Real-time Background Subtraction via L1 Norm Tensor Decomposition

Taehyeon Kim* and Yoonsik Choe†
* Yonsei University, Seoul, Korea
E-mail: pyomu@yonsei.ac.kr Tel/Fax: +82-10-2702-7671
† Yonsei University, Seoul, Korea
E-mail: yschoe@yonsei.ac.kr

***Abstract*—Currently, background subtraction is being actively studied in many image processing applications. Nuclear Norm Minimization (NNM) and Weighted Nuclear Norm Minimization (WNNM) are commonly used background subtraction methods based on Robust Principal Component Analysis (RPCA). However, these techniques approximate the RPCA rank function and take the form of an iterative optimization algorithm. Therefore, due to the approximation, the NNM solution can not converge if the number of frames is small. In addition, the NNM and WNNM processing times are delayed because of their iterative optimization schemes. Thus, NNM and WNNM are not suitable for real-time background subtraction. In order to overcome these limitations, this paper presents a real-time background subtraction method using tensor decomposition in accordance with the recent tensor analysis research trend. In this study, we used the closed form TUCKER2 decomposition solution to omit the iterative process while retaining the L1 norm of the RPCA rank function. This proposed method allows for convergence even when the number of frames is small. Compared to NNM and WNNM, the proposed method reduces the processing time by more than 80 times and has a higher precision even when the number of frames are less than 10.**

I. Introduction

Along with computer vision and video technology, real-time background subtraction methodology is regarded as an important technology that has been the focus of many studies since the early 20th century. Real-time background subtraction applications include video surveillance, detection of moving objects, human motion capture, etc. To fulfill the needs of these applications, many background subtraction methods such as the statistical method [7, 8], the sparse method [9], the robust subspace method [2,11,12], the tensor deocmposition method [17] and the neural network method [4, 12] have been proposed.

In this paper, we propose a Robust Principal Component Analysis (RPCA) based background subtraction method which starts with the assumption that the video frame (X) is the sum of the background image (low-rank matrix A) and the foreground image (sparse matrix E). However, minimizing the rank-function is an NP-hard problem.

To overcome this problem, Nuclear Norm Minimization (NNM) has been proposed in order to treat the rank function in a tractable manner [2]. However, NNM simply minimizes the sum of the singular values obtained through Singular Value Decomposition (SVD) for approximation. Therefore, the solution does not converge if the number of frames is small, and the processing time is longer due to the iterative optimization algorithm.

A modified NNM called the Weighted Nuclear Norm Minimization (WNNM) minimizes the sum of the weighted singular values by multiplying weights according to the importance of each singular value obtained through SVD [3]. WNNM is highly accurate even when the number of frames is small since it uses a close approximation of the L0 norm, the actual signal form. However, WNNM is not only defined as a nonconvex problem but also is an iterative algorithm; so the processing time is even longer than NNM.

To reduce the number of iterations, technique were also proposed to utilize the tensor decomposition [17]. However, all of the methods above are based on approximation method and therefore essentially require the iteration process which has limitation on real-time process

Therefore, the proposed method is a real-time background subtraction processing method based on L1 norm exact solution by using tensor decomposition [1]. Since there are no iteration process to get optimal solution, this method is highly accurate and has short processing time even when the number of frames is small.

The proposed scheme uses the L1 norm is used since RPCA uses the absolute value of the singular value. And TUCKER decomposition of 3-way tensor constructed from video frames is simplified to the TUCKER2 decomposition, thus, L1 norm TUCKER2 decomposition is used to perform the background subtraction. As a result, the separate iteration based approximation process to provide a feasible rank function solution can be omitted, by directly utilizing the closed form L1 norm calculation.

Since the proposed background subtraction method is optimized using an exact closed form solution, the number of iteration for optimization is small. Therefore, the proposed method is not only highly accurate when the number of frames is small but also needs short processing time that enables real-time processing.

II. Related Methods

A. Background Subtraction Through Nuclear Norm Minimization

As previously mentioned, background subtraction can be performed using RPCA. However, the direct minimization of the rank function is a NP-hard problem. In order to solve this problem, the NNM scheme, which approximates the rank function using the nuclear norm, has been proposed [2,6]. The NNM formula is as follows:

$$min_{A,E}||A||_* + \lambda||E||_1 \qquad s.t. X = A + E \tag{1}$$

where $||.||_*$ is the nuclear norm, which is the sum of the singular values obtained through SVD; and λ is the regularizing parameter. NNM is a tractable convex optimization using the nuclear norm. However, since the nuclear norm is merely the sum of the singular values without considering the significance of each singular value, NNM only converges when there are a sufficient number of frames. In addition, NNM is an iterative optimization algorithm, which is not suitable for real-time processing.

B. Background Subtraction Through Weighted Nuclear Norm Minimization

Since NNM uses the nuclear norm, accuracy is reduced when the number of frames is small. In order to overcome this drawback, WNNM has been proposed [3]. The WNNM equation is as follows:

$$min_{A,E}||A||_{w,*} + \lambda||E||_1 \qquad s.t. X = A + E \tag{2}$$

where $||.||_{w,*}$ is the weighted nuclear norm, which is the sum of the results obtained by multiplying each singular value by a weight according to the importance of the singular value; and λ is the regularizing parameter. WNNM yields an approximation closer to the L0 norm leading to performance improvement since the actual background and foreground images are in the L0 norm form. Therefore, compared to NNM, WNNM is highly accurate even with a small number of frames. However, WNNM is a nonconvex optimization problem. As a result, the processing time is even longer than NNM since more iterations are required.

III. Proposed Algorithm

Tensor is a multi-dimensional array, also called N-way array, where N is the dimension of the tensor [10]. The TUCKER decomposition is a form of high-order principal component analysis. It disintegrates a tensor into a core tensor multiplied by a matrix along each mode. In a three-way tensor case, there are three factor-matrices and one core tensor [15]. If n-way tensor is composed of a (n-1)way tensor, the TUCKER decomposition can be simplified to the TUCKER2 decomposition. The reason is that one of the factor matrices can be considered as an identity matrix. In the background subtraction task, since the input tensor (3-way tensor) consists of video frames(2-way tensor), the TUCKER2 decomposition is applicable.

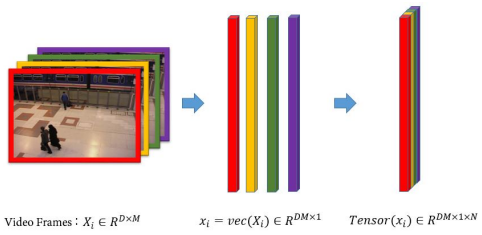

Fig. 1. Construct 3-way tensor through set of n video frames

When performing the background subtraction with the proposed method, the background image in the video frame is assumed to have a rank of 1 since the background image is a low rank component that exists at all times. Therefore, a 3-way tensor is constructed by changing the video frames into column vectors as shown in Fig. 1

The equation of the L1 norm TUCKER2 decomposition proposed in [5] is as follows:

$$max_{\substack{U\in R^{D\times d}; U^TU=I \\ V\in R^{M\times d}; V^TV=I}} \sum_{i=1}^{N} ||U^T X_i V||_1 \tag{3}$$

where X_i is i^{th} (n-1)-way tensor that constitutes the n-way tensor.

Therefore, from a background subtraction point of view, X_i is a video frame of size $D \times M$ where D and M are the width and height of the frame, respectively. U and V are left hand and right hand singular vectors, respectively. d is the rank of X_i, and $||,||_1$ is L1-norm. In (3), $U^T X_i V$ is a diagonal tensor in high-order SVD terms since it is X_i multiplied by the transpose of singular vectors. Therefore, the low-rank of X_i can be approximated by $UU^T X_i V V^T$. Since $X_i \in R^{D\times M}$ takes the form of column vector ($x_i \in R^{DM\times 1}$), (3) can be modified as follows:

$$max_{u\in R^{DM\times 1}; v\in R; ||u||_2=|v|=1} \sum_{i=1}^{N} |u^T x_i v| \tag{4}$$

where x_i is $vec(X_i) \in R^{DM\times 1}$, and u and v are left hand and right hand singular vectors, respectively. In (4), since x_i is a column vector, v can be omitted by keeping $v = 1$ for any u. Maximizing the sum of the absolute values of variables can be achieved by multiplying $+1$ if the variable is positive or multiplying -1 if variable is negative. Therefore, using above two properties, (4) can be changed as follows:

$$max_{b\in\{\pm 1\}^N} u^T (\sum_{i=1}^{N} b_i x_i) \tag{5}$$

where b is a vector composed only of +1 and -1, and b_i is a i^{th} component of b vector. According to [16], in order for (5) to have the maximum value, u become, by the matrix approximation optimality of SVD, the left hand dominant singular vector of $\sum_{i=1}^{N} b_i x_i$ respectively. Also, If

TABLE I
F2-SCORE RESULTS OF NNM, WNNM AND PROPOSED METHOD USING PETS2006 DATASET (576×720)

Frames	3	4	5	6	7	8	9	10
NNM-ALMM	0.1088	0.3967	0.4603	0.494	0.484	0.6871	0.7732	0.7732
WNNM-ALMM	0.591	0.512	**0.8163**	0.788	**0.7985**	0.7766	**0.7751**	0.7702
Ours	**0.6403**	**0.7757**	0.8137	**0.7908**	0.7916	**0.7778**	0.7741	**0.778**

TABLE II
PROCESSING TIME (*sec*) RESULTS OF NNM, WNNM AND PROPOSED METHOD USING PETS2006 DATASET (576×720)

Frames	3	4	5	6	7	8	9	10
NNM-ALMM	1.627	2.236	2.848	3.497	4.152	4.907	5.762	6.894
WNNM-ALMM	5.620	9.079	11.728	15.771	17.503	23.038	27.136	32.827
Ours	**0.054**	**0.092**	**0.145**	**0.27**	**0.535**	**0.977**	**2.012**	**4.033**

TABLE III
F2-SCORE RESULTS OF NNM, WNNM AND PROPOSED METHOD USING PEDESTRIANS DATASET (320×240)

Frames	3	4	5	6	7	8	9	10
NNM-ALMM	0.3322	0.4115	0.6632	0.6720	0.7388	0.7655	0.7742	0.7764
WNNM-ALMM	0.7671	0.7526	**0.789**	**0.7846**	**0.8090**	**0.8032**	**0.8007**	**0.7929**
Ours	**0.7702**	**0.785**	0.7705	0.7649	0.7661	0.7844	0.7852	0.7789

TABLE IV
PROCESSING TIME (*sec*) RESULTS OF NNM, WNNM AND PROPOSED METHOD USING PEDESTRIANS DATASET (320×240)

Frames	3	4	5	6	7	8	9	10
NNM-ALMM	0.439	0.504	0.631	0.734	0.864	1.006	1.138	1.310
WNNM-ALMM	0.834	1.335	1.605	2.201	1.752	3.639	4.179	5.335
Ours	**0.018**	**0.027**	**0.04**	**0.046**	**0.075**	**0.105**	**0.233**	**0.503**

$x_i \in R^{DM \times 1}$ is concatenated to make a tensor such as $X = [x_1, x_2, ..., x_N] \in R^{DM \times N}$, $\sum_{i=1}^{N} b_i x_i$ can be expressed as $X(b \otimes I_M)$ by simplified Kronecker matrix product. However, in this paper, M refers to the width of the frame; and because x_i is a column vector, it is changed to $X(b \otimes 1)$ as $M = 1$. Therefore, (5) can be changed as follows:

$$max_{b \in \{\pm 1\}^N} \sigma_{max}(X(b \otimes 1)) \tag{6}$$

where $\sigma_{max}(,)$ is the largest value among the singular values. Through (6), we use the L1 norm to perform background subtraction. As a result, unlike NNM and WNNM, the approximation using a surrogate norm is omitted and replaced by an optimization method that only uses a combination of b vector. This is equivalent to using an exact closed form solution guaranteeing high accuracy and a short processing time when the number of frames is small, which is important in real-time background subtraction.

However, when the video data is directly used, the configuration of the b vector is constituted by only positive side due to the nonnegative property of the image data. Therefore, the average frame is subtracted from each video frame so that the data is zero-centered. If the original video frame is represented by a_i, x_i in (4) is $x_i = a_i - \frac{1}{N} \sum_{i=1}^{N} a_i$, and N is the total number of video frames.

If the number of frames being processed at one time becomes large, the proposed method has a longer processing time than NNM and WNNM because the number of b vectors to be constituted is also getting large; and the accuracy is lower than that of WNNM, which approximates the L0 norm. However, real-time background subtraction processing conditions require high accuracy and short processing time when the number of frames is small. Thus, the effect of b on the processing time is insignificant; and the proposed method is more suitable for real-time processing than NNM and WNNM.

IV. EXPERIMENT RESULT

In order to compare the background subtraction real-time processing performance of the proposed method to NNM and WNNM, MATLAB was used on a machine with an intel core i5 3.20 GHz processor and 16 GB of RAM. The dataset utilized the PETS2006 dataset and the Pedestrians dataset in gray scale [13]. For the precise quantitative evaluation of the background subtraction methods, we used the F2 score using the foreground image extracted from the background image obtained from each method.

Since the processing time is an important criterion in real-time background subtraction, the experiment was conducted using NNM and WNNM along with the Augmented Lagrange

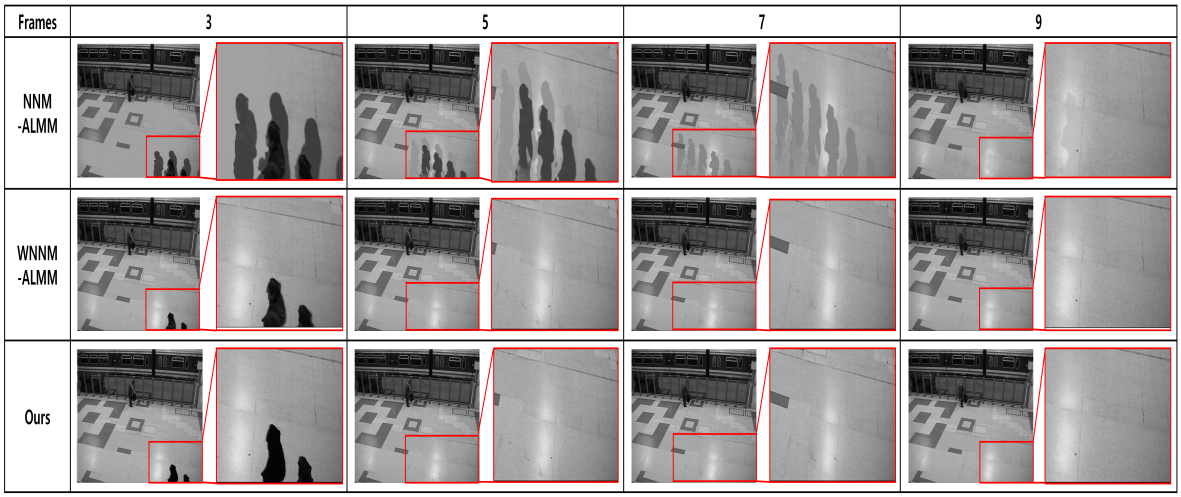

Fig. 2. Background Image results of NNM, WNNM and Proposed method using PETS2006 dataset (576×720)

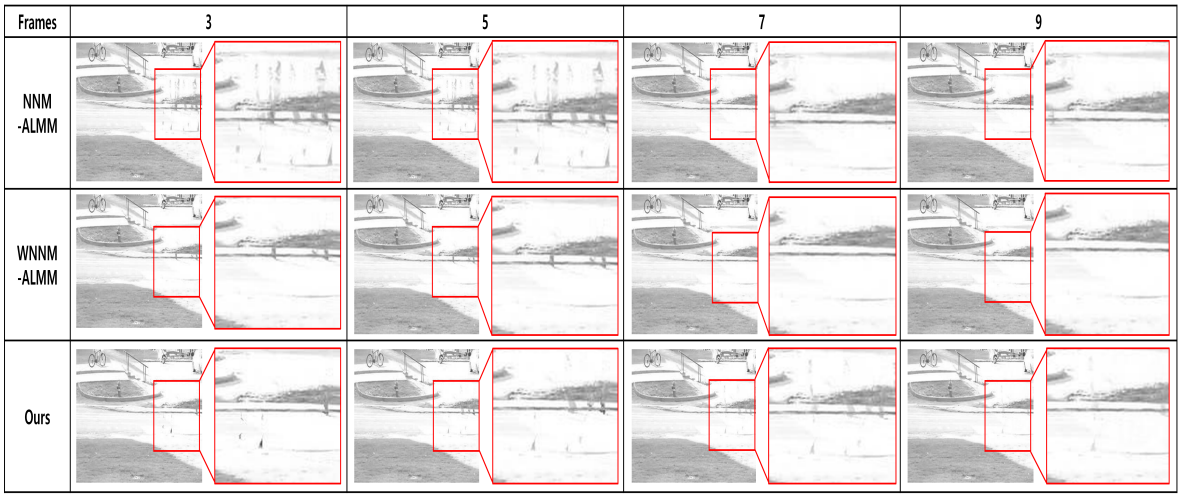

Fig. 3. Background Image results of NNM, WNNM and Proposed method using Pedestrians dataset (320×240)

Multiplier Method (ALMM) [14]. ALMM is one of the methods used to accelerate NNM and WNNM. In case of WNNM, we used $\sqrt{2 \times max(D^3, M^3)}$ as the weight in [3], where D and M are the height and width of the image, respectively.

Table I and Table II show the F2 scores and processing times derived from the PETS2006 dataset, respectively. WNNM has a high F2 score even when the number of frames is low. However, when compared with the proposed method, WNNM has more than 100 times longer processing time in experiments with only 3 frames. As mentioned earlier, NNM requires at least 9 frames to yield the high F2-score. Although the NNM processing time is less than WNNM's, NNM has more than 80 times longer processing time than the proposed method in experiments with only 3 frames.

Tables III and IV show the F2 scores and processing times derived from the Pedestrians dataset, respectively. The trends are similar to the PETS2006 dataset results.

As seen from the PETS2006 dataset experimental results in Fig. 2, the NNM method showed visible moving object traces when less than 7 frames were used due to the imperfect progression of the background subtraction. However, it was possible to extract the background image, excluding the overlapping areas between moving objects, with only 3 frames when using WNNM or the proposed method. In addition, when 5 or more frames are used, the background image was extracted leaving only small moving object traces which were difficult to visually recognize.

The Pedestrians dataset experimental results in Fig. 3 show the same trend as the PEST2006 dataset. However, there were more moving object traces using the proposed method than

WNNM due to the fact that WNNM can approximate the L0 norm closer.

Tables II and IV sow the processing time for each method. In case of the proposed method, the processing time approximately is doubled each time a frame is added. Therefore, as the number of frames continue to increase, WNNM and NNM will eventually process in shorter processing time than the proposed method. However, the final goal of realtime background subtraction processing is to achieve high accuracy in short processing time when the number of frames is small. Consequently, the proposed method is convinced to be a better method for real-time processing than NNM and WNNM.

V. Conclusions

This paper proposes a real-time background subtraction method using closed form solution of L1 norm tensor decomposition. Therefore, this method minimizes the approximation disadvantages of the conventional methods such as NNM and WNNM by using the exact solution of L1 norm tensor decomposition directly.

The experimental results confirmed that NNM has low accuracy due to the fact that the approximation uses the nuclear norm. In addition, WNNM had a longer processing time than the proposed method and NNM because WNNM is a nonconvex optimization technique based on an iteration process. Overall, the proposed method has shorter processing time and better accuracy than NNM and WNNM.

In conclusion, the experimental results using real-world data show that the proposed method is more suitable for real-time background subtraction processing than WNNM and NNM proving that exact solution of L1 norm tensor decomposition can be successfully used for real-time background subtraction.

Acknowledgements

This work was supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program(10073229, Development of 4K highresolution image based LSTM network deep learning process pattern recognition algorithm for real-time parts assembling of industrial robot for manufacturing) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea)

References

[1] Markopoulos, Panos P., Dimitris G. Chachlakis, and Evangelos E. Papalexakis. "The Exact Solution to Rank-1 L1-norm TUCKER2 Decomposition." *IEEE Signal Processing Letters* (2018).

[2] Candès, Emmanuel J., et al. "Robust principal component analysis?." *Journal of the ACM (JACM)* 58.3 (2011): 11.

[3] Gu, Shuhang, et al. "Weighted nuclear norm minimization and its applications to low level vision." *International Journal of Computer Vision* 121.2 (2017): 183-208.

[4] Culibrk, Dubravko, et al. "Neural network approach to background modeling for video object segmentation." *IEEE Transactions on Neural Networks* 18.6 (2007): 1614-1627.

[5] Pang, Yanwei, Xuelong Li, and Yuan Yuan. "Robust tensor analysis with L1-norm." *IEEE Transactions on Circuits and Systems for Video Technology* 20.2 (2010): 172-178.

[6] Wright, John, et al. "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization." *Advances in neural information processing systems*. 2009.

[7] Stauffer, Chris, and W. Eric L. Grimson. "Adaptive background mixture models for real-time tracking." *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*. Vol. 2. IEEE, 1999.

[8] Zivkovic, Zoran. "Improved adaptive Gaussian mixture model for background subtraction." *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on.* Vol. 2. IEEE, 2004.

[9] Huang, Junzhou, Xiaolei Huang, and Dimitris Metaxas. "Learning with dynamic group sparsity." *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009.

[10] Kolda, Tamara G., and Brett W. Bader. "Tensor decompositions and applications." *SIAM review* 51.3 (2009): 455-500.

[11] Shu, Xianbiao, Fatih Porikli, and Narendra Ahuja. "Robust orthonormal subspace learning: Efficient recovery of corrupted low-rank matrices." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.

[12] Zhou, Xiaowei, et al. "Low-rank modeling and its applications in image analysis." *ACM Computing Surveys (CSUR)* 47.2 (2015): 36.

[13] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P.Ishwar, CDnet 2014L An Expanded Change Detection Benchmark Dataset, in Proc. *IEEE Workshop on Change Detection (CDW-2014) at CVPR-2014*, pp. 389-394.2014

[14] Lin, Zhouchen, Minming Chen, and Yi Ma. "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices." *arXiv preprint arXiv:1009.5055* (2010).

[15] Tucker, Ledyard R. "Some mathematical notes on three-mode factor analysis." *Psychometrika* 31.3 (1966): 279-311.

[16] Van Loan, Charles F. "Matrix computations (Johns Hopkins studies in mathematical sciences)." (1996).

[17] Sobral, Andrews, et al. "Online stochastic tensor decomposition for background subtraction in multispectral video sequences." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2015.