MULTIPLE OBJECTS TRACKING BY AN EVIDENCE BASED APPROACH

Dr.M.Rajaiah, Dean Academics & HOD, Dept of CSE, Audisankara College of Engineering and Technology, Gudur.

Mr.D.V.Varaprasad, Associate Professor, Dept of CSE, Audisankara College of Engineering and Technology, Gudur.

Mr.SK.Shahul Hameed, UG Scholar, Dept of CSE, Audisankara College of Engineering and Technology, Gudur.

Mr.Y.Venkatesh, UG Scholar, Dept of CSE, Audisankara College of Engineering and Technology, Gudur.

Mr.U.Jagadeesh, UG Scholar, Dept of CSE, Audisankara College of Engineering and Technology, Gudur.

Mr.V.Rohith, UG Scholar, Dept of CSE, Audisankara College of Engineering and Technology, Gudur.

ABSTRACT: The issue of choosing appearance features for multiple object tracking (MOT) in urban scenes is addressed in this paper. Numerous features have been employed for MOT over the years. Whether some of them are superior to others is unclear, though. Colour histograms, histograms of oriented gradients, deep features from convolutional neural networks, and re-identification (ReID) features are examples of frequently used features. In this study, we evaluate the performance of these features in urban scene tracking scenarios to distinguish objects from a bounding box. Several affinity measures, including the Rank-1 counts, the cosine similarity, the L1, L2, and Bhattacharyya distances, are also evaluated for their effect on the discriminative power of the features. . Results from several datasets demonstrate that, regardless of the detector quality, features from ReID networks are the best at differentiating between instances. Colour histograms may be chosen in the absence of a ReID model if the detector has a good recall and few occlusions; otherwise, deep features are more resistant to detectors with lower recall. An picture's colour histogram shows how the colours are distributed throughout the image. The number of pixels in each type of colour and the various colour variations are displayed. Re-identification is a general term for **any process that re-establishes the relationship between data and the subject to which the data refer**.

1.INTRODUCTION:

Cities must figure out how to transfer people for daily activities in a safe and effective manner, among other issues.

Therefore, information regarding the movements of all road users is required. Different types of sensors, including video cameras with computer vision algorithms, can automatically gather this data. The primary objective is to identify and monitor every road user, a process known as multiple object tracking (MOT). This is only one of numerous examples of how MOT is used.

Modern MOT techniques sometimes use a technique known as "tracking-by-detection" [1]: they first identify items of interest, including moving automobiles or pedestrians, and then they connect the detections between frames to produce trajectories. Several variables, including appearance, geographical data, and motion, are used in the second phase [2]. MOT has been extensively explored [3]–[6], however there are still numerous problems that haven't been resolved, which lowers the effectiveness of the solutions. One of them is explaining how things look. Every tracked object should be distinct from the others while also taking into account the possibility that an object's appearance may change over time due to a change in viewpoint and fluctuations in illumination. Therefore, choosing the most discriminative features and figuring out how to compare them properly become two crucial steps in the visual appearance modelling process.

Should handcrafted features be used, or should one learn how an object looks? It can be challenging to determine whether a method is superior because of the feature that was chosen, the data association method, or the method to predict where the object should be in the future. This is because in MOT, several aspects, such as appearance, spatial, and motion information, are typically investigated at the same time.

Recently, Kornblith et al. [7] investigated the classification performance of models with good ImageNet [8], [9] performance. The comforting response is that the models perform better on ImageNet when compared to other datasets.. Can the same conclusion be made for a task that comes after, such as tracking? In fact, MOT faces unusual difficulties such the necessity for classification, deformations, altered lighting, occlusions, blur, etc. The ImageNet dataset lacks some of these difficulties. Additionally, the classification needed for tracking is more precise in order to identify every instance of an object. For example, models trained on

ImageNet succeed when they correctly identify people as people, but for the MOT task, these people must

must be distinguished from one another. This study examines how well-liked visual qualities represent things in MOT under diverse urban scene circumstances. These are some of the most common MOT scenarios and the subjects of the most common MOT datasets.

Therefore, the primary things of interest to describe are diverse cars and pedestrians. We solely concentrate on the visual appearance description and comparison for picture regions surrounded by bounding boxes in order to prevent interference from other MOT components (BB). In this paper, no spatial or motion data are employed. The findings imply that the optimal visual features for MOT tasks are re-identification (ReID) features.

When these features are not accessible, deep features may be employed instead, which performs better when objects are farther apart in time than the colour histogram. When the detector provides inaccurate BBs, the HOG properties deteriorate severely.

2.PROPOSED SYSTEM:

We attempted to link two bounding boxes referring to the same object throughout a video in order to assess the performance of 35 descriptor-affinity pairs (each pair composed of a feature and an affinity measure with the exception of pairs between a non histogram-based feature and the Bhattacharyya distance). In order to do that, we used a feature descriptor to characterise a BB-enclosed item that was extracted from a frame. We next choose another frame in which this object is present, describe all of the objects there with the same feature descriptor, and compute an affinity measure to choose the object in the second frame that is the most similar to the one in the first (Figure 1). We then use the ground truth to determine if the match is accurate. This gives us an more accurate result than the previous models.

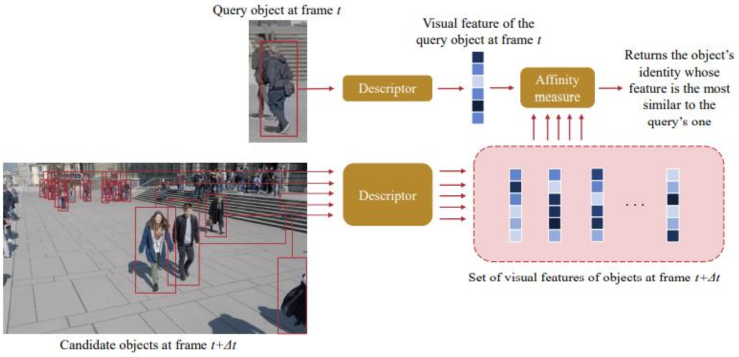
Query object at frame t
Visual feature of the query object at frame t
Descriptor
Affinity measure
Returns the object's identity whose feature is the most similar to the query's one
Descriptor
Set of visual features of objects at frame $t+\Delta t$
Candidate objects at frame $t+\Delta t$

Fig. 1. High-level explanation of our experimental methodology. From bounding boxes, a feature descriptor is calculated for both a query object and candidate matching objects in another frame. Then, the affinity measure is calculated for all query and candidate pairs, and the best match is returned and evaluated based on the ground truth.

3.LITERARTURE SURVEY:

For tracking, a wide range of appearance features have been applied. In this section, a few of them are briefly reviewed. The colour histogram is one of MOT's most well-liked features. Colour histograms were one of the tools utilised in the research of [10]–[12]. Colour histograms are integrated with other aspects including optical flow and a sparse representation in the work of Riahi et al. [10]. While a sparse representation reconstructs an image region using templates from the image regions of a model object and trivial templates, which contain just one non-zero value, optical flow determines the motion vector of pixels between two frames. If an object's visual appearance differs from the model object, many simple templates will be needed for the reconstruction. In the works of Zhu et al. [11] and Sun et al. [12], oriented gradient histograms are mixed with colour histograms (HOG). HOG concentrates on an object's texture, whereas the colour histogram concentrates on an object's overall colour appearance (spatial arrangement of the colours). HOG features frequently also capture the general shape of an item because they are calculated using gradient magnitudes as weights. A HOG feature can therefore be viewed as both a texture and a shape description.

For instance, the MOT techniques described in [11]–[13] rely on HOG. HOG is only utilised in conjunction with a Kalman filter in Heimbach et al[13] .'s study to anticipate object position. Deep features are frequently used in studies as general descriptors for MOT [14]–[18]. In the works of Tang et al. [16] and Sadeghian et al. [15], the object appearance is characterised using features from VGG-16, whereas Wang et al. [14] use a two-layer proprietary Convolutional Neural Network (CNN). Recently, class names (such as "vehicle," "pedestrian," and "bike") have also been employed as a rough description of an object's appearance [3]. Regarding [17], the authors used VGG-19, from which they collected several outputs from various layers. ReID traits are also found in recent publications [19]–[22]. These characteristics are calculated using a model that may be trained to determine if two detections from different angles are examples of the same object.

Unexpectedly, we were able to locate just one paper that compared attributes for MOT [23]. The features that were compared are the separation between the centre of gravity, the size of the BBs, and the correlation between object pixels because they go back to 1996.

4.RESULTS AND ANALYSIS:

General feature performance

To rank the descriptor-affinity pairs according to 24 "step configurations" and the sampling step, we condensed all the findings from the four datasets into four figures. We only presented the top five descriptor-affinity pairs for each case and dataset, along with the top model for any feature categories that did not make the top five.

The colour and hatching codes used in pictures 3, 4, 5, and 6 are described in Tables II and III.

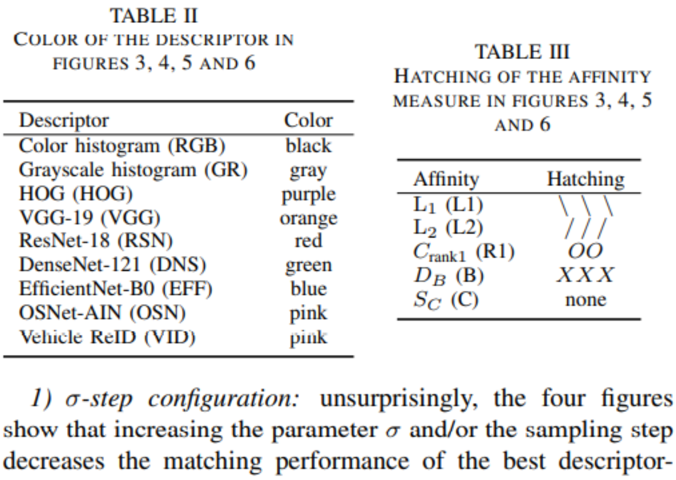

TABLE II
COLOR OF THE DESCRIPTOR IN FIGURES 3, 4, 5 AND 6

Descriptor	Color
Color histogram (RGB)	black
Grayscale histogram (GR)	gray
HOG (HOG)	purple
VGG-19 (VGG)	orange
ResNet-18 (RSN)	red
DenseNet-121 (DNS)	green
EfficientNet-B0 (EFF)	blue
OSNet-AIN (OSN)	pink
Vehicle ReID (VID)	pink

TABLE III
HATCHING OF THE AFFINITY MEASURE IN FIGURES 3, 4, 5 AND 6

Affinity	Hatching
L_1 (L1)	\ \ \
L_2 (L2)	/ / /
C_{rank1} (R1)	*OO*
D_B (B)	*XXX*
S_C (C)	none

1) σ-step configuration: unsurprisingly, the four figures show that increasing the parameter σ and/or the sampling step decreases the matching performance of the best descriptor-

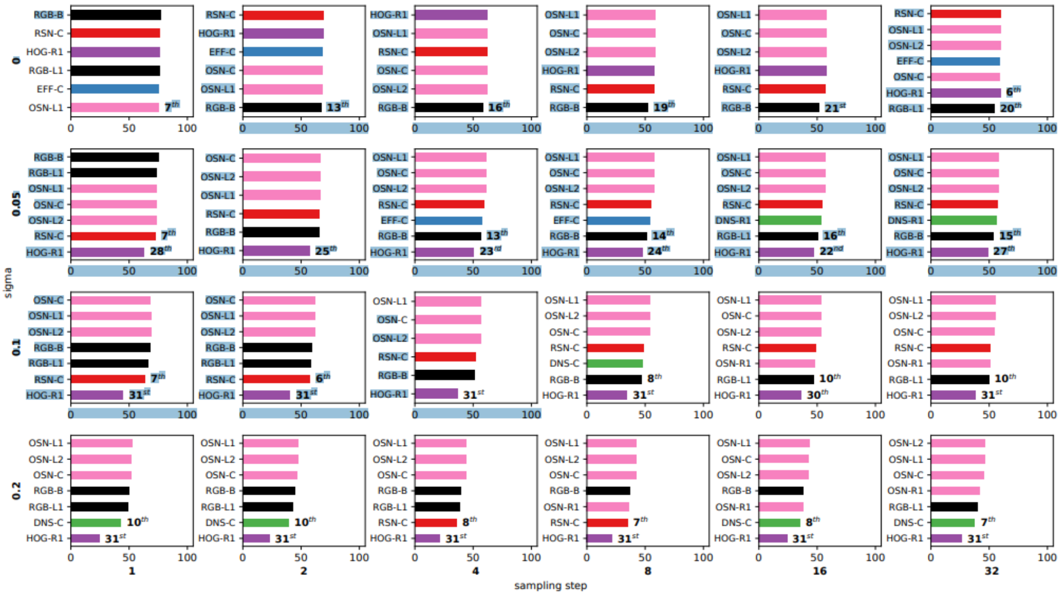

Fig. 3. Mean average precision on WildTrack of the five best descriptor-affinity for each configuration σ-step (when one category of descriptors is not in the top-5, the best result is added). See Tables II and III for the color and hatching codes used in the figure. Best viewed in color.

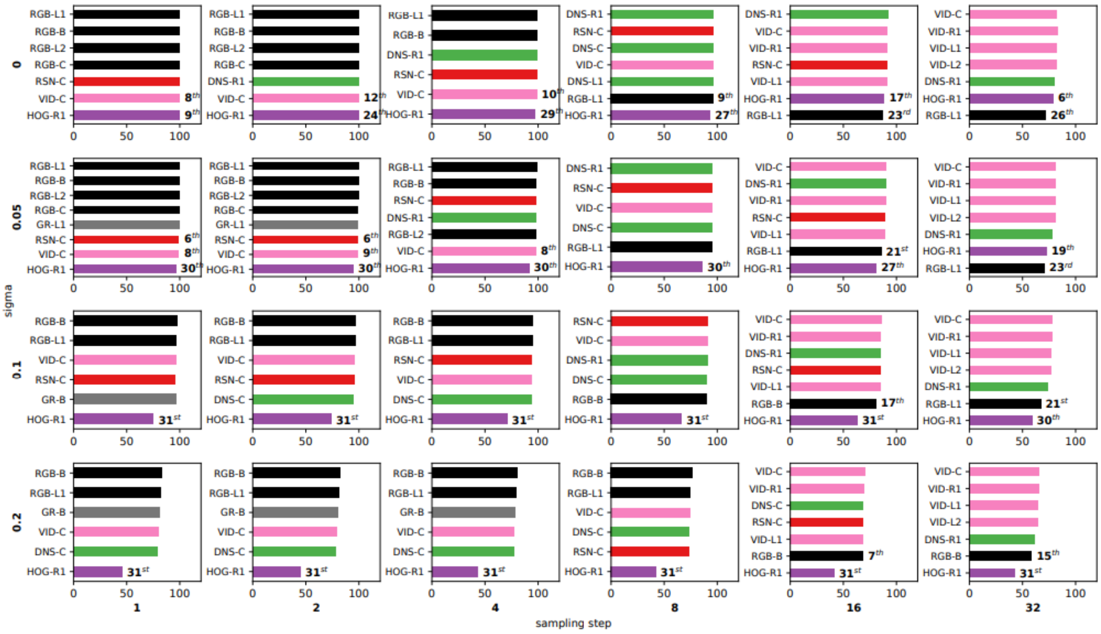

Fig. 5. Mean average precision on DETRAC of the five best descriptor-affinity for each configuration σ-step (when one category of descriptors is not in the top-5, the best result is added). See Tables II and III for the color and hatching codes used in the figure. Best viewed in color.

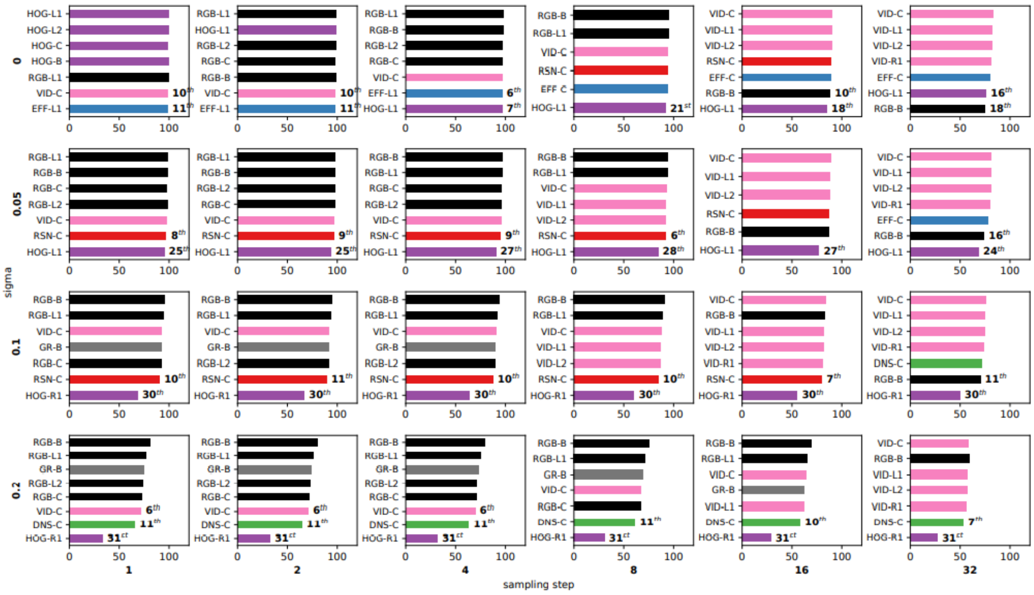

Fig. 6. Mean average precision on UAVDT of the five best descriptor-affinity for each configuration σ-step (when one category of descriptors is not in the top-5, the best result is added). See Tables II and III for the color and hatching codes used in the figure. Best viewed in color.

2) **Color and grayscale histograms:** Color histograms are competitive features for a low sampling step and a low, especially when coupled with the Bhattacharyya distance, in particular for datasets that focus on automobiles. This model nearly always ranks among the top five for these configurations, depending on the datasets. Because of Wild Tracks low framerate, pedestrians cannot be distinguished when their BBs are separated by more than two seconds, so the objects shouldn't be obscured for too long or the detector needs to have a good recall. This was explained by the fact that their hue appeared to change drastically between these two frames. However, these models consistently placed in the top five when the BBs do not completely encircle the object (= 0.2).

3) **Histograms of oriented gradients (HOG):** When the BB coordinates match the ground truth, HOG is an useful appearance feature descriptor. The performance of HOG drastically declines once they become erratic. No matter the dataset, the HOG descriptor is one of the best models if the BB match the ground truth (when it does not, the absolute difference from the best model is minimal). However, as rises, the candidate's performance drops on average to position 30 out of 35. The difference between this trait and others is substantial in the event of extremely erratic BBs. This is a result of the way HOG is built: feature vectors are calculated over cells that are 8 by 8 pixels.

4)**Models based on CNNs** are competitive when the sampling step is high and moderate. A CNN-based model frequently ranks among the top five models when is less than 0.1 and the sampling step is greater than 8. VGG-19 descriptor-based features are incomparable to the three previous CNN-based models because they never place in the top 5. Additionally, cosine similarity is not always a reliable indicator of affinity.

5)**ReID models:** Regardless of how well the detector performs, OSNet-AIN is typically the greatest visual feature for monitoring pedestrians. This model comes in top with either L1, L2 distances or cosine similarity in practically all combinations. This model is made to extract useful instance-specific properties from photos since it is trained to distinguish pedestrians. So, even if the image's BB is damaged, it can still discriminate between people. When is more than 0.1 or when the sample step is greater than 4, the model from [31] is among the top five for automobiles. The absolute value of the departure from the best model is minimal for BBs that are closer to the ground truth. The ideal affinity metric for this descriptor is cosine similarity.

B. Feature performance according to size of object:

The size of the objects may have an impact on the choice of the visual feature descriptor, in addition to the detector's capacity to accurately forecast the positions of the BBs and prevent missed detections.

In MOT, smaller objects are frequently the most challenging targets to monitor. However, it is unknown how BB size affects aesthetic traits.

On the UAVDT dataset with low occlusions, Figure 7 shows the average precision in relation to the query object size. The hardest configuration (0.2-32) where differences are more significant corresponds to the configuration -step chosen.

The only distance used for a fair comparison is the L2 distance.

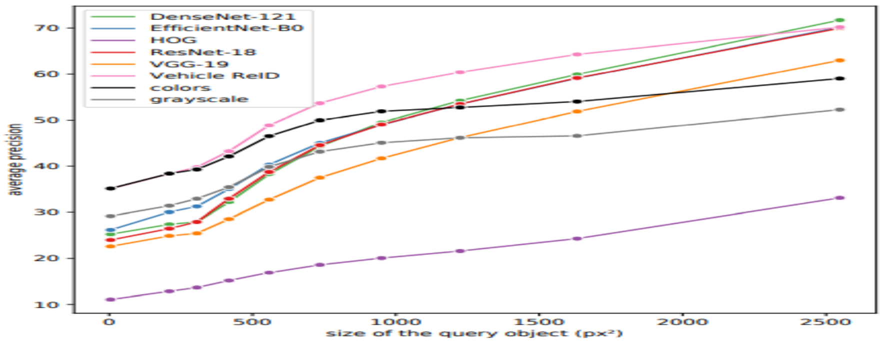

Fig. 7. Average precision according to the query object size, with $\sigma = 0.2$, sampling step at 32 and the L_2 distance on UAVDT, computed at each decile. Best viewed in color.

First off, it is always easier to find the right match for any characteristic when the query object is larger. For the smallest objects (about smaller than 250 pixels2 in area), where it is challenging to extract semantics, RGB-histograms are among the finest visual features. However, ReID features deliver the optimum performance for larger objects. The results from the evaluated CNN-based models are thus comparable to ReID, with the exception of VGG-19, which performs worse. This shows that improving performance on ImageNet does not always translate into improved features for MOT. With the exception of Wild track due to its modest scale in terms of the amount of data available, similar findings may be reached from other datasets (cf appendix).

CONCLUSION:

In the context of MOT in urban environments, we compared a number of feature descriptors in this research. Our research demonstrates that features behave differently depending on the bounding box quality. Regardless of the detector's effectiveness, ReID features in combination with cosine similarity are one of the best descriptors for pedestrians and automobiles. When the boxes are not too noisy, colour histograms using the Bhattacharyya distance are competitive in the absence of these models. But as soon as the bounding boxes get more noisy, these approaches cannot compete with deep features. Additionally, the size of the object affects the choice of visual features: in challenging situations, ReID features stand out particularly on medium-sized objects when compared to RGB-histograms and contemporary deep features.

REFERENCES:

[1] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, X. Zhao, and T.-K. Kim, “Multiple Object Tracking: A Literature Review,” arXiv:1409.7618, May 2017.

[2] S. Gladh, M. Danelljan, F. S. Khan, and M. Felsberg, “Deep Motion Features for Visual Tracking,” in International Conference on Pattern Recognition (ICPR), 2016.

[3] H.-L. Ooi, G.-A. Bilodeau, N. Saunier, and D.-A. Beaupre, “Multiple ´ Object Tracking in Urban Traffic Scenes with a Multiclass Object Detector,” in International Symposium on Visual Computing (ISVC), 2018.

[4] J.-P. Jodoin, G.-A. Bilodeau, and N. Saunier, “Tracking All Road Users at Multimodal Urban Traffic Intersections,” IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 11, pp. 3241–3251, Nov. 2016.

[5] Y. Yang and G.-A. Bilodeau, “Multiple Object Tracking with Kernelized Correlation Filters in Urban Mixed Traffic,” in Conference on Computer and Robot Vision (CRV), May 2017.

[6] L. Leal-Taixe, A. Milan, I. Reid, S. Roth, and K. Schindler, “MOTChal- ´ lenge 2015: Towards a Benchmark for Multi-Target Tracking,” arXiv: 1504.01942, Apr. 2015.

[7] S. Kornblith, J. Shlens, and Q. V. Le, “Do Better ImageNet Models Transfer Better?” in CVPR, 2019.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, L. Kai, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in CVPR, 2009.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[10] D. Riahi and G.-A. Bilodeau, “Multiple object tracking based on sparse generative appearance modeling,” in IEEE International Conference on Image Processing (ICIP), 2015.

[11] D. Zhu, H. Sun, and N. Yang, “A real-time and robust approach for short-term multiple objects tracking,” in International Conference on Computer Science and Information Processing (CSIP), 2012. [12] L. Sun, G. Liu, and Y. Liu, “Multiple pedestrians tracking algorithm by incorporating histogram of oriented gradient detections,” IET Image Processing, vol. 7, no. 7, pp. 653–659, Oct. 2013.

AUTHOR PROFILES

Dr.M.Rajaiah, Currently working as an Dean Academics & HOD in the department of CSE at Audisankara College of Engineering and Technology, Gudur, Tirupathi(DT).He has published more than 35 papers in Web of Science, Scopus, UGC Journals.

Mr.D.V.Varaprasad, currently working as an associate professor in the department of CSE at Audisankara College of Engineering & Technology, Gudur, Tirupati(DT).

Mr.SK.Shahul Hameed currently pursuing B.tech in the stream of Computer Science & Engineering at Audisankara College Of Engineering & Technology.

Mr.Y.Venkatesh currently pursuing B.tech in the stream of Computer Science & Engineering at Audisankara College Of Engineering & Technology.

Mr.U.Jagadeesh currently pursuing B.tech in the stream of Computer Science & Engineering at Audisankara College Of Engineering & Technology.

Mr.V.Rohith currently pursuing B.tech in the stream of Computer Science & Engineering at Audisankara College Of Engineering & Technology.