Abnormal Activity Recognition Using Saliency and Spatio-Temporal Interest Point Detector: Second International Conference, ICSCS 2018, Kollam, India, April 19–20, 2018, Revised Sel...

Chapter · September 2018

DOI: 10.1007/978-981-13-1936-5_32

Abnormal Activity Recognition Using Saliency and Spatio-Temporal Interest Point Detector

Smriti H. Bhandari(✉) and Navnee S. Babar

Department of Computer Science and Engineering,
Walchand College of Engineering, Sangli, India
smriti_bhandari@yahoo.com, navni010190@gmail.com

Abstract. Detecting abnormal activities is a crucial research topic nowadays because of its wide variety of applications in such as security monitoring, video surveillance and healthcare applications. The proposed method is used to distinguish between normal and abnormal human activities. Two-dimensional visual saliency map is created from color video sequences and used for further processing. Selective spatio-temporal interest point (STIP) detector is used to extract interest point features from saliency. 3D Image gradients are calculated using intensity patches to describe STIPs and feature vector is computed by quantizing them. The activities are described finally using bag-of-features representation. Support Vector Machine is used as a classifier to distinguish between normal and abnormal activities. The performance of the system is evaluated using UR fall Dataset and dataset S provided by Le2i CNRS that shows significant accuracy.

Keywords: Abnormal activity recognition · Spatio-temporal interest points
Visual saliency

1 Introduction

Recognizing behavior of a human from videos is a challenging problem in many application areas, such as video surveillance and retrieval, security and health care involving computer vision and machine learning applications. To moniter the behaviour, daily activities and any other information of elderly, a close observation is necessary to protect them [1]. With increased population of older adults in society, there is an urgent need for assistive technologies in the home. Older adults face many difficulties while undergoing their daily activities because of age-related changes. Thus, to take care of older adults, it is very important to know if any unusual activities are there.

Abnormal activities of human are still difficult to recognize because they are not able to predict it prior and those types of strange events does not occur frequently [2]. It is important to identify an emerging medical condition prior to it gets critical. So, it becomes necessary to observe the activities of daily livings (ADLs) and seek for abnormal behaviour in daily life [3]. In this paper, the work is focused on to detect unusual or abnormal activities instead of considering regular activity recognition. "Abnormal activities" can be defined as "activities which are infrequent and not predicted in advance" [4].

I. Zelinka et al. (Eds.): ICSCS 2018, CCIS 837, pp. 291–301, 2018.
https://doi.org/10.1007/978-981-13-1936-5_32

In this work, a system is proposed to distinguish between normal and abnormal activities which uses 2D visual saliency map from color video sequences. Two datasets such as UR fall detection [5] and Dataset S by Le2i CNRS [6] are used to evaluate the performance. Selective STIP detector is used to find the interest points which are robust to the complex and moving background [7]. To extract relevent features is crucial step to detect and recognize the activity. Hence STIP detector is used which focuses on local spatio-temporal information and the performance is improved by retaining most repeatative, balanced and distinguishable STIPs for human subjects after elimination of undesired STIPs in background. Image gradients are computed after extracting selective interest points and then the gradients are quantized using spherical co-ordinate method to form a vector of final features. With the use of bag-of-features (BoFs) representation and support vector machine (SVM) performance is evaluated.

The rest of the paper is organized as follows. In Sect. 2 work related to abnormal activity detection is discussed. The proposed methodology is in Sect. 3. In Sect. 4 experimental results are presented with all the necessary details and discussion. Finally, the conclusion is provided in Sect. 5.

2 Related Work

Human activity recognition has given much attention especially from those who work in the field of machine learning and computer vision. Human activities are classified into normal and abnormal activities. To detect abnormal activities is main concern nowadays and lots of research is focused on identifying abnormalities in video surveillance. Abnormality detection belongs to video analysis which includes human activity detection and recognition. These systems are mainly classified into single-layered and hierarchical approaches as shown in Fig. 1. Single layered techniques are used to represent the activity directly based on image sequences and further classified into methods such as space-time and sequential approaches. Space-time approaches consider the activity in 3D volume and represent it in space-time features from given video sequences. These approaches are again divided based on features used from 3D

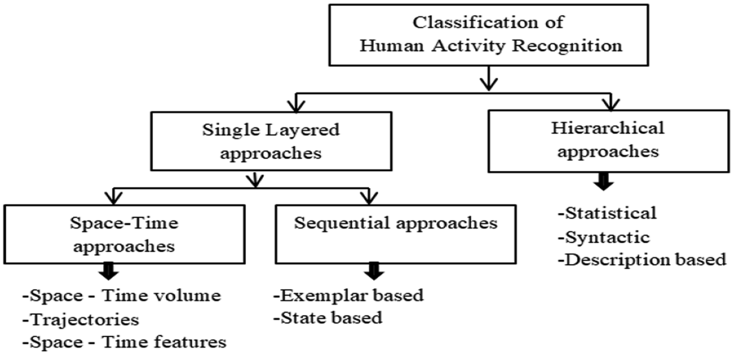


Fig. 1. Classification of human activity recognition system

volume [8]. Sequential approaches describe the activity in sequence of observations and can be further classified depending on the technique they use for recognition such as Exemplar or state based recognition [9].

Hierarchical techniques used for recognizing complex activities by analyzing the video into various feature descriptors [10]. Hierarchical approaches are categorized into statistical, syntactic and description based methodologies. Statistical strategies used to represent the high-level human activities by concatenating state based models hierarchically. Hidden Markov Model (HMM) is an example of such type of approach. Syntactic models use grammar based syntax and model the activities as strings of symbols [11]. Description based models describe the sub-event of activities to represent human activities. Abnormality or anomaly can be realized by using normal activities and depends on the approaches used to classify them. Three broad categories such as supervised, semi-supervised and un-supervised are used to construct the model [12–14].

The training data of normal and abnormal behavior is provided in case of supervised approaches to detect anomalous or unusual data. Semi-supervised approaches use only normal behavior to train the model and detect abnormality either automatically or through the training process. Un-supervised approaches do not need any kind of training [15]. These approaches work on some rule base or the conditions describing distinction between classes.

3 Proposed Methodology

This work proposes the method for abnormal activity detection in the home environment. Here we consider normal activities as daily activities of the person such as sitting on a chair, lying on the bed, reading, writing, picking up the fallen object, tightening shoelace, sweeping, cleaning, etc. Abnormal activities include forward fall, backward fall, fall from standing position, fall from a chair, etc. The overall methodology is depicted in Fig. 2.

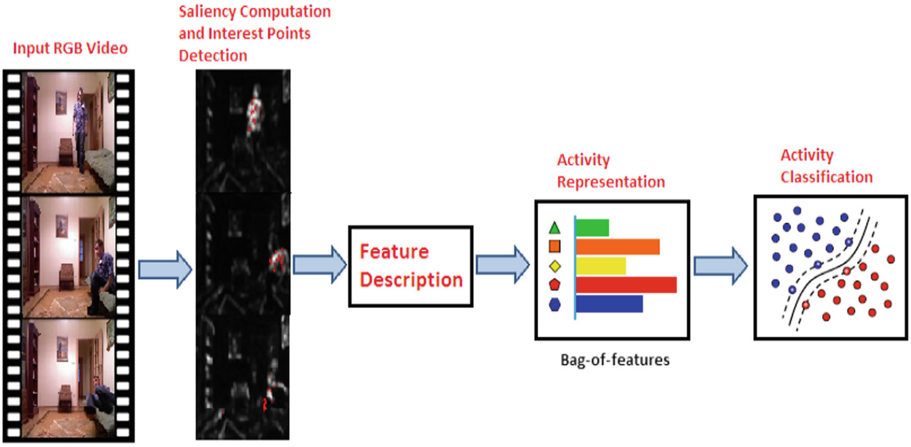


Fig. 2. Block diagram for activity recognition

Input videos are converted into frames. Further, we detect salient regions from each frame. The saliency of each frame is computed using DCT-based image signatures [16]. Once we find the salient image, next task is to find interest points those effectively contribute in describing the motion of the object. We use selective Spatio-Temporal Interest Point (STIP) detection method proposed by Chakraborty et al. [7]. Authors have claimed that performance is improved by retaining most repeatative, balanced and distinguishable STIPs for human subjects after removing undesired STIPs in background. However, in our experiments, we observed few unwanted background STIPs when the frames are processed directly for feature description. To remove or minimize unwanted background STIPs, we introduced intermediate step of saliency computation before interest point detection. Further, based on interest points, feature descriptor is formed for each frame in a video. The activity in the video is represented using a feature vector after application of bag-of-features technique. Finally, SVM classifier is used for classification of activities into two classes: normal and abnormal.

The detailed methodology is described in the following text.

The color visual data (i.e., color videos) are denoted as a sequence of 2D frames $\{I_1, \ldots, I_T\}$. The 2D frame at time instance t is denoted by $I_t = (x, y, i, t)$, $\forall t \in [1, T]$, where, x and y denote the spatial co-ordinates of pixel in the image frame; and i is the intensity value computed from its respective RGB values.

3.1 Saliency Computation

A part of an image that catches the attention of a viewer as it stands out from its neighborhood is called salient region. An object or a pixel in an image is referred as salient depending on the measure or quality by which it is distinguished from its surrounding. With respect to the application under consideration, we need to detect salient object; that is the moving foreground object in video under consideration. Hou et al. [16] used a binary, and holistic image descriptor called the "image signature" to highlight salient regions in the image. It is defined as the sign function of the Discrete Cosine Transform (DCT) of an image. We have used this method to compute salient image for further processing.

Let us assume that gray-scale images can be decomposed as:

$$\mathbf{x} = \mathbf{f} + \mathbf{b}, \quad \text{where}\, \mathbf{x}, \mathbf{f}, \mathbf{b} \in \Re^N \tag{1}$$

The image signature is defined as:

$$\begin{aligned} &\hat{\mathbf{x}} = \mathrm{DCT}(\mathbf{x}) \\ &\mathrm{Im}\, ageSignature(\mathbf{x}) = sign(\hat{\mathbf{x}}) \end{aligned} \tag{2}$$

Given an image which can be decomposed as in (1), the support of $\mathbf{f}$ can be taken as the sign of the mixture signal $\mathbf{x}$ in the transformed domain and then computing the reconstructed image in spatial domain using inverse DCT.

$$\bar{\mathbf{x}} = IDCT(sign(\hat{\mathbf{x}})) \tag{3}$$

Foreground of an image is assumed to be visually aparent and discernible with respect to its background, then we can form a saliency map **m** [17] by smoothing the squared reconstructed image as in (4)

$$\mathbf{m} = g * (\bar{\mathbf{x}} \circ \bar{\mathbf{x}}) \tag{4}$$

where g is the Gaussian kernel. '*' is convolution operator and 'o' is Hadamard (entrywise) product operator.

3.2 Interest Points Detection

STIP-based methods avoid temporal alignment problem. Also these methods exhibit invariance to geometric transformations. We have adopted Selective STIP detection method [7], in which surround suppression mask is used to remove undesired points in the background considering local and temporal constraints. This formulation makes the system robust to camera motion as well as background clutter.

3.3 Feature Description

For each interest point detected in a frame, we construct a patch $i_p(x, y, t)$ of size m x m. Then, spatio-temporal gradients are computed for intensity patch sequence along x, y, and t dimensions as:

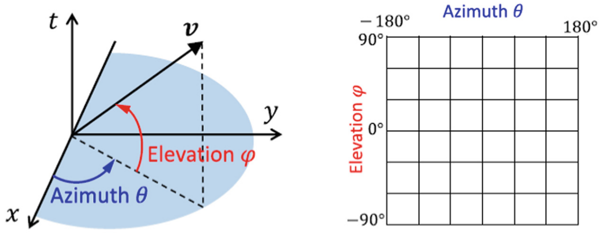
$$\nabla i_p = \left(\frac{\partial i_p}{\partial x}, \frac{\partial i_p}{\partial y}, \frac{\partial i_p}{\partial t} \right) \tag{5}$$

Here, we use 3D sobel operator [18] to compute gradient along each dimension.

The gradients of image patch sequence are quantized using a spherical coordinate-based scheme [19]. The azimuth angle $\theta(\nabla i_p)$ and elevation angle $\phi(\nabla i_p)$ are computed for each gradient vector obtained as per (5). This characterizes 3D orientations of image patch sequence in xyt space, as:

$$\theta(\nabla i_p) = \arctan\left(\frac{\partial i_p}{\partial y} \Big/ \frac{\partial i_p}{\partial x} \right), \qquad \phi(\nabla i_p) = \arctan\left(\frac{\partial i_p}{\partial t} \Big/ \sqrt{\frac{\partial^2 i_p}{\partial x} + \frac{\partial^2 i_p}{\partial y}} \right) \tag{6}$$

Feature descriptor is computed based on image gradient orientations in xyt space. Interest points provide visual cue to be used as features. As, orientation does not depend on its magnitude, it is not affected by changes in the illumination as well as noise. Thus, orientation quantization is a robust way for describing features as depicted in Fig. 3(b). Figure 3(a) shows orientation computation of azimuth and elevation angles. Further as shown in Fig. 3(b), these angles are quantized in bins. As an example, Fig. 3(b) shows subdivision of azimuth angle and elevation angle into six



(a) (b)

Fig. 3. 3D feature description using spherical coordinates. (a) Orientation computation. (b) Orientation quantization [19]

different bins, resulting in 36 bins of 1D histogram. By normalizing histogram of intensity gradient orientations h_i, final feature vector h is computed as given in (7).

$$h = \frac{h_i}{N_i} \tag{7}$$

where N_i is the total number of gradient orientations in h_i.

3.4 Activity Representation

We apply the standard BoFs representation to encode visual cues representing human activities. The BoF representation is based on a codebook or visual vocabulary. To construct vocabulary k-means clustering is used. Each cluster has a cluster center and is indexed by a visual word. Euclidean distance is used to assign every feature vector to its nearest visual word. Then, a video consisting of a sequence of color frames can be encoded as a histogram of visual word occurrences.

4 Results and Discussion

4.1 Dataset

As the real dataset containing abnormal activities of elderly people is difficult to acquire, two benchmark datasets such as Dataset UR-fall detection (URFD) and Dataset S provided by Le2i CNRS are used to estimate the performance of the proposed system to detect abnormal activity. Detail description of datasets are given below.

UR fall dataset [5]: The dataset consists of 70 (30 fall activities + 40 activities of daily living) RGB and depth videos. The proposed system uses only RGB videos for processing. Here, fall is considered as abnormal activity and activities of daily living are considered as normal activities to evaluate the performance. Normal activities such as walk, sit down, bend, lye on a bed and abnormal activities such as fall while sitting on a chair, fall while walking has been performed by five subjects.

Dataset S by Le2i CNRS [6]: The entire dataset contains total 221 videos out of which 126 are of abnormal activities and 95 are of normal activities. The original resolution of the frames in the dataset (640 × 480) is resized to 320 × 240 pixels for the analysis. Nine different subjects have performed various The dataset includes normal activities such as walk, sit down, stand up, bend, house keeping, move chair and abnormal activities as forward falls, falls when inappropriate sitting-down, loss of balance, stroke, etc.

4.2 Results

Experiments are performed with UR Fall dataset and Dataset S by Le2i CNRS. The results of intermediate steps as per methodology are shown in Figs. 4 and 5. SVM classifier is used for classification. Holdout method is used for dataset partitioning with 50% samples under both categories are used for training and remaining 50% are used for testing. Results are obtained by undergoing the methodology as explained in Sect. 3 and reported as with saliency. As mentioned in feature description we have computed orientation histograms to describe interest points. A group of angle bins controls the granularity of orientation histograms. The elevation angle ϕ is divided into 9 bins; whereas azimuth angle θ is divided into 18 cells. Thus, the resulting feature vector contains 162 elements. Size of the vocabulary for BoF activity representation is kept as 100. This parameter determines the size of final feature vector that describes the activity by encoding the original features obtained by STIPs.

The experiments are also carried out by omitting the step of saliency computation. For UR Fall dataset the accuracy is 100% for both normal as well as abnormal categories when STIPs are detected after saliency computation. However, it is observed that the accuracy obtained for correct classification is 86.67% for abnormal class when STIPs are computed by omitting saliency detection. The results are depicted in Fig. 6. The further experimentation of the results pointed out the possible reason for reduced accuracy. As shown in Fig. 4, for few of the frames in the video the interest points are detected in the background too which may mislead training as well as testing.

For Dataset S, the results are comparatively low than the results reported in [6]. In [6], the authors have reported the results separately for the dataset with different backgrounds. In the experiments, we have collectively used the videos to train and test the dataset and obtained the results irrespective of the room (background) used for acquiring the dataset. Further, in case of results with saliency for dataset S, though the system provides false alarms, those can be accepted to some extent as we are much concerned about correctly detecting abnormal activities. For dataset S, the accuracy for correct classification of abnormal activities is 96.83%, i.e. out of 63 abnormal activities tested, 61 are correctly recognized as abnormal and 2 activities are wrongly reported as normal. The results carried out in experimentation are depicted graphically in Fig. 6.

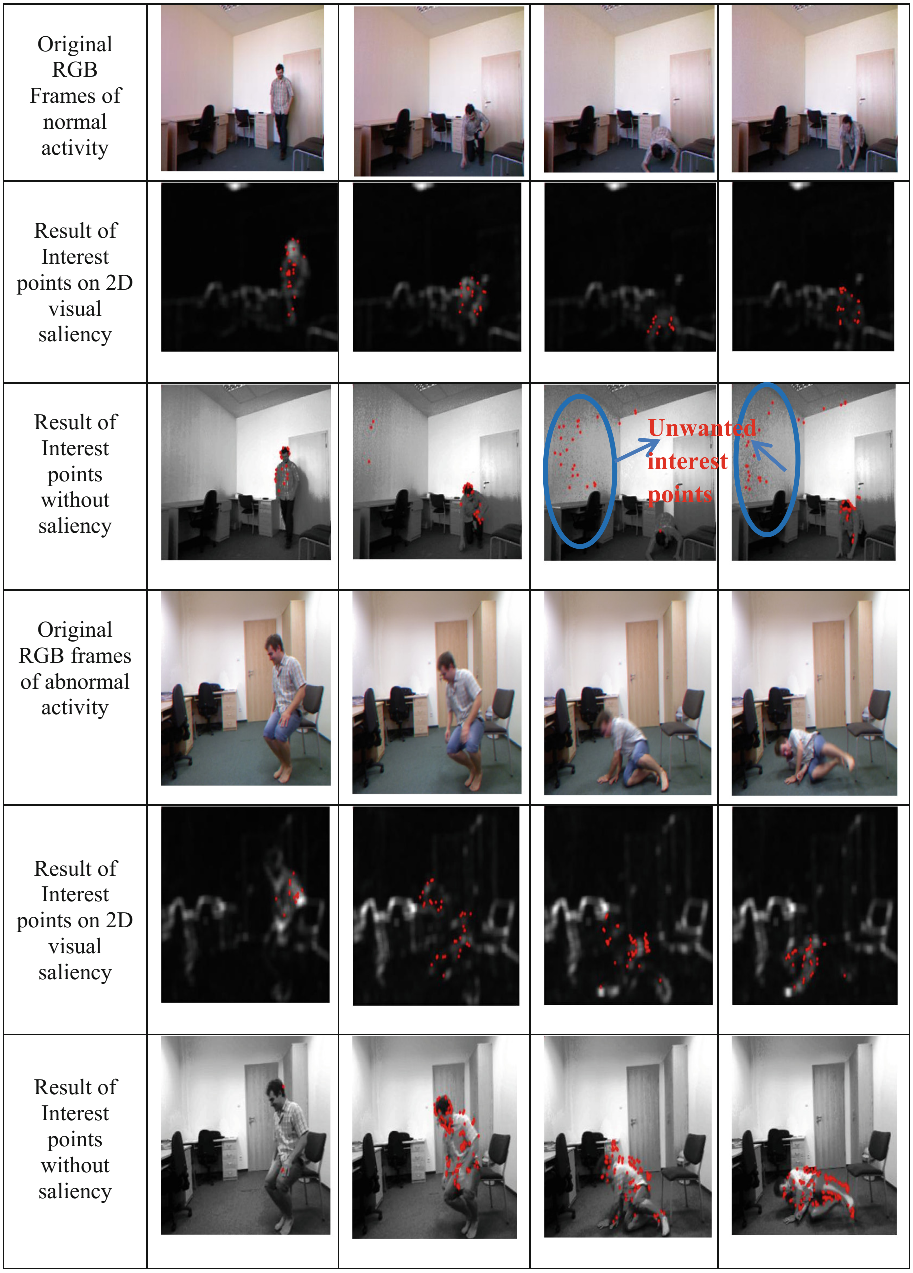

Original RGB Frames of normal activity
Result of Interest points on 2D visual saliency
Result of Interest points without saliency
Unwanted interest points
Original RGB frames of abnormal activity
Result of Interest points on 2D visual saliency
Result of Interest points without saliency

Fig. 4. Results of interest points detection on normal and abnormal activities of UR fall dataset

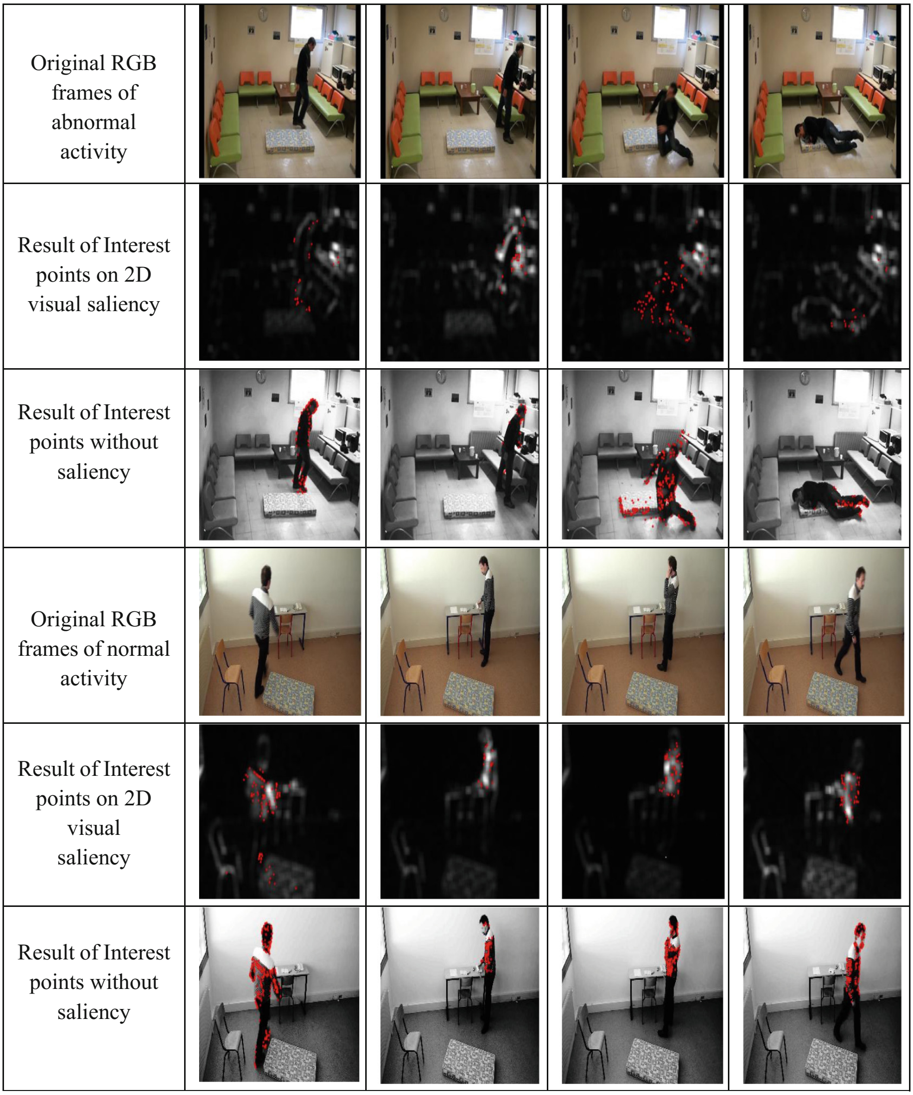

Original RGB frames of abnormal activity
Result of Interest points on 2D visual saliency
Result of Interest points without saliency
Original RGB frames of normal activity
Result of Interest points on 2D visual saliency
Result of Interest points without saliency

Fig. 5. Results of interest points detection on normal and abnormal activities of Dataset S

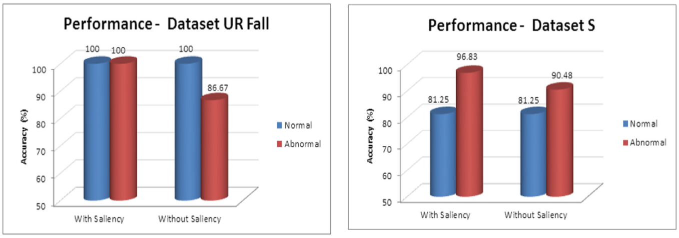


Fig. 6. Performance of proposed system for activity recognition

5 Conclusion

We have presented a method for detecting abnormal activities in the home environment. This is an attempt towards the aim of building a support system assisting elderly people living alone. The proposed system first computes salient regions from frames and then detects Spatio-Temporal Interest Points for the description of motion in the frame. BoF technique is used for building vocabulary for activity description. SVM is used as a classifier to recognize whether the activity is normal or abnormal. As our aim is to report abnormal activity of elderly people correctly, we found the results of our system encouraging. For UR Fall dataset the system is giving 100% accuracy whereas in case of Dataset S by Le2i CNRS, though a system is giving false alarms, the rate of correct classification for abnormal activities is 96.83%. The work is being extended with experimentation on our own dataset including more abnormal activities related to elderly health issues. Also, the attempts will be made to improve the accuracy of abnormal activities as well as reducing the false alarms by varying the parameters, size of codebook etc.

Acknowledgements. The authors acknowledge Department of Science and Technology, Government of India for financial support vide Reference No. SR/CSRI/84/2015 under Cognitive Science Research Initiative (CSRI) to carry out this work.

References

1. Mathur, G., Bundele, M.: Research on intelligent video surveillance techniques for suspicious activity detection critical review. In: IEEE International Conference on Recent Advances and Innovations in Engineering 2016 (ICRAIE-2016), 23–25 December, Jaipur (2016)
2. Tong, Y., Chen, R., Gao, J.: Hidden state conditional random field for abnormal activity recognition in smart homes. Entropy **17**, 1358–1378 (2015)
3. Wang, C., Zheng, Q., Peng, Y., De, D., Song, W.-Z.: Distributed abnormal activity detection in smart environments. Int. J. Distrib. Sensor Netw. **2014**, Article ID 283197, 1–15 (2014)
4. Hua, D.H., Zhang, X.-X., Yinc, J., Zhenga, V.W., Yang, Q.: Abnormal activity recognition based on HDP-HMM models. In: International Joint Conference on Artificial Intelligence (2009)
5. Kwolek, B., Kepski, M.: Human fall detection on embedded platform using depth maps and wireless accelerometer. Comput. Methods Programs Biomed. **117**(3), 489–501 (2014)
6. Charfi, I., Mitéran, J., Dubois, J., Atri, M., Tourki, R.: Optimised spatio-temporal descriptors for real-time fall detection: comparison of SVM and Adaboost based classification. J. Electron. Imaging (JEI) **22**(4), 17 (2013)
7. Chakraborty, B., Holte, M.B., Moeslund, T.B., Gonzàlez, J.: Selective spatio-temporal interest points. Comput. Vis. Image Underst. **116**, 396–410 (2012)
8. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. Int. J. Comput. Vis. **103**, 60–79 (2013)
9. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: a review. ACM Comput. Surv. **43**, 1–43 (2011)

10. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: Proceedings of IEEE International Conference on Computer Vision, Rio de Janeiro, pp. 1–8 (2007)
11. Tsai, W., Fu, K.S.: Attributed grammar-a tool for combining syntactic and statistical approaches to pattern recognition. SMC **10**, 873–885 (1980)
12. Brax, C., Niklasson, L., Smedberg, M.: Finding behavioural anomalies in public areas using video surveillance data. In: Proceedings of 11th International Conference on Information Fusion (2008)
13. Zhang, D., Gatica-Perez, D., Bengio, S., McCowan, I.: Semi-supervised adapted HMMs for unusual event detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 611–618 (2005)
14. Xiang, T., Gong, S.: Video behavior profiling for anomaly detection. IEEE Pattern Anal. Mach. Intell. **30**, 893–908 (2008)
15. Beddiar, D.R., Nini, B.: Vision based abnormal human activities recognition: an overview. In: 8th International Conference on Information Technology (2017)
16. Hou, X., Harel, J., Koch, C.: Image signature: highlighting sparse salient regions. IEEE Trans. Pattern Anal. Mach. Intell. **34**(1), 194–201 (2012)
17. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
18. Sun, B., Sang, N., Wang, Y., Zheng, Q.: Motion detection based on biological correlation model. In: Zhang, L., Lu, B.-L., Kwok, J. (eds.) ISNN 2010. LNCS, vol. 6064, pp. 214–221. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13318-3_28
19. Zhang, H., Parker, L.E.: CoDe4D: color-depth local spatio-temporal features for human activity recognition from RGB-D videos. IEEE Trans. Circuits Syst. Video Technol. **26**(3), 541–555 (2016)