Robust and Persistent Visual Tracking-by-Detection for Robotic Vision Systems

Yao Yeboah, Zhuliang Yu, and Wei Wu

***Abstract*—Visual object tracking in robotics applications comes riddled with challenges associated with almost incessant object motion, robot motion or in certain cases, both object and robot motion that may not be necessarily correlated. This problem is further compounded by appearance variations which result from scale and pose variations, rendering the establishment of robust online tracking schemes a challenging task. The paper presents and extension of the CSK tracker via a effectively incorporation of depth features and an improved model update scheme into a single tracking framework. In realizing this framework, feed-forward and feedback strategies are introduced into the CSK tracking scheme allowing for the seamless incorporation of depth features which are extracted on a per frame basis. Additionally, coherency is achieved between the depth and RGB feature spaces via a coupling scheme which is applied towards warping the depth and RGB spaces on a per frame basis. An intermediary stage between object detection and model update is further introduced towards intelligent and adaptive model and classifier parameter update schemes. The contributions of the paper are manifold. Firstly, the paper achieves an efficient incorporation of depth features into the CSK tracker towards classifier robustification; Secondly, intelligent and adaptive classifier and model parameter update strategies are achieved towards robust tracking by means of feed-back and feed-forward strategies; Finally, a coupling scheme allows for warping to be achieved between RBG and depth space thereby facilitating robustness in the tracker without introducing additional computational overhead and significantly trading off classifier speed. Experimental results suggest that the proposed scheme achieves tracking robustness in situations of partial occlusion and this offers a means by which the CSK tracker could be robustified towards feasibility in robotics applications. The scheme also allows the Circulant Tracker to operate at speeds feasible in online robotic applications.**

***Index Terms*—Robust tracking, tracking-by-detection, tracking for robotics, visual tracking.**

I. Introduction

Visual object tracking offers a wide range of practical applications that span across various and diverse fields. Such applications include but are not limited to Human Computer Interaction (HCI) [1], bio-medical imaging [2], [3] and surveillance [4]. In visual tracking, the tracking task could either be limited to domain specific objects such as faces [5], hands [6], humans [7] or even vehicles [8]. These domain specific trackers have achieved some remarkable success and this could partly be attributed to the feasibility of offline target modeling and training schemes towards the online detection and tracking task. However, in real world scenarios including robotic vision systems, the target to be tracked could remain undefined until the tracker has been initialized with the target object parameters (e.g., shape, size, colour and depth). This rules out the feasibility of applying offline training schemes as such an approach would require tremendously large volumes of highly diversified training data, a task that is practically infeasible. This therefore requires the development of trackers with the capability of tracking an arbitrary object from the instance of initialization. In such schemes, once the tracker has been assigned an initialized state in one video frame, the goal of the tracker is then to estimate the location and state of the in object subsequent frames. It quickly becomes clear that attaining such a tracking goal would require an online modeling of the target object, coupled with online model update strategies and ultimately, online training of a classifier towards the tracking of the object in subsequent frames. This has given rise to tracking-by-detection [9] which tackles tracking by treating the tracking task as a detection task spanning across all the frames that constitute the video sequence. Tracking-by-detection strategies have dealt with target representation [10], [11], appearance modeling [12], [13] and motion modeling [13] with some components being merged in certain cases [14]. While tracking-by-detection is a comparatively new paradigm in tracking, some success has been achieved in [14]-[18]. A majority of these tracking-by-detection algorithms treat object detection as a binary problem and hence rely on the application of discriminative classification approaches towards the realization of feasible object detection schemes. In such schemes, each successful detection of the object provides a premise for the extraction of relevant features to facilitate the training of a classifier towards the detection of the object in future instances. Some various machine learning and pattern recognition schemes that have been applied towards tracking-by-detection include but have not been limited to support vector machines (SVM) [19], K-means [20], structured output SVM [15], ranking SVM [21], Bayes Classifier [22], and boosting [23]. Based on the tracking components that a tracking algorithm emphasizes one, the overall tracking performance tends to excel in certain aspects while falling short in other aspects [24]. While some trackers apply learning techniques towards the efficient encoding of background information into the tracking scheme [15], others apply this information towards the development of explicit

Manuscript received February 11, 2016; revised June 1, 2016. This work was supported in part by the Western Transportation Construction Technology Project of the Ministry of Transport, P.R.China under Grant B1110210.

Yao Yeboah is with the Department of Electrical and Computer Engineering, South China University of Technology, Tianhe, Guangzhou, P.R. China (e-mail: mail@ yaoyeboah.com).

Zhuliang Yu and Wei Wu are with the School of Automation Technology, Tianhe, Guangzhou, P.R. China (e-mail: zlyu@scut.edu.cn, auweiwu@scut.edu.cn).

context information [17]. Such schemes have allowed the robustness of tracking to be improved significantly. In order to allow for trackers to adapt to partial changes in the target object, some techniques have relied upon local models [13], [25]. Dense sampling schemes [15]-[17] have also be shown to allow trackers to effectively tackle rapid object motion due to their ability to expand the search domain and hence discriminative models become effective in sorting out clutter that may exist between the object and the background [24]. The impact of these various schemes on overall tracking performance as well as their applicability to domain-specific tracking problems is an area that merits further research.

In the domain of robotic vision, object tracking is increasingly attracting research attention due to its vast application potential. In this domain however, tracking challenges increase manifold due to an almost incessant object motion, robot motion or in some cases, both object and robot motion that may not necessarily be correlated. The problem is further compounded by the appearance variation of target objects, which results from scale and pose variations, rendering the establishments of reliable appearances models a challenging task [26]. Additionally, deployment environment of robotic vision systems could be riddled with illumination variations, camera perturbations and other forms of random interferences that could result in tracking drift or even complete target loss [27]. This establishes the premise for efficient and robust trackers with the capability of providing resolution speeds that match the demands and challenges of the robotics domain. Among the current state-of-the-art tracking-by-detection algorithms, the Circulant Structure Kernel [18] offers the highest speed which has been attributed to its proposed circulant tracking scheme. Although the CSK tracker achieves simple, fast and efficient tracking, it suffers from an inability to recognize and tackle partial occlusions [28]. This problem is further escalated by its naïve approach towards model update and online classifier training. The result of this shortcoming is a degradation of the object model in the conditions of partial occlusion and hence a gradual drift in tracker performance. While most classical and state-of-the-art algorithms operate with 2D data which is in fact a mapping of 3D information into a 2D framework and hence leads to a dimensionality reduction and hence loss of crucial information, the rapid development and reduction in the cost of Time of Flight (ToF) sensing technology has lead to a growth in 3D vision systems. While most of these ToF sensors are limited in range, a significant number of robotic vision systems are confined to indoor environments where the range limitations become insignificant. This paper proposes an extension of the CSK tracking framework through an incorporation scheme with depth information and an improved model updating strategy. Feed forward and feedback strategies are introduced into the CSK framework allowing for the incorporation of depth features which are extracted on a frame bases. In order to establish coherency between depth and RGB features within a single frame, a coupling scheme is applied in warping together the depth space and RGB space features on a frame basis. As the paper demonstrates, the depth information is robust to partial occlusion situations and this offers a means by which the CSK tracker can be reinforced in situations of partial occlusions. Towards the achievement of this goal, the paper proposes an intermediary stage between object detection and model update. This intermediary stage is capable of determining partial object occlusion which allows for much more intelligent and adaptive model and classifier parameter update schemes. Experimental results obtained on various scenarios demonstrate the effectiveness of this algorithm in maintaining object model integrity and alleviating classifier drift in occlusion situations through an adaptive learning scheme. This offers a means by which persistent tracking can be attained without a significant trade-off of speed and hence the algorithm is shown to satisfy the requirements of robotic vision systems. The rest of the paper is thus organized. Section II presents an in-depth literature review of the related work pertaining to the research effort presented here in the paper. This is followed by Section III where the proposed algorithm is presented in both theoretical and mathematical frameworks. In this section the modified circulant tracking scheme with an adaptive online learning strategy along with the depth incorporation mechanism coupled with the feedback strategy towards adaptive tracking are presented. Experimental results achieved with the proposed algorithm as well as results from comparison experiments are presented in the Section IV. The paper concludes in the Section V.

II. Related Work

While some object tracking algorithms have succeeded with offline appearance modeling [29], [9] a considerable number of them have adopted online modeling schemes [14], [30]. This later category of algorithms offer higher feasibility to a much wider range of applications due to their versatility and ability to target arbitrary objects. In order to adapt to the object online, these methods apply on-the-fly appearance modeling and classifier training schemes that result in a trade off between classifier robustness and tracking speed. These tracking-by-detection methods have therefore excelled in achieving efficiency in arbitrary object tracking but suffered from some problems including drift and vulnerabilities associated with abrupt object appearance changes as well as illumination changes, [24], [17]. Some methods have attempted to robustify performance and mitigate drift through the incorporation of context information into the tracking framework [17]. By exploiting so-called distractors and supporters, the method proposed in [17] is able to overcome problems associated with tracking in unconstrained situations. While the method succeeds in overcoming occlusion and abrupt motion by taking advantage of context, fast appearance changes and articulated objects still prove challenging. By stressing on robust model update and the incorporation of a so called forgetting factor, the method in [11] is capable of tackling certain intrinsic (eg. Pose variation, shape deformation etc.) and extrinsic (eg. Illumination change, camera motion, occlusion etc.) appearance variabilities. While the method handles these challenges efficiently, it still suffers from occasional drifts and does not achieve significant tracking speeds. By taking a decomposition approach to tracking, the method proposed in [31] provides a scheme that achieves the efficient design of multiple motion and

observation models into multiple basic trackers. These basic trackers are then combined into a single compound tracker and by leveraging and strengths and weaknesses of the basic trackers, the resulting tracker is capable of achieving efficient performance in unconstrained video sequences. This decomposition scheme has however been associated with a performance bottleneck that has been directly linked with its adoption of sparse principal component analysis [24]. While occlusion has been well tackled by state-of-the-art trackers [15], [16], where dense sample and local sparse representations have been adopted, the area still merits research effort as these same trackers become vulnerable to background clutter leaving them open to degradation when initialized with larger scales [24].

Due to the complexity of the environments where robots are deployed, the need for tracking schemes with the capability to learn and adapt online while remaining robust to background clutter and appearance variations continues to rise. Recent methods have relied on the incorporation of depth information into detection schemes. The method proposed in [32] applies an RGB-D scheme in hand detection and localization towards design and realization of a natural interaction system for robots. This method however relies on a shape matching scheme and therefore limits itself to domain specific objects. An RGB-D fusion scheme is proposed in [33] towards online tracking-by-detection. This work however stresses on fusion rather than tracker design and therefore adopts classical mean shift tracking towards realizing the proposed online RGB-D tracker. More related to the work discussed here in this paper is the method proposed in [34] which achieves arbitrary object tracking using an RGB-D incorporation framework. This framework applies compressive tracking on RGB frames while relying on a variance ratio features shift (VR-V) tracker for depth frames. While this coupling scheme is able to partially overcome occlusions and illumination changes, the parallel executions of two primary trackers introduces an obvious bottleneck in tracking speed which could limit the method only to situations without speed constraints.

III. Proposed Algorithm

Drawing from the tracking efficiency and speed of the CSK tracker, as well as the robustness and sensitivity of depth features to background clutter and occlusion, this paper proposes an RGB-D tracking method with the capability of adapting to illumination changes and target occlusion via an adaptive online model update and classifier parameter adjustment scheme. The method is built upon the CSK tracker and succeeds in overcoming its vulnerability to occlusion and accumulation of faulty detections due to naïve model updates which eventually lead to tracking drift and target loss in certain situations. By subsuming depth information into the tracking framework, trackers can be realized via RGB-D combination schemes which possess the capability to overcome some basic challenges faced with tracking-by-detection algorithms. However, since adopting tracking schemes that are originally designed for RGB sequences and transferring them to depth sequences is a practically infeasible task, there is the need for the design of feature extraction schemes that operate exclusively on depth frames in order to realize such RGB-D tracking methods. Local Ternary Patterns (LTP), histogram oriented gradients (HOG) and Histogram of Oriented Vectors (HONV) have been some of the feature extraction schemes that have been adopted towards depth sequences [35], [36]. The proposed algorithm as illustrated in the Fig. 1 applies feed forward and feedback schemes towards a realization of efficient RGB-D fusion thereby ensuring a coherency in the tracking framework. Key contributions of the proposed scheme lie in the following: efficient incorporation of depth features into the CSK classifier training scheme, adaptive classifier and model parameter update strategies towards robust tracking and an efficient feedback strategy towards occlusion-robust tracking with the CSK tracker.

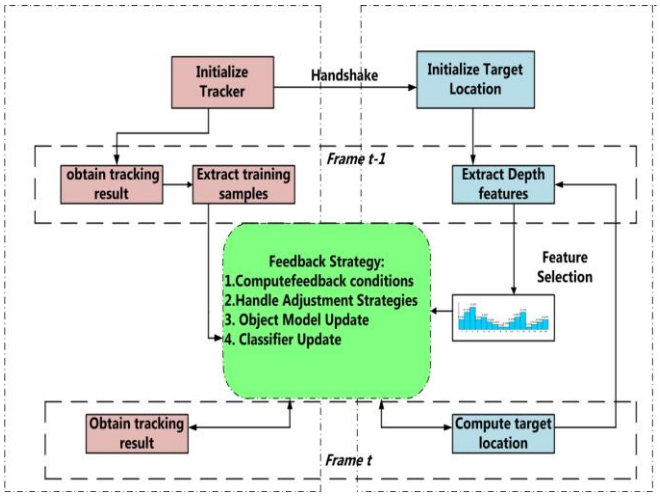


Fig. 1. General framework of the proposed tracking approach.

A. Circulant Structure Kernel Tracking with Adaptive Online Learning

The CSK tracker which is originally proposed in [18] has the capability of achieving efficient tracking performance with a lightweight and high speed approach. The tracker achieves the highest tracking speed according to current evaluation schemes [24]. This speed has been attributed to its efficient exploitation of the circulant structure that emerges when a periodic selection of the local image patch is conducted. We refer to [18]in providing a brief overview of the algorithm. Once candidate patches are selected, a classifier is trained by means of a single image patch x of a fixed size, $M \times N$, which centers around the object of interest [18]. At the classifier training stage, cyclic shifts of the patch are also considered such that $(m,n) \in \{0,...,M-1\} \times \{0,...,N-1\}$ Labeling of these cyclically shifted patches is performed by means of a Gaussian function. Training is achieved in this circulant tracking scheme by means of finding the parameters that minimize the regularized risk according to the generalized form expressed in (1).

$$\min_{w,b} \sum_{i=1}^{m} L(y_i, f(x_i)) + \lambda \|w\|^2 \quad (1)$$

The interested reader is referred to the original work [18]

for detailed reading. While this tracking scheme suffices in achieving efficient tracking-by-detection results, the tracker suffers drawbacks. Apart from short-comings in the presence of occlusion, which this paper proposes to tackle through the incorporation of depth features into the tracking framework, the CSK tracker also suffers from an inability to robustly update object appearance overtime. In original form, the CSK adopts a target object model $\hat{x}$ along with a transformed classifier coefficient A. The algorithm then adopts a simple linear interpolation scheme: $\alpha^p = (1-\gamma)\alpha^{p-1} + a\gamma$, towards updating the classifier parameters, where p and γ represent the current frame index and the learning rate parameter respectively. The problem with such a scheme is two-fold. Firstly, this basic linear interpolation scheme fails to consider all the previous appearances or representations of the target object and hence, crucial appearance information which could increase the adaptability of the object model is discarded over time. This has already been highlighted in recent studies [37]. Secondly, the learning rate of the classifier remains fixed and lacks the flexibility and adaptability required of the various tracking conditions that may be encountered in real-world scenarios. Both problems are further compounded by the fact that the classifier parameter update scheme operates at a constant rate on a per-frame basis, exposing the tracker to degradation when the object is occluded or misclassified in a particular frame. Therefore, while the incorporation of robust features such as depth could improve tracking performance and render such a circulant tracking scheme feasible in the robotics domain, we first need to address this parameter update scheme in an attempt to robustify the tracker before the incorporation of depth information and the feedback strategy proposed in this paper could succeed in significantly improving performance.

The first drawback of the parameter update scheme is addressed in a manner similar to the work in [37]. Instead of discarding previous appearance representations of the target, we adopt all representations by means of a weighted average quadratic error which considers all frames of the target sequence in the same manner as shared in [37]. A fixed $\alpha_k \geq 0$ is applied as the weight for each frame represented by k. This yields a cost function:

$$\vartheta = \sum_{k=1}^{p} \alpha_k \Big(\sum_{m,n} \left| \left\langle \phi(x_{m,n}^k), w^k \right\rangle - y^k(m,n) \right|^2 + \lambda(w^k, w^k) \Big) \quad (2)$$

The fourier transformed kernel is denoted as $U_x^k = F\{u_x^j\}$ and the weights of the frames, α_k, are adjusted by means of the learning rate parameter γ. Finally the object appearance model, $\hat{x}^p$, is updated as:

$$\hat{x}^p = (1-\gamma)\hat{x}^{p-1} + \gamma x^p \quad (3)$$

The classifier parameters are also updated in a similar manner as illustrated in the (4):

$$\alpha_N^p = (1-\gamma)\alpha_N^{p-1} + \gamma Y^p U_x^p \quad (4a)$$

$$\alpha_D^p = (1-\gamma)\alpha_D^{p-1} + \gamma U_x^p U_x^p (U_x^p + \lambda) \quad (4b)$$

This model update scheme exploits the appearance of the object over all frames without explicitly storing all previous appearance models. This update scheme makes use of problem domain information towards robustifying the tracker and addresses the first parameter update drawback of the CSK algorithm. At this point it becomes clear that although the tracker enriches itself by learning from all previous models of the object, the weights of all previous appearances as well as the learning rate parameter remain fixed throughout the sequence. This further highlights the second drawback of the CSK model and classifier parameter update scheme. Towards resolving this second drawback and further improving the adaptability of the tracker to the target scene dynamics, this paper proposes a feedback strategy that incorporates depth features into the tracking scheme. While the feedback strategy aims at enhancing adaptability of the tracking in online scenarios, the incorporation of depth features improves the robustness of the tracker to occlusion and this combination scheme renders the tracker more adaptive and persistent even in challenging tracking scenarios. The second drawback is addressed with the adaptive classifier training and parameter update strategies in the section below by means of depth incorporation and a feedback mechanism.

B. Depth Incorporation and Feedback Strategy towards Adaptive Occlusion-Robust tracking

Our choice of depth information as a reinforcement mechanism in occlusion situations is motivated by the hypothesis that an object in a scene possesses a uniform distribution of depth values which causes an abrupt change in the associated depth values of the object in situations of partial occlusion in the same manner in which edges are detected in RGB frames. These abrupt changes occur around the regions in the object where occlusion occurs as illustrated in the Fig. 2. Additionally, adaptive online tracking is achieved through an intelligent fusion scheme between RGB and depth features which are extracted on a frame bases and applied towards intelligent model update and classifier training. Since depth features are further robust to illumination changes, this combination scheme offers the potential to robustify the basic CSK tracker not only to partial occlusion situations but also to illumination variations.

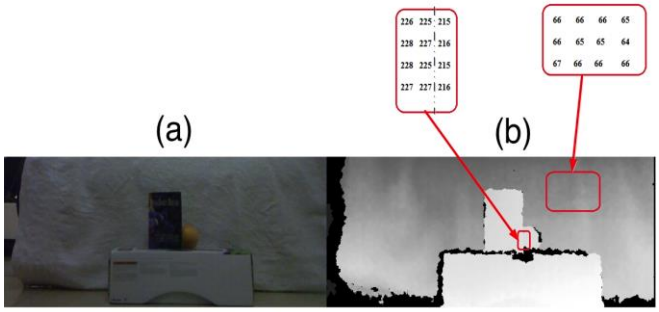


Fig. 2. Robustness of depth to partial occlusions. When the object is occluded, the depth values of its pixels change abruptly.

The incorporation of depth features into already existing color trackers is an ongoing research effort. However, most of these schemes propose to run parallel trackers in both depth

and color space and then proceed to rely on mechanisms in leveraging both tracking performances into one compound tracker. In this paper however, in order to maintain the speed and simplicity of the CSK tracker, we propose to rely upon a simple core tracker while extracting features from the depth space and applying them towards adaptive classifier training via robust feedback mechanisms.

Through the incorporation of depth information into the CSK tracking framework, an adaptive feedback strategy is proposed in this paper towards robustifying tracking performance. The feedback strategy is designed to facilitate the tracker and allow it in establishing a sense of occlusion and overcoming it in an adaptive and intuitive manner. The strategy achieves an online coupling of the trackers performance in both depth and color feature spaces which enables the overall tracking framework to achieve an adaptive classifier and model parameter update strategy within the CSK tracking framework.

Within the scope of this paper, depth features are extracted in the form of depth histograms which we represent as D_{hist}. While there exist other forms of depth features covered in recent literature, the experimental results presented demonstrate the depth histograms are sufficient in illustrating the efficiency if the proposed tracking scheme. In order to allow the CSK tracking to become occlusion aware and possess the capability of efficiently tackling this challenge which is inevitable in real-world scenarios especially in robotic applications, the feedback strategy proposed performs an analyses of the trackers performance within each frame. This approach allows for the depth similarities of the object within each depth frame to be computed and applied towards occlusion detection. The extracted depth histograms are applied towards occlusion detection through implementation with the Bhattacharyya coeffecient [38] as a supporting feature within the depth space in an ad-hoc manner. The Battacharyya distance computed between consecutive depth frames represents an approximate measure of the degree of overlap that exists between the subjective depth histograms and is illustrated in (5).

$$D_{hist}(D_t, D_{t-1}) = \sum_{u=1}^{N} \sqrt{H_{D_t}(u) H_{D_{t-1}}(u)} \quad (5)$$

where, D_{hist} denotes the Bhattacharyya distance between the two normalized depth histograms D_t and D_{t-1}; N represents the number of bins within the histograms. This distance is a measure applied towards the detection of instances in which the object appearance has undergone significant changes resulting from partial occlusion or in certain cases, environmental factors. However, in a majority of robotic applications in indoor environments, environmental factors remain constrained and this parameter bears a strong correlation with occlusion.

This measure of overlap offers a means by which partial occlusion can intuitively be detected within depth space and hence the coefficient allows the object's depth representation across all frames of a sequence to be applied towards the detection of partial occlusion instances.

This forms a crucial component within the feedback strategy that applies this occlusion awareness towards adaptive parameter update and search strategy within the CSK framework. A detailed presentation of the feedback strategy with its constituent parameters and transitions is illustrated in the Table I.

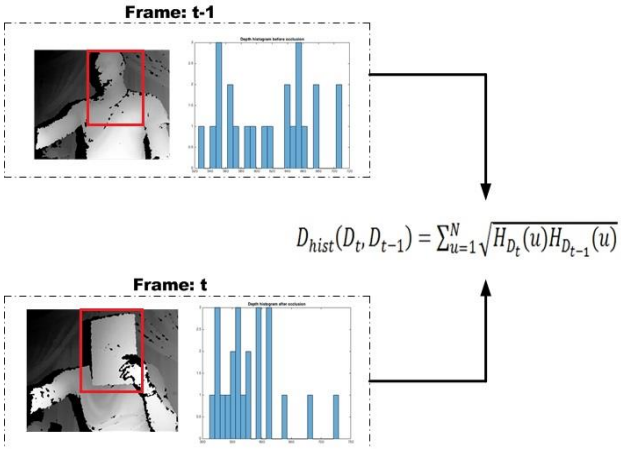


Fig. 3. The Battachryya distance towards occlusion detection in depth space.

In the Table I, D_{hist} represents the measure of similarity computed with the Battachryya distance between the $t-1^{th}$ and t^{th} depth frames, I_t represents the tracking results at the t^{th} frame, r represents the sampling radius defined by the bounding box size and η remains the learning rate parameter. λ and δ represent feedback adjustment parameters. According to the proposed feedback strategy, the case one suggests a consistency in the tracking results as well as a high degree of similarity in the depth histograms measured by the Battachryya distance and coefficient. This causes the feedback strategy to retain the sampling radius as well as the learning rate parameter accompanied with a classifier update. In cases 2 and 4, the feedback strategy suggests that there exists a minimal level of inconsistency in either the depth histograms of the object or the tracking results between two consecutive frames. This implies that the target object is either varying in appearance or undergoing rapid motion. In such cases the strategy gradually increases the search radius as well as the learning rate parameter of the classifier in order to allow the tracker to rapidly adapt to the object's changes over time. On other hand, in cases 3 and 5 where a large discrepancy exists between the depth histograms of the object or the tracking results between consecutive frames, there exists a high likelihood that a partial occlusion has occurred or the object appearance has altered significantly. Additionally, these cases could imply that the object model has been corrupted. This causes the feedback strategy to significantly increase the sampling radius of the tracker while holding off on model and classifier updates in order not to degrade the object model or classifier. Finally, the case 6 represents a scenario where the Battachryya distance between the depth histograms is far greater than an upper threshold while the tracker's results between consecutive frames are significantly inconsistent signifying that the object is either completely lost due to full occlusion or has exited the scene. In this case as well, in order to retain classifier and

model integrity, model and classifier updates are suspended while detection is reinitialized using the present object model in a dense sampling manner.

TABLE I: AN ILLUSTRATION OF THE FEEDBACK STRATEGY AND ITS ASSOCIATION PARAMETERS AND TRANSITIONS

Case	Condition	Adjustment Strategy	Classifier State
1	$D_{hist} < \delta_0$; $I_{t-1} - I_t < \lambda_0$	$r = r_0$; $\gamma = \gamma_0$	Update
2	$D_{hist} < \delta_0$; $\lambda_0 < I_{t-1} - I_t < \lambda_1$	$r = 1.2r_0$; $\gamma = 1.2\gamma_0$	Update
3	$D_{hist} < \delta_0$; $I_{t-1} - I_t > \lambda_1$	$r = 1.5r_0$; $\gamma = \gamma_0$	Hold
4	$\delta_0 < D_{hist} < \delta_1$; $I_{t-1} - I_t > \lambda_0$	$r = 1.2r_0$; $\gamma = 1.2\gamma_0$	Update
5	$D_{hist} < \delta_0$; $I_{t-1} - I_t < \lambda_0$	$r = 1.5r_0$; $\gamma = \gamma_0$	Hold
6	$D_{hist} > \delta_1$; $I_{t-1} - I_t > \lambda_1$	$r = r_0$; $\gamma = \gamma_0$	Hold

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Experimental Setup

Since the proposed tracking approach is aimed at satisfying real-time requirements of robotic applications, the robustness and efficiency of the approach are verified through both computer-based and robotic-based experiments. Both categories of the experiments are designed to validate the robustness of the proposed circulant tracking scheme to partial occlusion in scenarios that are characteristic and representative of robotic applications. Ideally, the first step in the verification phase of the proposed tracker would be the benchmarking of it performance against well established online tracking-by-detection algorithms including but not limited to STRUCK [15], TLD [16] and the convention CSK tracker [18]. However, this benchmarking is inhibited by the fact that these well established state-of-the art trackers are designed to operate on RGB or intensity frames while the proposed tracking approach based on an improved CSK tracker is designed to operate through a fusion scheme between RGB and depth features. For this reason, the most unbiased comparison scheme between the proposed scheme and existing trackers would be to test the proposed scheme on a fusion of RGB+D frames while the state-of-the-art trackers are only tested on the RGB sequences while excluding the accompanying depth frames. Furthermore, online performance of the proposed scheme is validated in computer-based and robot-based experiments with proposed future extensions of the validation process though robot-based experiments. The computer-based experiments are conducted as follows:

1) Computing platform: 2.8 GHz Intel Core i7 with 8 GB RAM @1067 MHz)
2) Data Acquisition Scheme: RGB+D frames obtained via the Kinect sensor

In conducting experiments towards qualitative and quantitative validation of the performance of the proposed circulant tracking scheme, video sequences containing both RGB and depth information were captured with the Kinect. The video benchmarking sequences applied towards the verification of the proposed scheme are designed pose specific challenges to the tracking task in a manner in line with the targeted application domain of robotics. Table II presents the individual sequences with the accompanying challenges they simulate.

TABLE II: BENCHMARK SEQUENCES AND THEIR ASSOCIATED CHALLENGES

Sequence	Associated Challenges
1. Stick-it-Note	Partial Occlusion, Rapid Motion
2. Mug	Partial Occlusion, Random Motion, Illumination Changes
3. Notebook	Partial Occlusion, In-plane Rotation
4. Palm	Out-of-view, Random Motion, Partial Occlusion
5. Magazine	Partial Occlusion, Full Occlusion, Fast Motion

Since the proposed scheme is aimed at robust tracking specifically targeting difficulties posed by partial occlusion, all the benchmarking sequences are designed to incorporate this challenge with at least one other form of tracking challenge. The proposed scheme is firstly compared with existing state-of-the-art trackers. In this initial comparison, the ground truth of the targets is applied towards computing the success rate, a core metric in comparing the performances of the trackers. The success rate is calculated according to (6) below:

$$\text{success} = \frac{\text{area}(BBR_G \cap BBR_T)}{\text{area}(BBR_G \cup BBR_T)} \quad (6)$$

where BBR_G represents the bounding box region of the ground truth and BBR_T represents the bounding box region of the tracker. The success rates of the selected trackers in comparison with the proposed algorithm are illustrated in Table III.

B. Experimental Results

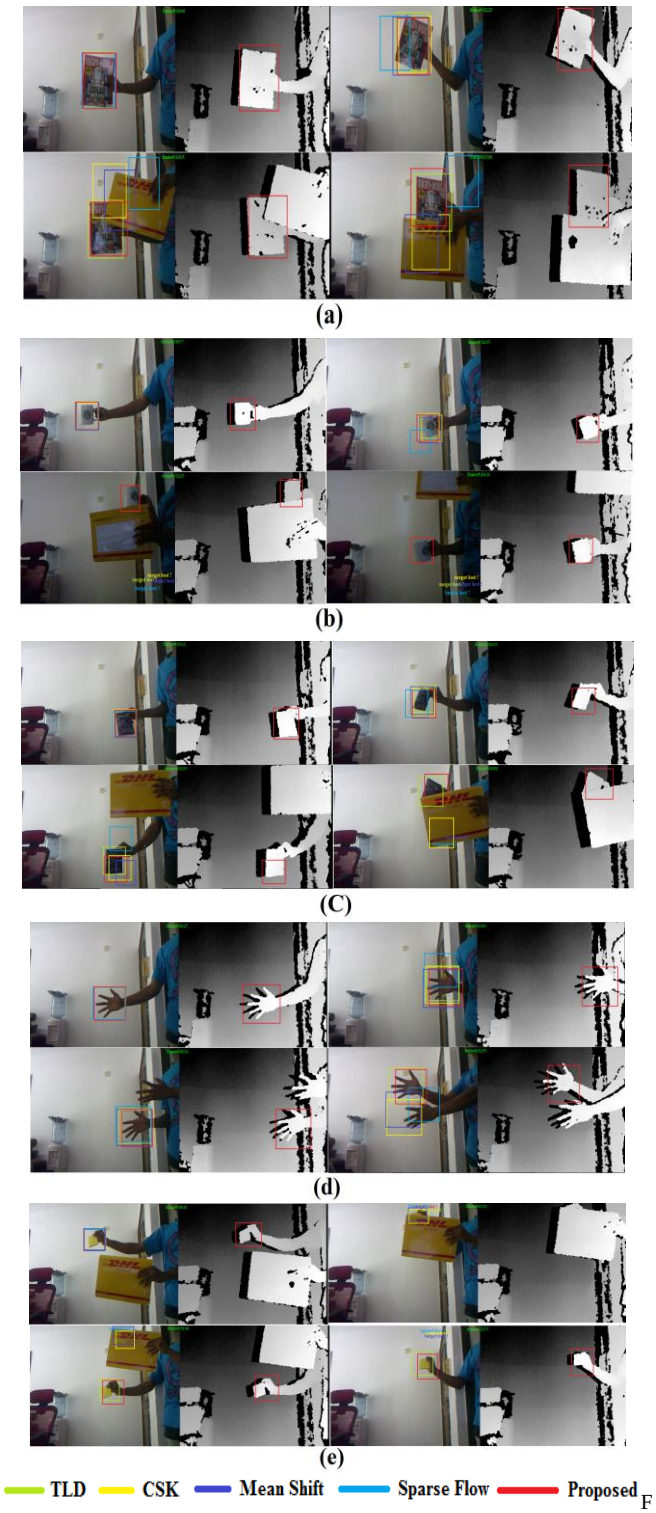
The results illustrated in the Table III indicate that the proposed algorithm is capable to maintaining a stable tracking performance and robustness to various levels of target occlusion. Additionally, due to the incorporation of a depth features, the tracker is further robustified against significant illumination changes characteristic of most real-world situations.

TABLE III: PROPOSED ALGORITHM VS. STATE-OF-THE-ART: SUCCESS RATE (ALL TRACKERS EXECUTED IN REAL-TIME WITH KINECT SENSOR)

Sequence/ Frame count	CSK	TLD	Mean Shift	Sparse Flow	Proposed
Stick-it-note/400	22.8	32.7	18.1	15.2	27.3
Mug/450	15.3	23.9	15.8	13.2	31.7
Notebook/400	24.5	38.3	22.7	18.0	26.2
Palm/400	24.6	36.1	25.2	20.5	32.9
Magazine/300	18.0	32.7	16.3	13.8	25.5

Comparatively, the tracker outperforms the traditional CSK tracker across most sequences but it is necessary to point

out that most of these outperformances are directly linked with frames where occlusion and drastic illumination changes occur. The TLD tracker achieves a more robust performance across all sequences except the "Mug" sequence where drastic illumination changes cause the all trackers except the proposed scheme to lose the target completely. In terms of tracking speed, the CSK maintains the highest and most stable speed due to its lightweight and simple kernel structure. The computational complexity of the proposed scheme however does not impede upon its real-time performance requirements.



ig. 4. Tracking results of the proposed tracking scheme in comparison with state-of-the-art trackers on real world scenarios. (a) "Magazine" scenario (b) "Mug" scenario (c) "Notebook" scenario (d) "Palm" scenario (e) "Stick-it-note" scenario.

All the sequences applied in the testing of the proposed scheme contain varying levels of occlusion as well as varying rates of motion and rotation. These sequences are not only designed to verify the robustness of the proposed tracking scheme but to also highlight its relative strengths and weaknesses in comparison with some already well-established trackers. In order to ensure fairness in comparison, all trackers are initialized in synchrony both spatially and temporally. As illustrated in the tracking results in Fig. 5, the traditional CSK tracker although, fast and efficient in most scenarios, undoubtedly suffers from an inability to ascertain whether or not a target ahs been partially or fully occluded, hence is prone to classifier degradation in such situations. Throughout all the sequences, this remains the major area in which the proposed RGB+D circulant tracking scheme outperforms the traditional CSK. The most outstanding performance is demonstrated by the TLD (Predator)which maintains the target in both partial and full occlusion situations through a decoupling of the tracking, learning and detection tasks. On the other hand, the most unstable performance is exhibited by the Sparse flow tracker. The Sparse Flow tracker which adopts the Lucas-Kanade method is resolving the flow within a local neighbourhood is highly unstable in most of the applied test sequences due to its reliance upon a 2-dimensional flow computation while assuming that the flow within the selected local neighbourhood remains a constant. The experimental results have shown that in cases of partial and full occlusion, the proposed tracking scheme indeed outperforms the traditional CSK tracker as well as the Mean Shift and Spare Flow trackers. This performance is attributed to the tracker's ability to incorporate depth features into the classfier's feature space and therefore since depth features offer a means of detecting occlusion by means of the proposed feedback strategy, the proposed tracker is equipped to effectively detect occlusions of various intensities. In situations of occlusion, unlike the traditional CSK, the proposed tracker holds off on model updates and hence is capable of avoiding degradations of the target model and classifier. This allows the proposed tracking scheme to come second only to the TLD tracker which is also capable of maintaining model and classifier integrity by simultaneously tracking the target, learning its dynamic appearance and detecting it constantly across all frames.

Due to the incorporation of depth features, the proposed tracking scheme is able to outperform all other trackers on the "Mug" video sequence illustrated in Fig. 4b. The mug sequence varies from all other sequences in that, it is the only sequence in which illumination changes are introduced into the tracking challenge. As illustrated in the results, all trackers are able to maintain the target up to frame #0300 when a significant change in illumination is introduced into the scene. At this point, the CSK, TLD, Sparse Flow and Mean Shift trackers all fail to detect the target although the target maintained a smooth and non-erratic motion. Trackers such as the TLD, which are equipped to reinitialize themselves in such cases still failed to reacquire the target due to the drastic change in illumination and the absence of prior learning data in previous training sequences. The proposed tracking scheme however succeeds in maintaining the target even

when full occlusion occurs at frame #0325 in the presence of relatively poor illumination conditions. This further highlights a contribution of this paper and the need for more illumination-robust trackers since constant and sufficient illumination are usually not guaranteed in most real world scenarios. As shown in the Fig. 4c, in plane rotation does little to impede the performance of the proposed tracking scheme. In terms of tracking speed however, the CSK remains the highest performing tracker with tracking speeds that even supersede that of the TLD tracker.

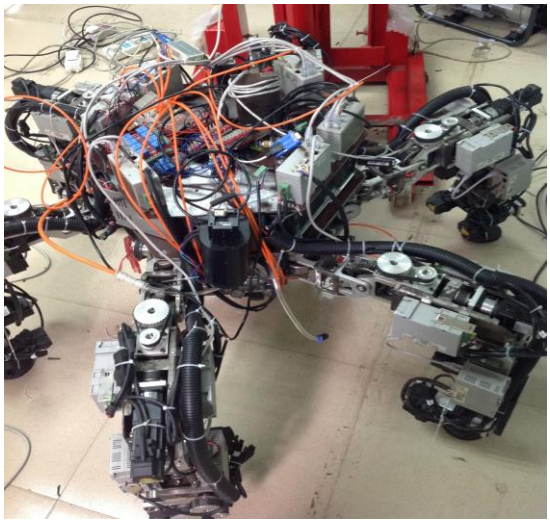


Fig. 5. Prototype 6-legged spider robot adopted towards robot-based experiments.

V. Conclusion

Drawing from the inadequacies of the CSK tracker and the lack of sufficient robust trackers towards robotic real-time applications, this paper proposes a robust and persistent tracking-by-detection scheme which is built upon the core CSK tracking. The CSK tracker is selected for core tracking because its lightweight and simple kernel structure allows for the tracker to achieve the most remarkable tracking speeds amongst the state-of-the-art and this is real-time performance is crucial in robotic application. Additionally, drawing from the robustness of depth features to various degrees of illumination changes as well as the capability of the feature to maintain a sensitivity to occlusion, a new tracking scheme is proposed to draw from the strength of the CSK while mitigating the problem of naïve classifier and model updates via an adaptive feedback mechanism. This approach allows the tracking scheme to draw from depth features and apply them towards a more refined, adaptive and robust tracking framework.

The proposed scheme is implemented and its performance is verified using the Kinect sensor. Experimental evaluation and comparison with state-of-the-art indicated that the proposed scheme is robust against various levels of occlusion-related challenges and outperforms state-of-the-art in terms of robustness to occlusion and drastic illumination changes. Furthermore, experimental results suggest that the significant increase in the computational complexity of the proposed scheme compared with the traditional CSK does not impede upon its real-time performance.

Since the proposed tracking scheme aims to address challenges associated with robust real-time tracking of arbitrary target for robotic applications, the future extension of the work presented in this paper will be two-fold:

1) The experimental evaluation of the proposed scheme will be extended to robot-based verification experiments aimed at subjecting the scheme to real-life challenges outside the constraints a lab-controlled environment. The target platform in these robot-based experiments will be conducted upon a prototype six-legged spider robot designed towards semi-autonomous structural inspection and maintenance of civil structures.
2) While the proposed scheme effectively combines RGB and depth features into a single robust tracking framework via an intuitive feedback strategy, there still exists the potential to adopt a more systematic approach towards the incorporation of depth and RGB features. Such an approach would operate the circulant tracker directly upon and RGBD tensor rather than in the decentralized manner discussed in this paper. The potential of such a scheme remains unexplored to the best of our knowledge and opens the path for future extensions of this work.

Acknowledgment

The authors gratefully acknowledge the support of Western Transportation Construction Technology Project under the Ministry of Transport of the Peoples' Republic of China.

References

[1] C. Chang and W. H. Tsai, "Vision-based tracking and interpretation of human leg movement for virtual reality applications," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no.1, pp. 9-24, Jan. 2001.
[2] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 34, no.4, pp. 1-45, Dec. 2006.
[3] K. Cannons, "A review of visual tracking," Technical Report CSE-2008-07, CA: York University, 2008.
[4] J. Zhu, Y. Lao, and Y. F. Zheng, "Object tracking in structured environments for video surveillance applications," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 20, no. 2, pp. 223-235, Feb. 2010.
[5] S. Bairchfield, "Elliptical head tracking using intensity gradients and color histograms," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 1998, pp. 232-237.
[6] L. Sun, G. Liu, and Y. Liu, "3D hand tracking with head mounted gaze-directed camera," *IEEE Journal Sensors*, vol. 14, no. 5, pp. 1380-1390, May 2014.
[7] M. Israd and J. Maccormick, "Bramble: A Bayesian multiple-blob tracker," in *Proc. IEEE International Conference on Computer Vision*, 2001, vol. 2, pp. 232-237.
[8] Z. A. Bakar, R. Samad, D. Pebrianti, and N. L. Y. Aan, "Real-time rotation invariant hand tracking using 3D data," in *Proc. IEEE International Conf. on Control System, Computing and Engineering*, pp. 49-495, Nov. 2014.
[9] S. Avidan, "Support Vector Tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 1064-1072, Aug. 2004.
[10] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi, "Robust online appearance models for visual tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 25, no. 10, pp. 1296-1311, Oct. 2003.
[11] D. Ross, J. Lim, R. S. Lin, and M. H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1, pp. 125-141, May 2008.
[12] J. Yang, B. Price, X. Shen, Z. Lin, and J, Yuan, "Fast Appearance Modeling for Automatic Primary Video Object Segmentation," *IEEE Trans. on Image Processing*, vol. 25, no. 2, pp. 503-515, Feb. 2016.

[13] X. Jia, H. Lu, and M. H. Yang, "Visual tracking via adaptive structural local sparse appearance model," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 1822-1829.
[14] B. Babenko, M. H. Yang, and S. Belongie, "Visual Tracking with Online Multiple Instance Learning," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 983-990.
[15] S. Hare, A. Saffari, and P. H. S. Torr, "Struck: Structure output tracking with kernels," in *Proc. IEEE International Conference on Computer Vision*, Nov. 2011, pp. 263-270.
[16] Z. Kalal, J. Matas, and K. Mikolajczyk, "P-N learning: Bootsrapping binary classifiers by structural constraints," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 49-56.
[17] T. B. Dinh, N. Vo, and G. Midioni, "Context tracker: Exploiting supporters and distracters in unconstrained environments," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2011, pp. 1177-1184.
[18] F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. 12th European Conference on Computer Vision*, Oct. 2012, vol. 4, pp. 702-715.
[19] S. Avidan, "Support vector tracking," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2001, vol. 1, pp. 184-191.
[20] Y. Qi, K. Suzuki, H. Wu, and Q. Chen, "EK-means tracker: A pixel-wise tracking algorithm using kinect", in *Proc. Third Chinese Conference on Intelligent Visual Surveillance*, Beijing, China, Dec. 1-2, 2011.
[21] Y. Bai and M. Tang, "Robust tracking via weakly supervised rankings SVM," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 1854-1861.
[22] X. Chen and J. Wu, "Scalable compressive tracking based on motion," in *Proc. IEEE International Conference on Robotics and Biometrics*, Shenzhen, China, Dec. 12-14, 2013.
[23] S. Avidan, "Ensemble tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 261-271, Feb. 2007.
[24] Y. Wu, J. Lim, and M. H. Yang, "Online object tracking: A Benchmark," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 2411-2418.
[25] W. Zhong, H. Lu, and M. H. Yang, "Robust object tracking via sparsity-based collaborative model," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 1838-1845.
[26] K. Wang, Y. Liu, and L. Liu, "Visual servoing trajectory tracking of nonholonomic mobile robots without direct position measurement," *IEEE Trans. Robot*, vol. 30, no. 4, pp. 1026-1035, Aug. 2014.
[27] K. Wang, Y. Liu, L. Liu, "Visual servoing based trajectory tracking of underactuated water surface robots without direct position measurement," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, Chicago, IL, USA, Sept. 12-18, 2014.
[28] O. Akin and K. Mikolajczyk, "Online learning and detection with Part-based circulant structure," in *Proc. International Conference on Pattern Recognition*, Aug. 2014, pp. 4229-4233.
[29] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments based tracking using the integral histogram," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2006, pp. 798-805.
[30] M. Rohrbach, M. Stark, G. Szarvas, I. Guravych, and B. Schiele, "What helps where and why? Semantic relatedness for knowledge transfer," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 910-917.
[31] J. Kwom and K. M. Lee, "Visual Tracking Decomposition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 1269-1276.
[32] D. Xu, X. Wu, Y. L. Chen, and Y. Xu, "RGB-D Hand detection and localization," *Journal of Intelligent & Robotic Systems*, vol. 77, pp. 583-596, 2015.
[33] B. Amit and W. Michael, "RGB-D Fusion with mean shift tracking," *Dynamic 3D Imaging*, Berlin: Springer, 2009, vol. 5742, pp. 58-69.
[34] S. Pan, L. Shi, and S. Guo, "A kinect-based real-time compressive tracking prototype system for amphibious spherical robots," *IEEE Journal Sensors*, vol. 15, pp. 8232-8252, April 2015.
[35] P. Wang, S. Ma, and Y. Shen, "Performance study of feature descriptors for human detection on depth map," *International Journal of Modeling, Simulation, and Scientific Computing*, vol. 5, pp. 13-29, May 2014.
[36] L. Spinello and K. Arras, "People detection in RGB-D data," in *Proc. Of IEEE/RSJ International Conference on Intelligent Robots and Systems*, San Franscisco, CA, USA, Sept. 25-30, 2011.
[37] D. Martin, W. Joost *et al*., "Adaptive color attributes for real-time visual tracking," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 1090-1097.
[38] D. Comanicu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2000, vol. 2, pp. 142-149.

Yao Yeboah received the B.Eng. in the field of electronic information engineering from the Huazhong University of Science and Technology, Wuhan, China in 2011 and the M.Eng. from the South China University of Technology, Guangzhou, China in 2013. He is currently pursing his Ph.D. in the Department of Electrical and Computer Engineering, School of Automation Science and Engineering, South China University of Technology. His research interests include pattern recognition, intelligent systems and robotic vision.

Zhuling Yu received the BSEE in 1995 and the MSEE in 1998, both in electronic engineering from the Nanjing University of Aeronautics and Astronautics, Nanjing, China and the Ph.D. in 2006 from Nanyang Technological University, Singapore. He joined the Center for Signal Processing, Nanyang Technological University in 2000 as a research engineer, then as a group leader. In 2008, he joined the College of Automation Science and Engineering, South China University of Technology, China. He was promoted to be a full professor in 2009. His research interests include signal processing, machine learning, computer vision and applications in biomedical engineering and robotics.

Wei Wu received the Ph.D. degree in the field of control theory and control engineering from the Huazhong University of Science and Technology, Wuhan, China in 2000. He is currently a professor in the School of Automation Science and Engineering of South China University of Technology, Guangzhou, China. His research interests include intelligent control systems, robotic control and engineering, pattern recognition and intelligent systems.