A Novel Weighted Semi-Supervised Clustering Algorithm and its Application in Image Segmentation

Zhaofeng Li[1], Lanqi Liu[2]

1 College of Information Engineering
Henan Institute of science and technology
Xinxiang, Henan, China
2 Henan University of Economics and Law
Zhengzhou, Henan 450046, China

Email: Zhfengli@126.com

***Abstract* — In this paper we propose a novel weighted semi-supervised clustering algorithm and then study on how to apply it in the problem of image segmentation. We explain how to obtain weights of the semi-supervised clustering algorithm using the number of unlabeled data samples and the number of data samples. After defining the data sample weights, the next task is to obtain the cluster labels by optimizing a contingency matrix, and then cluster labels can be randomly initialized and updated in an iterative mode. Afterwards, based on the weighted semi-supervised clustering algorithm, novel image segmentation is given. In our proposed method, each image is represented by a D-dimensional random vector, and each pixel is drawn independently from the mixture density. Next, experts are invited to evaluate the quality of image segmentation using the training dataset. Afterwards, each segment of the images in the training dataset is assigned a class label. Then, a one-to-one mapping function between mixture model components and the segment classes is defined. Furthermore, each pixel of an image is assigned to the labelled component of the mixture which has the highest posterior probability. Finally, the image segmentation result can be obtained using the class label of the mixture component. Experimental results demonstrate that the proposed image segmentation algorithm can achieve both high segmentation accuracy and low computation cost.**

Keywords - Semi-supervised clustering, Image segmentation, Gaussian mixture model, Mixture density, Contingency matrix.

I. INTRODUCTION

Image segmentation denotes a key problem in computer vision, and it is also refers to a process of allocating a label to each pixel in an image [1,2]. After segmenting an image, pixels with the same label may share similar characteristics. On the other hand, image segmentation can be regarded as partitioning of a specific image into several non-overlapping regions. That is, it is known as a low-level image processing technique which transforms an image into several regions for high-level image description according to visual features, salient objects, and background scenes [3-5]. Image segmentation refers to one of the most key and challenging tasks in computer vision and pattern recognition, which is also very helpful for many application fields, such as object detection, object recognition, image retrieval, scene analysis, medical image processing, and video surveillance.

However, different segmentation approaches solve the same image in different modes and then different quality of segmentation results is obtained. Furthermore, parameters of image segmentation algorithms may be by the user in advance. But, each image may have its own optimal parameters, and fixed parameter settings are not suitable [6,7]. To obtain high quality segmentation results, appropriate parameters are needed for image segmentation algorithm. Considering there are no unsupervised algorithms for best parameters choosing, in this paper, we introduce the semi-supervised learning in image segmentation.

As is well known that semi-supervised learning can combine prior knowledge in the unsupervised approach, and then the computing effectiveness can be promoted greatly without requiring a complete training dataset. Hence, this paper aims to enhance image segmentation results utilizing incomplete training datasets, that is, the proposed algorithm only need less manual analysis by experts than a supervised algorithm.

The rest of this paper is organized as follows. The next section gives an overview of the related works about semi-supervised clustering. Section 3 provides a weighted semi-supervised clustering algorithm. Afterwards, section 4 explains how to use the proposed weighted semi-supervised clustering algorithm for image segmentation. In section 5, experiments are conducted to make performance evaluation. Finally, the concluding section draws the findings from the analyses of the proposed algorithm.

II. RELATED WORKS

The existing former image segmentation methods mainly focus on the following categories, such as thresholding [8], clustering [9-12], graph cuts [13], region split and merge[14], matrix decomposition based approaches [15] and partial differential equation based algorithm. Up to now, in the field of image segmentation, a wide variety of surveys have been done.

Different from the above methods, in this paper, we focus on using Semi-Supervised learning in image segmentation. In recent years, Semi-Supervised learning has been proved to be an effective approach, and has been successfully utilized in many fields as follows.

Meng et al. proposed a generalized organization of Heterogeneous Fusion Adaptive Resonance Theory, and is can be used in co-clustering of large-scale web multimedia documents. By extending the two-channel Heterogeneous Fusion ART to multiple channels, this theory is designed to solve multimedia data with an arbitrarily rich level of meta-information.

Calandriello et al. presented a new semi-supervised clustering algorithm using the information-maximization principle. Moreover, the proposed approach is computationally efficient because the clustering solution may achieve analytically through eigende composition.

He et al. illustrated a new semi-supervised non-negative matrix factorization algorithm. This algorithm exploits the local structure of the data characterized by the graph Laplacian, and then the label information as the fitting constraints to learn is used as well.

Baghshah et al. utilized a spectral embedding to learn a square matrix, in which the number of rows is the number of dimensions in the embedded space. Hence, this algorithm can achieve higher scalability than other methods.

Yu et al. propose a feature selection based semi-supervised cluster ensemble framework for tumor clustering from bio-molecular data. Ben et al. solved the group extraction of LinkedIn users based on their profiles utilizing a semi-supervised clustering method with quantitative constraints ranking. Anand et al. proposed a semi-supervised algorithm for kernel mean shift clustering, and this algorithm introduce pairwise constraints to implement the clustering process. Wang et al. presented a new framework by using the constrained pairwise data points and its neighbors and then it can naturally tackle constraint conflicts.

However, the above studies have not considered the weights in semi-supervised clustering. Therefore, in the following parts, we will discuss how to design a weighted semi-supervised clustering algorithm.

III. WEIGHTED SEMI-SUPERVISED CLUSTERING ALGORITHM

Clustering technology can classify similar data points into same group. Assuming that $X=\{X_i, i=1,2,\cdots,n\}$, and X_i is a data sample. The aim of clustering is to obtain a partition $C=\{C_1,C_2,\cdots,C_m\}$, and the following conditions should be followed.

$$X=\bigcup_{i=1}^{m} C_i, C_i \neq \varnothing, i \in \{1,2,\cdots,\mathrm{m}\} \tag{1}$$

$$C_i \bigcap C_j = \varnothing, i, j \in \{1,2,\cdots,m\}, i \neq j \tag{2}$$

Assuming that there are N data sample points $X=\{x_1,x_2,\cdots,x_N\}$ with labels $\{l_1,l_2,\cdots,l_N\}$ in K clusters. For a given cluster C_k, it has a prototype c_k and a category label l_k. Values of category labels for all the data samples have the range of $l_i \in \{0,1,2,\cdots,Q\}, \forall i=1,2,\cdots,N$, in which the value l_i of an unlabeled data sample is equal to zero. The weights of the semi-supervised clustering algorithm are defined by the following equation.

$$w^c(i)=\begin{cases}1, & if\ l_i=0\\ 1+\lambda\cdot\dfrac{Num(l=0)}{Num(l=c)}, & if\ l_i=c\\ 0, & if\ l_i\neq c\end{cases} \tag{3}$$

where parameter λ is always equal or larger than zero, $Num(l=0)$ means the number of unlabeled data samples and $Num(l=c)$ represents the number of data samples which are labeled by c. Afterwards, for each cluster label l_k, weight of the sample with index i to a given cluster is defined as:

$$\sigma_k(i)=w_{l_k}(i) \tag{4}$$

After defining the data sample weights, the next task is to obtain the cluster labels (denoted as $\hat{l} \in \{1,2,\cdots,Q\}^K$). Next, the contingency matrix is given as follows.

$$M_{c,k}=\left\{x_i \middle| k=\arg\max_{p=1,2,\cdots,K} f_p(i), l_i=c, i\in\{1,2,\cdots,N\}\right\} \tag{5}$$

Through optimizing the above contingency matrix, labels of cluster could be randomly initialized and then be updated in an iterative way as follows:

$$\tilde{l}_k=\arg\max_{c\in\{1,2,\cdots,Q\}} \Delta F\left(\hat{l},\tilde{l},\tilde{l}_k,c\right) \tag{6}$$

where ΔF is the objective function we aim to maximize. Afterwards, weighs of sample belonged to clusters and cluster labels $\hat{l}$ can be obtained.

IV. THE PROPOSED IMAGE SEGMENTATION METHOD

Supposing that I refers to an image, and it is represented by a set of pixels $I=\{x^1,x^2,\cdots,x^n\}$. Moreover, each image is denoted by a d-dimensional random vector $x^i=(x_1,x_2,\cdots,x_d)$, where n is the number of pixels in the image. Furthermore, each image region r satisfies the density function $p_r(x|\theta_r)$, $r\in\{1,2,\cdots,k\}$, where k is the total number of regions in the given image. Hence, each pixel is drawn independently from the mixture density which is calculated as follows:

$$p\left(x^i|\xi\right)=\sum_{r=1}^{k}\tau_r\cdot p_r\left(x^i|\theta_r\right) \tag{7}$$

where τ_r means the mixture mixing rates which are not negative, and $p_r\left(x^i|\theta_r\right)$ denotes the component density which is corresponding to region r and ξ refers to a set of mixture parameters.

$$\xi=\left\{\theta_1,\theta_2,\cdots,\theta_k,\tau_1,\tau_2,\cdots,\tau_k\right\} \tag{8}$$

where τ_i is the mixing proportion, which is related to the prior probability of each pixel in the i^{th} class.

Using the above weighted semi-supervised clustering algorithm, some samples are randomly from the dataset. Afterwards, we invite some expert to evaluate the clustering results. If an image is segmented with high quality, experts allocate a label to each segment of this given image. Therefore, each segment of the images which are belonged to the training dataset has a category label. Furthermore, we suppose one to one mapping function between mixture model components and the segment categories. Symbol c_j represents the j^{th} class. For the clusters of training image dataset, mixture parameters $\xi_j=\{u_j,\Sigma_j,\pi_j\}$ of category c_j can be obtained using its allocated cluster pixels as follows:

$$\pi_j=\frac{|n_j|}{n} \tag{9}$$

$$\mu_j=\frac{\sum_{x^i\in c_j} x^i}{|n_j|} \tag{10}$$

$$\Sigma_j=\frac{\sum_{x^i\in c_j}\left(x^i-\mu_j\right)\left(x^i-\mu_j\right)^T}{|n_j|} \tag{11}$$

where the symbol $|n_j|$ represents the number of pixels which is belonged to the j^{th} category

Next, for a new image to be segmented, parameters $\xi_j,\forall_j$ should be calculated in advance as the initial parameters of the Gaussian Mixture Model to smooth convergence of EM algorithm. Considering the model parameters are calculated from a correctly segmented image, parameter ξ_j can make the EM convergence to achieve the global maxima.

In the end, using the above method, each pixels of an image to be segmented is allocated to the labelled component of the mixture that holds the highest posterior probability. Then, the image segmentation result can be obtained via the class label of the mixture component which has been allocated to the segmentation cluster.

V. EXPERIMENT

In this section, we conduct a series of experiments to make performance evaluation. To make performance comparison, several image segmentation methods are utilized, such as PWT, DTIB, LGLD, Ncut. PWT (parzen window estimate) refers to a image segmentation method combines histogram with the Parzen window technique to estimate the spatial probability distribution of gray-level image values, and then propose a new criterion function. DTIB means a novel double-threshold image binarization approach using the edge and intensity information. LGLD is a new transition region extraction and thresholding

algorithm based on gray level difference. Ncut is a typical image segmentation algorithm, which regards the image segmentation problem as a graph partitioning problem, and then design a novel global criterion (denoted as the normalized cut) for segmenting the graph.

The dataset we used in this paper is selected from paper, four types infrared images ("Tank", "Sailboat", "Airplane", and "Moon") with a small single object were exploited as testing samples. To objectively evaluate quality of image segmentation, some standard measures are used, which are 1) ME, 2) FPR, and 3) FNR. ME denotes the mis-classification error rate, and it is defined as follows:

$$ME = 1 - \frac{|B_o \cap B_T| + |F_o \cap F_T|}{|B_o| + |F_o|} \tag{12}$$

where symbols B_o and F_o mean the background and the foreground for the ground truth image. Meanwhile, B_T and F_T denote the background and the foreground in the segmentation result respectively. Particularly, lower value of ME denotes the better quality of image segmentation.

FPR refers to the rate of number of background pixels which are wrongly classified to the total number of background pixels in the correctly segmented image. FNR represents the ratio of number of wrongly classified foreground pixels to the total number of foreground pixels in the in the correctly segmented image. Based on the above definitions, FPR and FNR are defined as follows.

$$FPR = \frac{|B_o \cap F_T|}{|B_o|} \tag{13}$$

$$FNR = \frac{|F_o \cap B_T|}{|F_o|} \tag{14}$$

Integrating the performance given by FPR and FNR, the quality of image segmentation can be demonstrated.

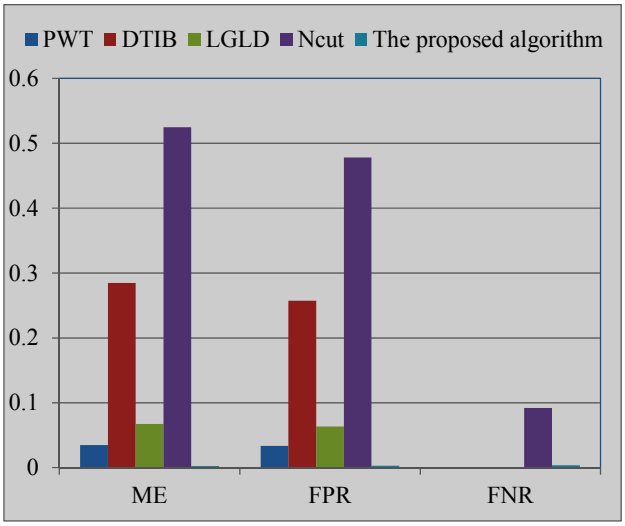


Figure 1. Performance comparison for different methods using "Tank"

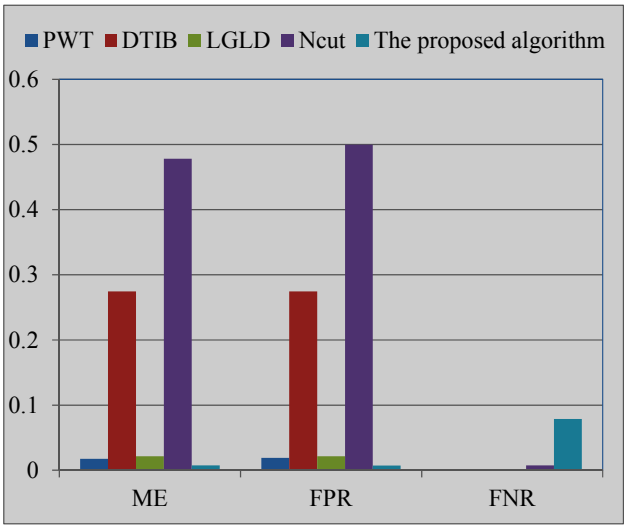


Figure 2. Performance comparison for different methods using "Sailboat"

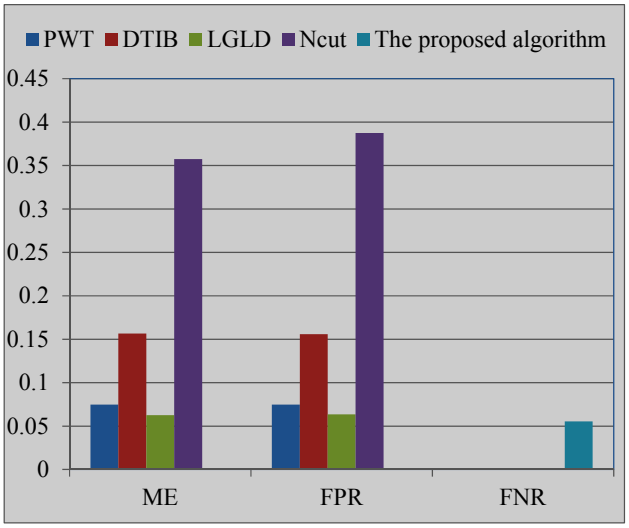


Figure 3. Performance comparison for different methods using "Airplane"

To represent the overall performance of the given four types of images, we average all the experimental results in Fig.1 to Fig.4 (shown in Table.1).

Integrating the above experimental results, we can see that for the above four types of images, our proposed algorithm perform best under the ME and FPR metric. For the FNR metric, our proposed is very close to the optimal performance. Therefore, the conclusions can be drawn that the proposed is the most effective approach for image segmentation.

On the other hand, apart from the segmentation quality, time cost is another important factor for image segmentation. Furthermore, real-time segmentation is another goal of our work. To illustrate the computation cost of each method, we use the same computing platform to ensure fairness. The Thinkpad T430 is used as the experimental platform, which is made up of Core i5-3320M (2 Kerne, 2.60GHz, 3MB cache) processor, NVIDIA Optimus Graphics (NVS 5400M, orderbar mit 1 oder 2GB), and 4GB memory.

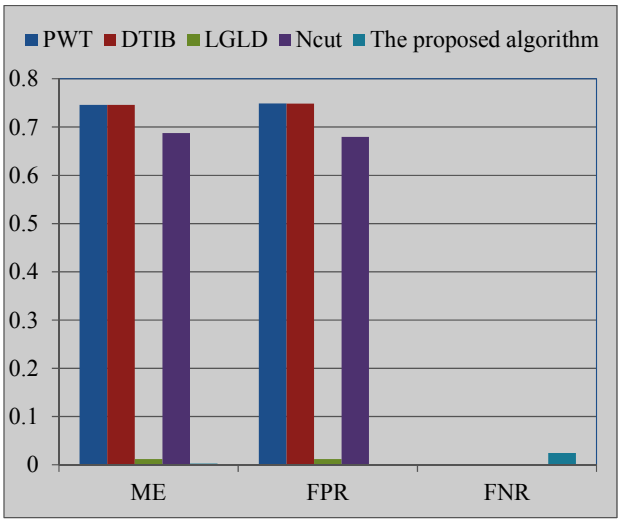


Figure 4. Performance comparison for different methods using "Moon"

TABLE I OVERALL PERFORMANCE OF THE ABOVE FOUR TYPES IMAGES

	PWT	**DTIB**	**LGLD**	**Ncut**	**The proposed algorithm**
ME	0.218215	0.365443	0.040877	0.511888	0.003369
FPR	0.218905	0.35907	0.040139	0.511224	0.002995
FNR	0	0	0	0.02485	0.040561

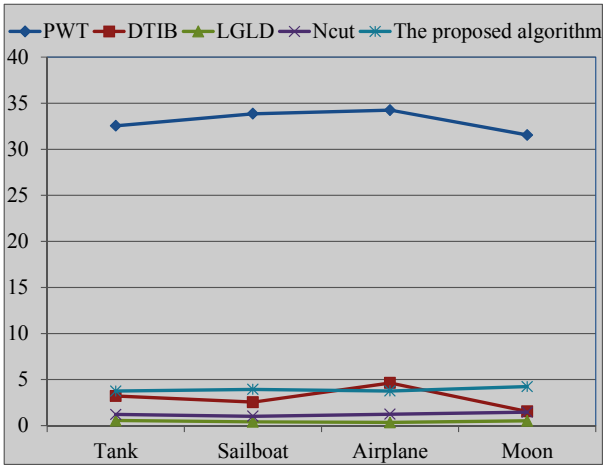


Figure 5. Computation cost of different methods

As is shown in Fig. 5, PWT requires the highest time among all the methods, and the computation time of LGLD is the lowest. The time cost of our proposed algorithm is satisfied, because the performance of our algorithm is very close to the optimal one. The reason lies in that the proposed algorithm is belonged to semi-supervised learning, and the training time is lower than the supervised learning.

Supervised learning requires a large scale of labelled training dataset, which needs high computation time, meanwhile, unsupervised learning cannot solve the problem of local traps. Furthermore, our proposed semi-supervised algorithm introduces prior knowledge to the unsupervised learning, and then it can promote the segmentation quality without labelled training dataset.

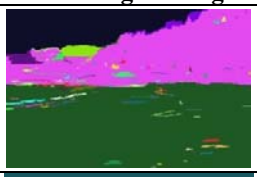
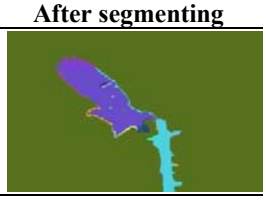
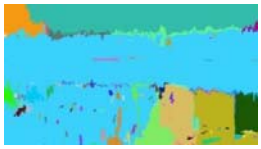
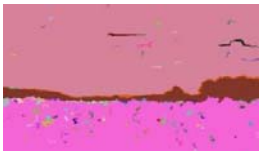
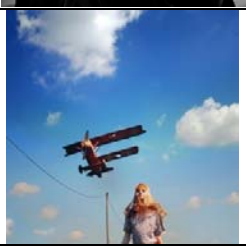
To make the performance description more clearly, eight examples of the image segmentation are illustrated in Table.2.

Table. 2 shows that our proposed algorithm can effectively segment images, especially it can precisely locate salient objects in images.

VI. CONCLUSIONS

This paper focuses on the weighted semi-supervised clustering algorithm and utilizes it in image segmentation. The main innovations of this paper lie in that 1) we introduce the weights in traditional semi-supervised clustering algorithm, and obtain the cluster labels by optimizing a contingency matrix, and 2) each image is represented by a D-dimensional random vector, and each pixel is drawn independently from the mixture density. Using the proposed weighted semi-supervised clustering algorithm, image segmentation result can be obtained through the class label of the mixture component. Experimental results demonstrate the effectiveness and efficiency of the proposed algorithm.

TABLE II EXAMPLES OF IMAGE SEGMENTATION RESULTS.

Original image	After segmenting	Original image	After segmenting
			
			
			
			

REFERENCE

[1] Reboucas Filho Pedro Pedrosa, Cortez Paulo Cesar, da Silva Barros Antonio Carlos, de Albuquerque Victor Hugo C., "Novel Adaptive Balloon Active Contour Method based on internal force for image segmentation - A systematic evaluation on synthetic and real images", Expert Systems with Applications, vol. 41, No.17, pp. 7707-7721, 2014.

[2] Liao Xiangli, Xu Hongbo, Zhou Yicong, Li Kunqian, Tao Wenbing, Guo Qiuju, Liu Liman, "Automatic image segmentation using salient key point extraction and star shape prior", Signal Processing, vol. 105, pp. 122-136, 2014.

[3] Zeng Shan, Huang Rui, Kang Zhen, Sang Nong, "Image segmentation using spectral clustering of Gaussian mixture models", NEUROCOMPUTING, vol. 144, pp. 346-356, 2014.

[4] Qi Chengming, "Maximum Entropy for Image Segmentation based on an Adaptive Particle Swarm Optimization", Applied Mathematics & Information Sciences, vol. 8, No. 6, pp. 3129-3135, 2014.

[5] Mignotte Max, "A label field fusion model with a variation of information estimator for image segmentation", Information Fusion, vol. 20, pp. 7-20, 2014.

[6] Chen Yin, Cremers Armin B., Cao Zhiguo, "Interactive color image segmentation via iterative evidential labeling", Information Fusion, vol. 20, pp. 292-304, 2014.

[7] Wang Xuchu, Shan Jinxiao, Niu Yanmin, Tan Liwen, Zhang Shao-Xiang, "Enhanced distance regularization for re-initialization free level set evolution with application to image segmentation", NEUROCOMPUTING, vol. 141, pp. 223-235, 2014.

[8] Gao Xinbo, Fu Rong, Li Xuelong, Tao Dacheng, Zhang Beichen, Yang Huigen, "Aurora image segmentation by combining patch and texture thresholding", Computer Vision and Image Understanding, vol. 115, no. 3, pp. 390-402, 2011.

[9] Ahmed Mohamed N, Yamany Sameh M, Mohamed Nevin, Farag Aly A, Moriarty Thomas, "A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data", IEEE Transactions on Medical Imaging, vol. 21, No. 3, 193-199, 2002.

[10] F. Chung-Hoon Rhee, "Uncertain fuzzy clustering: insights and recommendations", in: IEEE Computational Intelligence Magazine, pp. 44-56, 2007.

[11] Gong Maoguo, Liang Yan, Shi Jiao, Ma Wenping, Ma Jingjing, "Fuzzy c-means clustering with local information and kernel metric for image segmentation", IEEE Transactions on Image Processing, vol. 22, No. 2, pp. 573-584, 2013.

[12] Krinidis Stelios, Chatzis Vassilios, "A robust fuzzy local information C-means clustering algorithm", IEEE Transactions on Image Processing, vol. 19, No. 5, pp. 1328-1337, 2010.

[13] B. Peng, L. Zhang, D. Zhang, J. Yang, "Image segmentation by iterated region merging with localized graph cuts", Pattern Recognition, vol. 44, no. 10-11, pp. 2527-2538, 2011.

[14] D. Chaudhuri, A. Agrawal, "Split-and-merge procedure for image segmentation using bimodality detection approach", Defence Science Journal, vol. 60, no. 3, pp. 290-301, 2010.

[15] T. Zhou, D. Tao, "GoDec: randomized low-rank & sparse matrix decomposition in noisy case", in: International Conference on Machine Learning, vol. 9, no. 10, pp. 33–40, 2011.