and REPTree classifiers. In order to evaluate the performance we calculate Precision (P), Recall (R) and F1 measure.

$$Precision(P) = \frac{TP}{(TP+FP)} \tag{5}$$

$$Recall(R) = \frac{TP}{(TP+FN)} \tag{6}$$

$$F1 = 2P \times \frac{R}{(P+R)} \tag{7}$$

where, TP is the total number of true positive prediction and FP is the total number of false positive prediction for the specific class. Total number of True and False negative predictions of a specific class is given by TN and FN. Tables 1 and 2 shows the results obtained for the extracted features of UT-Interaction dataset fed to RandomForest and REPTree. Performance result shows that the features extracted performed well with RandomForest when compared with REPTree.

5.4. Performance Evaluation

Tables 3 and 4 depicts the statistical measures for the Tree classifier and these measures vivid the limitations of them. The confusion matrices of Tables 1 and 2 were used to calculate the statistics. However Tree classifier produce only satisfactory results on UT-Interaction dataset. Average performance reported using RandomForest in Table 3 shows that Precision(P) = 86.5%, Recall(R) = 85.6%, F1 measure = 85.8%. On the other hand Table 4 enumerates the average performance using REPTree classifier with Precision (P) = 66.11%, Recall (R) = 65.68%, F1-measure = 65.81%. Analyzing the performance metrics of the table action recognition is better illustrated using RandomForest when compared with REPTree classifier.

Table 1: Confusion matrix (%) of UT-Interaction Dataset Using RandomForest classifier (85.64%)

Class	Handshake	Hug	Kick	Point	Punch	Push
Handshake	**83.85**	9.36	0.89	3.83	0.53	1.51
Hug	3.75	**88.5**	0.22	5.1	0.75	1.65
Kick	3.78	8.01	**83.5**	2.29	0.22	2.17
Point	2.19	4.13	0.61	**90.76**	1.14	1.14
Punch	3	7.78	0.44	3.33	**84.09**	1.33
Push	4.49	9.4	0.62	3.76	0.62	**81.08**

Table 2: Confusion matrix (%) of UT-Interaction Dataset Using RepTree classifier (66.06%)

Class	Handshake	Hug	Kick	Point	Punch	Push
Handshake	**68.95**	9.45	4.72	5.97	5.53	5.35
Hug	8.48	**69.27**	4.5	7.21	4.13	6.38
Kick	9.04	8.24	**64.71**	5.38	5.72	6.87
Point	9.67	7.56	5.54	**68.16**	3.43	5.62
Punch	8.12	9.78	5.56	6.67	**63.18**	6.67
Push	9.82	11.38	7.21	6.58	5.32	**59.66**

Table 3: Performance measure (%) of UT-Interaction Dataset using RandomForest classifier

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Handshake	83.9	3.4	84.1	83.9	84.0
Hug	88.5	7.7	75.5	88.5	81.5
Kick	83.5	0.6	96.0	83.5	89.3
Point	90.8	3.8	84.0	90.8	87.2
Punch	84.1	0.7	95.3	84.1	89.4

Table 4: Performance measure (%) of UT-Interaction Dataset using REPTree classifier

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Handshake	69.0	9.0	62.2	69.0	65.4
Hug	69.3	9.2	66.7	69.3	67.9
Kick	64.7	5.4	65.7	64.7	65.2
Point	68.2	6.4	69.9	68.2	69.0
Punch	63.2	4.7	68.8	63.2	65.9

6. Conclusion and Future work

The method presented above is used for action recognition in video surveillance using the proposed 80 dimension feature extraction LTBDP pattern. UT Interaction Dataset are used for action recognition like handshake, hug, kick, point, punch, push. The interest points are extracted from the difference image and further depth information is obtained from our proposed feature extraction method.

Tree classifiers like Random Forest and REPTree are used for further process of classification. Experiments results show that the system gives an accuracy of 85.64% with RandomForest.

References

[1] Bobick, Aaron F. "Movement, activity and action: the role of knowledge in the perception of motion." *Philosophical Transactions of the Royal Society B: Biological Sciences* Vol.352, No. 1358 (1997), pp. 1257-1265.

[2] Laptev, Ivan, and Patrick Pérez. "Retrieving actions in movies." *In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on,* pp. 1-8. IEEE, 2007.

[3] Turaga, Pavan, Rama Chellappa, Venkatramana S. Subrahmanian, and Octavian Udrea. "Machine recognition of human activities: A survey." *IEEE Transactions on Circuits and Systems for Video technology* Vol.18, No.11 (2008), pp.1473-1488.

[4] Ryoo, Michael S., and Jake K. Aggarwal. "Semantic representation and recognition of continued and recursive human activities." *International journal of computer vision* Vol.82, No.1 (2009), pp. 1-24.

[5] Poppe, Ronald. "A survey on vision-based human action recognition." *Image and vision computing* Vol.28, No. 6 (2010): 976-990.

[6] Aggarwal, Jake K., and Michael S. Ryoo. "Human activity analysis: A review." *ACM Computing Surveys (CSUR)* Vol.43, No. 3 (2011), pp. 16.

[7] Vishwakarma, Sarvesh, and Anupam Agrawal. "A survey on activity recognition and behavior understanding in video surveillance." *The Visual Computer* Vol. 29, No.10 (2013), pp. 983-1009.

[8] Ziaeefard, Maryam, and Robert Bergevin. "Semantic human activity recognition: a literature review." *Pattern Recognition* Vol.48, No.8 (2015), pp. 2329-2345.

[9] Geetha, M. Kalaiselvi, J. Arunnehru, and A. Geetha. "Early Recognition of Suspicious Activity for Crime Prevention." *Emerging Technologies in Intelligent Applications for Image and Video Processing* Vol.205 (2016).

[10] Gupta, Abhinav, Aniruddha Kembhavi, and Larry S. Davis. "Observing human-object interactions: Using spatial and functional compatibility for recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol.31, No.10 (2009), pp. 1775-1789.

[11] Yao, Bangpeng, and Li Fei-Fei. "Grouplet: A structured image representation for recognizing human and object interactions." *In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on,* pp. 9-16. IEEE, 2010.

[12] Delaitre, Vincent, Ivan Laptev, and Josef Sivic. "Recognizing human actions in still images: a study of bag-of-features and part-based representations." *In BMVC 2010-21st British Machine Vision Conference.* 2010.

[13] Datta, Ankur, Mubarak Shah, and N. Da Vitoria Lobo. "Person-on-person violence detection in video data." *In Pattern Recognition, 2002. Proceedings. 16th International Conference on,* Vol. 1, pp. 433-438. IEEE, 2002.

[14] Park, Sangho, and Jake K. Aggarwal. "Simultaneous tracking of multiple body parts of interacting persons." *Computer Vision and Image Understanding* Vol.102, No.1 (2006), pp. 1-21.

[15] Bloom, Victoria, Vasileios Argyriou, and Dimitrios Makris. "Linear latent low dimensional space for online early action recognition and prediction." *Pattern Recognition* Vol.72 (2017), pp. 532-547.

[16] Marín-Jiménez, Manuel J., Enrique Yeguas, and Nicolás Pérez De La Blanca. "Exploring STIP-based models for recognizing human interactions in TV videos." *Pattern Recognition Letters* Vol.34, No.15 (2013), pp. 1819-1828.

[17] Sener, Fadime, and Nazli Ikizler-Cinbis. "Two-person interaction recognition via spatial multiple instance embedding." *Journal of Visual Communication and Image Representation* Vol.32 (2015), pp. 63-73.

[18] Harris, Chris, and Mike Stephens. "A combined corner and edge detector." *In Alvey vision conference,* Vol. 15, No.50, (1988), pp. 10-5244.

[19] Breiman, Leo. "Random forests." *Machine learning* Vol.45, No.1 (2001), pp. 5-32.

[20] Arunnehru, J., and M. Kalaiselvi Geetha. "Difference intensity distance group pattern for recognizing actions in video using Support Vector Machines." *Pattern Recognition and Image Analysis* Vol.26, No.4 (2016), pp. 688-696.

[21] Lepetit, Vincent, and Pascal Fua. "Keypoint recognition using randomized trees." *IEEE transactions on pattern analysis and machine intelligence* Vol.28, No.9 (2006), pp. 1465-1479.

[22] Ozuysal, Mustafa, Pascal Fua, and Vincent Lepetit. "Fast keypoint recognition in ten lines of code." *In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on,* pp. 1-8. Ieee, 2007.