Group Activity Recognition Using Deep Autoencoder with Temporal Context Descriptor

S. A. Vahora
CSPIT, Charusat University, Changa, Gujarat, India.
safvan465@gmail.com
and
N. C. Chauhan
A. D. Patel Institute of Technology, New V. V. Nagar, Gujarat, India.
narendracchauhan@gmail.com

In this paper, we propose a novel method for group activity recognition in the video sequence. The problem of recognizing group activity requires information about individual person action, interaction potential, social cues bonding relationship of the people in the context region and analysis of this context region over a time period. We propose a deep architecture model, stacked deep autoencoder to provide a high-level representation of a group activity context descriptor, build at the top of local level human action pose feature. These local and global level representations of the group activity analyzed over a time period to build robust temporal group activity context descriptor. Our experimental results show the efficiency of the proposed approach over a benchmark collective activity dataset.

Keywords: Group Activity Recognition, Deep Autoencoder, Context Descriptor, Temporal Information.

1. INTRODUCTION

Regardless of abundant research in the field of computer vision to observe human activities in surveillance videos, there are as yet challenging issues and constraints. Many previous types of research have presented high interest in recognizing actions carried out with the aid of a single human or complex human activity in the video (Tran, Gala, Kakadiaris, and Shah, 2014), (Aggarwal and Ryoo, 2011). In any case, in everyday applications, group activity recognition is a challenging and critical because of its technical hitches and realistic necessities in broad daylight places like airline terminals, railroad station, sub-ways, transport stop and so on (Vahora and Chauhan, 2017). Here, communication among individuals and context environment pick up a solid semantic importance. The significance of context environment has first developed in scene understanding (Biederman, 1981), (Hoiem, Efros, and Hebert, 2006), signal processing (Vinciarelli, Pantic, and Bourlard, 2009).

Encouraged by these views, in this work we consider the issue of recognizing group human activities usually referred as collective activities. Collective activities can be better comprehended and differentiated by breaking down the social context of the individual person, framed by closer-by individuals, rather than spotting one individual person for a period. Furthermore, the social context of the individual analysis over a number of temporal sequences in the video provides better comprehended and disambiguated results rather than observing a single frame sequence at a period. For an example, collective activities include, a group of people standing in a queue at the food parlour or being associated with a discussion of the passage and so on. In fig. 1 we present two sample video frames depicting the importance of context information in group activity recognition. In the event, considered single human without context as highlighted in the figure, the information we infer that might be limited to a particular activity, as fig. 1(a) and fig. 1(b) both looks similar, by looking towards left direction and standing action, but in fact person in fig. 1(a) and fig. 1(b) involved in a different group activity as talking and queuing respectively. Our approach uses context information of each person within context region as an independent group descriptor to recognize group activity.

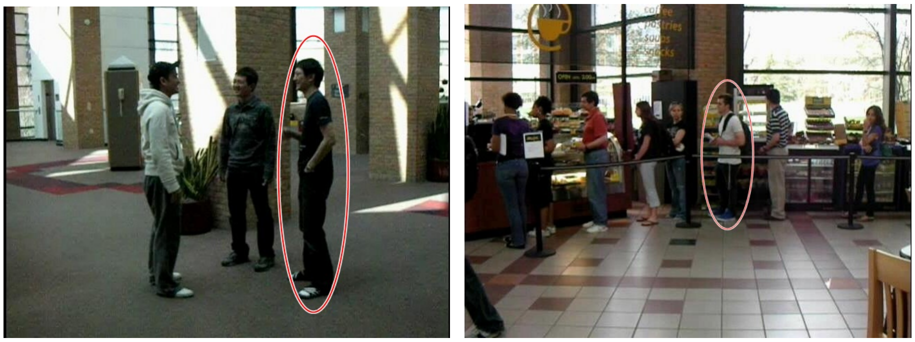

(a) (b)

Figure 1. Importance of context information : (a)talking video frame (b)queuing video frame

The main contributions of this paper are:

(1) A group activity context descriptor: we propose a spatiotemporal multi-level group activity context descriptor using a deep autoencoder that describes the behaviour property of an individual person, behaviour properties of person's framed by close-by individuals to recognize group activity. It also appends the information of the framed by close-by individuals over a time period to build effective group context descriptor.

(2) Deep architecture model: we propose a deep architecture model, stacked deep autoencoder to provide a high-level representation of a group activity context descriptor, build at the top of local level human action pose feature. These local and global level representation of the group activity analyzed over a time period to build robust temporal group activity context descriptor.

The rest of the paper is structured as follows. In section 2, we present the related work. In section 3, review the schema of our proposed model for group activity recognition. In section 4, we present experimental setup, result analysis, visual results and comparison with state-of-the-art performance and lastly, in section 5, we summarize our paper and conclude.

2. RELATED WORK

Context-based visual recognition has obtained an awful lot attention these days. Most of the work in context is in scene recognition, scene understanding, signal processing, human activity recognition and many more. Toward acknowledging a kind of characteristic feature descriptor utilized for group activity recognition, handcrafted feature descriptor and self-learned feature descriptor categorized (Vahora and Chauhan, 2017). Handcrafted feature descriptor formed by utilizing low-level feature like histogram of oriented gradient (HOG) (Dalal and Triggs, 2005), spatiotemporal local (STL), scale-invariant feature transform (SIFT), deformable part model (DPM), etc. Besides these, descriptor models can also be categorized as they use context model - action context (AC), relative action context (RAC), interaction model - distance based attraction and repulsion model or collective approach. Learned features are acquired by providing training and preparing models for testing along with validation.

AC descriptor proposed by Lan et al. (Lan, Wang, Mori, and Robinovitch, 2012), binds the action score of focal individual person and action score of individuals within context region connected to a focal person. With the use of AC descriptor and the multi-scale relationship between individuals within context region provided as potential input for recognizing group activity proposed by Kaneko et al. (Kaneko, Shimosaka, Odashima, Fukui, and Sato, 2012a). Viewpoint invariant RAC descriptor, captures the pose of focal individual and relative pose of persons within context region (Kaneko, Shimosaka, Odashima, Fukui, and Sato, 2012b). Moreover, in this direction, combining the outcome of AC and RAC descriptor with the multi-scale relationship of individuals within context region proposed by Kaneko et al. (Kaneko, Shimosaka, Odashima, Fukui, and Sato, 2014). These group descriptors are fed to the SVM classifier for group activity recognition. The hybrid version of structure level and feature level methodologies (Lan, Wang, Yang, Robinovitch, and Mori, 2012) described as person-person social interaction derived from the individual person action and action context descriptor from action person score respectively. Structure-based group activity recognition function learns over input, intermediate discrete latent variables and output. AC description of each individual person, the input of the model, provide to the number of multiple non-linear functions generated by using gradient ascent and prepare model using boosted hidden conditional random fields (HCRFs) algorithm (Hajimirsadeghi and Mori, 2015) as it converges. Furthermore, the use of cardinality based multi-instance, kernel over the bags (video) (Hajimirsadeghi, Yan, Vahdat, and Mori, 2015) representation, with SVM classification configuration provides measurable high accuracy.

Choi et al. (Choi, Shahid, and Savarese, 2009) proposed STL descriptor based on the histogram to capture histogram of a number of individuals with the pose in different directions, covered by focal individuals over a time period. Chain model for group activity recognition over a time period wraps space-time descriptor bag of the right detections (BORDs) (Amer and Todorovic, 2011) which removes noisy people from the group, captures histogram of person action pose in the space-time domain entered at a specific point in the video. The group descriptor actually encodes the group view include mean and variance of human speed for each activity by dividing the quantization space into four different regions. This also encodes the orientation of the person, represent with multilevel SVM classification configuration to recognize group behaviour (Noceti and Odone, 2014). Tran et al. (Tran et al., 2014) proposed local group activity descriptor to encode human speed and orientation together with the person-person interaction potential for group activity recognition. Novel attraction repulsion feature (ARF) based on the relative distance between individuals over a time period used to recognize group activity within the said region as a group interaction zone (Kim, Cho, and Lee, 2014). By dividing the image frame into NxN grid cells with human motion interaction analysis, Nabi et al. (Nabi, Bue, and Murino, 2013) proposed poselet activation pattern over time (TPOS) descriptor using clustering methods for group activity recognition. The collection of neighbouring hypersurface normals of a depth sequence represented as polynormal to capture shape and motion information, proposed by Yang et al. (Yang and Tian, 2017). The proposed design scheme aggregate low-level polynormals and divide a depth video into space-time cells.

Deep learning models such as deep neural network, deep auto-encoder, convolutional neural network (CNN), deep recurrent neural network (RNN) recently used in numerous applications of video summarization, object categorization, human activity recognition and shown impressive performance. Donahue et al. (Donahue, Hendricks, Guadarrama, Rohrbach, Venugopalan, Darrell, and Saenko, 2015) proposed long term recurrent convolutional network (LRCN) model for large-scale human activity recognition. Here, an image fed as input to the CNN model and extracted features from CNN are provided to the stack of RNN model such as long short term memory (LSTM) to infer variable length output using caffe library (Jia, Shelhamer, Donahue, Karayev, Long, Girshick, Guadarrama, and Darrell, 2014) on UCF101 videos (Soomro, Zamir, and Shah, 2012). Deng et al.(Deng, Zhai, Chen, Liu, Muralidharan, Roshtkhari, and Mori, 2015) proposed scene, action and pose level CNN model followed by two-step message passing by considering dependencies between classes to infer group activity. Furthermore, AC descriptor is used as a global feature along with this learned feature. The pre-trained CNN model AlexNet network used in this framework for fine-tuning, which is already trained on ImageNet dataset (Krizhevsky, Sutskever, and Hinton, 2017). Ibrahim et al. (Ibrahim, Muralidharan, Deng, Vahdat, and Mori,

2016) proposed a multi-level CNN-LSTM framework, to recognize group activity, as extracted feature fc7 from AlexNet CNN model served in person level LSTM model, an aggregated person level feature of group persons served to the group level LSTM model followed by softmax classifier. The spatiotemporal long-term motion using sequential deep trajectory descriptor (sDTD) proposed by Shi et al. captures standing spatial feature, short-term motion and long-term motion from video sequences (Shi, Tian, Wang, and Huang, 2017). To learn an efficient illustration of long-term motion, dense trajectories are projected into two-dimensional plane followed by CNN-RNN network.

The representational capabilities of the spatiotemporal group activity context descriptors investigated by different approaches found in present literature are limited. Based on the study and review of stacked deep autoencoder, they reveal higher representational capability for encoding complex events. In this paper, we propose a deep architecture model, stacked deep autoencoder to exploit higher level representation from complex event such as group activity recognition.

3. PROPOSED MODEL

In this paper, our goal is to recognize human activities performed by multiple persons in a group in the surveillance video. In a short outline of our model, the first step is preprocessing step which identify persons in a video sequence. Next, action pose features of individual persons within their context region from the bounding box of each person are extracted. These extracted features are passed through a deep architecture that has the high expressive ability. These extracted group activity context features generated from individual persons pose and action along with similar features from the temporal sequence of $2\beta+1$ number of video frames which provide the proposed temporal group activity context descriptor. The overall architecture of the model is presented in fig. 2. Moreover, the technical details of our proposed model presented in the following sections.

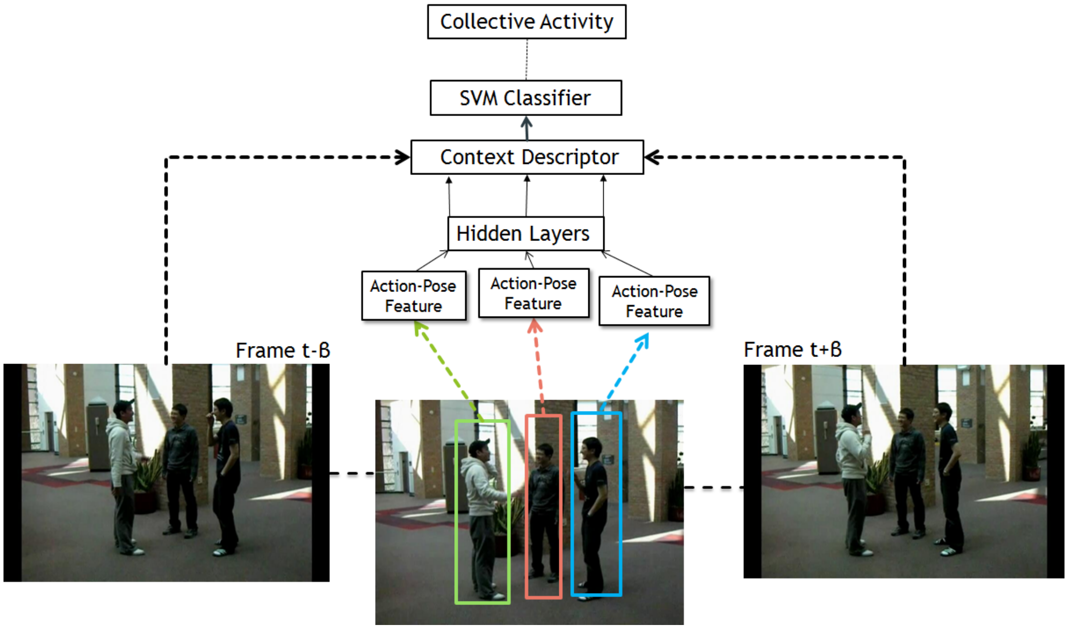


Figure 2. Temporal activity context descriptor model architecture

3.1 Model Formulation

Deep learning architecture has shown the promising result in finding a hidden representation of images, signals, natural language in a number of vision applications. These models provide a flexible and general features, that avoid manually designing of domain-specific features (Zeng, Yu, Wang, Li, and Tao, 2017), (Hou, Nie, Li, Yi, and Wu, 2014), (Zhu, You, Chen, Tao, Ou, Jiang, and Zou, 2015). Transformation of high-dimensional information to low-dimensional codes is effectively carried out by the multi-layer neural network. Weight adjustment in autoencoder can be performed by gradient descent methods. The deep autoencoder provides better results compared to different handcrafted dimensionality reduction methods such as principal component analysis (PCA), independent component analysis (ICA), etc. used in numerous vision applications (Hinton, 2006).

3.1.1 *Deep Autoencoder.* In this paper, we proposed a framework with the stacked deep autoencoder. The goal of autoencoder is to provide an effective and compact representation of the input by maintaining the important information (Bengio, 2009). This property of autoencoder helps us to build a compact representation of group context descriptor. Training objective of the autoencoder is to minimize the reconstruction error for a set of samples given as $Y = [y_1, y_2, ..., y_n]$ where $y_i \in R^d$, defined as

$$\iota = \sum_i (y_i - \hat{y_i})^2 \tag{1}$$

where y_i and $\hat{y_i}$ are the original input and the reconstructed input, respectively. The hidden layer implies encoding process and decoding process

$$\begin{cases} h_i = f(wy_i + b) \\ \hat{y} = f(w'h_i + b') \end{cases} \tag{2}$$

where $h_i \in R^n$ is the compact representation, w and w' represent the weight matrices for encoding and decoding layers, b and b' denote the bias terms. $f(.)$ is the activation function, which we use the logistic sigmoid function in this paper.

$$f(x) = \frac{1}{1 + e^{-z}} \tag{3}$$

Group activity context feature for each person in the group is developed as the person action pose feature of each and every person extracted and represented as x_i. This high dimensional feature of every person provided to stacked deep autoencoder model. This deep model has a three-layer deep autoencoder as shown in fig. 3. The output of this three-stage deep model provides a higher level representation of human action pose, to build robust feature vector for the group activity context. The output of the one deep autoencoder fed as input to the next autoencoder. The intrinsic representation of the input feature x_i described as

$$x_{i1} = f(w_1 * x_i + b_1) \tag{4}$$

$$x_{i2} = f(w_2 * x_{i1} + b_2) \tag{5}$$

$$x_{i3} = f(w_3 * x_{i2} + b_3). \tag{6}$$

Reconstruction of the encoded feature described as

$$\hat{x_i} = f({w_1}' * x_{i1} + {b_1}') \tag{7}$$

$$\hat{x_{i1}} = f({w_2}' * x_{i2} + {b_2}') \tag{8}$$

$$\hat{x_{i2}} = f({w_3}' * x_{i3} + {b_3}'). \tag{9}$$

These parameters $(w_1, w_2, w_3, b_1, b_2, b_3)$ are used to characterize the robust person action features to build a group activity context feature. This deep autoencoder model trained using softmax layer using training data. The hidden layer dimension of deep autoencoder are $h_1 = 1000$, $h_2 = 500$ and $h_3 = 100$ used in this experiment.

3.1.2 *Temporal Group Activity Context Descriptor.* The enriched active person pose feature y_i produced using the deep autoencoder to build group activity context descriptor. The context region can be divided into M equally separable sub-context region with angle θ in each direction. The group context descriptor over a focal person C represented as $(M + 1)$ x K dimension vector

$$C_i = [D_{ci}, D_{1i}, D_{2i}, ..., D_{Mi}] \tag{10}$$

$$= [y_{1c}, \max_{j \varepsilon N_1(i)} y_{1j}, ..., \max_{j \varepsilon N_1(i)} y_{nj}, ..., \max_{j \varepsilon N_M(i)} y_{1j}, ..., \max_{j \varepsilon N_M(i)} y_{nj}] \tag{11}$$

where D_{mi} is enriched person action pose feature of a person in the m^{th} sub-context region reference to i^{th} focal person. $N_m(i)$ defines the index of i^{th} person in the m^{th} sub-context region. The context descriptor of the focal person F_c, at time t denoted as C_i. In our experiment, the dimension of the last hidden layer of the deep model h_3 is 100, that represents the K dimensional vector. The temporal activity context descriptor of i^{th} person at time t over a time period $2\beta + 1$, as $[t - \beta, t - \beta + 1, ..., t, ..., t + \beta - 1, t + \beta]$ mathematically expressed as

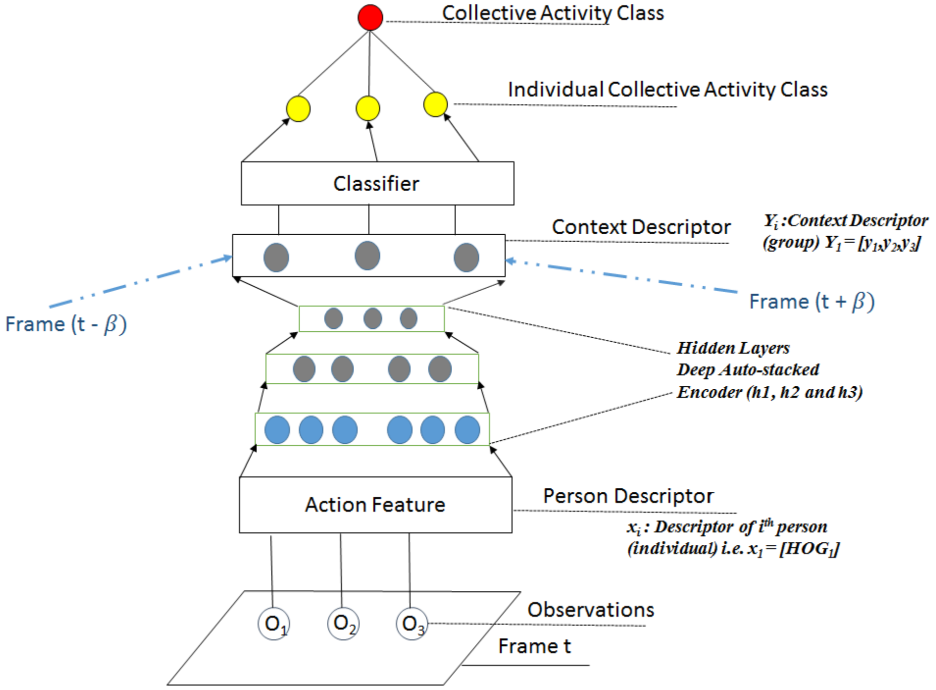

Collective Activity Class
Individual Collective Activity Class
Classifier
Context Descriptor
Y_i :Context Descriptor (group) $Y_1 = [y_1, y_2, y_3]$
Frame (t - β)
Frame (t + β)
Hidden Layers Deep Auto-stacked Encoder (h1, h2 and h3)
Action Feature
Person Descriptor
x_i : Descriptor of i^{th} person (individual) i.e. $x_1 = [HOG_1]$
O_1
O_2
O_3
Observations
Frame t

Figure 3. Illustration of feature extraction using deep autoencoder

$$F_{it} = C_{i(t-\beta)} \oplus C_{i(t-\beta+1)} \oplus ... \oplus C_{i(t)} \oplus ... \oplus C_{i(t+\beta-1)} \oplus C_{i(t+\beta)} \tag{12}$$

In this equation, C_{it} is group activity context descriptor of i^{th} person at time t derived using the proposed deep model. We concatenate these features over a time period (represented by $\oplus$) to obtain temporal representation of a group activity context descriptor F_{it} of i^{th} person at time t. Next, we use multi-class SVM classifier to train temporal group activity context descriptor F_i associated with the group activity label L^n, where n is the total number of group activity classes. In the experiment, we use radial basis function kernel implementation of SVM using LIBSVM (Chang and Lin, 2011). The score return by the SVM for each of the activity class of each person within the context region use to classify the group activity.

In inference, group activity label $\hat{y}$ is estimated using the maximum a posteriori (MAP) estimation

$$\hat{y} = \arg\max_{y \varepsilon L^n} P(y/x) \tag{13}$$

The distribution of $p(y/x)$ is computed for all group activity class $L = (l_1, l_2, ..., l_n)$ of each person in the context region. This probability of persons in the each sub-context region along with focal person is provided by the multi-class SVM.

4. EXPERIMENT AND RESULTS

4.1 Dataset

In this section, we evaluated the overall performance of proposed model and compared with the several published baseline methods over a benchmark collective activity dataset(Choi et al., 2009). This dataset is found appropriate for our assessment, as it includes activities performed by a group of people in natural environment. Most of the previous work reported on KTH (Blank, Gorelick, Shechtman, Irani, and Basri, 2005), Weizman (Schuldt, Laptev, and Caputo, 2004) datasets, which focuses on actions performed by a single person. The collective activity dataset is widely used in computer vision tasks for evaluating the performance of group activity recognition consisting of five classes such as talking, crossing, waiting, queuing and walking. The video recordings of this dataset includes recording from a hand held camera at a low edge to view under the realistic condition like occlusion, camera shivering, and background clutter. Ground truth of the dataset is available. Every frame of the video is labeled with the human pose, group activity of frame and person with rectangle bounding box information. A group activity label to the frame is assigned based on the majority of what people are doing in the scene.

4.2 Experiment and Analysis

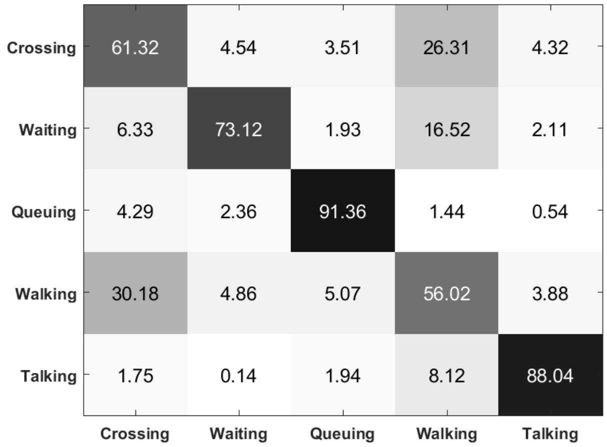
The overall goal is to assess the overall performance of the proposed model and compare it with the state-of-the-art methods. We split train/test dataset in the same approach provided by (Hajimirsadeghi et al., 2015), on benchmark collective activity dataset (Choi et al., 2009). We have evaluated the performance for different value of sub-context region and the best result obtained with sub-context region N = 3, using RBF kernel of SVM. The results of the proposed model in the form of confusion matrix is shown in fig. 4. The confusion matrix presents that walking and crossing are closely connected group activities as misclassification between these two group activities is quite high. The visual results obtained using proposed model is shown in fig. 5. Here, the red labels are ground truth class label in the video frame and green labels are predicted label using proposed approach.

The comparative results of our approach to the baseline methods are presented in Table I. The comparison of our proposed method with with baseline methods indicate that our approach is better than most other methods except methods proposed by Kaneko et al. (Kaneko et al., 2014) and Tran et al. (Tran et al., 2014). Kaneko et al. (Kaneko et al., 2014) embed multi-scale relationship as position, size motion an time sequence along with group context descriptor. However, our results are nearly similar to their results. Tran et al. (Tran et al., 2014) embed human action, motion along with relative distance information for group activity recognition which provides better results compared to our approach. Our approach has not considered relative distance information in the group descriptor, inclusion of which may further enhance our results. Overall an increase in higher level information feature descriptor raises the performance of the group activity recognition system. Furthermore, in our work we have considered single

Table I: Comparison of our method with previously published works

Approaches	Average Accuracy
Kaneko et. al. (Kaneko et al., 2012a)	72.20
Kaneko et. al. (Kaneko et al., 2014)	74.70
Wongun et. al. (Choi et al., 2009)	65.90
Lan et. al. (Lan et al., 2012)	68.20
Kaneko et. al. (Kaneko et al., 2012b)	73.20
Nabi et. al. (Nabi et al., 2013)	72.30
Tran et. al. (Tran et al., 2014)	78.75
Proposed Approach	74.31

group activity in a video sequence by classifying group activity in which majority of the group persons involved. This can be further extended by considering multiple group activity if present in a video sequence by classifying each sub-group activity.

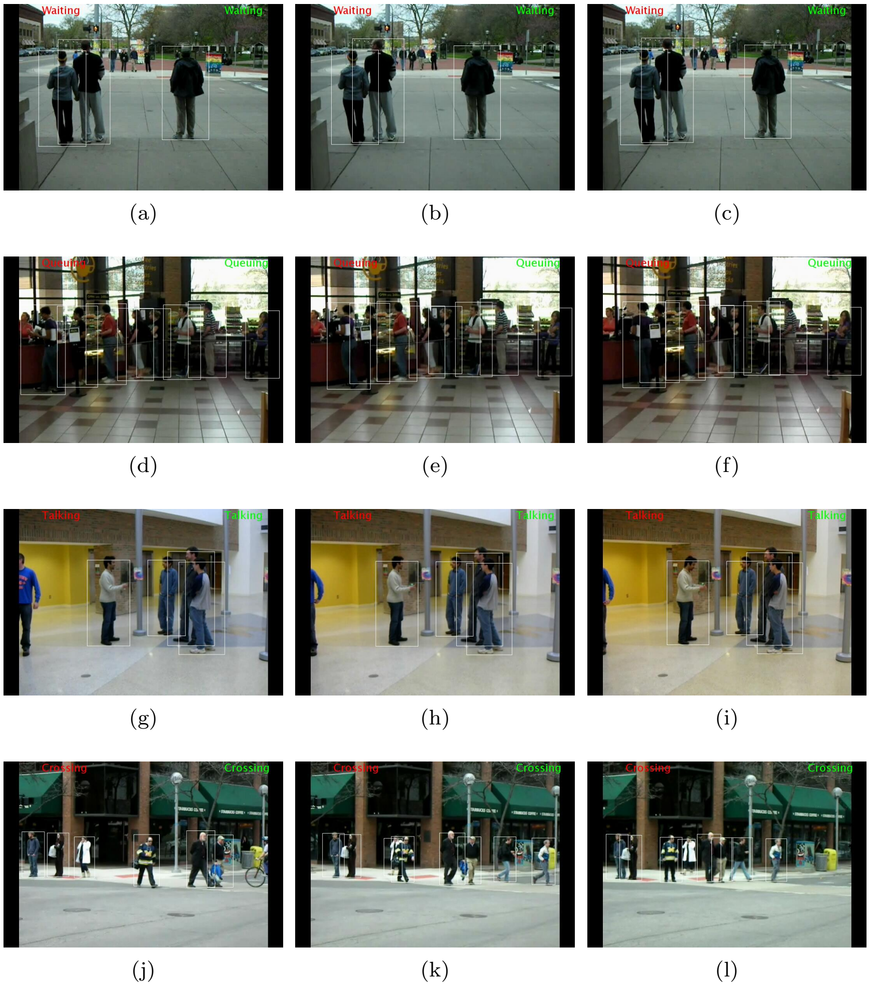


	Crossing	Waiting	Queuing	Walking	Talking
Crossing	61.32	4.54	3.51	26.31	4.32
Waiting	6.33	73.12	1.93	16.52	2.11
Queuing	4.29	2.36	91.36	1.44	0.54
Walking	30.18	4.86	5.07	56.02	3.88
Talking	1.75	0.14	1.94	8.12	88.04

Figure 4. Confusion matrix for the collective activity dataset using proposed model

5. CONCLUSION

In this paper, we propose a deep architecture model to build context descriptor which provides a high-level representation of the handcrafted human action pose feature. The stacked deep autoencoder builds global context descriptor that successfully captures a meaningful contextual representation of the group by considering the person-person interaction within the group context region. In this model, group context descriptor is encoded over a time period to get a temporal group activity descriptor. This descriptor encodes motion information about every person in the group rather than individual human motion information. Experiment results show that our model improves the result of group activity recognition than most of the state-of-the-art methods over benchmark collective activity dataset. Our future work focuses on inclusion of background scene information in the group activity descriptor to effectively differentiate walking and crossing group activity.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 5. Results visualization of temporal group activity context descriptor Model. Red tags are ground truth, Green tags are predicted label

References

Aggarwal, J. and Ryoo, M. 2011. Human activity analysis. *ACM Computing Surveys 43,* 3 (apr), 1–43.

Amer, M. R. and Todorovic, S. 2011. A chains model for localizing participants of group activities in videos. In *2011 International Conference on Computer Vision*. IEEE.

Bengio, Y. 2009. Learning deep architectures for AI. *Foundations and Trends® in Machine Learning 2,* 1, 1–127.

Biederman, I. 1981. *On the semantics of a glance at a scene.*

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. 2005. Actions as space-time shapes. In *Tenth IEEE International Conference on Computer Vision (ICCV 05) Volume 1*. IEEE.

Chang, C.-C. and Lin, C.-J. 2011. LIBSVM. *ACM Transactions on Intelligent Systems and Technology 2,* 3 (apr), 1–27.

Choi, W., Shahid, K., and Savarese, S. 2009. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. IEEE.

Dalal, N. and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *2005*

IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 05). IEEE.

Deng, Z., Zhai, M., Chen, L., Liu, Y., Muralidharan, S., Roshtkhari, M. J., and Mori, G. 2015. Deep structured models for group activity recognition. In *Procedings of the British Machine Vision Conference 2015*. British Machine Vision Association.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Darrell, T., and Saenko, K. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Hajimirsadeghi, H. and Mori, G. 2015. Learning ensembles of potential functions for structured prediction with latent variables. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE.

Hajimirsadeghi, H., Yan, W., Vahdat, A., and Mori, G. 2015. Visual recognition by counting instances: A multi-instance cardinality potential kernel. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Hinton, G. E. 2006. Reducing the dimensionality of data with neural networks. *Science 313,* 5786 (jul), 504–507.

Hoiem, D., Efros, A., and Hebert, M. 2006. Putting objects in perspective. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR 06)*. IEEE.

Hou, C., Nie, F., Li, X., Yi, D., and Wu, Y. 2014. Joint embedding learning and sparse regression: A framework for unsupervised feature selection. *IEEE Transactions on Cybernetics 44,* 6 (jun), 793–804.

Ibrahim, M. S., Muralidharan, S., Deng, Z., Vahdat, A., and Mori, G. 2016. A hierarchical deep temporal model for group activity recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. 2014. Caffe. In *Proceedings of the ACM International Conference on Multimedia - MM 14*. ACM Press.

Kaneko, T., Shimosaka, M., Odashima, S., Fukui, R., and Sato, T. 2012a. Consistent collective activity recognition with fully connected crfs. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. 2792–2795.

Kaneko, T., Shimosaka, M., Odashima, S., Fukui, R., and Sato, T. 2012b. Viewpoint invariant collective activity recognition with relative action context. In *Computer Vision – ECCV 2012. Workshops and Demonstrations*. Springer Berlin Heidelberg, 253–262.

Kaneko, T., Shimosaka, M., Odashima, S., Fukui, R., and Sato, T. 2014. A fully connected model for consistent collective activity recognition in videos. *Pattern Recognition Letters 43*, 109–118.

Kim, Y.-J., Cho, N.-G., and Lee, S.-W. 2014. Group activity recognition with group interaction zone. In *2014 22nd International Conference on Pattern Recognition*. IEEE.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2017. ImageNet classification with deep convolutional neural networks. *Communications of the ACM 60,* 6 (may), 84–90.

Lan, T., Wang, Y., Mori, G., and Robinovitch, S. N. 2012. Retrieving actions in group contexts. In *Trends and Topics in Computer Vision*, K. N. Kutulakos, Ed. Springer Berlin Heidelberg, Berlin, Heidelberg, 181–194.

Lan, T., Wang, Y., Yang, W., Robinovitch, S. N., and Mori, G. 2012. Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34,* 8 (aug), 1549–1562.

Nabi, M., Bue, A. D., and Murino, V. 2013. Temporal poselets for collective activity detection and recognition. In *2013 IEEE International Conference on Computer Vision Workshops*. IEEE.

Noceti, N. and Odone, F. 2014. Humans in groups: The importance of contextual information for understanding collective activities. *Pattern Recognition 47,* 11 (nov), 3535–3551.

Schuldt, C., Laptev, I., and Caputo, B. 2004. Recognizing human actions: a local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* IEEE.

Shi, Y., Tian, Y., Wang, Y., and Huang, T. 2017. Sequential deep trajectory descriptor for action recognition with three-stream CNN. *IEEE Transactions on Multimedia 19,* 7 (jul), 1510–1520.

Soomro, K., Zamir, A. R., and Shah, M. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Tran, K., Gala, A., Kakadiaris, I., and Shah, S. 2014. Activity analysis in crowded environments using social cues for group discovery and human interaction modeling. *Pattern Recognition Letters 44*, 49–57.

Vahora, S. A. and Chauhan, N. C. 2017. A comprehensive study of group activity recognition methods in video. *Indian Journal of Science and Technology 10,* 23 (feb), 1–11.

Vinciarelli, A., Pantic, M., and Bourlard, H. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing 27,* 12 (nov), 1743–1759.

Yang, X. and Tian, Y. 2017. Super normal vector for human activity recognition with depth cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence 39,* 5 (may), 1028–1039.

Zeng, K., Yu, J., Wang, R., Li, C., and Tao, D. 2017. Coupled deep autoencoder for single image super-resolution. *IEEE Transactions on Cybernetics 47,* 1 (jan), 27–37.

Zhu, Z., You, X., Chen, C. P., Tao, D., Ou, W., Jiang, X., and Zou, J. 2015. An adaptive hybrid pattern for noise-robust texture analysis. *Pattern Recognition 48,* 8 (aug), 2592–2608.

Mr. S A Vahora pursuing his Ph.D. from CSPIT, Charusat University, Changa, India. He has received M.E. degree in computer engineering from Gujarat Technological University, India, 2011 and B.E. degree in Information Technology from Sardar Patel University, India, 2009. He is at present working as an Assistant Professor at Department of Information Technology, Vishwakarma Government Engineering College, Gujarat, India. His research interest spans computer vision, image processing and machine learning. He has published research paper in prestigious conferences and journals in the field of computer vision and machine learning

Dr. N C Chauhan received the B.E. and M.E. degrees in Computer Engineering from Birla Vishwakarma Mahavidyalaya (BVM) Engineering College, Anand, Gujarat, India in 2001 and 2005 respectively. He earned his PhD degree from Department of Electronics and Computer Engineering, Indian Institute of Technology Roorkee, Roorkee, India in 2010. He is at present working as a Professor at Department of Information Technology, A D Patel Institute of Technology, Anand, Gujarat, India. His research interest includes AI and Soft Computing, Data Mining, Image and Video Processing, and Approximate Modeling and Optimizations. He has more than 40 research publications to his credit. He is principle author of a book titled ”Soft Computing Methods for Microwave and Millimeter-wave Design Problems” (Springer, 2012).