International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Identifying Human-Object Interactions in Motionless Images by Modeling the Mutual Context of Objects and Human Poses

A. N. Bhagat[1], N. B. Pokale[2]

[1]Department of Computer Engineering, TSSM,s Bhivrabai Sawant College Of Engineering and Research, Narhe, Pune, Maharashtra, India.

[2]Associate Professor, Department of Computer Engineering, TSSM,s Bhivrabai Sawant College Of Engineering and Research, Narhe, Pune, Maharashtra, India.

Abstract: ***The two most challenging problem in computer vision is to detect an object in clutter view and estimate articulated human body from 2D images. While the difficulty occurs during the activities which involves human objects interaction, since the related items tend to be very small or moderately visible and the human body parts are often self occluded. We examine that the human pose and the object are share joint background to each other, i.e. if we recognize one, make easy to recognize the other. In this paper, we proposed the mutual context model for jointly modeling object and poses of human body in the human object interaction activity. In this method, recognition of object provide the strong prior for better human pose estimation, this human pose estimation enhance the efficiency of detecting the objects that interrelate with the human.***

Keywords: Mutual Context, Human pose detection and estimation, Human-Object Interactions, Image Indexing, K-means clustering

1. Introduction

Recently more attention is given to assisting the visual recognition by using the context. In the human visual system for recognition context plays an important role, this can be concluded from the psychology experiments.[3], [4] The contexts has been used in the problems like detecting and recognizing an object [1], [2], [5], recognizing the scene[7], action classification[8], and segmentation of an image[6]. The concept of using the context is good one, an interesting inspection describes that the most of the context information devoted relatively little for boosting the performance in recognition task. In recent Pascal VOC challenge [10], distinction among the context based and sliding window based methods for detecting an object is only within a small margin 3 to 4 percent [9], [11].

Because of the absence of strong context from this reason the account gives the relatively small margin. While it is nice to detect vehicle in the context of roads, powerful vehicle detector can nevertheless detect vehicles with high efficiency no matter whether the cars are on the road or not. Indeed for the human visual system detecting visual system detecting visual abnormality out of context is crucial for survival and social activities [12].

Here in this paper, we proposed the mutual context among the objects and humans in human object interaction movement since each can expedite the recognition of other. In this mutual context model especially two contextual information's are considered. One of this is co occurrence context model, in this model the co occurrence statistics among the items and specific types of human poses within each activity. The atomic pose means the types of poses of the human [14], we can also said it is a dictionary of poses of human in which the poses of human can be represent by the same atomic pose compare to the similar arrangement of the body parts. Another one is which we considered is spatial context, it geographical relation among the objects and distinct human body parts. Here we show that proposed algorithm considerably improve the performance of both object detection and human pose assessment on a six class sports dataset. Instead of object detection and post estimation together, the proposed method achieved higher accuracy in classifying human object interface activities.

In [15] and [16], it has been demonstrated that humans have the improved observation of human expression when the objects are presented and vice versa. In [17], authors approved the relationship among the spatial and functional object, and poses of human is the human object interface activities. In the proposed work, we absolutely model these relationships so that the recognition of objects and human poses can equally benefit from each other. From this approaches the proposed method considerably distinct from the existing method of the activity recognition, in which the process of activity recognition is considered as a pure image or video classification problem [18], [19], [20], [21], without detailed analysis of the objects and human poses that are involved in these activities.

The rest of the paper can be organized in the following manner: in section II we discussed about the related work done for finding the objects and poses of humans in the human object interface activity. In section III we discussed about the proposed work like the algorithms, and system architecture. In section IV we discussed about the experiment result and in section V we discussed the conclusion and future scope.

2. Related Work

Since from many years the study has been done the human pose estimation and object detection in the computer vision. Some of the methods of pose estimation method model the parts of the human body in the tree structure and uses the pictorial structure method [22], [23] for the adequate inference. The concept of pictorial structures and its

derivation [24], [25], [26], [27], it works very nice on the picture with the clean environment. This method improves the pose estimation performance in the composite scene like the TV show. In order to capture more composite human body articulations, there are some non tree model also proposed [28], [29]. Lately a system has been built named as real time human pose estimation system in which after applying random forest [30] method to the depth images [31]. Still, the most challenging difficulty faced when the human body parts are exceedingly uttered and occluded is the human pose estimation on the 2D images.

One of the most flourishing strategies for detecting an object is the sliding window. Some of the technique has been proposed for avoiding thoroughly searching the image [32], [33]. While the most of the detector are based on the sliding windows, more current work has tried to mix context to get improved performance. Moreover, the performance is enhanced by the comparatively small margin in some of the methods.

For developing an object detection or human pose estimation scheme which normally apply to all situation this is the out of scope of this paper. Rather than this we target on the role of context in these problems. The proposed work is motivated by the number of previous work which have used context in the vision task. In most of this work one type of scene information serve as contextual facilitation to a main recognition problem. Suppose one example, ground planes and horizons can help to clarify pedestrian recognition. Specially for helping the recognition of other items the object context has been widely used, the recital of the human pose estimation can be improved from the object context [34], [35]. Task like the motion capture [36] and inferring surface contact [37] have been treated by the poses of human.

In the proposed work, we analyze the mutual context among the object detection and human pose estimation rather than simply pleasuring one task as the main recognition problem and other one as the contextual facilitation. Our method allows the two tasks serve as a context for each other, so that the recognition performance of both tasks is improved.

In the recent years the attention is given to recognizing the human activities in the motionless images. While many works treat the task as an image categorization problem, more and more people have tried to obtain a detailed understanding of the humans and the objects as well as their interactions. In [38], after identifying the faces of human and gestures of human the method of action recognition is carried out. In [39], the performance of poses of human and items in humans are evaluated. In [40], the author used the term ‘visual phrases’ i.e. the communication among the humans and objects are learning in the discriminative way. The proposed work takes additional steps by unambiguously modeling the poses of human, their objects and the mutual context in the human object interface activities. Additionally, for obtaining the accurate the result we test the recital of the poses of humans, recognition of objects, and the categorization activity in the different domain together with the poses of people in the sports.

The prior version of this paper was depicted in [41] and protracted in [14]. Described model in this paper is based on [14] which can be different from [41] in some aspects: 1. In [41] the each activity should be model separately for the interaction of human objects and in our model which is based on [14] we learn an overall relationship among the distinct activities, objects and poses of humans. 2. In the proposed paper we can deal with the conditions where human can be intermingle with number of objects but in [41] there is limitation of one human and one object interaction. Additionally in this paper we test the dataset of sports for examine the performance of the method, in this sport dataset the interaction between the human and the objects are occurred in the large scale.

3. Proposed Work

In the proposed work we design the model for detecting an objects and poses of human in the human object interface activities. The following system architecture of shows the flow of the proposed work.

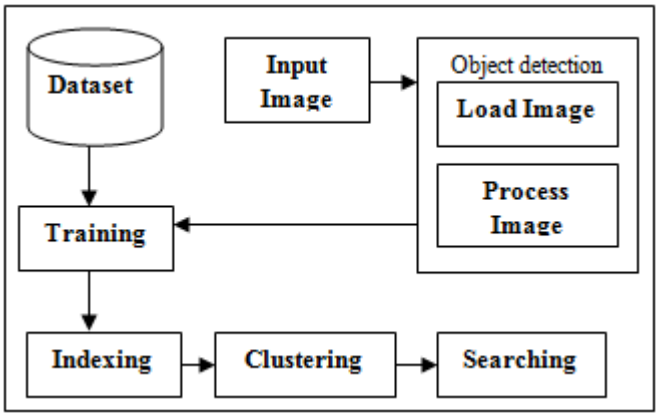


Figure 1: System Architecture

Number of input images are loaded and the processing is done to the on all the input image. The process of training, indexing and clustering is done on the input image and saved in the database. Now the new input image loaded and the training, indexing clustering and searching is done. Now we will see in detail each modules of the proposed work.

1. Indexing

In the indexing, training of an image is done. Indexing provides some values to images and according to that values the images gets clustered.

2. Clustering

In the clustering the trained data are placed in the related group without knowing having the advanced knowledge about the group definition. Simple in clustering the partition of a set of data into a group is done. There are number of clustering algorithms are exists like k-means clustering, expectation maximization clustering.

3. Searching

The process of searching is done on the trained image. In the database the training dataset of more than 1000 of images are saved, when the user give the input image for object detection the training of an input image is done after the

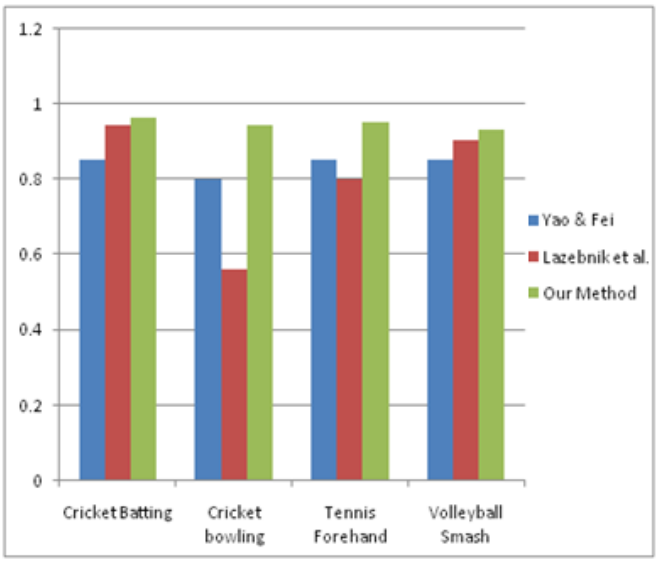
searching is done. We search the exact match of the input image with the image saved in the database.

4. Results

In this section the comparison of the proposed work with the existing method is done. Our method is compared with the two existing method of the author Yao & Fei and Lazebnik et al. method. Now we will see the Comparison table and Comparison for the proposed model.

Table 1: Comparison Table

Method	Yao & Fei	Lazebnik et al.	Our Method
Cricket Batting	0.85	0.94	0.96
Cricket bowling	0.80	0.56	0.94
Tennis Forehand	0.85	0.80	0.95
Volleyball Smash	0.85	0.90	0.93



Graph 1: Comparison Graph

The Table 1shows the comparison table between the two existing methods and the proposed methods. For the comparison four dataset are used, the dataset user are cricket batting, cricket bowling, Tennis Forehead and Volleyball Smash. We use the images of this database in the different pose for detecting an object. As we see in the comparison table the accuracy of searching in the Yao and Fei method is from 0.80 to 0.85 and in Lazebink et al method the accuracy result is ranging from 0.80 to 0.90 and in our proposed method 0.90 to 0.99. Hence from the comparison it is conclude that the accuracy of the proposed method is more than the existing method.

In the Graph 1: the graph is plot from the values of the comparison table. In the x axis the different dataset use for the comparison is given, the dataset used are cricket batting, cricket bowling, and Tennis forehead. And volleyball Smash. In the y axis the different values are taken as an input from the 0 to 1.2.

5. Conclusion

In this paper, we indulge objects and poses of human as the context of each other in distinct human object interface activity classes. In the proposed work we promote the conditional random field model which learns the co–occurrence context and the geographical context among the objects and the poses of humans. The implementation results shows that the designed model appreciably exceed other state of the art methods in both the problems. Here we study a new problem in which the context among the objects and poses of humans can appreciably improve the recognition performance. However, the context plays the vital roles in the many situations like detecting the remote near the TV. It would be valuable to propose the computer vision technique which uses the context in such a situation.

In proposed work we take the dataset of sport images and extract the images and perform the further operation. The accuracy of searching the image is more than 90%. In future we take the dataset of videos of the sports, and extract this videos into frames and perform the further operation of training, clustering and searching the best match. When we take video as a input dataset, the accuracy of searching the best match is more than 90% as comparing to the existing method whose of searching the best match is only 60 to 70%.

References

[1] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, “Objects in context,” in *International Conference on Computer Vision*, 2007.
[2] G. Heitz and D. Koller, “Learning spatial context: Using stuff to find things,” in *European Conference on Computer Vision*, 2008.
[3] I. Biederman, R. Mezzanotte, and J. Rabinowitz, “Scene perception: Detecting and judging objects undergoing relational violations,” *Cognitive Psychology*, vol. 14, pp. 143–177, 1982.
[4] A. Oliva and A. Torralba, “The role of context in object recognition,” *Trends in Cognitive Sciences*, vol. 11, no. 12, pp. 520–527, 2007.
[5] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert, “An empirical study of context in object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[6] J. Shotton, J. Winn, C. Rother, and A. Criminisi, “TextonBoost: joint appearance, shape and context modeling for multi-class object recognition and segmentation,” in *European Conference on Computer Vision*, 2006.
[7] K. Murphy, A. Torralba, and W. Freeman, “Using the forest to see the trees: a graphical model relating features, objects, and scenes, “in *Advances in Neural Information Processing Systems*, 2003.
[8] M. Marszalek, I. Laptev, and C. Schmid, “Actions in context,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[9] C. Desai, D. Ramanan, and C. Fowlkes, Discriminative models for multi-class object layout,” in *International Conference on Computer Vision*, 2009.
[10] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman, “The PASCAL VOC2008 Results.”
[11] H. Harzallah, F. Jurie, and C. Schmid, “Combining efficient object localization and image classification,” in *International Conference on Computer Vision*, 2009.

[12] J. Henderson, "Human gaze control during real-world scene perception," *Trends in Cognitive Sciences*, vol. 7, no. 11, pp. 498–504, 2003.
[13] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *ECCV Workshop on Statistical Learning in Computer Vision*, 2004.
[14] B. Yao, A. Khosla, and L. Fei-Fei, "Classifying actions and measuring action similarity by modeling the mutual context of objects and human poses," in *International Conference on Machine Learning*, 2011.
[15] D. Bub and M. Masson, "Gestural knowledge evoked by objects as part of conceptual representations," *Aphasiology*, vol. 20, pp. 1112–1124, 2006.
[16] H. Helbig, M. Graf, and M. Kiefer, "The role of action representation in visual object," *Experimental Brain Research*, vol. 174, pp. 221–228, 2006.
[17] P. Bach, G. Knoblich, T. Gunter, A. Friederici, and W. Prinz, "Action comprehension: Deriving spatial and functional relations," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 31, no. 3, pp. 465–479, 2005.
[18] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *International Conference on Computer Vision*, 2003.
[19] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2/3, pp. 107–123, 2005.
[20] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[21] J. Niebles, C. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *European Conference on Computer Vision*, 2010.
[22] P. Felzenszwalb and D. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.
[23] D. Ramanan, "Learning to parse images of articulated objects," in *Advances in Neural Information Processing Systems*, 2006.
[24] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[25] M. Eichner and V. Ferrari, "Better appearance models for pictorial structures," in *British Machine Vision Conference*, 2009.
[26] B. Sapp, A. Toshev, and B. Taskar, "Cascade models for articulated pose estimation," in *European Conference on Computer Vision*, 2010.
[27] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixture-of-parts," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
[28] X. Ren, A. Berg, and J. Malik, "Recovering human body configurations use pairwise constraints between parts," in *International Conference on Computer Vision*, 2005.
[29] Y. Wang and G. Mori, "Multiple tree models for occlusion and spatial constraints in human pose estimation," in *European Conference on Computer Vision*, 2008.
[30] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1,pp. 5–32, 2001.
[31] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
[32] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2001.
[33] C. Lampert, M. Blaschko, and T. Hofmann, "Beyond sliding windows: object localization by efficient sub window search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
[34] A. Gupta, T. Chen, F. Chen, D. Kimber, and L. Davis, "Context and observation driven latent variable model for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
[35] H. Kjellstrom, D. Kragic, and M. Black, "Tracking people interacts with objects," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
[36] B. Rosenhahn, C. Schmaltz, T. Brox, J. Weickert, and H.-P. Seidel, "Staying well grounded in marker less motion capture," in *Symposium of the German Association for Pattern Recognition*, 2008.
[37] M. Brubaker, L. Sigal, and D. Fleet, "Estimating contact dynamics," in *International Conference on Computer Vision*, 2009.
[38] L. Jie, B. Caputo, and V. Ferrari, "Who's doing what: Joint modeling of names and verbs for simultaneous face and pose annotation," in *Advances in Neural Information Processing Systems*, 2009.
[39] V. Singh, F. Khan, and R. Nevatia, "Multiple pose context trees for estimating human pose in object context," in *CVPR Workshop on Structural Models in Computer Vision*, 2010.
[40] M. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
[41] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.