International Journal of Computer Engineering in Research Trends
Multidisciplinary, Open Access, Peer-Reviewed and fully refereed
Research Paper Volume-6, Issue-7 ,2019 Regular Edition E-ISSN: 2349-7084

Object Identification Using Weakly Supervised Semantic Segmentation

P.S.Gunde[1], Dr. S.K.Shirgave[2]

[1] PG Scholar, Dept of Computer Science and Engineering, DKTE Society's Textile & Engineering Institute, Ichalkaranji (An Autonomous Institute), India.
[2] Associate Professor, Dept of Computer Science and Engineering, DKTE Society's Textile & Engineering Institute, Ichalkaranji (An Autonomous Institute), India.

e-mail: priyankagunde0192@gmail.com, skshirgave@gmail.com

Abstract:- Image segmentation is referred to as one of the most important processes of image processing. Image segmentation is the technique of dividing or partitioning an image into parts, called segments. It is mostly useful for applications like image compression or object recognition because for these types of applications, it is inefficient to process the whole image. So, image segmentation is used to segment the parts from the image for further processing. Semantic image segmentation is a vast area for computer vision and machine learning researchers. Many vision applications need accurate and efficient image segmentation and segment classification mechanisms for assessing the visual contents and perform real-time decision making. In this paper, we recommend conditional random field (CRF) based framework for weakly supervised semantic segmentation. First merging super pixels into large pieces and use these pieces for further use to identify objects. The pieces from all the training images are gathered and associated with appropriate semantic labels by CRF. In the case of testing, by using the potential energy of each piece merged from super pixels are compare with piece library. For results, we use commonly used the dataset for image segmentation is MSRC-21 and VOC 2012 with state-of-art.

Keywords: Conditional random field (CRF), Generate super pixels, Merge super pixels, Constructing piece library, Semantic label.

---

1. Introduction

The human visual system can know an image by recognizing objects and their backgrounds. Similarly, the machine can also understand the image by using semantic segmentation. Semantic segmentation solves the problem of assigning the label to every pixel in the image from pre-defined categories. The natural objects may generate many images with various appearances, pose, viewpoints, complicated background and limited access to training data, etc. One of the solutions for all the above problem is to explore segmentation methods with less supervision.

Image-level annotation is more convenient to obtain than pixel-level ground truth. So it is appropriate for weakly supervised semantic segmentation. The Conditional Random Field (CRF) is mainly used for semantic segmentation. The meaning can be interpreted as an image area that is visually similar or partially close to the same semantic label, and a different area gives various labels. To make spatial and visual features of the image and the correlation between semantic tags, weakly supervised semantic segmentation framework based on CRF can be proposed.

The existing weakly supervised semantic segmentation methods are mostly designed to train and rest on the same database to achieve excellent performance. The database shares some semantic categories, which results in lack of universality. Due to this problem, the database can be disassembled into object classes.

In the CRF framework, superpixels are first merged chunks because superpixels are the root component that contains the semantic information. The piece amount is set to be less than two times of semantic label amount. This would significantly reduce the computational cost and make the framework train faster. Then fragments from all the images are collected together to form piece pool. Finally, the repository of that class is constructed by associating semantic tags with fragments by CRF.

2. Related Work

In the survey of semantic segmentation, based on their supervision, the segmentation methods can be divided into three slices: fully supervised, unsupervised, and weakly supervised.

Fully supervised Methods: In the past few years, semantic image segmentation has often been considered a fully supervised task [1]-[3] . L. Ladicky, C. Russell, et al.[1] examines the use of co-occurrences statistic in the likelihood model. The work in H. Lu, G. Fang, X. Shao, X. Li [2] and Y.-L. Houet al. [4] proposed the human segmentation in images and videos. Yuan et al. [5] for traffic sign detection developed graph-based ranking and segmentation algorithm. The methods are all built based on having enough pixel-wise annotated samples for training. J. Carreira and C. Sminchisescu [3] proposed to generate hypothesis by solving a sequence of constrained parametric min-cut problems and rank plausible ones for the spatial extent of objects. Chen et al. [6] explained the multi-instance object occlusions in segmentation. J. Wang, and A. L. Yuille [7] proposed a novel algorithm for semantic segmentation for the animal. It also deals with the multiclass problem. This approach is typically not applicable for general application.

Unsupervised Methods: Unsupervised semantic segmentation methods that utilize image data without any annotation for training. Zhang, J. E. Fritts, S. A. Goldman [8], Wang, Q. Huang, M. Ovsjanikov [10]. Note: different from image segmentation and unsupervised methods on another field, these works care for the category of each pixel but in an unsupervised manner. H. Zhang, J. E. Fritts, and S. A. Goldman [1], Csurka and F. Perronnin [9] for performance development proposed label correlation in semantic segmentation is well established. Without sufficient utilization of image level annotation, the unsupervised methods tend to suffer from the under-constrained nature inherently and consequently impair their robustness towards variation.

Weakly supervised Methods: In this approach [11]-[17] often require image level annotation for training. A. Vezhnevets, V. Ferrari [11] developed weakly supervised semantic segmentation only classes they contain, not their position in the image. For weakly supervised segmentation, they projected the Multi Image Model (MIM) the same pixels in images shares label are likely to fit into the same classes.Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu [12] proposed weakly-supervised dual clustering method is proposed, which uses spectral clustering and discriminative clustering. It is used to work together with image segmentation and tag alignment. Zang et al. [13] proposed probabilistic graphlet cut to efficiently utilize the distribution of spatially structured superpixel sets from image-level and [15] learning the semantic association between the graphlets. K. Zhang, W. Zhang, Y. Zheng and X. Xue [14] a new method for evaluating classification models using sparse rebuilding and iterative merging algorithm was developed to obtain the best parameters. Pinheiro and R. Collobert [16] develop a model based on a convolution neural network (CNN), useful pixels are added for classification during training. L Zang et al. [17] iteratively updated a pool of region proposals and assign them labels by training convolution networks. It makes use of object bounding boxes as supervision.

3. Methodology

System Architecture

Figure 1 shows the architecture of the Semantic Image Segmentation, which consists of four modules.

- Superpixel generation
- Merging superpixels into pieces
- Constructing piece library
- Inference of testing

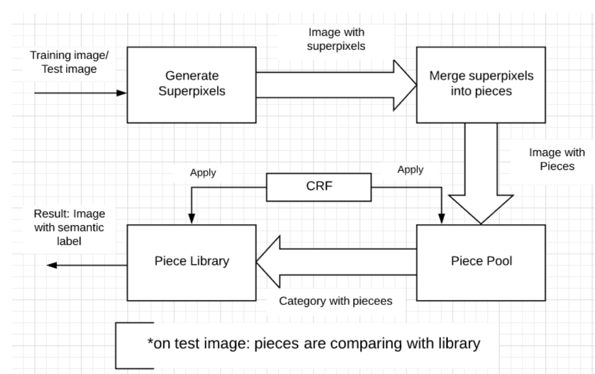

Figure 1. System architecture

Figure 1 shows the system architecture of the weakly supervised semantic segmentation. The system includes training images and test images. On the training image, first superpixels are generated on the same model. These superpixels are then merged into pieces, provided that each piece corresponds to only one semantic label. The second of all image pieces are collected into a pool. Finally, each piece is associated with an appropriate semantic name by integrating priors from its neighborhood and semantic label correlation. Thus, the piece library is constructed.

When testing an image, first creates the superpixels form test image. Combine it into large pieces. Second, provide semantic labels for each piece by using piece library.

Description of each module
Module 1: Super pixel Generation

Simple Linear Iterative Clustering (SLIC) generates superpixels by clustering pixels based on their color similarity and proximity in the image plane. A 5 dimensional **[labxy]** space is used for clustering. For smaller color distances, the CIELAB color space is considered to be permanently uniform. It is not recommended to use Euclidean distance in the 5D, so a new distance metric that takes into account the superpixels size.

Distance Measure: SLIC takes as input the desired number of approximately equally-sized superpixels K. So each superpixel will have approximately N/K pixels. Therefore, for superpixels of the same size, there will be a superpixel center at each grid spacing $S = \sqrt{N/K}$. Select K superpixel cluster centers $C_k = [l_k, a_k, b_k, x_k, y_k]$, where k = [1, K] at regular grid spacing S. Since the spatial extent of any cluster is approximately S^2, it can be assumed that pixels associated with this cluster are located within 2S x 2S region around the center of the superpixel in the XY plane. The Euclidean distance in the CIELAB color space makes a scene for small distances. If the spatial pixel distances exceed the perceptual color distance limit, they begin to exceed pixel color similarity. The distance measure D_s is defined as follows

$$\boldsymbol{d_{lab}} = \sqrt{((\boldsymbol{l_k} - \boldsymbol{l_i})^2 + (\boldsymbol{a_k} - \boldsymbol{a_i})^2 + (\boldsymbol{b_k} - \boldsymbol{b_i})^2)}$$

$$\boldsymbol{d_{xy}} = \sqrt{((\boldsymbol{x_k} - \boldsymbol{x_i})^2 + (\boldsymbol{y_k} - \boldsymbol{y_i})^2)}$$

$$\boldsymbol{D_s} = \boldsymbol{d_{lab}} + (\boldsymbol{m}/\boldsymbol{S}) * \boldsymbol{d_{xy}} \text{.........} \textbf{(Eq. 1) [18]}$$

where D_s is the sum of the lab distance and the xy plane distance normalized by the grid spacing S. Introducing the variable m in D_s allow to control the compactness of superpixel. The larger the value of m, the greater spatial proximity is emphasis and the more compact the cluster. This value can be in the range [1, 20]. The value of m is set to 20.

*1) **Algorithm***

It first samples the cluster centers of K regular intervals and then moves them to seed position corresponding to the lowest gradient position in the 3 × 3 neighborhood. This is done to avoid placing them at the edge and reducing the chances of selecting noisy pixels. Image gradient is calculated as

$$G(x, y) = \|I(x+1, y) - I(x-1, y)\| + \|I(x, y+1) - I(x, y-1)\|$$

Where I(x, y) is the lab vector corresponding to the pixel at position (x, y), and ||.|| is the L2 norm. This takes into account both color and intensity information. Each pixel in the image is associated with the nearest cluster center, with its search area overlapping the pixel. After all the pixels are associated with the nearest cluster center, a new center is calculated as the average labxy vector of all the pixels belonging to the cluster.

Module 2: Merging Superpixels Into Pieces

Suppose, $X = [x_1, \ldots\ldots, x_n] \in R^{m \times n}$ is an image with n superpixels and x_i is the m- dimensional feature descriptor of the i^{th}superpixels. The corresponding category label of this superpixels are denoted by $y = [y_1, \ldots\ldots y_n] \in R^{m \times n}$ where $y_i \in \{1, \ldots. L\}$ with L represents the total number of an object category. For training image weakly supervised problem, the superpixel y is no longer available. Instead of these image level labels, denoted by $l = [l_1, \ldots. l_L]$ where $l_i \in \{0,1\}$, and $l_i = 1$ indicates the presence of category i in the image while $l_i = 0$ indicates absence. Here, l must be noisy or only partially provided.

Now built a graph $G(V, E)$on the image and use CRF models to merge superpixels into pieces, where V refers to the set of nodes and E the edges. A CRF models the conditional posterior distribution of labels as a Gibbs distribution

$$P(y|X, \theta) = \frac{1}{Z} exp(-E(y, X, \theta))$$

where θ is the parameters, Z denotes the normalization term, and E(..) is the energy function defined as sum of potentials of all cliques in the graph G. The energy function is further described as follows

$$E(y, X, \theta) = \sum_{i \in V} \varphi_u (y_i, X, \theta) + \sum_{(i,j) \in E} \varphi_p (y_i, y_j, X, \theta)$$

where φ_u is the unary potential modeling the cost of assigning label y_i to node x_i and φ_p is the pairwise potential modeling the cost of assigning a pair of labels (y_i, y_j) to pair of connected nodes (x_i, x_j). The objective is to search for the best label assignment that maximizes the conditional probability, based on the minimization of energy function

$$y^* = arg \max_y P(y|X, \theta) = \arg \min_y E(y, X, \theta)$$

The unary potential for image pieces merging is indicated byφ_u^l. To formulate it cluster all the superpixels in an image into K groups by existing algorithm K-means. Image level labels determine k: K ≥ ||*l*|| and it can be set K=2||*l*||. The unary potential is formulated as follows: Eq. (2)

$$\varphi_u^I(z_i, x_i) = ||x_i - c_{z_i}||$$

where $z_i \in \{1, \dots\dots, K\}$ is the label indicating which image piece does x_i belong and c denotes the corresponding cluster center. The pairwise potential is in the form of, Eq. (3) [21]

$$\varphi_p^I(z_i, z_j, x_i, x_j) = \lambda_1 I(z_i \neq z_j) exp\left(-\frac{||x_i - x_j||}{\delta}\right)$$

where λ_1 weights the contribution of the pairwise potential, $I(\dots)$is an indicator function that equals 1 if the input is true and 0 otherwise, and δ is the parameter of Gaussian kernel. Set δ =1 for the entire employed Gaussian kernel. The amount of pieces P in the image might be less than that of cluster K.

Module 3: Constructing Piece Library

After merging pieces on each image, gather them together to form a piece pool, where each piece is to be associated with the most appropriate semantic label. Each piece is denoted by its center feature c in the pool is regarded as a node. All the superpixels in one piece share same semantic labels$\in \{1, \dots\dots L\}$. The CRF model contributes to assigning closely related semantic labels to the similar piece while assigning diverse labels to disparate pieces. At the same time, initialize the semantic label of each piece with its image-level label *l*, to control divergence between the assigned label and its priors. The unary potential for label mapping is formulated as, Eq.(4) [19]

$$\varphi_u^L(s_i, c_i) = \exp(-\frac{l_i(s_i)}{Z} \sum_{c_j \in N(c_i)} l_j(s_i))$$

where $l_i(s_i)$ indicates the s_i xth element of l_i and Z is for normalization. The $N(c_i)$ represents neighborhood of c_i containing similar pieces with c_i. It can also be obtained from the k-means algorithm. Li is binary and might be missing or incorrect. To solve this problem, replace zeros in l_i with $\in (0 < \in < 1)$ and becomes l_i'. Here the E is used to control the confidence of labels not corresponding to piecec_i. Thus it enhances the robustness of the model.

The exploitation of label correlation that assists in label mapping becomes vital, for the location of each label is unknown. To take full advantage of semantic label correlation, integrate both co-occurrence statistics and label similarity into the pairwise potential φ_p^L . Let $L = [l_1, \dots l_N]^T \in R^{N \times L}$ be the category of labels of all images in the training set with N indicating the total number of images. The label co-occurrences matrix A is symmetric whose entry can be formulated by

$$A(i,j) = \frac{count(i \cap j)}{count(i \cup j)}$$

where $count(\dots)$ is the count of input, $i \cap j$indicates the co-occurrence of l_i and l_j, and $i \cup j$is the union set. Suppose $L_{\langle i \rangle}$ is the i^{th}column of $\boldsymbol{L}$, then $L_{\langle i \rangle} \in R^N$can be regarded as a type of feature vector of labell_i. Hence, the entry of the label similarity matrix B is formulated as,

$$B(i,j) = \frac{L_{\langle i \rangle} . L_{\langle j \rangle}}{|L_i||L_j|}$$

The same as the co-occurrence matrix A, the similarity matrix B is also symmetric. Then, the pairwise potential is formulated as, Eq.(5) [19]

$$\varphi_p^L(s_i, s_j, c_i, c_j) = \frac{\lambda_2}{A(s_i, s_j)}\left(1 - B(c_i, c_j)\right) I(s_i \neq s_j) \exp(-\frac{\|c_i - c_j\|}{\delta})$$

By minimizing the energy function, each piece is associated with a semantic label. These pieces and labels make up the piece library, which is quite convenient to enlarge.

Algorithm: CRF-based image piece learning framework algorithm.

Input: N images over-segmented into superpixels $\{X^i\}_{i=1}^N$ and their image-level ground truth $\{l^i\}_{i=1}^N$, piece number in each image K

Output: piece library containing piece centers $C = \{c_i\}_{i=1}^P$and their semantic labels $\{s_i\}_{i=1}^P$

1) $C^0 = \emptyset$
2) Permute training data randomly
3) For $i = 1 \rightarrow N$do
4) Merging superpixels in image X^i into K pieces $\{c_j^i\}_{j=1}^K$by CRF Eq. (2) and Eq. (3)
5) Update piece center set $C^i = C^{i-1} \cup \{c_j^i\}_{j=1}^K$
6) End for
7) Converging all the pieces to form piece pool
8) Calculating label co-occurrences statics matrix A and label similarity matrix B
9) Constructing piece library by associating c_i with s_i by CRF using Eq. (4) and Eq. (5)

Module 4: Inference of Testing

As mentioned above, for a test image, the image-level label remains unavailable. The same as the training images, first each test image is over-segmented into superpixels $\{x_i\}$ and graph $G'$is built upon them. The CRF adopted for testing borrows elements from the previously used model: the graph structure of $G'$and the pairwise potential for the label mapping. As for the unary potential, formulate it based on the piece library out of the training phase, which can be written as

$$\varphi_u^T(y_i, x_i) = \frac{\frac{1}{N_{y_i}} \sum_{c_j \in N(x_i)} \|x_i - c_j\| I(s_j = y_i)}{\sum_{y_i=1}^{L} \frac{1}{N_{y_i}} \sum_{c_j \in N(x_i)} \|x_i - c_j\| I(s_j = y_i)}$$

Considering the existence of few mislabeled pieces in the training and to avoid harm from them, the neighborhood $c_j \in N(x_i)$ is obtained by setting an adapted threshold η to $(x_i - c_j)$. In experiments, empirically set $\eta = mean(\|x_i - c_j\|)$ and N_{y_i} is the number of activated pieces with $s = y_i$ in the neighborhood.

The inference for a test image is to seek for the optimal solution that satisfying

$$y^* = arg\min_y \varphi_u^T(y_i, x_i) + \sum_{(i,j)\in E} \varphi_p^T(y_i, y_j, x_i, x_j)$$

Finally, for input test image y^* is the semantic label.

Python libraries such as mathplots, PIL, pillow, scikit-image, numpy, scipy, cv2, pystruct, lensers etc these all libraries are used to implement the proposed system.

4. Results and Discussion

The experiments is carry on commonly used datasets for semantic segmentation: PASCAL VOC 2012 and MSRC-21. The widely used per-class accuracy and per-class intersection -over-union (IoU) is used for performance measurements on pixel level. Per-class accuracy is defined as [#TP/(#TP + #FN)]Per-class intersection-over-union (IoU) defined as [#TP/(#TP + #FN + #FP)]where #TP, #FN, and #FP are the number of true positives, false negatives, and false positives, respectively.

Table 5.2.1 Result evaluation

Categories	Ours Percentage
Animal	90%
Non-Animal	90%

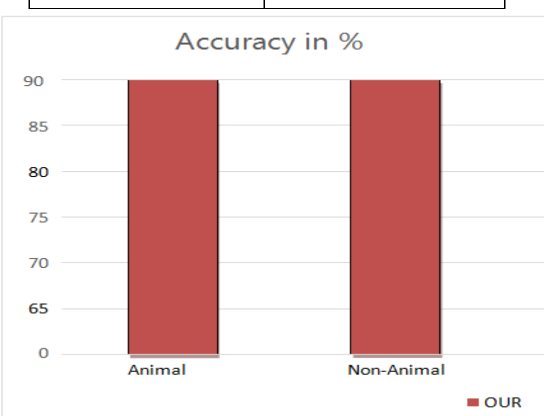

Figure 2. Evaluation Graph

The above graph shows the Accuracy with respect to animal and non-animal categories. The performance evaluation done on the commonly used dataset for segmentation is MSRC-21 and VOC 2012. For evaluation purpose, standard Animal dataset images are used.

5. Conclusion and Future Scope

The input image is given to the system. The system generates superpixels by using the SLIC algorithm. The superpixels are used to merging into pieces for constructing the piece pool. This merging is performed by using the CRF. The generated pieces are used for building a piece library. This library is used by every image given by the user to the system. The popular CRF is used for segmentation purpose to gets the labeling to every superpixel in the image. By considering the superpixels similarity and region occupied by each superpixel they were merged. Finally, the Merged regions are comparing with ground truth images and give semantic label with respect to the categories. In the future, we can implement more strong CRF methods for image semantic segmentation by combining different methods with deep learning. We can improve the accuracy of our project.

6.References

[1] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr, "Graph cut based inference with co-occurrence statistics," in Proc. Eur. Conf. Comput.Vis., Heraklion, Greece, 2010, pp. 239–253.

[2] H. Lu, G. Fang, X. Shao, and X. Li, "Segmenting human from photo images based on a coarse-to-fine scheme," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 42, no. 3, pp. 889–899, Jun. 2012.

[3] J. Carreira and C. Sminchisescu, "CPMC: Automatic object segmentation using constrained parametric min-cuts," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 7, pp. 1312–1328, Jul. 2012.

[4] Y.-L. Hou and G. K. H. Pang, "Multicue-based crowd segmentation using appearance and motion," IEEE Trans. Syst., Man, Cybern., Syst., vol. 43, no. 2, pp. 356–369, Mar. 2013.

[5] X. Yuan, J. Guo, X. Hao, and H. Chen, "Traffic sign detection via graphbased ranking and segmentation algorithms," IEEE Trans. Syst., Man, Cybern., Syst., vol. 45, no. 12, pp. 1509–1521, Dec. 2015.

[6] Y.-T. Chen, X. Liu, and M.-H. Yang, "Multi-instance object segmentation with occlusion handling," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Boston, MA, USA, 2015, pp. 3470–3478.

[7] J. Wang and A. L. Yuille, "Semantic part segmentation using compositional model combining shape and

appearance," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Boston, MA, USA, 2015, pp. 1788–1797.

[8] H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," Comput. Vis. Image Understand., vol. 110, no. 2, pp. 260–280, 2008.

[9] G. Csurka and F. Perronnin, "A simple high performance approach to semantic segmentation," in Proc. Brit. Mach. Vis. Conf., Leeds, U.K., 2008, pp. 1–10.

[10] F. Wang, Q. Huang, M. Ovsjanikov, and L. J. Guibas, "Unsupervised multi-class joint image segmentation," in Proc. IEEE Conf. Comput. Vis.Pattern Recognit., Columbus, OH, USA, 2014, pp. 3142–3149.

[11] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly supervised semantic segmentation with a multi-image model," in Proc. IEEE Int.Conf. Comput. Vis., Barcelona, Spain, 2011, pp. 643–650.

[12] Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu, "Weakly-supervised dual clustering for image semantic segmentation," in Proc. IEEE Conf. Comput.Vis. Pattern Recognit., Portland, OR, USA, 2013, pp. 2075–2082.

[13] L. Zhang et al., "Probabilistic graphlet cut: Exploiting spatial structure cue for weakly supervised image segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Portland, OR, USA, 2013, pp. 1908–1915.

[14] K. Zhang, W. Zhang, Y. Zheng, and X. Xue, "Sparse reconstruction for weakly supervised semantic segmentation," in Proc. Int. Joint Conf.Artif. Intell., Beijing, China, 2013, pp. 1889–1895.

[15] L. Zhang et al., "A probabilistic associative model for segmenting weakly supervised images," IEEE Trans. Image Process., vol. 23, no. 9,pp. 4150–4159, Sep. 2014.

[16] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in Proc. IEEE Conf. Comput. Vis.Pattern Recognit., Boston, MA, USA, 2015, pp. 1713–1721.

[17] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in Proc. IEEE Int. Conf. Comput. Vis., Santiago, Chile, 2015,pp. 1635–1643

[18] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine S□usstrunk, SLIC Superpixels, EPFL Technical Report 149300, June 2010.

[19] Yi Li, Yanqing Guo, *Member, IEEE*, Yueying Kao, and Ran He "Image Piece Learning for Weakly Supervised Semantic Segmentation" IEEE transaction on systems, man, and cybernetics: systems, vol. 47, april 2017.

Authors Profile

Priyanka S. Gunde pursed Bachelor of Engineering from SIT College of Engineering, Yadrav, India University, in year 2013, She is currently pursuing Master of Technology from DKTE's TEI, (An Autonomous Institute), Ichalkaranji, India. Her research work focuses on recommendation systems.

Suresh Shirgave is currently Associate Professor of Computer Science and Engineering at DKTESociety's Textile and Engineering Institute, Ichalkaranji, Maharashatra, India. He has published many research papers in national and international conferences and Journals. His research interests include data mining, Web usage mining, recommendation systems, social networks and Internet security.

