# 다중증거기반 확률적 실시간 사람 추적

김현우*

요약

본 논문은 스테레오 카메라를 통해 입력되는 위에서 내려다보는 영상 정보를 해석하여 실시간으로 사람을 추적하는 알고리즘을 제안한다. 위에서 내려다보는 영상에서의 추적은 물체간의 가려짐 현상을 최소화하여 효과적이나, 시각적인 정보가 약하고 스테레오 카메라를 통한 거리 측정의 정확도가 떨어진다. 본 논문에서는 이를 극복하기 위해서 확률을 기반으로 한 2단계 알고리즘을 제안하고자 한다. 1단계에서 거리측정 정보를 사용하여 결정론적 서치를 통해 추적하고자 하는 물체의 대략적인 위치 및 스케일을 빠르게 찾고, 다음 단계에서는 확률적인 방식에서 2차원 상의 파라미터를 정밀하게 찾아낸다. 마지막으로, 제안된 알고리즘이 사람출입시스템에 실제 적용된 실험 결과를 통해 알고리즘의 속도와 정확도를 보여준다.

# Real-time Tracking of Multiple Persons using Multi-modal Information

Hyun-Woo Kim*

ABSTRACT

This paper presents a real-time tracking method of multiple persons using stereo from a top-down view. Tracking in a top-down view is effective to overcome occlusion problem among multiple persons, but visual cues are weak and depth estimation isn't accurate enough especially when wide-view lenses are used to cover large field of view. We approach the problem with a two-step probabilistic method: deterministic and stochastic searches. The deterministic search reduces the candidate locations/scales of the tracking objects by maximizing a 1-D rotation/scale-invariant depth similarity. Then, a customized particle filter is followed to finely estimate the locations/scales of the object by a 2-D search in a stochastic manner. The depth similarities from the previous deterministic search are incorporated into the stochastic particle filter as an importance function. Finally, the proposed algorithm has been tested on various situations and evaluated quantitatively in a person counting scenario.

Key Words : person tracking, visual surveillance, particle filter, data fusion, probabilistic tracking

---

* 한독미디어대학원대학교(✉hwkim@kgit.ac.kr)
· 제1저자(First Author) : 김현우 · 교신저자(Correspondent Author) : 김현우
· 접수일(2010년 6월 9일), 수정일(1차 : 2010년 7월 7일), 게재확정일(2010년 7월 12일)

## I. Introduction

Video analysis of human activity has been studied forseveral decades in computer vision society with manyapplications, such as video surveillance, content based video service, virtual reality, customer relationship management, biometrics, and intelligent interface. Recently, social demands (e.g., aging population and personal security) and emerging new computing environments (e.g., smart home and ubiquitous computing) have been accelerating its related researches and developments.

Visual tracking of multiple persons is a fundamental component of human analysis. It provides moving trajectories of multiple persons and their body parts as well, and is used as a key input for human activity analysis.

To track persons, a lot of different approaches in different camera configurations have been developed. Recently, the probabilistic tracking approaches have been spotlighted because of its effectiveness to fuse multiple observations in its probabilistic framework. The methods can be classified into the deterministic searching and the stochastic searching methods.

The deterministic searching methods are known to be fast to track objects, while it can be applied when the motion can be modeled by a Gaussian. Comaniciu and Meer [1] presented a color-based tracking method, called the "Mean Shift" algorithm. Tracked objects are modeled as a color probability distribution, and in the tracking stage the position and scale of the object were estimated by searching candidate regions with a metric derived form the Bhattacharyya coefficient. Kang et al. [2] modeled motion and appearance of moving objects separately using two probabilistic models. The tracking was performed by the maximization of a joint probability model.

The stochastic searching methods extend the motion model into non-Gaussian motion in the presence of complex background clutters. Recently, the particle filer has been developed and used by a lot of researchers. Isard and Blake [3] introduced the Condensation (Conditional density propagation) algorithm in computer vision society as the transfer of the particle filter. Owing to the probability approximation using particles, the particle filter can handle any motion model. It was found to be very robust in the presence of background clutters.

Nait-Charif and McKenna [4] proposed the ILW (Iterative Likelihood Weighting) scheme, achieving accurate tracking even when the modeling of motion dynamics is poor. In the PETS-ICVS 2003 workshop, they presented outstanding results on the same video data set given by the committee. However, the stochastic approaches using particle filter can be slow without clever handling of the particle numbers and modeling customized to the individual problems.

Recently, Tarek Yahiaoui et al. [5] developed a people counting system based on dense and close stereovision using a novel stereo matching algorithm. This has been done in laboratories, but our work focus on industrial applications by overcoming challenging environmental variations.

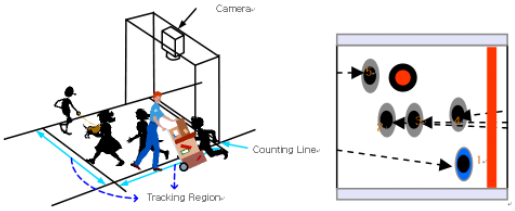



그림 1. 사람 계수 시스템 개요
Fig. 1 Visual tracking from a top-down view.

In this paper, we combine the advantages of both approaches without forcing the Gaussian modeling of motion. A 1-D deterministic search dramatically reduces candidate regions with a confidence measure, and a simple and accurate 2-D stochastic search is followed. Specially, we focus on the real-time visual tracking of multiple persons from a top-down view, and apply it to count the number of enter/exit persons in doorway. The typical situation is depicted in Figure 1. The left and the right images show a camera setting and a view from the camera, respectively.

## II. Review of Particle Filter

Suppose that $\mathbf{x}_k$ and $\mathbf{D}_k$ denote the state vector and the measurement vector at the discrete time $k$ respectively. According to the Bayesian theory and probability propagation theory, a tracking framework can be formulated by

$$p(\mathbf{x}_0) = p_0 \quad (1)$$

$$p(\mathbf{x}_k \mid \mathbf{D}_{k-1}) = \int p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) p(\mathbf{x}_{k-1} \mid \mathbf{D}_{k-1}) d\mathbf{x}_{k-1} \quad (2)$$

$$p(\mathbf{x}_k \mid \mathbf{D}_k) \propto p(\mathbf{y}_k \mid \mathbf{x}_k) p(\mathbf{x}_k \mid \mathbf{D}_{k-1}) \quad (3)$$

Here, Equations (1), (2), and (3) correspond to initialization, propagation (prediction), and update stages in visual tracking, respectively. Since generally the motion probabilities cannot be represented by parametric, deterministic models, they are implemented by particles (samples) using Monte-Carlo simulation [6]. This technique is called the "particle filter,"because the probability is represented by particles instead of a set of parameters.

## III. Probabilistic Tracking

We adopt the particle filter to implement our tracking algorithm but several parts are modified for a fast and accurate tracking. To do that many things should be specified, e.g., state/measurement model, state transition model, the number of samples, etc. We incorporate a rotation-invariant visual cue into the framework for a fast deterministic search providing self-adaptation of the sample number.

Our algorithm uses a two-step-search approach. The first step is a deterministic search using a rotation-invariant depth cue, and a stochastic search using a particle filter follows. This approach accelerates tracking speed without resorting to local minima. The search steps are implemented in the propagation (prediction) and update stages of the particle filter, respectively. Including initialization, the proposed algorithm consists of three stages, and the overall flow is as shown in Figure 2.

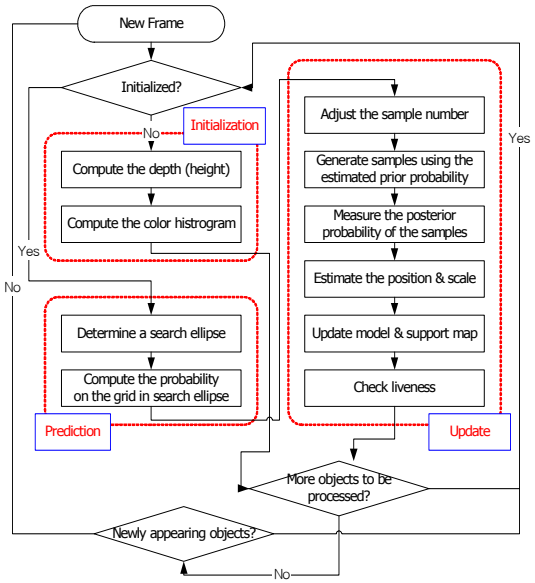



그림 2. 시스템 흐름도

Fig. 2 The overall flow.

### 3.1 Initialization Stage

We assume persons are detected each frame. The detection is done using stereo-based segmentation [7] and is tuned to segment the head-shoulder parts of persons because it is the most invariant part among the frames from a top-down view.

Among the detected persons, non-overlapping persons with the currently tracked ones are determined as newly appearing persons, and they are initialized. In this stage, the 1-D depth map and color histogram of the detected persons are stored, denoted by $\mathbf{y}_0^{depth}$ and $\mathbf{y}_0^{color}$ in a vector form, respectively. The 1-D depth map is computed by the average of depths. First, the detected region is normalized by a square with the length of the average of the width and height of the rectangular region, and the depths in the same distance are averaged. We consider the 1-D depth map as a depth feature of the detected person. When the persons are segmented, it can be considered to be invariant under rotation change and insensitive to scale as well. Practically, even non-symmetric segmentation worked reasonably.

As an initial state, the center position and the scale is stored, denoted by $\mathbf{x}_0 = \{x_0, y_0, 1.0\}$ where $x_0$ and $y_0$ means the x and y position of the center and scale is chosen as 1.0 (a reference) with the width and height of the initial region. Assuming uniform distribution, in Equation (1), $p_0$ set to be $1/N_0$, where $N_0$ is the number of initial particles specified by users.

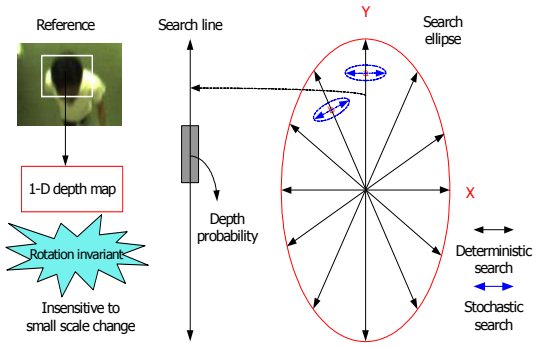



그림 3. 결정론적 서치 및 깊이 확률

Fig. 3 Deterministic search and depth probability.

### 3.2. Propagation Stage

In this stage, the particles are propagated to predict the location/scale of the tracked object. Generally, the propagation is modeled according to the known/learned motion dynamics of the object. In noisy environments, however, the previous observation is not accurate enoughto predict the next

locations. Moreover, complex motion dynamics including crossing of objects and agile motion is not easy to model. Some researchers have been tried to learn the dynamic model using examples, but it is not easy and time-consuming job to learn all possible situations.

We propose a deterministic search using a semi-rotation/scale-invariant feature. Instead of assuming an unrealistic, simple dynamic model, the model is substituted by a depth probability computed by 1-D search. Because it uses the current measure, it may be included into a part of the update stage. However, it can be considered as a propagation (prediction) stage because it models the motion dynamics and is followed by random propagation. Note that this is very similar to importance sampling in ICONDENSATION [8]. Therefore, Equation (2) is approximated by

$$p(\mathbf{x}_k \mid \mathbf{D}_{k-1}) \approx \int p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) p(\mathbf{x}_{k-1} \mid \mathbf{D}_{k-1}, \widetilde{\mathbf{D}}_k) d\mathbf{x}_{k-1} \quad (4)$$

where $\widetilde{\mathbf{D}}_k$ denotes the partial observation from the current frame.

This search is done in sampled positions on 12 directions in an ellipsoidal neighboring region with a center of the previous estimated position $\mathbf{x}_{k-1}$ ignoring the scale factor. The depth probability is computed by the average of the correlation with a reference depth map, and it is represented by

$$p(\mathbf{x}_{k-1} \mid \mathbf{D}_{k-1}, \widetilde{\mathbf{D}}_k) = \left(\mathbf{y}_k^{depth}\right)^T \mathbf{y}_0^{depth}, \quad (5)$$

where $\mathbf{y}_k^{depth}$ denotes the depth map around the sample position $\mathbf{x}_{k-1}$ in time $k$. Based on the computed depth probability, Equation (5) gives the searching candidates around the position with high depth probability. The method is depicted in Figure 4.

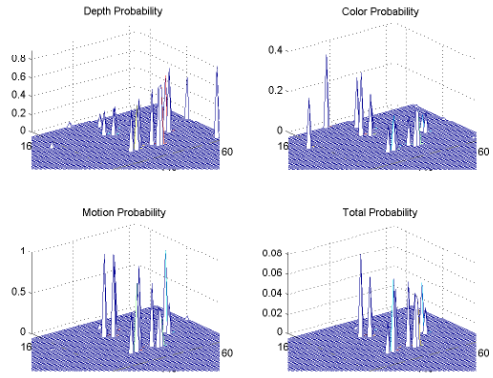



그림 4. 다중 증거 융합
Fig. 4 Multiple cue fusion.

More beauty of this approach is that the number of particles can be automatically adjusted based on the statistics computed from the depth probability. The number of particles is determined by multiplying a specified max number $N$ of particles by the ratio $r$. The ratio is compute by

$$r = \frac{\int (\mathbf{x}_{k-1} - \bar{\mathbf{x}}_{k-1})^T p(\mathbf{x}_{k-1} \mid \mathbf{D}_k)(\mathbf{x}_{k-1} - \bar{\mathbf{x}}_{k-1}) d\mathbf{x}_{k-1}}{\int (\mathbf{x}_{k-1} - \bar{\mathbf{x}}_{k-1})^T (\mathbf{x}_{k-1} - \bar{\mathbf{x}}_{k-1}) d\mathbf{x}_{k-1}} \quad .(6)$$

Physically, it represents the ratio between the position variance under the depth distribution and that under uniform distribution. As the computed distribution has uniform, it becomes 1, while as the

computed distribution has a delta function (ideal measurement), it becomes 0. That is, when the 1-D search gives more localized position result, less particles is needed.

### 3.3. Update Stage

The update stage is straightforward thanks to the previous propagation stage. The main issue is what kinds of visual cues are effective and efficient for tracking from a top-down view. We found depth cue $\mathbf{y}_k^{depth}$ and motion cue $\mathbf{y}_k^{motion}$ are very distinguishable from background and other objects, and color cue $\mathbf{y}_k^{color}$ can be used to distinguish a person from other persons. The total likelihood is the multiplication of those probabilities and it is represented by

$$p(\mathbf{y}_k \mid \mathbf{x}_k) = p(\mathbf{y}_k^{depth} \mid \mathbf{x}_k) \cdot p(\mathbf{y}_k^{color} \mid \mathbf{x}_k) \cdot p(\mathbf{y}_k^{motion} \mid \mathbf{x}_k) \quad (7)$$

The depth probabilities are computed by the average of the distances between the observed depths and the reference depth of the initial region. The motion probabilities are computed by the average of the distances between the observed pixels and the reference pixels of background images, considering pixel variances. For the color probability, color histogram similarity is adapted [9]. To exclude the effect by background and noise, each point is multiplied by the corresponding motion probability. Finally, the position and scale are determined by the average of the total probability (Equation (3)).

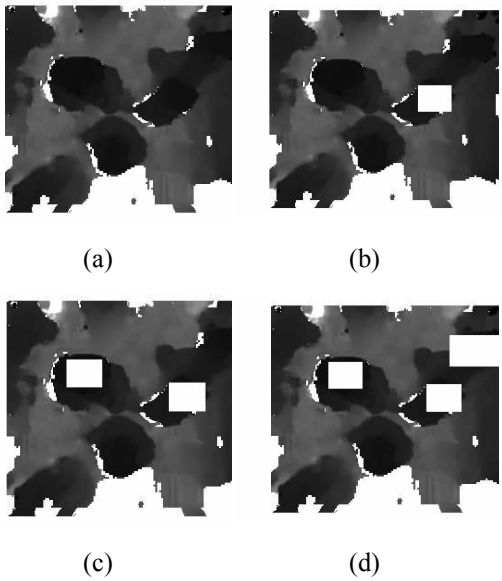
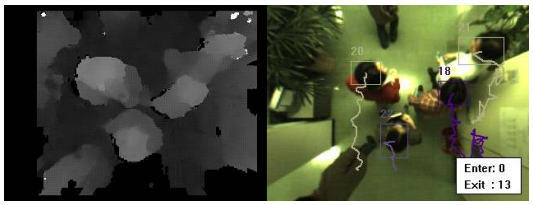



그림 5. 마스크 처리 영역의 예
Fig. 5 An example of support mask management. (a) The disparity image. (b), (c), and (d) The images after the first, second, and last persons are tracked respectively. The corresponding frame is shown in Figure 6(b).

## IV. Multi-Object Tracking

Tracking from a top-down view makes the multi-object tracking problem easy and fast because there is little occlusion by other persons. That is, since there is little possibility where multiple persons occupysame position, the extra problem related to multiple persons is how to prevent multiple persons'occupancy at a time in the same position. It is done by managing a binary mask, which stores the already occupied regions by other persons, and we call the mask a "support mask."First, the map with the same size of the image is set to 1. Then, a person tracked using a single tracker and the estimated position occupied by the person is set to 0 to be

masked. For the next person, the masked region is not considered in likelihood estimation, so more than one person cannot exist in the sample position at the same time. Note that it allows small overlaps between persons to tolerate the estimation error during single object tracking. An example of the support mask overlaid on a depth image is shown in Figure 5. A snapshot of an experiment is shown in Figure 6.




(a) (b)

그림 6. 시스템 동작 예

Fig. 6 A snapshot of an experimental result. (a) The estimated depth using stereo. (b) The tracked trajectories.

## V. Experimental Results

We attached a stereo camera, STH-MDCS-Cfrom VIDERE Design, on the ceiling at the 2.6 m far away from the floor, and 3.5mm lenses give a large field of view, 5m× 4m, with disadvantage of large image distortion. From the system, we captured 140 scenes with 15 min run time for each from a doorway and hallway in our lab, and the scenes are classified into several situations for evaluation purpose. The numbers of collected scenes with respect to the situations are presented in Figure 6. They are classified in two classes, basic and complex situations. The basis situations include normal pass-by persons with various moving velocities and directions, and the complex situations include several complex situations such as pause, U-turn, carrying on objects, etc. To evaluate the performance in real time, although they are processed in offline, our evaluation software simulates online execution by simulating running time. We used a Pentium 4 PC with a 2.7G Hz CPU.

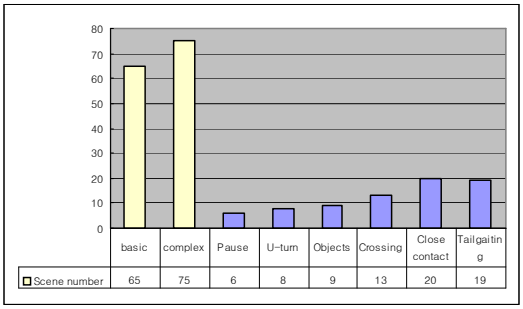



그림 7. 상황별 데이터 분포

Fig. 7 The numbers of the collected scenes with respect to the situations.

The performance of the tracking algorithm is evaluated by the difference between the estimated value and the ground truth value of the remaining person inside room, i.e., the difference between the count of entering persons and that of exiting persons normalized by the total number of entering and exiting count. It is represented by

$$Err = \frac{\sum_i \left| enter_i^{true} - exit_i^{true} - enter_i^{est} + exit_i^{est} \right|}{\sum_i person_i^{true}} \quad (8)$$

where $enter_i^{true}$ and $exit_i^{true}$ denote the ground truth counts of enter and the exit at the scene $i$, respectively, and $enter_i^{est}$ and $exit_i^{est}$ represents the corresponding computed counts. The number of

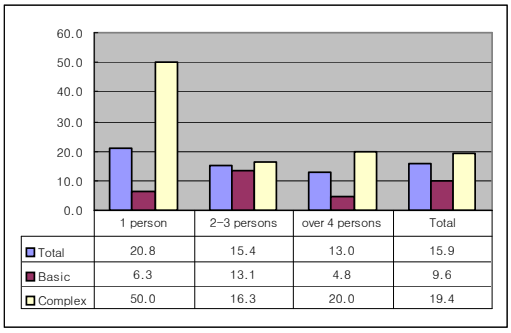
persons $person_i^{true}$ is computed by the max number of the scene $i$, represented by

$$person^{true} = enter^{true} + exit^{true} + pause^{true} + u\_turn^{true} \quad (9)$$

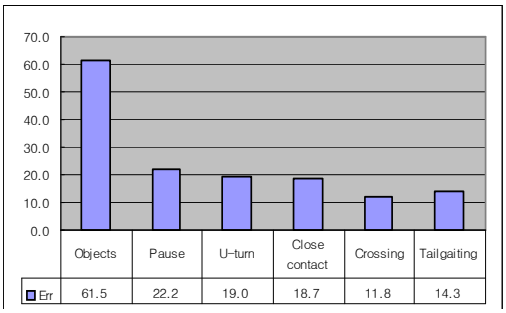
where denotes $pause_i^{true}$ and $u\_turn_i^{est}$ denotes the number of persons with pause and u-turns, respectively.

It took 25 mm seconds to track one person, and3-4 frames per second were handled including detection and stereo processing. The resulting error rates are presented in Figure 7. The error rates for 65 basis situations and 75 complex situations was 9.6% and 19.4%, and that of the total 140 scenes was 15.9 %. These error rates do not exactly correspond to the exact tracking failure rate because the scenes are captured separately each by each, not from a continuous video streams. We may say the tracker is successful for the basic situations and some complex situations. The strange high error rate in complex situation with 1 person is because of the division by the small number of maximum persons. Additionally, many errors came from the extremely poor accuracy in the image boundary because the data set is collected without considering the stereo performance.

To show the performance in complex scenarios, the error rates are described in each situation. Figure 7(b) shows which situation is the most difficult to track: carrying on objects, long pause during walking/running, U-turns, close contact, crossing, and tailing gating in order. We say tracker failed in the hardest situation, carrying on objects. Specifically, carrying on objects with comparable heights with humans failed and it could be excluded by the sophisticated person detector in the future.




(a)




(b)

그림 8. 에러율

Fig. 8. The resulting error rate. (a) The error rate graph with respect to the person number and large-scale classification. (b) The error rate graph for the complex cases.

## IV. Conclusions

We proposed a two-step probabilistic tracking method of multiple objects using the particle filter framework. The 1-D deterministic search dramatically extends the search region and reduces the search regions in the following 2-D stochastic search in a robust way. Although we evaluated the algorithm in the context of person counting from a

top-down view, it can be applied different contexts easily. For example, we can track human faces with fast 1-D search step with skin color feature. Experimental results showed the trajectories from the tracker are robust enough to be used for counting pass-by persons.

## 참고문헌


[1] D. Comaniciu, and P. Meer, "*Mean Shift: A Robust Approach toward Feature Space Analysis,*" IEEE Trans. Pattern Analysis Machine Intelligence., 24(5): 603-619, 2002

[2] Jinman Kang, Isaac Cohen, and Gerard Medioni, "*Continuous Tracking Within and Across Camera Stremas*," CVPR 2003.

[3] M. Isard, and A. Blake, "*Condensation-conditional density propagation for visual tracking*," International Journal of Computer Vision, 29(1): 5-28, 1998.

[4] H. Nait-Charif, and S. J. McKenna, "*Head Tracking and Action Recognition in a Smart Meeting Room*," IEEEInternational Workshop on Performance Evaluation of Tracking and Surveillance, Graz, Austria, 31 March 2003.

[5] Tarek Yahiaoui, Cyril Meurie, Louahdi Khoudour, François Cabestaing: A People Counting System Based on Dense and Close Stereovision. ICISP 2008: 59-66.

[6] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "*A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking*," IEEE Tran. On Signal Processing, 50(2): 174-188, Feb 2002.

[7] David Beymer, "People Counting using Stereo," IEEE Workshop on Human Motion, Austin, TX, December 7-8, 2000.

[8] M. Isard and A. Blake, "*ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework*," ECCV98.

[9] P. Pérez, C. Hue. Vermaak and M. Gangnet, "*Color-based probabilistic tracking,*" European Conference on Computer Vision, Copenhagen, Denmark, June 2002.


## 감사의 글

이 논문은 일부2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행됨(2010-0004359).


**김현우(Hyunwoo Kim)**

1994 한양대학교 전자통신공학과(공학사)
1996 포항공과대학교 전자전기공학과 (공학석사)
2001 포항공과대학교 전자전기공학과 (공학박사)
2001~2007 삼성종합기술원/삼성전자 책임연구원
2007~2008 ㈜한독산학협동단지 책임연구원
2008~2009 성균관대학교 연구부교수
2009~현재 한독미디어대학원대학교(KGIT) 조교수
※ 관심분야 : 컴퓨터 비전, 인지로보틱스,증강현실