Journal of Engg. Research **Vol. 5 No. (2) June 2017 pp.** *159-169*

دور ميزة الوظيفة في الاعتراف التفاعل سيارة البشري

هينا أرشد*، خرام خورشيد* ومحمد هارون يوسف**

* قسم الهندسة الكهربائية، معهد تكنولوجيا الفضاء، 1، إسلام آباد السريع ، إسلام أباد، باكستان 44000

** قسم هندسة الحاسوب، جامعة الهندسة والتكنولوجيا، تاكسيلا، 47050، باكستان

الخلاصة

تقترح هذه الورقة تحسنا في حقيبة من الكلمات (القوس) نموذج عن طريق إدخال موقف المكاني من الميزات في تمثيل التفاعل سيارة البشرية مرحلة التفاعل للاعتراف. تدرج مواقف المكانية من الميزات جنبا إلى جنب مع واصف ميزة الحصول على المعلومات الهيكلية اللازمة لتصنيف نوع مختلف من المركبات تفاعل الإنسان بشكل صحيح. القوس يفتقر إلى المعلومات، فضلا عن العلاقة البنيوية بين الميزات الزمانية المكانية. هذا يجعل نهج القوس تهت، في شكله الأساسي، مركبة غير فعالة للاعتراف التفاعل البشري. النهج المقترح يحسن تمثيل القوس من خلال التأكيد على دور الفضاء لشغل وظائف التفاعل وميزة الزمانية المكانية، ويتضمن العلاقة بين الميزات. و-تم اختبار هذا النهج عن حالة مجموعة البيانات الفن ودقة 92.8 ٪ منجز. وقد تم-تبين أن نهجنا ينفذ حقيبة من الكلمات نهج ودولة أخرى من الأعمال الفنية.

Role of Feature Position in recognition of Human Vehicle Interaction

Hina Arshad*, Khurram Khurshid* and M Haroon Yousaf**

* *Department of Electrical Engineering, Institute of Space Technology, 1, Islamabad Highway, Islamabad, 44000. Pakistan.*

** *Department of Computer Engineering, University of Engineering & Technology, Taxila, 47050, Pakistan.*

**Corresponding Author: hinaarsshad@gmail.com*

ABSTRACT

This paper proposes an improvement in the Bag of Words (BoW) model by introducing spatial position of features in interaction representation stage for Human Vehicle Interaction Recognition. The spatial positions of features are incorporated along with feature descriptor to obtain the structural information required to correctly classify different kinds of human vehicle interaction. BoW lacks structural information as well as the relationship between spatiotemporal features. This renders this BoW approach, in its basic form, ineffective for human vehicle interaction recognition. The proposed approach enhances the BoW representation by emphasizing the role of spatial feature positions for interaction and incorporates the spatial and temporal relationship between the features. This approach has been tested on the state of the art dataset and achieved accuracy of 92.8%. It has been found that our approach out-performs the Bag of Words approach and other state of the art work.

Keywords: Bag of words (BoW); human-vehicle-interaction-recognition; spatiotemporal features; spatial positions.

INTRODUCTION

With the increasing demand for security and surveillance, human vehicle interaction has emerged as a pivotal research area in computer vision. This problem of visually recognizing human vehicle interaction in real environment is more challenging than simple activity recognition. It involves many key challenges such as scale variations, occlusions, illumination variations, view point changes, etc. In order to provide the recognition of human vehicle interaction in real environment, there should be an effective interaction representation and classification strategy. From simple gesture recognition to complex interaction recognition, BoW approach has been frequently used for activity or interaction representation. However, BoW suffers from lack of structure information in its representation. This leads to lower recognition performance. There have been a few efforts made to improve bag of words model such as spatial pyramid matching (Lazebniket al., 2006) or approaches that tend to include spatiotemporal context information (Wu et al., 2011) or local feature distribution (Bregonzio et al., 2012). We propose a new approach to modify the BoW representation to work effectively for human vehicle interaction recognition. In context of human vehicle interaction recognition, our work is the first one to report improvement in BoW model.

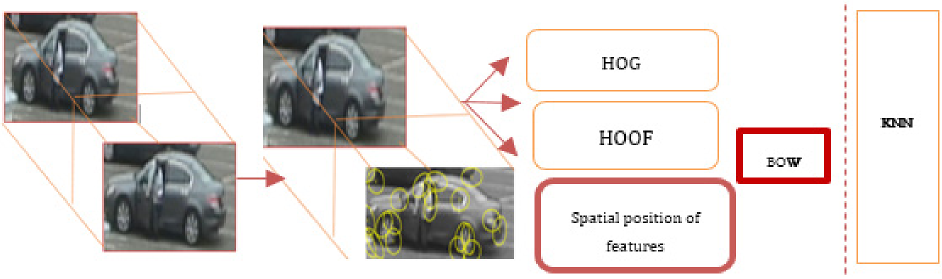


Fig. 1. Flow chart of our proposed approach

The conventional analysis of video involves tracking and motion estimation. However, tracking and segmentation is difficult in scenarios, where the background is non-stationary and appearance model changes with the evolution of frames. For this reason, we utilized the key places and key frames of the video. These key components carry most of the information and represents what is happening in the videos. On the basis of this, we extracted the key space time points that saliently characterize the interactions of video. In our approach, features are extracted from the video using spatial and temporal interest point detectors. For this task, we use 3D cuboid approach to extract the human vehicle interaction features. In order to describe these points, we employed histogram of oriented gradient (HOG) to capture the spatially discriminant patterns and histogram of oriented optical flow (HOOF) to describe the motion patterns associated with human vehicle interaction. To make the BoW representation enriched with more information, we introduced the concept of spatial positions of features to be incorporated in the final descriptor. These descriptors are then subjected to K means clustering for codebook learning and then histograms of features are obtained to provide feature representation. For classification purpose, we used K Nearest Neighbors (KNN) to recognize the human vehicle interaction. The flow chart of our proposed approach has been presented in Figure 1 and its detailed description is provided in coming sections in detail.

The aim of this work is bi-fold. Firstly, it provides an improvement in the BoW representation. Secondly, our work shows how BoW can serve to recognize human vehicle interaction with better classification results.

The paper is organized to provide literature review in the next section. After that it discusses the proposed approach in detail. The experimental setup has been drafted under the fourth main heading. Thereafter, recognition results and efficiency of approach are provided. Finally, the conclusion has been drawn at the end of the paper.

LITERATURE REVIEW

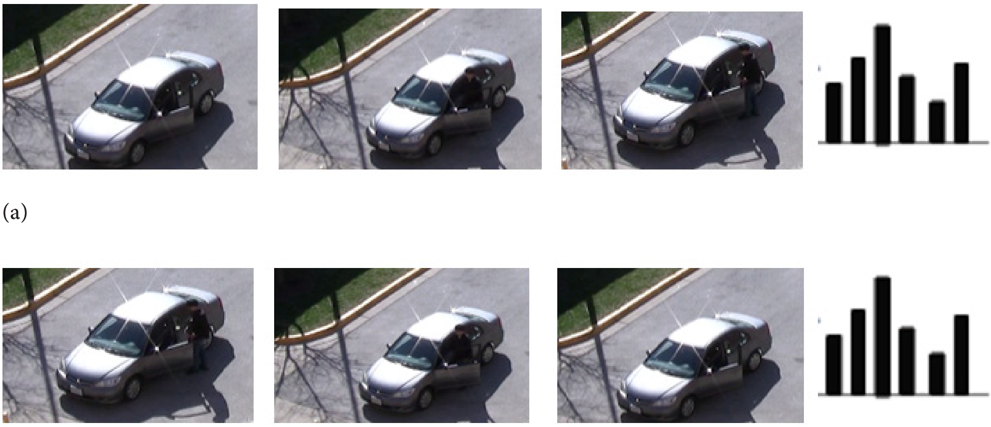
To recognize interaction in videos, there is a need for developing an effective activity representation. For interaction recognition in natural scenes or complex environment, inclusion of structural information in activity representation plays an important role. Here the literature review of activity representation is described, keeping in view which type of structural information is considered to be added and how this structural information can be incorporated in interaction representations.

In the spatial representation, the spatial configuration of parts is modeled. The spatial configuration of parts is modeled on low level features. This part based representation makes use of the technique as applied in part base representation of objects in still images (Usman et al., 2016). The approach of Wang & Mori (2013) involves developing frame level, hidden part model with basis as local motion features. They processed the video frame by frame. They used their part based model on each frame and employed majority voting to analyze the content of video. In the same way, Tian et al. (2013) created deformable part model (DPM) for video that organizes the discriminative parts over time. This is based on motion and local appearance described using HOG3D. This spatial representation has spatial structure that is enough to distinguish the activities with parts and diverse appearances. However, it fails to distinguish the patterns of the same parts, but in different temporal sequence as for example, the person getting in and getting out has same parts but constitutes different temporal sequence.

Another representation that is adopted by research community involves temporal representation. In temporal based representation, there are sequential-temporal representation and exemplar-temporal representation. In sequential temporal representation, activity progression is captured using sequence of hidden states that are inferred from motion or appearance observations. Yamato et al. (1992) proposed Hidden Markov Models (HMM) of an activity. This HMM of an activity observes the series of appearance symbols over the entire video frames. This model is tuned by making it to learn appearance symbols. Once it gets tuned, it assigns higher probability to the series of symbols which closely matches the learned series of symbols. Further extending this model, Tang et al. (2012) modeled the duration of each and every state of HMM during temporal evolution of the video. These temporal models or representation are quite robust to shift in time as well as time variance in progression of activities. However, such representation lacks the spatial structural that provides expressive power for activity representation. In exemplar temporal representation, temporal composition of a video is represented using sequence of templates. These template methods have their basis in low level features. These series of templates are rigid with little flexibility for variations in activity length. Efros et al. (2003) made use of the same technique and developed motion descriptors on every frame of track and calculated its cross correlation score with samples of activity database. The sample with best matching score was used to represent the activity. Brendel & Todorovic (2011) developed a model with a bit flexibility that created exemplars by keeping track of regions of activity-track with distinct motion and feature patterns. However, the limitation of this approach lies in its insufficient generalization to the samples that do not have close match with any templates. This can be improved by adding spatial information and spatial relations. Vahdat et al. (2011) incorporated spatial relations in exemplar based approach and developed key pose sequence model.

Key component representation has been introduced to represent the complex activity and interaction. In this kind of representation, activity is expressed as discrete series of discriminative components that are based on motion or appearance features. Niebels et al. (2010) developed key component model using pooled HOF and HOG features. These are extracted on interest points and span over several frames. Rapti & Sigal (2013) also made use of key component representation and introduced frame level key poses. However, this approach is blind to some spatial relationships. Making use of both spatial and temporal compositions of video provides the activity representation with intensive expressive power. For instance, Intelli & Bobick (2001) used spatial and temporal cues associated with football players and football to represent the activity such as football game.

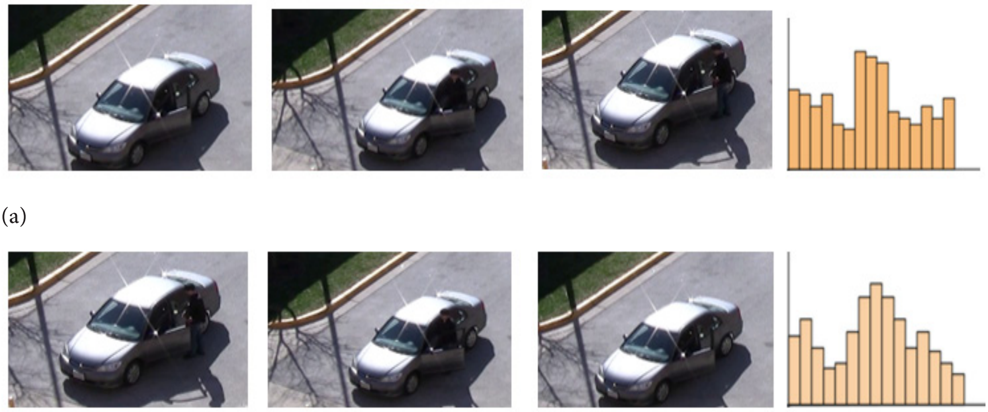
The most popular representation is the BoW representation. In this form of activity representation, low level features are extracted over the entire video. These features are then aggregated in histogram form. This results in lack of spatial and temporal structure information in the interaction model. In this context, Schuldt et al. (2004) and Niebels et al. (2010) used this approach. Schuldt et al. (2004) used the technique of extracting motion features for primitive events in the video and then aggregated motion and appearance information in the form of jets of histograms. They clustered these local motion and appearance features and utilized this clusters to make vocabulary of words.



(a)

(b)

Fig. 2 (a) & (b).Shows that the BoW model represent getting into the car and getting out of the car as the same interaction with almost same histograms.



(a)

(b)

Fig. 3 (a) & (b). Shows BoW representation based on our approach. Our approach differentiated getting into the car and getting out of the car as the different interactions with different histograms

These were then used to get the BoW representation of video. In the same way, Niebels et al. (2010) identified complex motion regions and described them using gradient descriptor. Afterwards, the clustering is carried out to make vocabulary of these features, which is then shaped into BoW representation. However, the histogram based representation or the bag of words representation throws away the discriminative structure information. In our approach, we equip this BoW representation with structural information by incorporating spatial positions of features along with their descriptors.

PROPOSED APPROACH

In our approach, the BoW representation is subjected to improvement by introducing structural information. The detailed description of our approach has been provided in the following sub sections.

Extraction of spatiotemporal cuboids

The salient features are detected by employing the spatial and temporal interest point detector. For this, the frames are filtered at various space time scales. We employed spatial Gaussian with the Gabor filtering (Dollar et al., 2005). Dollar et al found that this Gabor filtering method overcomes the issues of sparsity of STIP points and provide stable interest points. The Gabor filters used are given in Equation (1) and Equation (2).

$$h_1(t;\tau,\omega) = -\cos(2\pi\tau\omega)\, e^{-\frac{t^2}{\tau^2}} \qquad (1)$$

$$h_2(t;\tau,\omega) = -sin(2\pi\tau\omega)e^{-\frac{t^2}{\tau^2}} \qquad (2)$$

Where h_1 and h_2 denote the quadrature1D Gabor filters pair, where h_1 is even and h_2 is odd. Here, $\omega = 4/\,\tau$ and corresponds to the temporal scale. The output of these filters is used to compute the response function. The energy of the filter response is computed. The points having maximum energy in the local volume are obtained as center of cuboids. The spatial temporal points are obtained along with their positions. These points are subjected to more robust description to characterize the discriminant spatial and temporal occurrences in the frames. In conventional approaches, the positions of these features are discarded and only descriptors are obtained around these points to get the final descriptors. However we will take into account the spatial position of these features in the final descriptor.

Feature descriptors

To achieve R-HOG based spatially discriminative description of human vehicle interaction; firstly the image is divided into number of cells. For each cell, first step is to compute the gradient values. The kernels used to find the gradients are given in Equation (3) and Equation (4),where denotes the kernel in x direction and denotes the kernel in y direction. These kernels provide the first order gradients of image that are needed to capture the silhouette, contour and texture information required for HOG feature computation. The employment of these kernels also results in providing illumination variation.

$$M_x = \begin{bmatrix} -1 & 0 & 1 \end{bmatrix} \qquad (3)$$

$$M_y = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \qquad (4)$$

Using these, image gradient and orientation are calculated. In the second step, the binning of orientations in each cell is done. This is one of the basic parameters, as it leads to the creation of orientation based channel of histogram. The angles range from 0-360 degrees is categorized into bins. For this, we employed the approach of binning the orientations form 0-180 degrees. Any orientation value that is negative is placed at the location of orientation value plus 180 degrees. This binning provides the ability to distinguish between transitions in images such as from dark to light or from dark to dark. In order to capture the finer details of orientations, the number of these bins can be increased accordingly. Using this binning, histograms are created for each cell. For catering the contrast and illumination changes, the magnitude of the gradients is subjected to normalization. Normalization is obtained by grouping cells into blocks. By employing this, a feature vector is obtained to represent the spatially discriminant patterns. This feature vector length varies as we change the parameters of number of bins, number of cells, number of blocks and block overlap.

For the sake of capturing the characteristic motion features, HOOF is computed. HOOF is preferred over the state of the art optical flow method, as optical flow is susceptible to scale changes and movement directionality. The optical flow magnitude is greater in zoomed cases as compared to others. In our scenario of human vehicle interaction, directionality of motion and scale matter. Therefore, HOOF is applied to get the scale and directionality invariant profile of motion for interaction. For this, the optical flow is computed on every frame with interaction. The flow vectors, thus obtained, are subjected to binning according to the primary angle from horizontal-axis. Each binned flow vector is weighted keeping in view its magnitude. In this way, each optical flow vector with direction in the range Φ will put its contribution to the summation in bin. The direction range is given as below

$$-\frac{\pi}{2} + \pi\frac{b-1}{B} \leq \Phi < -\frac{\pi}{2} + \frac{\pi b}{B} \qquad (5)$$

Where, B is the total number of bins, Φ is the direction range of flow vector and b is a bin out of the total number of bins B. The histograms are finally normalized to 1.The number of bin B is a parameter that can be set according to the requirement. For our proposed approach, we used 30 numbers of bins, B, for better recognition results.

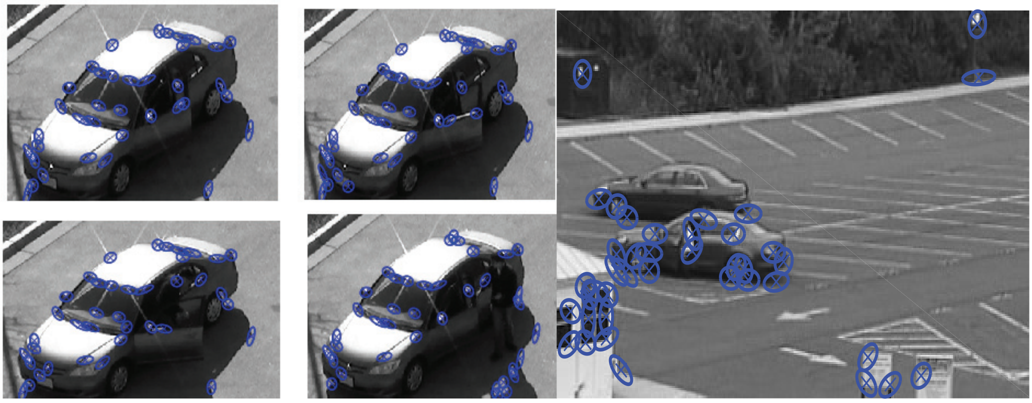


Fig. 4. Computation of salient points using 3D Cuboid Detector

Feature position in BoW representation

Once the descriptors are obtained, the BoW model approach is used to represent the interaction. Unlike the BoW approach, we introduce the BoW to be built upon the descriptors along with the feature spatial position. The role of feature spatial position is crucial for recognition of interaction that needs to be distinguished according to temporal sequence configuration. Every histogram is incorporated with position features. This changes the histogram representation for interactions. The person getting in and getting out of the car under the same scenario would have the same histogram representation. However, when position features are incorporated, these histograms can be distinguished and characterized by both interactions differently. The incorporation of feature position provides discriminative power to the histogram representation. This is elaborated in Figure 2 and Figure 3.

The features are extracted from the frames. These HOG and HOOF along with feature positions are used to learn the visual vocabulary. For this, K means clustering is used. In dictionary learning, K means tend to reduce the sum of squared Euclidean distances, i.e. the distances between the extracted discriminative points and their most neighboring centers of clusters. It can be represented as:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2 \qquad (6)$$

Where, the term $||x_i^{(j)} - c_j||^2$ shows the measure of distance between the data point$x_i^{(j)}$ and the cluster center$_j$,n corresponds to the total number of data points and k is the total number of centers. The K means clustering is started with initializing the K-clusters centers. Each data point is assigned to the neighboring cluster center. Each center is re-computed by taking the average of the data points and the newly assigned data point. This changes the cluster center. This keeps on iterating until there is no further change in K-clusters centers.

This unsupervised method is adopted to learn the visual dictionary for human vehicle interaction. Each centroid of cluster generated by K means is called a code vector. A code vector is also referred as visual word. The interaction is represented by the histograms of code vectors. Once the dictionary is obtained, it is used for quantization of features. Vector quantizer gets a feature vector and then maps it according to the index of closest code vector in the dictionary.

EXPERIMENTAL SETUP

Our approach has been tested on the state of the art video dataset, VIRAT video dataset Ground Release 2. This dataset has been recorded in real conditions such as parking lot. It does not contain the made up scenarios involving volunteer actors. Rather it is recorded in real scenarios with varying light condition, occlusions, view point changes, etc. We targeted two human vehicle interactions such as getting into the car and getting out of the car to be recognized by our approach. The similarity in the spatial patterns and varied configuration in temporal order of these patterns make it difficult to be recognized by simple BoW. We employed our approach on these two activities and compared it with simple BoW model. The experiments have been carried out on the Intel core 5 within MATLAB platform. We have based our experimental testing on the parking lot videos including VIRAT S 0401, VIRAT S 0000 and VIRAT S 0502. The duration of these videos ranges between 2 to15 minutes.

Different tests were done to evaluate the effect of the parameters: spatial and temporal scales, number of HOG bins, HOOF bins, and vocabulary size. The final evaluation is made by keeping in view the test results and the state of the art literature. For extracting spatial and temporal interest points, firstly, the region of interest containing the interactions is generated. 3D spatiotemporal Cuboids are extracted to reduce the noise level and redundancy. For this experiment $\omega = 4/\tau$, scales of 2 and 4 are chosen to be spatial and temporal scales. These scales provide the ability to capture the information and do not add much computational burden. The computed spatial and temporal interest points have been shown in Figure 4.The HOG is computed around the extracted points. Number of bins for histograms of oriented gradient is kept 8. This number of bins is selected by keeping in view the effect of choosing bin numbers on miss rate and false positive. To equip the BoW model with motion features, HOOF is applied to describe the extracted spatio temporal points. The LK optical flow is implemented and then HOOF is built upon them with 30 numbers of bins that are found to provide better results. The positions of features are also connected with descriptors using average pooling. The size of vocabulary is kept 500. Larger size of vocabulary provides better performance but at the cost of storage space and computational efficiency. Furthermore, literature depicts that for any database, there should be some optimized vocabulary size. We have found that keeping the size of vocabulary to be 500 provides suitable performance keeping in view the factors such as computation time, storage space and recognition results. Keeping the size up to 700 would have enhanced the recognition; however, it was not computation efficient. For the classification purpose we employed KNN classifier. The aim of choosing KNN is to check the performance of our approach with a baseline classifier.

RECOGNITION RESULTS

The overall accuracy of the recognition for getting into the vehicle and getting out of the vehicle has been found to be 92.8 % on the video set of Ground Release 2. These include all those frames that had noise, clutter, camera jitter, light variations and occlusions. We used our approach to calculate the accuracy of recognition for each case and compared it with the accuracy obtained using BoW +KNN as shown in Table 1.

Table 1. Comparison of accuracy of interaction recognition with Bow and our approach

Interaction to be recognized	Simple BoW+KNN	Our approach
Person getting into the car	38.4%	93.8%
Person getting out of the car	34.4%	91.8%

We have also compared our approach with the baseline method of BoW and that of Reddy et al. (2012).These results have been tabulated in Table 2. The accuracy has been calculated using $(TP + TN)/(TP + TN + FP + FN)$. Where, TP corresponds to true positive, TN is true negative, FP denotes false positive and FN depicts false negative. It can be clearly seen that with simple BoW, there is low accuracy. The accuracy achieved by our model clearly depicts that incorporation of feature position that has long been discarded in tradition BoW model leads to better representation and thus recognition.

Table 2. Comparative results

Method	Recognition Performance (Accuracy)
BoW representation	36.7%
STHOG based BoW Representation (Reddy *et al.*2012)	44.7%
Our Approach	92.8%

The results approach promise recognition of human vehicle interaction in complex scenarios. It has been found that the incorporation of structural information via using feature spatial position enhances the discriminative power of interaction representation. The role of spatial features position is to enhance and capture the essence of interactions. It provides space and time relationship by integrating it in representation model of BoW. In an attempt to recognize the interaction by BoW, we found that the histogram excludes the structural information by throwing away feature position. This approach is more suitable for the interactions that are same but varies in spatial and temporal order. However, incorporating the feature position increases the computational demand for representation and recognition of human vehicle interactions.

EFFICIENCY OF OUR APPROACH

There is always a tradeoff between computational efficiency and accuracy of any scheme employed for action recognition. In our approach, there are 4 main components that make the approach complex. These include computation of STIP, computation of feature vector, BOW representation and recognition. However the selection of these components and their parameters are made by keeping in view the tradeoff between recognition accuracy, complexity and computation time. Each component's average computation time is computed for all kinds of scenarios such as cluttered, occluded, noisy views and is tabulated in Table 3. The approach has been implemented on MATLAB installed on Intel Core i5 CPU platform with 2.4GHz and a 4GB RAM.

Table 3. Average Computation time of each component of our framework

Components	Average Computation Time (ms)
STIP	2/frame
Feature vector (HOG/HOOF)	12/interaction sample
BoW representation	3 /interaction sample
Recognition	1.8/interaction sample

CONCLUSION

Our approach provides an effective representation of human vehicle interaction that aids in improving the recognition performance. The overall inference lies in the effective role of considering the spatial position of features along with the feature descriptors. The positions of features help in discriminating the interactions that are comprised of same spatial occurrences but in different time order. As clearly shown in Table 1, the accuracy performance increases by approximately 50%. The miss rates are high in case of simple BoW or Reddy et al. (2012) approach for both types of interactions. In case of efficiency, literature shows our approach does quite well with less computational time and better accuracy of recognition. Increasing complexity of components, increase the recognition performance, but at the cost of computational burden. However, choosing new classification techniques such as SVM with our approach of BoW may increase the recognition performances. Our work can be extended to recognize other human vehicle interactions by choosing different coding technique for visual words and classifier for the recognition purpose to bring further improvement in the recognition performance. It can also be automated from the first step, by including detection of interactions approach. Moreover, this approach is highly applicable to test other kind of interactions within the BoW model.

ACKNOWLEDGMENT

The authors are highly thankful to the Institute of Space Technology, Islamabad, Pakistan for the research support.

REFERENCES

Bregonzio, M., Xiang, T. & Gong, S. 2012. Fusing appearance and distribution information of interest points for action recognition, Pattern Recognition, **45** (3):1220–1234.

Brendel, W. & Todorovic, S. 2011. Learning spatiotemporal graphs of human activities."In Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 778-785.

Dollár, P., Rabaud, V., Cottrell, G. & Belongie, S. 2005. Behavior recognition via sparse spatio-temporal features, in Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS '05), pp. 65–72.

Lazebnik, S., Schmid, C. & Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, Vol. 2, pp. 2169-2178.

Efros, A.A., Berg, A.C., Mori, G. & Malik, J. 2003. Recognizing action at a distance. In: International Conference on Computer Vision.

Intille, S.S. & Bobick, A.F. 2001. Recognizing planned, multiperson action. Computer Vision and Image Understanding, **81**(3), pp. 414-445.

Lazebnik, S., Schmid, C. & Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on (Vol. 2, pp. 2169-2178).

Niebels, J.C., Chen, C.W. & Fei-Fei, L. 2010. Modeling temporal structure of decomposable motion segments for activity classification. In Computer Vision–ECCV 2010, pp. 392-405.Springer Berlin Heidelberg.

Niebles, J.C., Wang, H. & Fei-Fei, L. 2008. Unsupervised learning of human action categories using spatial-temporal words. In Proc. of BMVC.

Reddy, K.K., Cuntoor, N., Perera, A. & Hoogs, A. 2012. Human action recognition in large-scale datasets using histogram of spatiotemporal gradients." In Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on, pp. 106-111

Raptis, M. & Sigal, L. 2013. Poselet key-framing: A model for human activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2650-2657.

Schüldt, C., Laptev, I. & Caputo, B. 2004. Recognizing human actions: local SVM approach. In: International Conference on Pattern Recognition.

Tang, K., Fei-Fei, L. & Koller, D. 2012. Learning latent temporal structure for complex event detection. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 1250-1257

Tian, Y., Sukthankar, R. & Shah, M. 2013. Spatiotemporal deformable part models for action detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2642-2649.

Babri, U. M., Tariq, M. & Khurshid, K. 2016. Feature Based Correspondence: A comparative study on Image Matching Algorithms, International Journal of Advanced Computer Science and Applications, 7(3), pp. 206-210.

Vahdat, A., Gao, B., Ranjbar, M. & Mori, G. 2011. A discriminative key pose sequence model for recognizing human interactions. In Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pp. 1729-1736

Wang, Y. Mori, G. 2013. Hidden part models for human action recognition: Probabilistic vs. max-margin. IEEE Transactions on Pattern Analysis and Machine Intelligence, **33**(7): 1310–1323.

Wu, X., Xu, D., Duan, L. Luo, J. 2011. Action recognition using context and appearance distribution features," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11), pp. 489–496.

Yamato, J., Ohya, J. & Ishii, K. 1992. Recognizing human action in time-sequential images using hidden markov model. In Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92. IEEE Computer Society Conference on, pp. 379-385.

Submitted: 20/April/2016

Revised : 11/Nov/2016

Accepted : 29/Nov/2016