ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

IJIREEICE

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

nCORETech 2017

LBS College of Engineering, Kasaragod

Vol. 5, Special Issue 1, March 2017

Multiple Object Detection, Identification and Tracking Based On Local Sparse Representation

Ashna.M.D[1], Nishidha T[2]

M.Tech Student, Department of ECE, KMCT College of Engineering, Calicut, India [1]

HOD, Department of ECE, KMCT College of Engineering, Calicut, India [2]

Abstract: Sparse representation has been successfully applied to visual tracking for finding the suitable candidate by using the target templates. But most of the sparse representation based trackers only consider the holistic representation of the target object and do not make the full use of the sparse coefficients to discriminate between target and background. Hence may fail with more possibility, when there is similar object or occlusion in the scene. This paper studies the visual tracking problem in video sequences and presents a sparse tracker using coarse and fine dictionaries. This representation exploits both partial and structural information of the target based on averaging and alignment-pooling method. The similarity obtained by pooling across the local patches helps not only locate the target more accurately but also handle partial occlusion. Object detection, identification and tracking are the three main objectives of this paper. For object/person detection a superpixel based face detection algorithm is used here that is followed by moment-based matching and isosceles triangle matching. Object tracking can be done by using extended Kalman filtering method. In addition, this method employs a template update strategy which combines incremental subspace learning and local sparse representation. This strategy adapts the template to the appearance change of the target with less possibility of drifting and reduces the influence of the occluded target template as well. The proposed algorithm is superior in accuracy and it has better robustness against to partial and full occlusion.

Keywords: Object tracking, sparse coding, averaging, alignment-pooling, occlusion detection.

I. INTRODUCTION

Visual tracking is an important and necessary task in the field of computer vision and industrial applications like vehicle navigation, human-computer interface, video indexing, security, and automated video surveillance etc. The proliferation of high-end computers, the availability of high quality video cameras, and the increasing need for automated video analysis have generated a great deal of interest in visual tracking algorithms. The state of this art has advanced significantly in the past 30 years. Visual Tracking is actually the process of locating, identifying, and determining the dynamic configuration of one or many moving objects in each frame of one or several cameras. While numerous methods are proposed for visual tracking, it remains a difficult problem in a dynamic environment due to the presence of 3D pose variations, partial and full object occlusions, complex background clutters, illumination changes, and view point variations.

In this paper, an important and efficient tracking method is proposed that is mainly based on two factors. That is sparse coding and coarse and fine scale dictionaries. Here image patches in a fixed spatial layout from the target are extracted and then encoded with the help of dictionary technique. This dictionary is formed by using patches from multiple target templates. Sparse representation has been successfully applied to visual tracking for encoding the image patches different candidate regions. So each target candidate can be represents as the linear combination of a few dictionary elements. We observe that sparse coding of local image patches with a spatial layout contains both spatial and partial information of the target object. The similarity measure is obtained by proposed averaging and alignment-pooling method across the local patches within one candidate region. This helps locate the target more accurately and handle partial occlusion.

To make the representation more distinctive and robust, we compute the likelihood of candidate regions based on the combination of patches extracted with coarse-and-fine strategy by using the dictionary. For person/object detection in tracking a superpixel algorithm is used in this paper that is followed by moment-based matching and isosceles triangle matching. The dictionary for local sparse coding updated sequentially based on incremental subspace learning method. An additional algorithm is used for occlusion detection in order to update template more accurately without including occluding pixels. The update method ensures the proposed method tracks the object against appearance variations caused by pose change and illumination change. Extended kalman filtering is also used here for multiple object tracking. The framework of this work is shown in Fig.1.

IJIREEICE
ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526
International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
nCORETech 2017
LBS College of Engineering, Kasaragod
Vol. 5, Special Issue 1, March 2017

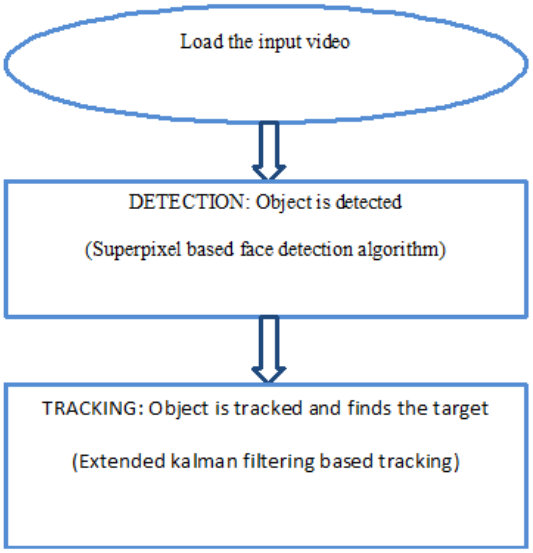

Fig.1. Basic framework of the system

II. RELATED WORK AND CONTEXT

Tracking algorithms can be grouped as either generative or discriminative. Discriminative methods define visual tracking as a classification problem which mainly aims to distinguish foreground objects from background regions by exploiting visual information from both the classes. But generative methods formulate the tracking problem as searching for the region that is most similar to the target model. A visual tracking algorithm can be decomposed into three components: an appearance model captures the visual characteristics of the target and evaluate the similarity between observed samples and the model. Where, the motion model locates the target between successive frames with certain motion hypothesis. An optimization strategy associates the appearance model with the motion model and finds the most likely location in the current frame.

Recently, tracking methods based on sparse representations have been proposed. Mei, Ling and Mei et al exploits a number of holistic templates of a target object as and then determine the most likely regions by solving the l_1 minimization problem. Most of these methods use holistic representation schemes and hence do not perform well when target objects are heavily occluded. Adam et al. propose a fragment-based tracking method which divides a target template into different patches where each one is tracked locally by measuring region similarity. Liu et al. proposed a local appearance model that is based on histograms of sparse coefficients and the mean-shift algorithm is used for object tracking. In many tracking algorithms, local sparse representation schemes are employed to model object appearance. Due to the representation of local patches, these methods perform well when target objects are heavily occluded. In addition, mean shift algorithms and voting maps are used to track target objects efficiently.

Our work is similar to some of the recent tracking algorithms, both use sparse coding to model local appearance of the object and also the same template updating strategy. However we use a superpixel based algorithm for face detection and multiple object tracking can be done by using extended kalman filters. Instead of using fixed templates or dictionary atoms learned from the first frame, we update the proposed appearance model adaptively. Furthermore, an occlusion detection scheme is proposed here to solve the problem where pixels belonging to the background regions are inadvertently included by straightforward model update schemes.

III. STRUCTURAL LOCAL SPARSE APPEARANCE MODEL

Most of the tracking methods use either multiple holistic templates or local appearance models to represent the target. In this paper, we develop coarse and fine structural local appearance models based on sparse representation of both multiple templates and local appearance models. Fig.2 shows the block diagram of feature formation in this work.

A. Object-Specific Dictionary

In this paper, for object classification and detection, a dictionary technique is used with two types of dictionaries. One is constructed by using coarse scale image patches and the other is constructed by using fine scale image patches. A dictionary learned from local image patches or SIFT features of a large dataset is expected to contain distinctive patterns of objects from numerous classes.

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

IJIREEICE

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

nCORETech 2017

LBS College of Engineering, Kasaragod

Vol. 5, Special Issue 1, March 2017

First, a set of n templates are collected to model the object appearance and thereby to construct the dictionary, i.e., $T = [T_1, T_2, \ldots\ldots, T_n]$ where, T_i is actually the image observation of the target object. For a given set of n templates, we extract a set of local image patches of fixed spatial layout inside the target region. These local patches are used as the dictionary atoms to encode the image patches inside the possible candidate regions, ie., $D = [d_1, d_2, \ldots\ldots d_{(n\times N)}]$, where d is the dimension of the image patch vector, and N is the number of local patches sampled within the target region. l_2 normalization is applied on the vectorized local image patches extracted from T to obtain each column in D**.** Each local patch represents one fixed part of the target object, and the local patches altogether represent the holistic structure of the target. For each target candidate region, we extract the corresponding local patches with $Y = [y_1, y_2, \ldots\ldots, y_N]$.

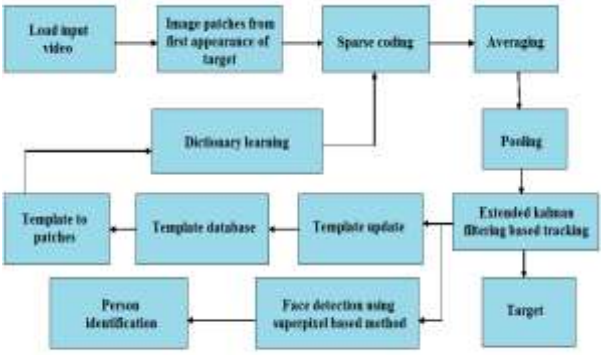

Fig.2. Block diagram of the proposed system

B. Sparse Coding

By using sparse coding, each local patch within the target region can be represented as the linear combination of a few basis elements in the dictionary by solving

$$\min_{b_i} \|y_i - Db_i\|_2^2 + \lambda\|b_i\|_1,\ b_i \geq 0 \qquad (1)$$

Where b_i is the corresponding sparse code of each of the extracted local patch y_i. $B = [b_1, b_2, \ldots\ldots, b_N]$ represents the sparse coefficients of patches within one candidate region. Another sparse coding method used in this paper is elastic net method. Elastic net method is based on ${}^{l_1}/_{l_2}$ regularization. Using ${}^{l_1}/_{l_2}$ method each patch is modelled by

$$\min_{b_i} \|y_i - Db_i\|_2^2 + \lambda_1\|b_i\|_1 + \frac{\lambda_2}{2}\|b_i\|_2^2,\ b_i \geq 0 \qquad (2)$$

l_1 regularization compute the sparse code of each of the image patch, but it also produces a large mean square error, while l_2 is able to produce small mean square error. So, these two regularization methods are combined here to take the advantage of both.

C. Averaging and Alignment Pooling

In the problem of visual tracking, the local pattern that is frequently occurring at the same position has an important role for robust representation and good alignment. These patterns are more distinctive than others because the template set contains the varying appearance of target object. Here, before pooling operation first we take the average of each of the sparse coded vector. For that purpose the encoding vector of a local patch is divided into several segments according to the template that each element of the vector corresponds to, i.e., $b_i^T = [b_i^{(1)T}, b_i^{(2)T}, \ldots\ldots b_i^{(n)T}]$. To enhance the stability of the sparse coding results, we use equally weighted averaging on different segments of encoding vectors. That is, these segmented coefficients are equally weighted to obtainv_i for i-th patch,

$$v_i = \frac{1}{C}\sum_{k=1}^{n} b_i^{(k)} \qquad (3)$$

All the vectors v_i of local patches in a candidate region form a square matrix V and are further processed by the proposed pooling method.

A single local patch of target object only captures the local appearance of the object. So, it is necessary to pool the averaged coefficient vectors for the modelling of the whole object. In this tracking method, alignment pooling method is used or directly concatenating these vectors, because it helps to eliminate irrelevant patches. After computing the average vector of each image patch, the local appearance of a patch with small variation is best described by the blocks

IJIREEICE
ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526
International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
nCORETech 2017
LBS College of Engineering, Kasaragod
Vol. 5, Special Issue 1, March 2017

at the same position of the templates. As per this, we use the diagonal elements of the square matrix V as the pooled feature.

$$f = \text{diag}(V) \quad (4)$$

Fig.3 shows the main steps of feature formation in this work.

After consistent local patterns are computed by the equally weighted averaging operation, this pooling method further aligns local patterns between target candidate and templates based on the locations of structural blocks. Fig.4 shows that the tracking result that aligns well with the target object has a higher score computed with the proposed pooling method than the poorly aligned one.

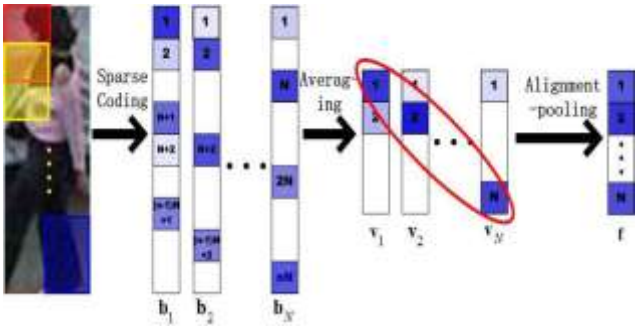

Fig.3. Local patches and feature vectors formed by averaging and alignment pooling

The alignment pooling operation enables the proposed appearance model to deal with partial occlusion. When occlusion occurs, encoding of the occluded local patches become dense due to significant appearance change. Thereby leading to smaller values in f. However, local patches which are not occluded have similar appearance to templates, and therefore can be described well by only using a few sparse coefficients and large values in f. The similarity between target candidate regions and the set of target templates computed in this way is still higher than the other candidates. That we can describe it as the good and bad candidate shown in Fig.4

.

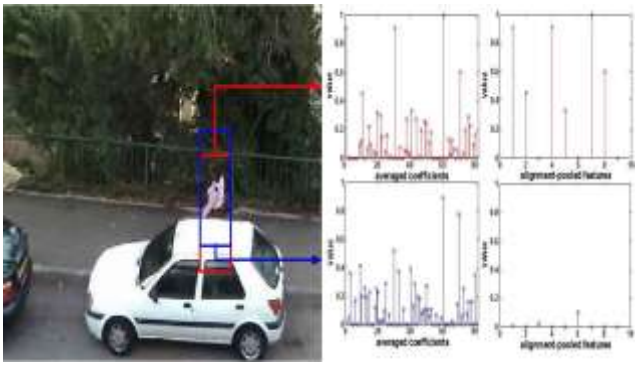

Fig.4. Comparison of the pooled features obtained by averaging and alignment-pooling for both good and bad candidates

D. Coarse and Fine Representations

To further improve tracking performance, we adopt a coarse and fine representation to model target appearance. Here, coarse and fine scale dictionaries are constructed by extracting image patches from multiple target templates. After computing the pooled features, system determines the coarse and fine scale similarities between candidate regions and target model. The appearance model constructed at fine scale distinguish foreground targets from background regions whereas, the appearance model at coarse scale facilitates for appearance change due to large deformation. The final likelihood of a target region is computed by the combination of similarities computed at both coarse and fine scales. We evaluate different weighting methods to compute the likelihood including similarities computed based on finer resolution patches are weighted more than those based on coarser ones; similarities computed based on coarser ones are weighted more; and they are equally weighted.

$$\eta_o = \sum_{m=1}^{M} \alpha^m \eta_o^m \quad (5)$$

Where η_o^m is the similarity computed on the m-th scale and α^m denotes the weight of the m-th scale.

IJIREEICE
ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526
International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
nCORETech 2017
LBS College of Engineering, Kasaragod
Vol. 5, Special Issue 1, March 2017

IV. OBJECT DETECTION

Face detection serves as a crucial step for a wide range of applications in computer vision. Given a still or a video image, face detection is to detect and localize an unknown number of faces. Face detection serves as a pre-processing step for most human-computer interaction techniques. In this paper, for object detection and identification, a superpixel based face detection algorithm is used that is followed by moment-based matching and isoceles matching shown in Fig.5. Five main modules are used to compose this algorithm. First, all the skin-like pixels in the input image are identified and are further grouped to form disjoint regions, each of which is a face candidate. We subsequently perform moment-based elliptic shape matching to discard those regions whose shapes are dissimilar to an ellipse. Afterwards, in each ellipse-shaped region, we search for two eyes and one mouth by constructing the Eyemap and the Mouthmap. Characteristics including the shape, colour and texture of human skin are put to use for locating the facial features. Finally, different rules are designed to disqualify spurious candidates.

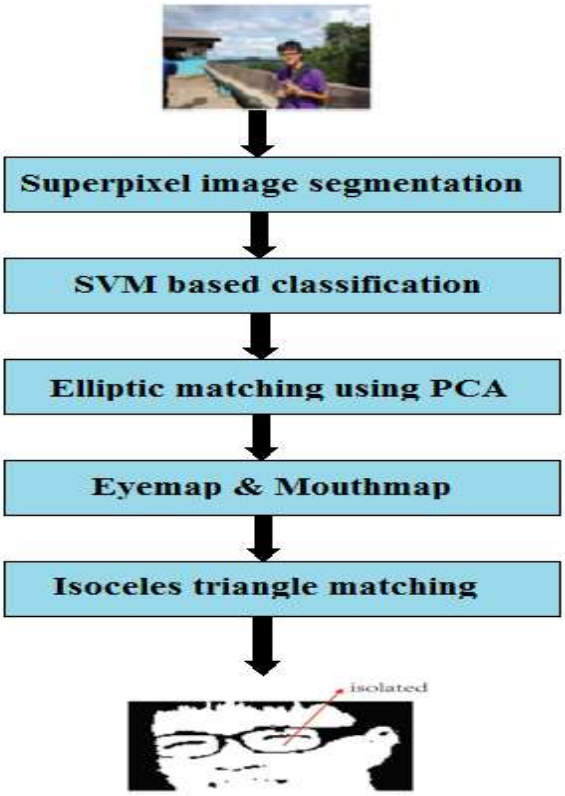

Fig.5. Superpixel based face detection algorithm

Here a novel approach is proposed that employs superpixel image segmentation followed by a detection step with the SVM classifier. Skin and non-skin superpixels can be viewed as two classes, and thereby implemented the SVM (Support Vector Machine) algorithm. The skin detection result is a binary map, the skinmap, where the white regions refer to skin zone and the black regions are non-skin areas. To remove false face candidates, we adopt Principal Component Analysis (PCA) to perform moment-based elliptic shape matching. Specifically, we use PCA to compute the major/minor axis of the region, which is equal to the length of the major/minor axis of the ellipse that has the same normalized second central moments as the region. Any region that is far from an ellipse is therefore excluded from face candidates.

Obtaining a pool of eye and mouth candidates, this method make use of the geometrical relations, skin texture as well as skin colour to phase out some false alarms. For every pair of two eyes and one mouth in each disjoint region, we examine whether or not it conforms to the rules. First set of rules are based on the assumption that the shape of the eyes-mouth triangle should be like an isosceles triangle where the angle α is equal to 45 degrees and θ is a right angle. However, one could plainly see that neither of the rules is rigid. The fact that the located mouth does not always lie on the centre of the true one exerts a significant influence on the angle α. Ideally, if the face is perfectly frontal, θ is nearly 90 degrees; however, as the head tilts, it θ is varied. Additionally, the second sets of rules are aimed at excluding non-human faces where we take into account human skin texture and skin colour.

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

IJIREEICE

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

nCORETech 2017

LBS College of Engineering, Kasaragod

Vol. 5, Special Issue 1, March 2017

V. EXTENDED KALMAN FILTERING BASED TRACKING

Initially, the Kalman Filter theory is only applicable to linear systems. Bucy and Sunahara proposed Extended Kalman filtering (EKF), to extend Kalman filtering theory to the nonlinear domain. EKF algorithm is developed on the basis of the standard Kalman filtering algorithm, its basic idea is: near the filter value, nonlinear systems using Taylor expansion formula will unfold, ignore the second order of high-order item above, the original system was changed into a linear system, and then, using the standard Kalman filtering algorithm model of linear filtering. In this paper we propose a nonlinear discrete mathematical estimation model to detect and tracking object based on EKF. In the extended Kalman filter, the state transition and observation models don't need to be linear functions of the state but may instead be differentiable functions.

Here dynamic model describes the temporal correlation of the target states between the consecutive frames. The state variables are assumed to be independent of each other. Observation model describes the likelihood of the observation at a state, which plays an important role in robust tracking. The observation model in this work is characterized by the similarity between the candidates and the set of target templates.

VI. EXPERIMENTAL RESULTS

The proposed algorithm is implemented in MATLAB and runs at one frame per second on an Intel Core i5 machine with 2GB memory. The regularization constant λ is set to 0.01, and λ_1 and λ_1 are respectively set to 0.05 and 0.01 in this experiment. This method normalize each target image patch to 32×32 pixels and extract local patches of size 16×16 and 8×8 within the target region with 8 pixels as step length. The threshold γ for occlusion detection is set to 0.5. We resize the target image patch to 32×32 pixels and extract overlapped 16×16 local patches within the target region with 8 pixels as the step length.

Here the tracking procedure is evaluated using a challenging dataset sequence of different types of atmosphere. In this woman sequence, we track a walking police woman in the street. The difficulty lies in that the woman is greatly occluded by the parked cars. Our tracker can follow the target accurately. In the woman sequence, the target object undergoes pose variation together with long-time partial occlusion. Based on the local patches and template update strategy, tracker focuses more on the upper body which remains almost the same though the lower body changes a lot or heavily occluded. It can successfully track the target throughout the entire sequence.

(a)

(b)

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

nCORETech 2017

LBS College of Engineering, Kasaragod

Vol. 5, Special Issue 1, March 2017

(c)

(d)

(e)

(f)

Fig.6. Tracking result of woman sequence

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

IJIREEICE

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

nCORETech 2017

LBS College of Engineering, Kasaragod

Vol. 5, Special Issue 1, March 2017

VII. CONCLUSION

We have proposed an accurate and efficient tracking method using coarse and fine scale dictionaries based on two stage sparse representations. Both the structural and partial information of the target objects are well used by the proposed algorithm. It exploits this by using averaging and alignment polling operations. In addition, sparse representation is used along with incremental subspace learning for template update. The template update scheme solves the problem of occlusion by removing the occluding pixels from the template set. The proposed algorithms are able to track targets more accurately and robustly under occlusion and clutters.

ACKNOWLEDGMENT

We thank the staffs and students in KMCT College of Engineering, Calicut for their comments and suggestions.

REFERENCES

[1] XuJia, Huchuan Lu and Ming-Husan Yang,"Visual Tracking via Coarse and Fine Structural Local Sparse Appearance Models", IEEE Transactions on Image processing, vol.25, no.10,October 2016.
[2] W.Zhong, H.Lu and M.H Yang, "Robust Object Tracking via Sparsity Based Collaborative model", IEEE Conference on Computer Vision and Pattern Recognition, pp. 1838-1845, June 2012.
[3] D. Comaniciu, V. Ramesh and P. Meer, "Kernal Based Object Tracking", IEEE Transactions on Pattern Analusis and Machine Intelligence, vol.25, no.5, pp. 564-573, May 2003.
[4] B. Liu, L. Yang, J. Huang, P. Meer, L. Gong and C. Kulikowski, "Robust and Fast Collaborative Tracking with Two Stage Sparse Optimization", Proceedings of the 11th European Conference on Computer Vision, pp. 624-637, 2010.
[5] D. A. Ross, J. Lim, R.S Lin and M.H Yang, "Incremental Learning for Robust Visual Tracking", International Journal for Computer Vision,vol.77, pp. 125-141, 2008.
[6] M. Yang, L. Zhang, J. Yang and D. Zhang, "Robust Sparse Coding for Face Recognition", IEEE Conference on Computer Vision and Pattern Recognition, pp. 625-632, June 2011.
[7] XuJia, Huchuan Lu and M. H. Yang, "Visual Tracking via Adaptive Structural Local Sparse Appearance Model", IEEE Conference on Computer Vision and Pattern Recognition, pp. 1822-1829, June 2012.
[8] S. Benhimane and E. Malis, "Homography Based 2D Visual Tracking and Servoing", The International Journal of Robotics Research, July 2007.
[9] W. Zhong, H. Lu and M. H. Yang, "Online Object Tracking: A Benchmark",IEEE Conference on Computer Vision and Pattern Recognition, pp. 1838-1845, June2012.
[10] S. Wang, H. Lu, F. Yang,"Superpixel Tracking", International Conference on Computer Vision, pp.777-783, 2011.
[11] J. Mairal, F. Bach, J. Ponce and G. Sapiro, "Online Learning for Matrix Factorization and Sparse Coding", Journal of Machine Learning Research, vol.11, pp. 19-60, 2010.
[12] K. He, J. Sun and X. Tang, "Guided Image Filtering", In Proceedings of 11th European Conference on Computer Vision, vol.1, pp. 1-14, 2010.
[13] J. Wright, Y. Yang, A. Ganesh,, S.S Sastry and Y. Ma, "Robust Face Recognition via Sparse Representation", IEEE Transactions on Pattern Analusis and Machine Intelligence, vol.31, no.2, pp. 210-227, February 2009.
[14] X. Mei and H. Ling, "Robust Visual Tracking Using l_1 Minimization", International Conference on Computer Vision, 2009.
[15] Shaung Wei and Kai, "Crowd Analyzer: A Collaborative Visual Analytic System", IEEE Conference on Visual Analytic Science and Technology(VAST), pp. 177-178, 2015.
[16] T. Zhang, B. Ghanem, S. Liu and N. Ahuja, "Robust Visual Tracking via Multi-Task Sparse Learning", IEEE Conference on Computer Vision and Pattern Recognition, pp. 2042-2049, June2012.
[17] Shengping Zhang, Hongxun Yao, Xin Sun and Xiusheng Lu, "Sparse Coding Based Visual Tracking: Review and Experimental Comparison", Elsevier, pp. 1772-1788, 2013. Available at: www.elsevier.com.

BIOGRAPHIES

Ashna.M.D is a M-Tech Digital Signal Processing student of KMCT College of Engineering, Calicut, India, degree in electronics and communication engineering from Government engineering college, Palakkad, India.

Nishidha T is now working as HOD in the Department of ECE, KMCT College of Engineering, Calicut, India. She completed her M-Tech in Digital Signal Processing from KMCT College of Engineering, Calicut, India.