I J C T A, 9(3), 2016, pp. 1671-1677

Egocentric Activity Recognition Using Bag of Visual Words

K.P. Sanal Kumar*, R. Bhavani and M. Rajaguru*****

Abstract: This paper presents an approach for recognizing activities using video from the egocentric setup. In this approach instead of using intermediate setup like object detection, pose estimation, modeling spatial distribution of visual words is implemented. The interactions are encoded by using Histogram oriented Pairwise Relation named (HOPR) between the visual words, orientations and alignments. A codebook is generated using a bag of visual words. This identifies the daily activities from the egocentric video.

Keywords: Egocentric, object detection, Histogram Oriented Pairwise Relation, codebook, Activity recognition,

1. INTRODUCTION

Activity recognition is a salient area of computer vision. Egocentric is thinking, only of oneself, without regard for the feelings or desires of others. This defines an egocentric activity recognition performed by using video from a first person view setup. In activity recognition, there is an interaction between objects and hands. Activity recognition is achieved to recognize the actions and wishes of one or more person's from a sequence of observations on the agent's actions and the conditions of environment.

In this paper, activities are recognized using video from a wearable camera (first person view). Activity recognition has received increasing attention due to its most important applications such as intelligent surveillance system, human computer interaction and smart monitoring system. Research scholars are now advancing from recognizing simple periodic actions like "cooking", "making tea", "vegetables cutting" to more critical and challenging activities involving multiple person and multiple objects, it has been increasing the interest in activity recognition from an first person (egocentric) approach using first person wearable camera's. These approaches are designed to differentiate the activities after fully regarding the entire sequence. Assuming each video contains a complete execution of single task activities. However, some features are alone used and often it is not enough for modeling the complex activities as the same action patterns can produce a various moment's patterns. For example, while making pizza one can pour water using one hand while the other hand was used for mixing and perform actions simultaneously using one hand.

2. RELATED WORK

A. Fathi. et al. presented a method to analyze daily activities, such as meal preparation, using video from an egocentric camera. Their method performed on the inference about activities, actions, hands, and objects. Daily activities are a challenging domain for activity recognition which were well-suited to an egocentric approach. In contrast to previous activity recognition methods, this approach did not require pre-trained detectors for objects and hands. Instead they demonstrated the ability to learn a hierarchical model of an activity by exploiting the consistent appearance of objects, hands, and actions that results from the egocentric context. They had shown that joint modeling of activities, actions, and objects led to superior performance

* Programmer Dept. of EEE Annamalai University, *Email: sanalprabha@yahoo.co.in*

** Professor, Department of Computer Science and Engineering Annamalai University, *Email: bhavaniaucse@gmail.com*

*** *PG Scholar* Dept. of CSE Annamalai University, *Email: gurubtechme@gmail.com*

in comparison to the case where they were considered independently. They introduced a novel representation of actions based on object-hand interactions and experimentally demonstrated the superior performance of our representation in comparison to standard activity representations such as bag of words. H. Bay et.al presented a novel scale-and rotation-invariant interest point detector and descriptor, coined SURF (Speedup Robust Features). It approximated or even outperformed previously proposed schemes with respect to repeatability, distinctiveness, and robustness, yet can be computed and compared much faster. This was achieved by relying on integral images for image convolutions; by building on the strengths of the leading existing detectors and descriptors (in case, using a Hessian matrix-based measure for the detector, and a distribution-based descriptor); and by simplifying these methods to the essential. This lead to a combination of novel detection, description, and matching steps. The paper presented experimental results on a standard evaluation set, as well as on imagery obtained in the context of a real-life object recognition application. Both shows SURF's strong performance. A. Behera. et.al presented a method for real-time monitoring of workflows in a constrained environment. The monitoring system should not only be able to recognize the current step but also provide instructions about the possible next steps in an ongoing workflow. In this paper, they addressed this issue by using a robust approach (HMM-pLSA) which relied on a Hidden Markov Model (HMM) and generative model such as probabilistic Latent Semantic Analysis (pLSA). The proposed method exploited the dynamics of the qualitative spatial relation between pairs of objects involved in a workflow. The novel view-invariant relational feature was based on distance and its rate of change in 3D space. The multiple pair-wise relational features were represented in a multi-dimensional relational state space using an HMM. The workflow monitoring task was inferred from the relational state space using pLSA on datasets, which consisted of workflow activities such as "hammering nails" and "driving screws". The approach is evaluated for both "off-line" (complete observation) and "online" (partial observation). The evaluation of this approach justifies the robustness of the technique in overcoming issues of noise evolving from object tracking and occlusions. A. Behera. et.al presented a novel approach for real-time egocentric activity recognition in which component atomic events were characterized in terms of binary relationships between parts of the body and manipulated objects. The key contribution was to summarize, within a histogram, the relationships that hold over a fixed time interval. This histogram was classified into one of a number of atomic events. The relationships encode both the types of body parts and objects involved (e.g. wrist, hammer) together with a quantized representation of their distance apart and the normalized rate of change in this distance. The quantization and classifier were both configured in a prior learning phase from training data. An activity was represented by a Markov model over atomic events. They shown the application of the method in the prediction of the next atomic event within a manual procedure (e.g. assembling a simple device) and the detection of deviations from an expected procedure. This could be used for example in training operators in the use or servicing of a piece of equipment, or the assembly of a device from components. They evaluated their approach ("Bag-of-Relations") on two datasets: "labelling and packaging bottles" and "hammering nails and driving screws", and shown superior performance to existing Bag-of-Features methods that worked with histograms derived from image features. Finally, they show that the combination of data from vision and inertial (IMU) sensors outperforms either modality alone.

3. PROPOSED TECHNIQUES

Egocentric activity recognition system consists of five major components. These are 1) Video to frame conversion, 2) Key points detection, 3) k-means clustering, 4) Codebook (vector quantization) generation, 5) Activity recognition. Proposed system given in the figure 1.

3.1. Videos to Frame Conversion

It is the initial stage of implementation the inputs are egocentric activity videos from egocentric database. These are videos converted into set of frames based on the frame rate, by video to jpg converter software.

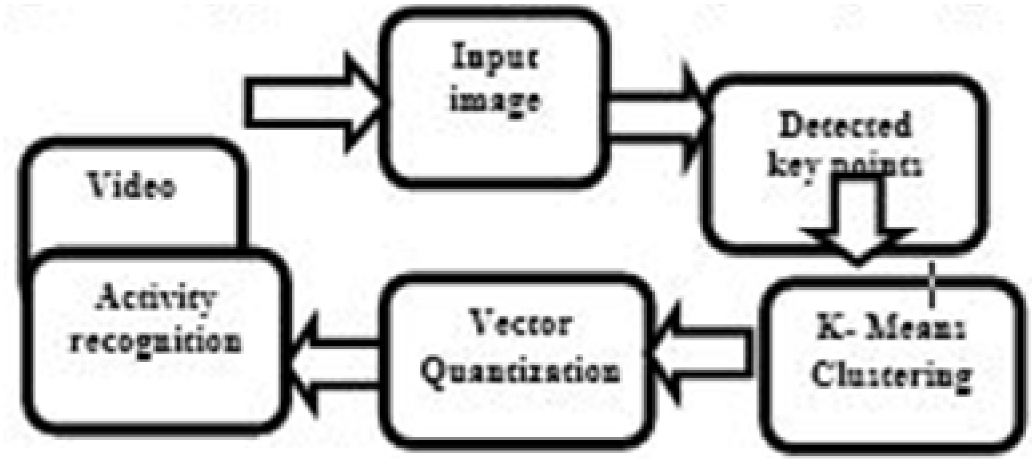

Figure 1: Block diagram

Figure 2: Dataset frames

In this paper 240 × 240, 320 × 240, 480 × 360, 1280 × 720 size frames are used. At each second one frame is extracted therefore finally 900 frames are extracted in a single video data set from the software. Later it was resized into 512 × 512 pixel size. Example Dataset frames are given in figure 2.

3.2. Key Points Detection

In this stage, key points are found from the image plane with the help of SURF feature. SURF feature is used as an object detector or blob detector based on the hessian matrix to an image plane. SURF is also use as the element of the hessian for selecting the scale, given a set point ρ = (x, y) in an image plane. Hessian matrix H(ρ, *í*) at point and scale í, is defined as follows:

$$H(\rho,)=\begin{bmatrix} L_{xx}(\rho,) \\ L_{xy}(\rho,) \\ L_{xy}(\rho,) \end{bmatrix}$$

where, L_{xx}(ρ, í) are the second order derivatives of the gray scale image. Detecting 300 key points in the image, all the key points are sorted by its strength. Strength will be defined as radius of the circle. Each key point is iterated from highest to lowest strength. Lowest strength key points are ignored from the set. The

Figure 3: SURF features

single task of finding correspondence between two images of same scene or object is part of most of the computer visions applications. Process of key point detection in the image is shown in the figure 3.

3.3. K-MEANS CLUSTERING

k-Means clustering is a process of vector quantization method, initially from image processing, that is cluster analysis in data mining. k—means clustering intent to barrier n perception into k-clusters with the nearest means, clusters are formed in training images by using k-means clustering. These clusters are formed based on center point. This process uses 5 clusters, where more clusters may lead to fail in calculating the nearest mean.

$$j = \sum \left\| x_i - c_i \right\|^2$$

Where $\|x_i – c_j\|^2$ is a chosen distance measure between a data point x_i^j and the cluster center c^j-is an indicator of the distance of the *n* data points from their respective cluster centers.

The algorithm is derived by the following steps:

Step 1: Place k-points into the image represented by the objects that are being grouped. These points represent the group centroids.

Step 2: Assign each object to the group that has the closest centroid.

Step 3: When all objects have been assigned, recalculate the positions of the k-centroids. Repeat Steps 2 and 3 until the centroids no longer move.

Step 4: This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

3.4. Code Book

Code book is a vector quantization method. It is a group of code words which are in matrix. The codebook is converted into matrix form separately at the dimensions of 256 × 256. All training datasets are available in codebook. Code words are encoded in codebook based on spatial distance. Encoding is performed by image1\1.jpg

256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256
256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256
256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256
256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256
256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256
256	256	256	256	256	256	256	256	256	256	256	255	254	256	256	256	256	256	256	256	256
256	256	256	256	256	256	256	256	256	255	250	250	253	256	256	256	256	256	256	256	255
256	256	256	256	256	256	256	256	256	255	253	254	256	256	256	256	256	256	256	256	255
256	256	256	256	256	256	256	255	255	252	255	256	256	256	256	255	253	254	252	212	158
256	256	256	256	256	256	255	253	252	256	256	256	256	256	256	254	198	73	80	126	151
256	256	256	256	256	254	253	254	256	256	256	256	256	252	253	246	153	179	138	60	51
256	256	256	256	256	255	255	255	256	256	256	256	256	253	216	92	73	106	188	203	125
256	256	256	256	256	256	256	256	256	256	255	252	255	158	101	210	152	78	39	116	177
256	256	256	256	256	256	256	256	256	256	255	255	238	150	100	61	174	196	123	110	118
256	256	256	256	256	256	256	256	256	256	253	224	72	150	191	120	43	64	179	223	119
256	256	256	256	256	256	256	256	256	256	255	125	143	58	27	144	176	75	34	97	140
256	256	256	256	256	256	256	256	255	252	225	72	61	145	172	41	65	183	212	145	60
256	256	256	256	256	256	256	255	253	235	18	203	256	148	15	187	190	44	73	182	163
256	256	256	256	256	256	256	256	253	84	167	229	100	87	151	74	138	198	140	83	138
256	256	256	256	256	256	256	256	202	97	138	43	166	92	84	173	82	40	137	213	234
256	256	256	256	256	255	255	255	211	164	81	92	66	175	83	39	179	154	114	114	51
256	256	256	256	256	254	255	233	113	84	206	21	174	71	193	118	77	138	227	42	158
256	256	256	256	256	256	253	214	199	102	139	125	78	171	112	137	64	98	39	96	21
256	256	256	256	256	256	181	118	112	195	125	151	74	143	105	79	70	59	61	91	104
256	256	256	256	256	254	194	198	147	132	194	104	169	86	91	34	155	64	109	132	62
256	256	256	256	256	251	245	122	171	87	149	127	78	95	127	235	76	80	126	100	52
256	256	256	256	256	253	173	106	63	224	70	87	76	122	190	42	112	113	70	96	82
256	256	256	256	256	252	149	70	230	54	117	53	121	104	89	90	165	23	136	35	132
256	256	256	256	256	253	228	159	37	161	40	130	43	116	49	154	56	116	78	140	48
256	256	256	256	256	256	96	41	221	44	188	60	137	27	117	142	110	48	126	73	175
256	256	256	256	256	244	17	221	83	130	110	121	40	110	57	70	192	61	118	108	217
256	256	256	256	256	240	195	128	68	140	93	39	130	49	170	73	106	161	82	131	185
256	256	256	256	256	252	138	76	96	110	30	165	40	144	78	92	61	202	25	67	169
256	256	256	256	256	173	108	108	112	140	118	84	105	83	112	87	44	169	68	88	145
256	256	256	256	256	256	205	91	235	46	167	62	55	23	145	74	58	153	140	119	116
256	256	256	256	256	126	98	229	72	182	38	135	86	59	146	83	23	167	115	167	41

Figure 4: Codebook matrix form

using interaction between visual words. The closeness matching property of vector quantization is powerful, notably for identifying the density of large and maximum dimension data, seeing that points are represented by the index of their closet centroid, commonly arising data have low error and rare data high error. Single image codebook matrix form shown in the fig 4.

3.4. Activity Recognition

Testing images are to be tested in codebook, the spatial distance of code word are calculated to classify what kind of images are tested. Interactions are encoded with their spatial distances. Distance is calculated by using Gaussian distance. Interactions are encoded using the HOPR named Histogram Oriented Pairwise Relation which is an enhancement technique. Finally the activity is recognized using the codebook representation.

4. EXPRIMENTAL RESULT AND ANALYSIS

SURF feature is used for feature extraction and 300 key points are selected using the trial and error method. Accuracy and the key points selection is shown in table 1. Accuracy will be decreased when the key points increased above 300.

The following table 2 defines the activity recognition rate by this approach.

Table 1
Accuracy of key points

Activity Types	*No. of test images*	*No of recognized*	*Recognition Rate*
Pizza making	10	9	80.00
Coffee making	10	8	80.00
Vegetable Cut	10	7	70.00
Packaging & Labeling	10	8	80.00
Cooking Activity	10	7	70.00

Table 2
Recognition percentage

Interesting points	*Accuracy*
50	80
100	86.5
200	87
300	89
400	87.5

5. SAMPLE SCREEN SHOTS

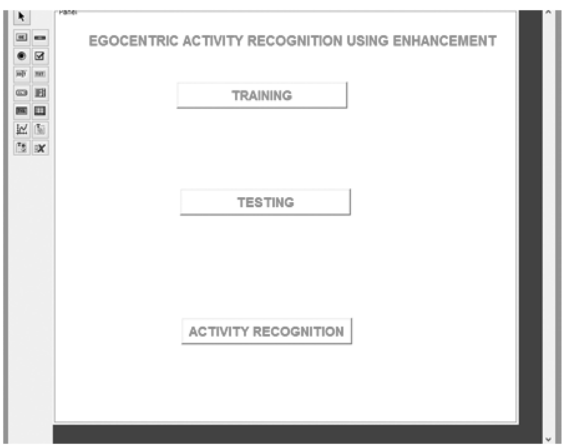

Figure 5: GUI Tool

Figure 6: Input frames

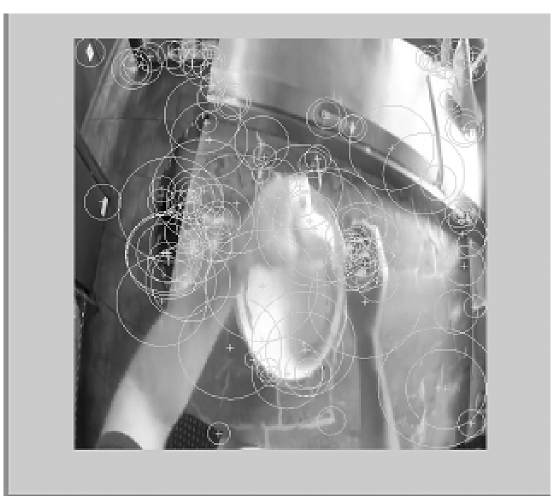

Figure 7: SURF features

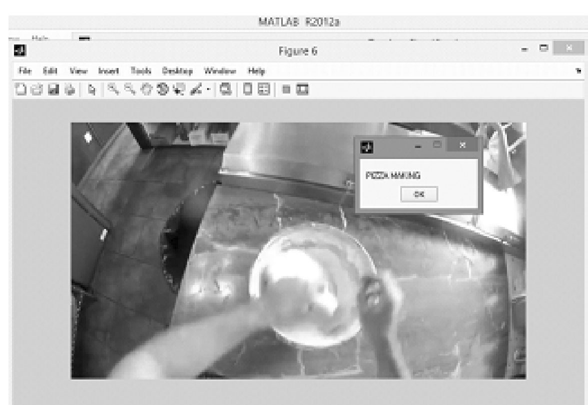

Figure 8: Activity Recognition

6. CONCLUSION AND FUTURE WORK

In this paper, video images are taken from GTEA & LEEDS Database. First the video is converted into frames. Each frame is resized in 512 × 512 size. Then using SURF Features key points are extracted from the images. These key points are clustered using k-means clustering algorithm and stored in a codebook using vector quantization method. When query images are given as input, the preprocessing work and feature points extraction are done in the query images. It has been classified using k-means clustering and the activities are recognized from the datasets. The Applications of Activity recognition are intelligent robots, monitoring system for children and elderly persons, intelligence surveillance systems, human-computer interaction and smart monitoring.

The future work can be done by implementing various enhancement techniques. The classifiers like svm, knn and other neural network models can be implemented for activity recognition.

References

[1] Behera A., Chapman M., Cohn G and Hogg (2012). "Egocentric Activity Recognition using Histograms of Oriented Pairwise Relations".

[2] Fathi, A., Farhadi, A., and Rehg, J. M. (2011a). Understanding egocentric activities. In ICCV, pages 407– 414.

[3] Bay, H., Tuytelaars, T., and Gool, L. V. (2006). SURF: Speeded up robust features. In ECCV, pages 404–417

[4] Behera, A., Cohn, A. G., and Hogg, D. C. (2012a). Workflow activity monitoring using dynamics of pair-wise qualitative spatial relations. In MMM, pages 196–209.

[5] Behera, A., Hogg, D. C., and Cohn, A. G. (2012b). Egocentric activity monitoring and recovery. In ACCV, pages 519–53.

[6] Matikainen, P., Hebert, M., and Sukthankar, R. (2010). Representing pairwise spatial and temporal relations for action recognition. In ECCV (1), pages 508–521.

[7] Shechtman, E. and Irani, M. (2007). Matching local self similarities across images and videos. In CVPR.

[8] Liu, W., Li, S., and Renz, J. (2009). Combining rcc-8 with qualitative direction calculi: Algorithms and complexity. In Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI), pages 854–859.

[9] Carneiro, G. and Lothey, D. (2006). Sparse flexible models of local features. In ECCV, pages 29–43.

[10] Lothey, D.G.(2004). Distinctive image features from scale invariant keypoints. International Journal of Computer Vision, 60(2):91–110.

[11] Fathi, A., Ren, X., and Rehg, J. M. (2011b). Learning to recognize objects in egocentric activities. In CVPR, pages 3281–3288.

[12] Niebles, J. C. and Li, F.F. (2007). A hierarchical model of shape and appearance for human action classification. In CVPR, pages 1–8.

[13] Ryoo, M. S. and Aggarwal, J. K. (2009). Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In ICCV, pages 1593– 1600.

[14] Laptev, I. and Lindeberg, T. (2003). Space-time interest points. In ICCV, pages 432–439.

[15] Sun, J., Wu, X., Yan, S., Cheong, L. F., Chua, T.-S., and Li, J. (2009). Hierarchical spatio-temporal context modeling for action recognition. In CVPR, pages 2004– 2011.