Journal of Digital Contents Society Vol. 17 No. 6 Dec. 2016(pp. 523-529)
https://doi.org/10.9728/dcs.2016.17.6.523

# Stereo Matching Algorithm Based on Fast Guided Image Filtering for 3-Dimensional Video Service

**Gwang-Soo Hong*, Byung-Gyu Kim****

## Abstract

Stereo matching algorithm is an essential part in computer vision and photography. Accuracy and computational complexity are challenges of stereo matching algorithm. Much research has been devoted to stereo matching based on cost volume filtering of matching costs. Local stereo matching based guided image filtering (GIF) has a computational complexity of $O(N)$, but is still not enough to provide real-time 3-dimensional (3-D) video services. The proposed algorithm concentrates reduction of computational complexity using the concept of fast guided image filter, which increase the speed up to $O(N/s^2)$ with a sub-sampling ratio $s$. Experimental results indicated that the proposed algorithm achieves effective local stereo matching as well as a fast execution time for 3-D video service.

Keywords : Stereo Matching Algorithm, Guided Image Filtering, Fast Guided Image Filtering, 3-D Video Service

# 3차원 비디오 서비스를 위한 고속 유도 영상 필터링 기반 스테레오 매칭 알고리즘

**홍광수*, 김병규****

## 요 약

스테레오 매칭 알고리즘은 컴퓨터 비전과 사진촬영에 필수적인 알고리즘으로 정확도와 복잡도는 스테레오 매칭 알고리즘의 주요 문제점이었다. 그 중에서도 복잡도가 높은 문제점을 극복하기 위해 비용 볼륨 필터링을 기반한 스테레오 매칭 알고리즘에 대한 많은 연구가 이루어졌다. 지역 스테레오 매칭 기술인 유도 영상 필터링 기술은 $O(N)$의 복잡도를 가지고 있지만, 여전히 실시간 3D 비디오 서비스를 제공하기에는 계산량이 많은 편이다. 따라서 본 논문에서 고속 유도 영상 필터링에 기반한 스테레오 매칭 알고리즘을 제안한다. 고속 유도 필터링은 서브샘플 비율 s에 따라 복잡도 $O(N/s^2)$을 가지는 알고리즘이다. 제안하는 알고리즘은 효과적인 스테레오 알고리즘을 성능을 보여줌과 동시에 3D 서비스를 위한 빠른 실행 시간을 보여준다.

키워드 : 스테레오 매칭 알고리즘, 유도 영상 필터링, 고속 유도 영상 필터링, 3차원 비디오 서비스

## 1. Introduction

Dense stereo matching algorithms are an essential area in computer vision and photography, and it plays a significant role in many 3-D applications. It is a popular technique for building a Depth of a scene observed from two slightly different viewpoints. Depth information can be obtained through disparity based on finding

※ Corresponding Author : Byung-Gyu Kim
Received : December 08, 2016
Revised : December 26, 2016
Accepted : Decembe 30, 2016
* Big data Utilization Research Center, Sookmyung Women's University.
email: gs.hong@vicl.sookmyung.ac.kr
** Dept. of IT engineering, Sookmyung Women's University.
Tel: +82-2-2077-7293 , Fax: +82-2-2077-7293
email: bg.kim@sm.ac.kr

correspondent pixels between a left image and a right image. Disparity $d$ is the difference in location of the image of the object between the reference and target images as left and right image, respectively. This process is called a stereo matching, which is one of the most active research topics in computer vision. However, the captured stereo pair are disturbed by noise, e.g. sensor noise, light variations, perspective distortion and uniform regions. Many approaches for development of stereo matching algorithms have been used to improve accuracy and fast execution.

A stereo matching algorithm can be classified into the two main groups of global and local approaches [1]. Global approaches, which achieve minimization of an energy function, aim to determine disparity values for all image pixels at once and usually return more accurate disparity values at the cost of a high degree of computational complexity.

Local approaches use intensity values located in a close neighborhood of pixels with use of a support window, leading to the smoothness assumption that all pixels within a support window have the same disparity. This leads to poor performance when the window contains disparity discontinuities.

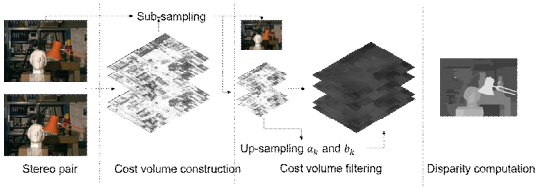
F. Tombari et al. [2] explicitly deployed a smoothness constraint within objects according to the segments. N. Y. Kwak proposed an object-based stereo matching method using segmented region[3]. Hosni et al. [4] proposed to compute the weights by the geodesic distance, which apply the foreground connectivity with low weights. Yang [5] proposed a non-local approach, in which the cost values are aggregated adaptively on a minimum spanning tree.

In recent years, cost aggregation is conducted by filtering on the cost-volume. K. J. Yoon et al. [6] used the bilateral filter weight inside a window. This approach provides weights for object boundaries that preserve edge characteristics. But the brute force implementations are of high computational complexity when the kernel window is large. Many accelerated bilateral weight aggregation methods have been proposed [7] using a sliding window technique. In this case, the aggregation process is independent of the support window size. A Honsi et al. [8] used this approach to increase the speed of a bilateral weight aggregation algorithm. C. Rhemann et al. [9] subsequently proposed local stereo matching as an efficient edge-aware local cost aggregation method using guided image filtering (GIF) [10] with time complexity $O(N)$.

Herein, a local stereo matching algorithm based on fast GIF [11] for improvement of the aggregation step with reduction of the time complexity from $O(N)$ to $O(N/s^2)$ is proposed. The contribution of this research is using a reduction of image dimensions for computation of linear coefficients at every disparity.

## 2. Proposed Algorithm

An overview of the proposed fast cost-volume filtering based approach is shown in (Figure 1). First, the cost volume is constructed based on pixel-wise matching cost computation functions.




(Figure 1) Overview of the proposed fast cost volume filtering approach

To obtain an initial cost volume, the matching cost function that combines truncated absolute differences (TAD) of the color and gradient at the matching pixel is

used. Then, each slice of the cost volume is independently filtered using the proposed method based on fast guided image filtering. Finally, the disparity of any pixel is simply chosen in a winner-take-all manner.

## 2.1 Cost Aggregation with Guided Filtering

Cost volume filtering approaches usually use a pixel-to-pixel comparison, which is sensitive to noise. Hence, each slice of cost volume $C$ has to be filtered. A guided image filter, designed to preserve edges, has linear time that is independent of the filter size and, thus, only depends on the number of image pixels.

GIF uses the guidance color image $G$ as the left image. Let $C'$ be the filtered cost volume. The basic idea of GIF is a local linear model between $G$ and $C'$ for each square window $\omega_k$ with a size of $k$ where the $k$-dependent filtered cost volume $C'_k$ is defined to be a linear model as:

$$C'_k(p) = a_k \cdot G(p) + b_k, \forall\ p \in \omega_k, \tag{1}$$

where $a_k$ and $b_k$ are a $3\times1$ coefficient vector and scalar, respectively. For determination of the linear coefficients $a_k$ and $b_k$, the cost function is defined to minimize the error between $C'$ and $G$ as:

$$E(a_k, b_k) = \sum_{p \in \omega_k} (((a_k \cdot G(p) + b_k) - C(p))^2 + \epsilon \cdot a_k^2), \tag{2}$$

where $\epsilon$ is a regularization parameter to prevent $a_k$ from being too large, and the criterion of flat and textured areas is controlled. The linear coefficients $a_k$ and $b_k$ are obtained when the cost function is solved as a Euler equation, as:

$$b_k = \overline{C_k} - a_k^T \cdot \begin{pmatrix} \mu_k^R \\ \mu_k^G \\ \mu_k^B \end{pmatrix}, \tag{4}$$

where $\mu^R$, $\mu^G$, and $\mu^B$ are mean images of

$$a_k = (\Sigma_k + \epsilon \cdot I_3)^{-1} \begin{pmatrix} \frac{1}{|\omega_k|}\Sigma_{p\in\omega_k} G^R(p)C(p) - \mu_k^R \overline{C_k} \\ \frac{1}{|\omega_k|}\Sigma_{p\in\omega_k} G^G(p)C(p) - \mu_k^G \overline{C_k} \\ \frac{1}{|\omega_k|}\Sigma_{p\in\omega_k} G^B(p)C(p) - \mu_k^B \overline{C_k} \end{pmatrix}, \tag{3}$$

each color channel of the guidance color image $G$. $\Sigma_k$ is the covariance matrix of the guidance color image $G$ in the window $\omega_k$, and $I_k$ is a $3\times3$ identity matrix.

The filtered cost volume $C'$ using the linear coefficients is defined as:

$$\begin{aligned} C'(p) &= \frac{1}{|\omega_k|}\Sigma_{p\in\omega_k}(a_k \cdot G(p) + b_k) \\ &= \overline{a_p} \cdot G(p) + \overline{b_p}, \end{aligned} \tag{5}$$

where $\overline{a_p}$ and $\overline{b_p}$ are the average of the linear coefficients $a_k$ and $b_k$ in the window $k$.

Algorithm 1 shows the overall procedure for cost volume filtering using GIF.

**Algorithm 1** Computation of the filtered cost volume.

**Procedure** CostVolumeFiltering $(G, C, \epsilon, k)$ as a guidance image $G$, cost volume $C$, window size $k$, and regularization parameter $\epsilon$
  **for** $d \in [d_{min}, d_{max}]$ **do**
    compute $\mu_k, \Sigma_{p\in\omega_k} G_s(p)C_s(p)$ at each channel, $\overline{C_k}$ and $\Sigma_k$
    compute $a_k$ and $b_k$
    compute $\overline{a}_k$ and $\overline{b}_k$
    compute $C'(p)$
  **end for**
**end procedure**

## 2.2 Cost Aggregation with Fast Guided Filtering

Disparity values obtained using GIF can provide good visual quality as well as a computational complexity of $O(N)$. The proposed cost volume filtering algorithm reduces computational complexity based on fast GIF from $O(N)$ to $O(N/s^2)$ for a sub-sampling ratio $s$. The algorithm uses linear coefficients $a_k$ and $b_k$ that are computed based on a sub-sampled slice of the cost

volume $C_s$ and a sub-sampled guidance color image $G_s$. Then, the linear coefficients are rewritten as:

$$a_{k/s} = (\Sigma_{k/s} + \epsilon \cdot I_{3\times 3})^{-1} \cdot \begin{pmatrix} \frac{1}{|\omega_{k/s}|}\Sigma_{p\in\omega_{k/s}} G^R(p)C(p) - \mu^R_{k/s}\overline{C}_{k/s} \\ \frac{1}{|\omega_{k/s}|}\Sigma_{p\in\omega_{k/s}} G^G(p)C(p) - \mu^G_{k/s}\overline{C}_{k/s} \\ \frac{1}{|\omega_{k/s}|}\Sigma_{p\in\omega_{k/s}} G^B(p)C(p) - \mu^B_{k/s}\overline{C}_{k/s} \end{pmatrix}, \quad (6)$$

$$b_{k/s} = \overline{C_{k/s}} - a^T_{k/s} \cdot \begin{pmatrix} \mu^R_{k/s} \\ \mu^G_{k/s} \\ \mu^B_{k/s} \end{pmatrix}, \quad (7)$$

where $k/s$ denotes the window size for sub-sampling. To obtain a filtered cost volume $C'$, sub-sampled linear coefficients are averaged. Up-sampled linear coefficients are calculated as:

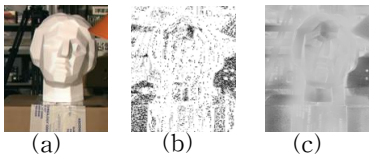
$$\bar{a}_p = F_{up-sampling}(\bar{a}_{k/s}, s), \quad (8)$$

$$\bar{b}_p = F_{up-sampling}(\bar{b}_{k/s}, s), \quad (9)$$

where $F_{up-sampling}$ is a function for obtaining up-sampling data.

The filtered cost volume $C'$ is computed using equation (5), and $\bar{a}_p$ and $\bar{b}_p$ are calculated based on up-sampled average values of linear coefficients obtained using equations (8) and (9).

Examples of cost volume slices filtered using both fast GIF and conventional GIF at disparity $d$ are shown in (Figure 2).



(a) (b) (c)

(Figure 2) A filtered cost volume slice (d= 10): (a) image patch, (b) cost volume slice, (c) cost volume slice based on fast GIF (s =2)

Cost volume slices constructed based on matching cost functions contain considerable amounts of noise. For filtering, the proposed cost volume filtering approach preserves edges, noise is removed, and halo artifacts are avoided.

**Algorithm 2** Fast computation of the filtered cost volume.

**Procedure** FastCostVolumeFiltering $(G, C, \epsilon, k, s)$ as a guidance image $G$, cost volume $C$, window size $k$, regularization parameter $\epsilon$, and sub-sampling ration $s$

compute $G_s = F_{sub-sampling}(G, s)$

**for** $d \in [d_{\min}, d_{\max}]$ **do**

compute $C_s = F_{sub-sampling}(C, s)$

compute $\mu_{k/s}, \Sigma_{p\in\omega_{k/s}} G_s(p)C_s(p)$ at each channel, $\overline{C}_{k/s}$ and $\Sigma_{k/s}$

compute $a_{k/s}$ and $b_{k/s}$

compute $\bar{a}_{k/s}$ and $\bar{b}_{k/s}$

compute $\bar{a}_p = F_{up-sampling}(\bar{a}_{k/s}, s)$ and $\bar{b}_p = F_{up-sampling}(\bar{b}_{k/s}, s)$

compute $C'(p)$

**end for**

**end procedure**

Algorithm 2 shows the overall procedure for computation of the filtered cost volume $C'$. For efficient implementation of the proposed algorithm, all summations in Algorithm 2 are computed using the box filter technique proposed by Crow [12].

Finally, after filtering the cost volume, a disparity map is obtained based on determination of the disparity $d$ of each pixel $p$ between a reference image and a target image using winner-take-all optimization [1] as:

$$d = \underset{k\in[d_{\min}, d_{\max}]}{\arg\min} C'(k) \quad (10)$$

where $d_{\min}$ and $d_{\max}$ are minimum and maximum disparity values, respectively.

The raw disparity map usually contains many outliers, especially near depth continuities and in occlusion regions. To

remove outliers, we use a weighted median filter [13], which may lead to artifact and removal of thin structures

| Algorithm | Tsukuba | | | Venus | | | Cones | | | Teddy | | | Avg. error |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | non occ | all | disc | non occ | all | disc | non occ | all | disc | non occ | all | disc | |
| CostAggr [9] | 0.24 | 2.77 | 8.4 | 1.58 | 2.78 | 16.2 | 3.06 | 11.8 | 8.5 | 7.82 | 16.4 | 18.3 | 8.30 |
| Proposed Alg. | 2.15 | 2.87 | 7.9 | 2.12 | 2.89 | 15.8 | 2.95 | 11.9 | 8.7 | 7.89 | 16.1 | 19.2 | 8.37 |

<Table 1> Quantitative comparison of CostAggr [9] based on GIF and the proposed algorithm based on fast GIF with different sub-sampling ratios ($s=2$)

## 3. Experimental Results

The proposed algorithm was verified using the Middlebury benchmark [14], which provides a collection of stereo pairs for development of stereo matching algorithms. The four standard data sets used were Tsukuba, Venus, Cones, and Teddy. The Middlebury benchmark defines three measures for evaluation of performance, including non-occluded (nonocc), all, and depth discontinuity (disc) regions. The parameters $\{k, T_c, T_g, \epsilon, \alpha\}=\{15,7,2,0.001,0.1\}$ were used and kept constant for all datasets.

Quantitative performance of the proposed algorithm with different sub-sampling ratios and CostAggr [9] values is shown in <Table 1> with percentages of "bad pixels" in measurements (nonocc, all, and disc), which is defined as:

$$B_{nonocc}=\frac{1}{N}\sum_{S\in nonocc}(|d(s)-d_\tau(s)|>\delta_d), \quad (11)$$

$$B_{all}=\frac{1}{N}\sum_{S\in all}(|d(s)-d_\tau(s)|>\delta_d), \quad (12)$$

$$B_{disc}=\frac{1}{N}\sum_{S\in disc}(|d(s)-d_\tau(s)|>\delta_d), \quad (13)$$

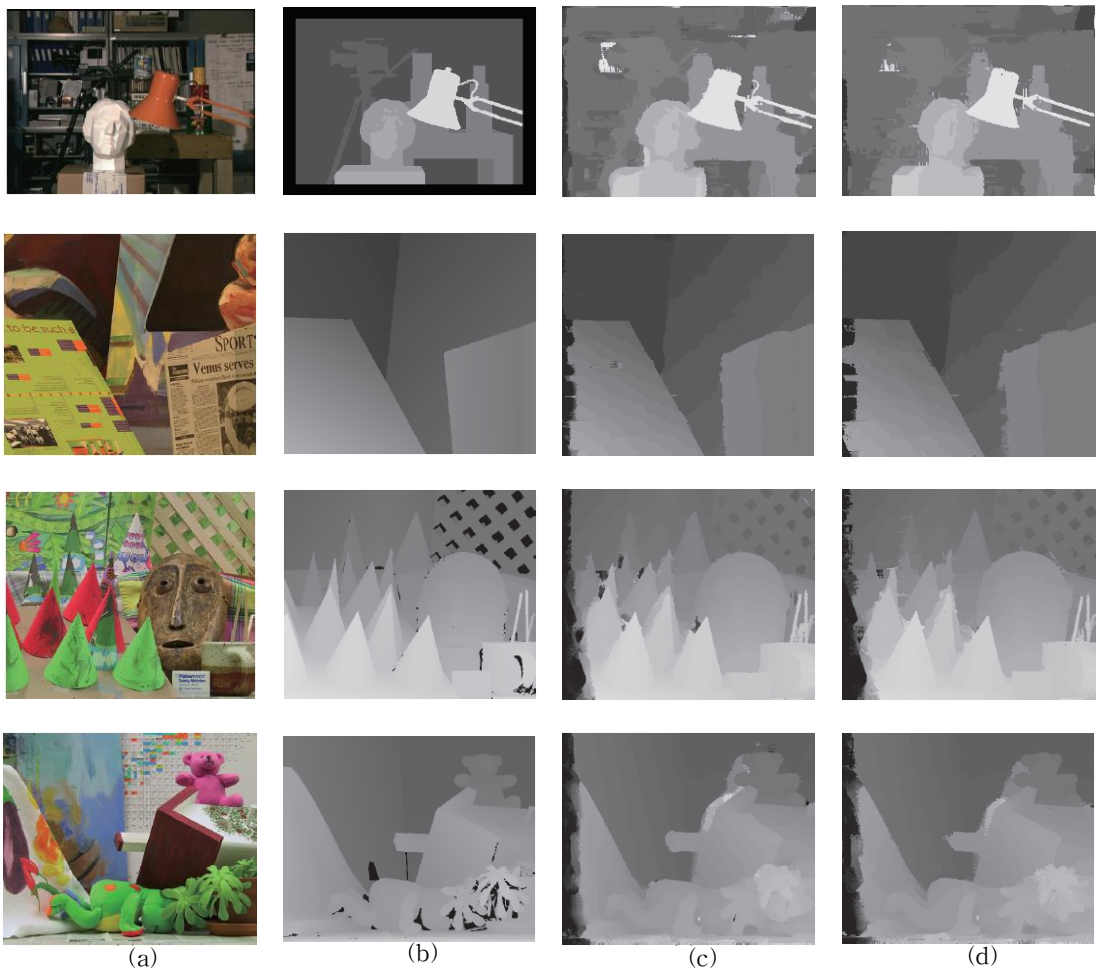
where $B_{nonocc}$, $B_{nonocc}$ and $B_{nonocc}$ are percentages of bad pixels, which is determined ground-truth image with disparity tolerance $\delta_d$.

<Table 1> shows the difference of average result about 0.07% compared with CostAggr [9] which was implemented ourself. We obtained similar quantitative results since the sub-sampling and up-sampling effects corrected the loss of disparity values by removing noise. The algorithm with a sub-sampling ratio of $s=2$ achieved quantitative performance similar to CostAggr [9].

Only the execution time of the aggregation performance for the left image was measured using a PC with an Intel i7 Core at 3.4 GHz. The proposed method was approximately $2\times$ than CostAggr [9] with sub-sampling ratios of $s=2$ in <Table 2>. Estimated disparity maps are shown for visual quality comparison in (Figure 3). Use of fast GIF with a sub-sampling ratio of $s=2$ achieved visual performance similar to CostAggr [9] because the sub-sampled matching cost for computation of linear coefficients removed noise.

| Image | Resolution | Disparity range | Time (Sec.) | |
|---|---|---|---|---|
| | | | Cost Aggr | Our Alg. |
| Tsukuba | 384×288 | 15 | 0.55 | 0.27 |
| Venus | 434×383 | 20 | 1.06 | 0.49 |
| Cones | 450×375 | 60 | 3.34 | 1.57 |
| Teddy | 450×375 | 60 | 3.33 | 1.55 |

<Table 2> Computational efficiency of the proposed algorithm with CostAggr [9] in the aggregation step.




(Figure 3) Qualitative results for disparity images: (a) stereo images, (b) ground-truth images, CostAggr [8], (d) proposed algorithm (s =2)

## 4. Conclusions

In this paper, a fast local stereo matching algorithm using fast GIF has been proposed that focuses on reduction of computation complexity for 3D video service. Experimental results showed the performance similar to CostAggr[9] with faster execution time (over 2 times) using the Middlebury benchmark images. The proposed algorithm achieved the reduced complexity of up to $O(N/s^2)$ with a sub-sampling ratio $s$, theoretically.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2016R1D1A1B04934750).

## References

[1] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” International Journal of Computer Vision, vol. 37, no. 1-3, pp. 7-42, April, 2002.

[2] F. etTombari, S. Mattoccia and L. Di Stefano, “Segmentation-based adaptive support for accurate stereo correspondence,” Proceedings of Advances in Image and Video Technology, pp. 427 - 438, 2007.

[3] N. Y. Kwak, “An object-based stereo matching method using block-based segmentation,” Journal of Digital Contents Society, vol. 5, no. 4, pp. 257-263, 2004.

[4] A. Hosni, M. Bleyer, M. Gelautz and C. Rhemann, “Local stereo matching using geodesic support weights,” Proceedings of IEEE International Conference on Image Processing, pp. 2093 - 2096, 2009.

[5] Q. Yang, “A non-local cost aggregation method for stereo matching,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1402 - 1409, 2008.

[6] K. J. Yoon and I. s. Kwoen, “Adaptive support-weight approach for correspondence search,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 4, pp. 650-656, April, 2006.

[7] M. Gerrits and P. Bekaert, “Local stereo matching with segmentation-based outlier rejection,” The 3rd Canadian Conference on Computer and Robot Vision, pp. 66-72, 2006.

[8] A. Hosni, M. Bleyer, and Margrit Gelautz, “Near realtime stereo with adaptive support weight approaches,” International Symposium on 3D Data Processing, Visualization and Transmission, pp. 1-8, 2010.

[9] C. Rhemann, A. Hosni, M. Bleyer, C. Rother and M. Gelautz, “Fast cost-volume filtering for visual correspondence and beyond,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 3017-3024, June, 2011.

[10] K. He, J. Sun and X. Tang, “Guided image filtering,” 11th European Conference on Computer Vision, pp. 1-14, 2010.

[11] K. he and J. sun, “fast guided filter,” arXiv: 1505.00996v1 [cs.CV] 5 May 2015.

[12] F.C. Crow, “Summed-area tables for texture mapping,” ACM SIGGRAPH Computer Graphics, pp. 207-212, 1984.

[13] Z. Ma, K. He, Y. Wei, J. Sun and E. Wu, “Constant time weighted median fitering for stereo matching and beyond,” IEEE International Conference on Computer Vision, pp. 49-56, December, 2013.

[14] Middlebury Stereo Vision Page, http://vision.middlebury.edu/stereo/

**홍 광 수**

2012년 : 선문대학교 컴퓨터공학과 석사
2013년~ 현재 : 선문대학교 컴퓨터공학과 박사과정

2016년～현 재: 숙명여자대학교 빅데이터활용연구센터 연구원
관심분야 : 스테레오 매칭 (Stereo Matching), 3D 비디오 코딩 (3D Video Coding), 딥러닝 (Deep learning) 등

**김 병 규**

1998년 : 한국과학기술원 전기및전자공학 석사
2004년 : 한국과학기술원 전기및 전자공학 박사

2004년～2008년: 한국전자통신연구원
2009년～2015년: 선문대학교 컴퓨터공학과 교수
2016년～현 재: 숙명여자대학교 IT공학과 교수
관심분야 : 비디오 신호 처리 (Video signal processing), 딥러닝 (Deep learning) 등