International Journal of Control, Automation, and Systems (2010) 8(5):926-939 http://www.springer.com/12555
DOI 10.1007/s12555-010-0501-4

Intelligent Visual Surveillance - A Survey

In Su Kim, Hong Seok Choi, Kwang Moo Yi, Jin Young Choi*, and Seong G. Kong

Abstract: Detection, tracking, and understanding of moving objects of interest in dynamic scenes have been active research areas in computer vision over the past decades. Intelligent visual surveillance (IVS) refers to an automated visual monitoring process that involves analysis and interpretation of object behaviors, as well as object detection and tracking, to understand the visual events of the scene. Main tasks of IVS include scene interpretation and wide area surveillance control. Scene interpretation aims at detecting and tracking moving objects in an image sequence and understanding their behaviors. In wide area surveillance control task, multiple cameras or agents are controlled in a cooperative manner to monitor tagged objects in motion. This paper reviews recent advances and future research directions of these tasks. This article consists of two parts: The first part surveys image enhancement, moving object detection and tracking, and motion behavior understanding. The second part reviews wide-area surveillance techniques based on the fusion of multiple visual sensors, camera calibration and cooperative camera systems.

Keywords: Behavior understanding, cooperative camera system, image interpretation, motion detection, object tracking, wide area surveillance.

1. INTRODUCTION

Intelligent visual surveillance improves conventional passive surveillance systems through automated object recognition and tracking, scene interpretation, and indexing/retrieval of visual events. Visual surveillance techniques have initiated a wide variety of applications in access control, person specific identification, anomaly detection and alarming in academic community as well as industry and government [1]. Large research projects on visual surveillance have driven realization of practical visual surveillance systems. Successful visual surveillance systems such as the Visual Surveillance and Monitoring (VSAM) [2], the Annotated Digital Video for Intelligent Surveillance and Optimized Retrieval (ADVISOR) [3], and the Smart Surveillance System of IBM [4] have been developed by combining computer vision, system engineering, and communication techniques.

Recently visual surveillance research focuses on intelligent visual surveillance (IVS) in a wide area, as a Third Generation Surveillance System (3GSS) [5] concept. Research trends in IVS can be divided largely into image interpretation and wide area surveillance control techniques. The goal of image interpretation is to extract high-level information of a visual event from a dynamic scene. Image interpretation often includes motion detection, object recognition, tracking, and behavior understanding. Recent studies in image interpretation focus on robust image processing techniques such as motion detection in situations with changes in illumination and weather, object tracking in scenarios with occlusion and non-rigid deformation and behavior understanding for human motion analysis. Wide area surveillance techniques expand the range of surveillance area to a broader territory. Until the Second Generation Surveillance System (2GSS) [5], visual surveillance system research was limited to local area surveillance using local Closed Circuit Television (CCTV) camera networks. Current research focus to widen the surveillance area is multiple sensor control and cooperative camera systems. More specifically, camera calibration and camera installation methods, which aim at reducing redundant camera installations, have been developed using multiple sensor control. In order to handle occlusion problem and broaden the surveillance area, techniques for integrating data are important issue in the cooperative camera system

This article reviews two major components of IVS systems: Image interpretation and Wide area surveillance control. The first part reviews research efforts related to image interpretation in IVS including motion detection, object tracking, and behavior understanding. The second part covers wide area surveillance control techniques and cooperative camera systems for distributed surveillance systems. Camera calibration and sensor installation are presented for the multiple sensor control.

Manuscript received February 26, 2010; accepted June 10, 2010. Recommended by Guest Co-Editor Mongi Abidi.

In Su Kim, Hong Seok Choi, Kwang Moo Yi, and Jin Young choi are with the Department of Electrical Engineering and Computer Science, Seoul National University, 599 Kwanak-ro, Gwanak-gu, Seoul 151-742, Korea (e-mails: {iskim, hschoi, kmyi, jychoi}@neuro.snu.ac.kr).

Seong G. Kong is with the Department of Electrical and Computer Engineering, Temple University, Philadelphia, PA 19122, U.S.A. (e-mail: skong@temple.edu).

* Corresponding author.

2. IMAGE INTERPRETATION

Image interpretation extracts high-level information on visual events from a sequence of scenes through image enhancement, motion detection, object tracking, and behavior understanding. At each individual image interpretation step, key issues are solving image occlusion problems, developing robust algorithms against illumination and weather changes in the scene, and reducing computation time to achieve real-time performance. This part includes surveys on each image interpretation module in IVS. The motion detection section covers background subtraction and motion detection using an active camera. The object tracking section reviews general tracking methods classified as point tracking, kernel tracking, and contour tracking methods, by their object representation methods. Finally, the behavior understanding section introduces research on human motion analysis.

2.1. Image enhancement

Image enhancement is to improve visual appearances of a scene captured in diverse environments. Image enhancement approaches can be divided into three major categories: Frequency-based, Histogram-based, and Transform-based approaches [6].

Frequency-based approaches decompose an image onto high- and low-frequency signals. Homomorphic filtering and unsharp mask filtering are some of widely used frequency based techniques. Homomorphic filtering is a Fourier Transform-based technique that enhances the contrast of an image by removing the low frequency and amplifying the high frequency in the frequency domain. Seow and Asrai [7] improve a color digital image using a neural network algorithm in the homomorphic system. Unsharp mask filtering enhances the edges by subtracting a smoothed version of an image from the original image and then adding the difference back to the original image. The edge extraction block in an unsharp image is often implemented with a linear highpass filter such as a discrete linear Laplacian operator. In [8], the linear filter is extended to the quadratic volterra (QV) filter inspired by the Weber-Feshner law [9]. Moreover The QV filter, which has poor performance in noisy environments, is extended to the quadratic weighted median (QWM) filter [10].

Histogram Equalization (HE) is commonly used method among histogram based image enhance approaches. Duan and Qiu [11] take the pixel distribution of the original image into account when performing the equalization process and control the degree of contrast enhancement. The Bi-Histogram Equalization (BBHE) [12] preserves the original brightness of an image to a certain extent, which is not possible in HE; however, an equal-area Dualistic Sub-Image Histogram Equalization (DSIHE) method [13] outperforms BBHE in brightness and image content (entropy) preservation. DSIHE can change the brightness to the level between the median and middle levels of the input image. The Minimum Mean Brightness Error Bi-Histogram Equalization (MMBEBHE) [14] can preserve the mean brightness which is not possible in BBHE and DSIHE. MMBEBHE separates the histogram using the threshold level which yields the minimum Absolute Mean Brightness Error (ABME). The Brightening Preserving Histogram Equalization with Maximum Entropy (BPHEME) [15] method finds the target histogram that maximizes the entropy, and then transforms the original histogram to the histogram of the target using histogram specification.

Transform-domain enhancement techniques enhance the image by manipulating the transform coefficients and mapping the image intensity data into a given transform domain using a transform function such as the discrete cosine transform (DCT), Discrete Fourier transform (DFT), wavelet [16] and other fast unitary transforms [17]. Transform-based techniques [18] can be used for illumination correction, night vision, and noise reduction. Kober [6] proposes a real time sliding discrete transform to enhance the local contrast of a noisy image. In this method, a minimum mean-square error estimator is derived and a fast recursive algorithm for computing the sliding transform is utilized. Arslan and Grigoryan [19] split the 2-D Fourier transform into different groups of sample. They separately process all splitting-signals [17] and then calculate and compose 2-D DFT of the processed image by the processed new splitting-signals. In [20], they propose a fast implementation of the alpha-rooting method by using one splitting-signal of the tensor representation with respect to the DFT. Agaian [21] proposes three methods for image enhancement: logarithmic transform histogram mapping and shaping, and logarithmic transform histogram shifting. Furthermore, they visualize the transform coefficient histogram and measure the overall contrast of the image.

2.2. Motion detection

Motion detection in IVS is to find moving target objects in an input image sequence. Conventional approaches for motion detection methods use background subtraction [22-25], temporal differencing [26], and optical flow [27]. With the growing use of active cameras in recent visual surveillance environments, motion detection algorithms using active cameras are becoming an important component for IVS. Motion detection using active cameras has been developed by making background mosaics [28], modified background subtraction methods which compensate for camera motion using optical flow [29], feature matching [30], and camera geometrical models [31] to detect moving objects by registering the current frame to the background image. This section reviews motion detection algorithms including background subtraction, temporal differencing and techniques for active cameras.

2.2.1 Background subtraction

Background subtraction is a widely used approach because of its accuracy and fast computation for detecting the foreground. In order to extract the foreground object, the background subtraction algorithm detects the difference between the current image and the

reference image, often called the "background image" or "background model." Recent background subtraction algorithms focused on robust background modeling and updating to adapt to varying illumination conditions between night and day, geometry reconfiguration of background structure, background change from weather change, and repetitive motion from clutter. Stauffer and Grimson [22] proposed a background subtraction method for modeling a multiple modal background distribution. They use a mixture of Gaussian models to construct the distribution of each pixel location. Usually, three to five mixtures of Gaussian models are used. Each pixel value can be modeled as several backgrounds using a mixture of Gaussian models to cover the motion of tree or gradual image change. If there is a matched Gaussian model with a current pixel value, then the current pixel is decided as the background and updated Gaussian model using the current pixel value. Otherwise, the pixel is classified as the foreground and the Gaussian model with the lowest weight is replaced by a new one centered in the current pixel value. Although the decision of the number of Gaussian models and initialization of the Gaussian model is ambiguous and background modeling may fail when drastic illumination change occurs, this approach can robustly model and update the background for multiple modal distribution such as motion of leaves and a gradually changing background. Moreover, computation speed is fast and does not require a relatively large memory.

Haritaoglu *et al*. [23] developed a statistical background modeling method by training background using pixel history. The background model is represented by the minimum (M) of pixel value, the maximum (N) of pixel value, and the maximum intensity difference (D) between frames observed during the training period. The current pixel is classified as a background when the difference between current pixel value and M, N is less than D; otherwise, the current pixel is classified as a foreground. A real-time surveillance technique [23] uses two different background update methods; the pixel-based update and the object-based update. The pixel-based update method updates the background periodically to adapt to illumination changes while the object-based update method updates the background to adapt to physical changes in the background scene. This statistical background model can adapt to illumination changes because of the training of historical pixel variance. Additionally, motion detection can be performed in real-time because of the simple computation manner in which background modeling and updating can be carried out.

Oliver *et al*. [24] propose a background subtraction method using the Principle Component Analysis (PCA). A background model called the "eigen-background" is created using PCA and eigen-decomposition. The foreground of the current pixel is detected by subtraction between the eigen-background and the projected image of current image. Horprasert *et al*. [25] presented a novel algorithm for detecting moving objects from a static background scene which contains shading and shadows using color images. Shadows and highlights have similar chromaticity with the background but brightness is different. Using this property, this background model improved the weakness of traditional background subtraction against local illumination change, such as shadows and highlights, as well as global illumination changes. However, in [25], the background model is designed under an assumption that the background scene is static. This proposed background model may suffer from dynamic background changes such as the entrance of a new background object. Therefore, improvement of the adaptive background update problem still remains.

2.2.2 Temporal differencing

Temporal differencing [26] uses the pixel-wise differences between two or three consecutive images in image sequences to extract a moving object. Temporal differencing is adaptive to dynamic environments and its computation for extracting a moving pixel is simple and fast. Generally, temporal differencing is not effective in extracting all the relevant pixels of a target object. There may be holes left inside moving objects, and it is sensitive to the threshold value when determining the changes within differences of consecutive images. Additionally, temporal differencing cannot handle an active camera environment without a camera motion compensation algorithm.

2.2.3 Motion detection on active camera platform

With the growth of active camera usage in recent visual surveillance environments, there are attempts to develop an active camera surveillance system. However, background subtraction or temporal differencing algorithms cannot be used directly to detect a moving motion in a moving active camera. Modified motion detection algorithms have been developed to register moving current images into background images. Modified motion detection algorithms can be classified into the following different approaches by their registering method: background mosaic approach, optical flow approach, feature matching approach, and camera geometrical model approach.

Bevilacqua *et al*. [28] propose a background mosaic method to extract a moving object in an active camera image. These papers present a real-time framework for making an image mosaic without camera information and scene geometrical information. In [28], once the background mosaic is constructed, the background subtraction can be used to detect moving objects by subtracting between the registered current image and the correspondent background region within the background mosaic. There are two main stages in making a background mosaic: image spatial registration and tonal alignment. At the image spatial registration stage, image registration is done by a camera motion estimation computed using corner point matching and projective model assumption with a subsample image in real-time. At the tonal alignment stage, the histogram specification technique is used to align the color around the image junction in order to overcome the errors arising from

photometric misalignments. This method is robust and is a real-time background subtraction algorithm on an active camera. Color background mosaic is made in real-time and does not need any camera parameter or scene information.

Cucchiara *et al*. [29] compute camera ego-motion using a dominant optical flow to detect moving objects on an active camera. In this paper, dominant camera motion is estimated by selecting the peak of direction histogram based on the angle of optical flow. Optical flow is computed using the Lucas-Kanade [32] method, and camera motion is modeled through the translation model. Motion detection is performed by aligning and temporal differencing between the current frame and the previous frame using the estimated camera motion. The direction histogram proposed in [30] makes the clustering step which aims to select the dominant optical flow faster than the existing complex and time-consuming way. Therefore, motion detection can be performed in real-time. However, this method needs assumptions that the background is dominant over the moving objects and camera motion can be approximated with pure translational model although camera moves little by little.

Micheloni and Foresti [30] propose a camera motion compensation algorithm which registers the current image to the background image using a feature tracking method. Shi and Tomasi [33] develop a feature extraction/selection method representing image se-quences, which aligns consecutive frames by estimating the best displacement between feature sets, assuming translation model. Moving objects are extracted through temporal differencing the aligned consecutive frames. In the paper, the motion detection algorithm on an active camera is presented using robust tractable feature matching. However, when it is impossible to select a set of tractable features, for example, the zoom is too high and the scene contains a wide moving object in a close up shot or uniform background, the camera motion compensation method proposed in [33] may fail to estimate camera motion.

Murray [31] built a motion tracking system that detects moving objects using an active camera. Camera motion is compensated by calculating the estimated pixel position using the camera's intrinsic parameter (focal length) and extrinsic parameter (pan and tilt angles), and moving pixels are segmented by the temporal differencing method. The paper presents a novel way to suppress the "ghost" in different images between consecutive images through the logical operations between the ghost image and edge image of the current frame. This way, the accurate estimation of the next pixel position using camera parameters can be achieved during motion tracking on an active camera; however, this paper needs an accurate measurement of camera parameters. Usually, there are many variances in the measurement of camera parameters (e.g., camera shaking by internal motor movement) in a surveillance environment. In order to compensate camera motion using this method, a method for handling the measurement noise of camera parameters should be included for robust performance.

2.3. Object tracking

The goal of object tracking is to find a moving object detected in motion detection stage from one frame to another in an image sequence. The performance of the high level image interpretation module such as behavior understanding is depends highly on the object tracking result. Difficulties in tracking an object can arise from abrupt object motion, changing appearance of object and scene, self-occlusion, occlusion by structure. Thus, these difficulties should be solved to track the target object accurately. In this section, we review the object tracking by classifying it as point tracking, kernel tracking, and contour tracking according to the object representation method [34].

2.3.1 Point tracking

The point tracking method represents the target being tracked by points which are detected in consecutive frame with the tracking procedure. Point representation of a target object has robustness to the changes of rotation, scale, and affine transform [35]. Point tracking can be classified into the deterministic and the statistical methods depending on the matching method used for finding point correspondences. The deterministic method uses proximity, maximum velocity, common motion, and rigidity constraints to match point correspondence. In [36], the point tracking method involving constraints has been proposed. On the other hand, the statistical method represents an object by the state-space of object parameters such as position, velocity, and size. When tracking is performed with state-space representation, the state is estimated using the dynamic model of state transition, and updated by the correction stage using the measurement from the image. Representative methods for estimating the dynamic model in statistical point tracking include the Kalman filter [37] and the particle filter [38]. The particle filter calculates state probability using the sequential importance sampling method and corrects state probability using the measurement. It can handle non-Gaussian state and non-Gaussian noise. Thus, a particle filter can track a point in a general environment. However, if the state and noise distribution follow the Gaussian distribution, then the Kalman filter provides a better optimal solution.

2.3.2 Kernel tracking

The kernel tracking represents a target object by a primitive object region such as a rectangular, ellipse, or circle, and tracking is performed by computing object motion from one frame to the next. Usually, the motion of the object is assumed in the form of a parametric model such as translation, conformal, and affine transform. Kernel tracking is a popular method, because it is robust to uncertain spatial deformations and its broad range of convergence. Kernel tracking can be classified into template model and appearance model.

The template model matches the target using a similarity measure between the template and a candidate

image. Rectangular and ellipse templates have been widely used to characterize the object, and histogram of the template and their intensity values are used to calculate the similarity score. Sum of squared differences (SSD) [39], normalized cross-correlation [23], and Battacharya coefficient [40] are popular similarity measures. The template model approach has been widely used because of its computational simplicity. The VSAM [2] system uses cross-correlation method to track the object detected by a motion detection module. In W4 [23], cross-correlation function is also used to track human body parts. In [39], the object tracker uses an SSD similarity function, and the mean-shift tracker [40] uses the Battacharya coefficient as a similarity measurement. Recently, there have been some attempts to make the computation speed of similarity function faster, or to reduce the search area of similarity measurement to shorten the computation time. In VSAM [2], the object tracker uses the sub-sampling method with motion information to reduce the computational cost in the template matching process. In [40], Comaniciu and Meer proposed a real-time object tracking method based on the mean-shift procedure, which can find the mode of a probability density function (PDF) through only a few iterations. The mean-shift tracker is a popular kernel tracking method, which uses a weighted color histogram to represent the object. This work is extended in [41], where the Linderbug's scale theory [42] is combined with the mean-shift tracker to solve the scale problem. In [43], a new object description method using a histogram is proposed to extend the description efficiency of the original mean-shift tracker.

In multi-view appearance-based kernel tracking, the appearance model of the target is trained using an offline learning machine, and the target object in the current frame is tracked by computing the classification score of the learning machine. Usually multi-view appearance model provides robust tracking performance in the changes of viewpoint. For example, Black and Jepson [44] proposed a subspace-based algorithm (so-called eigenspace) to compute the affine transformation using eigenvectors. Based on the eigenspace, object tracking is performed by estimating the affine parameters iteratively until the difference between the input image and the projected image is minimized. The eigenspace-based similarity in [44] provides a robust property for image distortion by illumination changes. Lim *et al*. [45] propose an incremental subspace-based tracking algorithm to update the appearance model, where object tracking is performed by using a particle filter and motion model (affine transform). Although this method can track the object in scenarios with illumination changes, the subspace update method does not consider occlusions. Avidan [46] presented an object tracking method by integrating the support vector machine (SVM) classifier into optical flow based tracker [47]. SVM is a general classification scheme to discriminate two classes (positive or negative class) by finding the best separating hyperplane which has the biggest margin between classes. In this paper, multi-view images of the target object are used for positive training samples, and negative training samples consists of all things that are not to be tracked (usually, background region). Object tracking is performed by maximizing SVM classification score over image regions. This method uses knowledge about the background object that makes the tracker more robust against complex background clutter image.

2.3.3 Contour tracking

Contour tracking method iteratively evolves an initial contour which represents the target object using an outline contour from the previous image to the next. Contour's representation ability provides an efficient tracking method to a target object with a complex shape and various changes of shape over time. Thus, recent studies have applied contour tracking method to the non-rigid object such as human tracking. In [48], Paragios developed a multiple object tracking algorithm using a geodesic active contour method and level set formulation scheme. Freund [49] proposed people tracking using a new active contour model based on the Kalman filter in spatio-velocity space. Isard [50] presented a contour tracking method based on the particle filter, as known as the Condensation algorithm. Condensation algorithm is the first application to use particle filtering for object tracking in the computer vision community. It can handle the non-Gaussian distribution of the state and the noise to overcome the limitations of the Kalman filter in a complex cluttered image, showing non-Gaussian distribution. Yilmaz [51] considered occlusion conditions and proposed an object tracking method using the active contour. In the paper, the contour evolution is performed by minimizing the energy function defined by the sum of image energy and shape energy. Image energy is defined by the color and texture around the contour band, while shape energy is defined by using past contour observation to cover occlusion problems. In the paper, energy function is minimized by the level set method [52]. Since this method does not use background subtraction to initialize contours, this method can be performed on a mobile camera. Also, robust tracking is possible for occlusion situations because of the shape model.

Contour tracking is generally better than kernel tracking when tracking objects with complex shape changes. However, the performance of contour tracking is sensitive to the initial contour. Therefore, contour tracking may fail to track the object when it encounters difficulties in extracting contours, such as when dealing with noisy images, blurred images, or low contrast images.

2.4. Behavior understanding: human motion analysis

The behavior understanding task is one of the representative high-level vision tasks in visual surveillance, which analyzes object behaviors and gives warnings to the human operator. Behavior understanding task is mainly focused on human motion analysis. Human motion analysis studies can be classified as gross level, intermediate level, and detailed level depending on

the analysis level of the detail [53]. At the gross level, individual people are represented as distinct moving bounding boxes or ellipses, and the pattern of the trajectories or motion patterns of these boxes or ellipses are analyzed to recognize human motion [54]. At the intermediate level, individual people are represented by their body parts such as head, torso, arms, and legs, and human motion analysis is performed by tracking and recognizing each body part [55]. At the detailed level, recognition of human activities is performed in terms of single body parts such as hand gesture recognition, face and head gesture recognition. Visual surveillance often employs gross and intermediate level recognition, and detailed level recognition mainly aims for developing the gesture based human-computer interfaces (HCI) [56]. In particular, hand gesture recognition has been studied by many researchers in HCI research. In the recent visual surveillance system, human body segmentation, defining basic motions, occlusion handling, and handling of time-ordered correspondence have been the focus of researches for accurate recognition of human motion. In this section, low-level tasks for modeling the human body and high-level tasks for recognizing human activity are reviewed.

2.4.1 Human body modeling - low level vision

Human body modeling can be divided into two approaches, the model-based approach and the appearance-based approach, based on whether a prior shape model is used or not. The model-based approach, which uses a prior shape model, can represent complex motion by efficiently integrating the human body model. However, the model-based approach usually requires additional processing steps of model selection and parameter estimation to fit the model to the input image. On the other hand, the appearance-based approach, which does not use a prior shape model, does not require additional steps but is sensitive to noise.

These two approaches commonly employ a stick figure, 2-D contour, and 3-D volumetric figure to represent a human body. The stick figure representation regards the human body as a composition of sticks and joints. VSAM [2,57] employs the stick figure representation, as known as the star skeleton, to analyze human gait such as running and walking, by using the cyclic motions of a stick figure. The 2-D contour figure representation regards the human body as a cardboard [58], ribbon [59], silhouette contour [60], and blob model [23]. A 3-D volumetric figure representation attempts to describe the detailed human body in 3-D space by using cylinders [61], generalized cones [62], and spheres [63]. From the stick figure to the 3-D volumetric figure, the complexity of the model increases along with the level of detail.

2.4.2. Human activity recognition - high level vision

Human activity recognition is a high-level task in human motion analysis. Human activity recognition can be divided into two approaches: (1) the general sequence matching approach which recognizes human activity by matching the pre-defined image sequences of human activity and the input image sequences and (2) the approach using prior knowledge for recognition. In the general sequence matching approach, general sequence classification schemes such as DTW, HMM, and DBN are widely used to cover variances of time interval between human activity sequences. In the approach using prior knowledge for recognition, human activity recognition is performed using rule-based inference, physical constraints, causal analysis, and syntactic analysis. In this section, we review general sequence matching approaches and approaches using prior knowledge for human activity recognition.

A. Human activity recognition using general sequence matching approaches

For human activity recognition using general sequence matching, human activity recognition is simply regarded as a classification problem of the time varying feature sequence of the human body. Therefore, general time varying feature classification schemes such as Dynamic Time Warping (DTW) [64], Hidden Markov Model (HMM) [65], and Dynamic Bayesian Networks (DBN) [66] have been widely used to recognize human activity. Additionally, temporal template matching and finite state machine approaches have also been developed for human activity recognition.

DTW has been widely used in speech recognition in the early days, which is a template-based dynamic programming matching technique that measures the similarities between two sequences using operations such as, deletion-insertion, compression expansion, and substitution of subsequences. The advantage of DTW is the conceptual simplicity and robust performance in the classification of time varying sequences. However, DTW lacks the consideration of interactions between nearby subsequences occurring in time. Bobick [67] proposed a gesture recognition method using DTW matching by defining gesture as a sequence of state. The algorithm proposed in [67] showed that the test sequences and reference sequences can be successfully matched even though they have different time scales.

HMM is a stochastic state machine for analyzing time-varying data with spatio-temporal variability. HMM is superior to DTW in handling uncertainty of consecutive data. Therefore, HMM has been widely applied for matching human motion sequences. In [68], HMM is used for human intention recognition and skill learning, and in [69], sign language algorithm is proposed using HMM. In VSAM [2], to recognize actions (e.g., object appearing, moving, stopping, and disappearing), interactions (e.g., near, moving away from, moving toward), and no interaction between humans, vehicles, and human groups, matching reference sequences and input sequences using HMM is performed. Oliver [70] proposed and compared two different learning architectures, namely, HMM and Coupled Hidden Markov Model (CHMM) for modeling people's behavior and interactions, such as following and meeting each other. CHMM is much more efficient and accurate than HMM.

A significant limitation of the HMM is that it cannot handle more than two independent processes efficiently [70]. To alleviate this problem, researchers have developed Dynamic Bayesian Network (DBN) as a generalization of HMM [66]. DBN is a Bayesian network that represents sequences of variables. Park [71] proposed a hierarchical Bayesian network to recognize two-people interactions such as pointing, punching, pushing, and hugging. In this architecture, the low level Bayesian network estimates the human body part poses, and the high level Bayesian network estimates the overall body poses. In [71], a hierarchical framework is used for representing multiple levels of event; from the body-part level, to the multiple-bodies level, and finally, to the video sequence level. Furthermore, occlusions occurring in human interactions are handled through Bayesian network inference.

Template matching recognizes human activity by comparing input sequences represented by a static shape pattern to a pre-stored activity prototype [72]. In [73], a temporal template using an accumulated image history is proposed to represent human activity. Temporal template matching using this template has conceptual simplicity and real-time computation. However, the proposed method can only recognize human activity when all motions in the image are incorporated into the temporal template. Thus, this method cannot handle inter-people interactions or human activity with occlusions. The advantage of template matching is its low complexity and simple implementation. However, it is usually more sensitive to noise and the change of the duration of the activity than DBN or HMM. Moreover, it is viewpoint dependent.

Finite State Machine (FSM) can be used to recognize human activity by representing human activity as a sequence of states. The state is defined by the representative static pose, and the state transition function is also predefined according to the specific application. The state transition function is the most important feature of FSM. Human activity recognition is performed by a matching tour of its state. In [74], FSM is used to recognize natural gesture. In addition, Bremod *et al*. [75] used hand-crafted deterministic automata to recognize airborne surveillance scenarios. However, selecting the optimal number of state and defining appropriate state transition remain a difficult issue.

B. Human activity recognition approaches using prior knowledge

The sequence matching approach for human activity recognition performs accurately at well-defined activity situations, but not for complex interaction or activities that have flexible representations. For these situations, it is hard to define a general motion sequence to allow the use of a general sequence matching approach. To overcome this limitation, human activity recognition approaches using prior knowledge have been studied. As a result, more universal schemes to recognize general situation of human behavior using contextual information have been found. In this section, human activity recognition research using prior knowledge is introduced. Several studies recognize human activity using scenarios that are set of rules manually constructed, namely, the rule-based inference approach. Intille and Bobick [76] built a rule network using a temporal graph to interpret American Football games.

Physical laws can be used as effective causal constraints at interaction recognition because every object in the world is under some physical law. Mann *et al*. [77] developed a universal scene understanding method using the kinematic and dynamic properties of the scene. In the paper, interactions between human and objects, for example, “lifting a can” and “pushing a can,” can be interpreted in terms of physical laws such as gravity and friction. In this way, they present a computational theory that can derive force-dynamic representations directly from camera input.

Although physical constraints provide useful causal constraints for human activity recognition, understanding human activity needs more abstract and meaningful schemes than pure physical constraints. Brand and Essa [78] proposed a recognition method for arm gestures, such as “lifting,” “pushing,” “resting,” and “opening,” using the kinematic and dynamic relationships of body parts. They formulate the knowledge about causal processes of body kinematics and dynamics in terms of position, velocity, and acceleration of wrists, elbows, and shoulders. For example, “the greatest acceleration of hands occurs at the beginning of different actions” can be formulated as segmentation constraint. Thus, these constraints can detect different motion types. In this way, complex human activity is interpreted using more generalized causal constraints.

Syntactic analysis uses contextual knowledge to recognize a visual event assumed to be composed of primitive prior knowledge. Recently, a grammatical approach has been used for behavior understanding. Ivanov and Bobick [79] described a stochastic parser to the detection and recognition of temporally extended behaviors and interactions between multiple agents. In this work, recognition of human activity is divided into two levels; the lower level performing temporal behavior detection step such as HMM and the higher level which uses the result of the lower level to recognize behavior by analyzing syntactic relations using the stochastic context-free parser proposed in [79]. In a similar way, Ayer and Shah [80] interpret human activity such as “entering room,” “opening cabinet,” and “picking up a phone,” in a static room.

3. WIDE AREA SURVEILLANCE CONTROL TECHNIQUES

Wide area surveillance control technique is a large-scale data analysis and management skill for covering a broad territory in IVS. Wide area surveillance control technique can be divided into multiple sensor control technique and cooperative camera systems. The multiple sensor control technique has been developed to cover a wide area with multiple sensors. Representative research

areas for multiple sensor control technique are camera calibration of various types of camera and efficient sensor installation. Cooperative camera system is another major research topic for wide area surveillance. In this part, we review these three topics related to wide area surveillance control technique.

3.1. Multiple sensor control techniques

The accuracy of image interpretation of a visual event is affected critically by the deployment of sensors and sensor parameter settings. Thus, in order to achieve a good performance, a good multiple sensor control scheme is fundamental and essential for IVS because a wide area surveillance system uses many sensors to cover a broad territory. A number of studies have been carried out for low cost sensor settings, such as camera self calibration and efficient sensor installation to reduce redundancy of sensor deployment.

3.1.1 Camera self calibration

Images from a camera can be different from the real-world scene because of distortions from the compound lens and A/D converter. Particularly, in occlusion situations, camera distortion directly causes an error in the interpretation algorithm. Thus, an efficient camera calibration algorithm is needed for an IVS system using multiple cameras. In a wide area surveillance environment, camera calibration cost is very high because of the large number of cameras is used. Thus, a self-calibration algorithm for a camera is important for wide area surveillance systems. Recent self-calibration researches widely use projective geometry constraints, camera motion constraints, and scene constraints to compute the camera's intrinsic parameter. Also, the parameter setting algorithms for adjusting camera zoom and focus are dealt within in camera calibration researches. Collins and Tsin [81] proposed an outdoor camera calibration method for active cameras. The intrinsic parameter is calculated without any information of the 3D scene using optic flow obtained by rotating and zooming the active camera. Extrinsic parameters are calculated by actively rotating the camera to sight a sparse set of surveyed landmarks over a virtual hemispherical field of view leading to a well-conditioned pose estimation problem. In [82], the camera calibration method for a PTZ camera deployed in a wide area is developed. The calibration method proposed in [82] models pan and tilt rotations as they occur around arbitrary axes in space. A survey of different techniques for camera calibration can be found in [83].

3.1.2 Sensor installation

The deployment of sensors has a great influence on the performance and the cost of the surveillance system. Redundant sensors increase the processing time and the cost of installation. On the other hand, lack of sensors may cause blind regions which reduce the reliability of a surveillance system. Thus, it is important to deploy sensors so that the configuration covers the entire area with the minimum number of sensors. In [84], an optimum algorithm for deploying multiple cameras in parking lots is proposed using field of view (FOV) overlapping constraints. The basic idea of the paper is to deploy one camera on a desired position, then another camera (camera 2) is installed to cover the 25~50% overlap region between the fields of view of camera 1 and camera 2. The rest of the cameras are placed one by one to keep the cameras from having an overlapping region (25~50%). This way, FOV overlapping constraints make the camera calibration more accurate.

3.2. Cooperative camera system

In order to perform visual surveillance with multiple sensors in a wide area, collecting and analyzing data from multiple sensors to obtain meaningful information is important. Recent researches for IVS are mainly focused on cooperative camera systems which integrate data from multiple sensors. In the cooperative camera system, the synchronization of cameras, finding corresponding objects in multiple sensors, and communication method for data transmission are the main issues of research. In [85], an indoor surveillance system consisted of a cooperative camera network (CCN) using a network of nodes, is proposed. In the system, each node is composed of a PTZ camera and a PC, and all nodes are connected to the central console to give information to the human operator. The CCN's purpose is to monitor potential shoplifters in department stores by reporting the presence of people, which is represented by an individual visual tag.

In [86], a surveillance system for parking lots has been developed using a cooperative camera system. The cooperative camera system is consisted of Active Camera Subsystems (ACS) and Static Camera Subsystems (SCS). Data fusion of each multi-tracker is performed using the Mahalanobis distance, and tracking is done through Kalman filtering. At the SCS, object detection and tracking is performed by integrating data from cameras. Once the SCS starts to track the object, the ACS selects the object to capture a high-resolution image. In [87], a human tracking system is proposed using two sets of stereo cameras in the living room. A stereo module is composed of each set of cameras connected to a PC, and a tracker module is composed of two sets of stereo modules connected to one PC. This system outputs the position and identity of a human in the room. The identity of the human is derived by calculating color histogram from the stereo module. Depth information and background subtraction result from the stereo module are used to track the human in the room. The system runs fast enough and tracks multiple people standing, walking, sitting, occluding, entering, and leaving the space. However, if there are people who wear clothes of similar color, the performance is reduced due to the poor color clustering.

In [88], a multi-tracking camera surveillance system for indoor environments is developed. The system divides the tracking task between the cameras by assigning the task to the camera that has better visibility of the object, taking occlusions into account. Each

camera module performs human tracking using the Kalman filter. However, this system has an assumption that FOV overlaps between cameras, thus an object which reappears would be tracked as a new object. MaKris *et al*. [89] propose an outdoor distributed multi-camera tracking system that can track across blind regions without camera calibration. Unsupervised probabilistic learning algorithm is developed to link different camera views using a large amount of observation. In this system, the camera network can automatically learn its structure as an initial step of "plug n play". Installation and tracking across "blind" regions of the network FOV can be supported by providing probabilistic estimates of the location and the time with which a target may reappear. Additionally, this system does not need manual camera calibration, which is a resource consuming process. Thus, the method proposed in [89] provides an efficient and low cost multiple camera system framework for tracking.

In [90], a multi sensor wide area surveillance system is proposed as a part of the VSAM project. This system provides object detection, tracking, and simple activity recognition using calibrated cameras. The distinctive feature of this cooperative camera system is that the localization of an object is determined by a ray intersection interaction with a full terrain model. Moreover, the tracking handoff and sensor slaving algorithm are developed for more robust object tracking. The handoff algorithm is to track objects seamlessly by coordinating multiple sensors with the 3-D real-world location of the object. The sensor slaving method keeps track of all objects in the scene while simultaneously gathering high-resolution. ADVISOR [3] is the metro station surveillance system using multiple cameras. In ADVISOR, human individual tracking, human group tracking, and human action recognition using predefined scenarios are performed in real-time to monitor metro stations. In particular, the crowd monitoring module can recognize crowd behavior such as "overcrowding and blocking of areas," "stationary objects and people," "congestion of pre-defined areas," and "counter-flow" by calculating crowd density.

In [91], a football player tracking system using multiple static cameras is proposed. The algorithm is consisted of two stages; the single camera tracking stage and the multi camera tracking stage. At the single camera stage, tracking is performed using an adaptive background method in the image plane. Data integration for single camera tracking is performed using the Kalman filter to estimate the position and velocity of player. However, to track the player accurately, the homography consistency of cameras must be maintained over time. Also, occlusion between players must be handled at single camera stage since the single camera stage is designed to output one measurement per player; thus, the occluded group is recognized as one player. Mittal and Davis [92] presented a system, called the M2 tracker, which is capable of segmenting, detecting, and tracking multiple people in a cluttered scene using multiple synchronized surveillance cameras located far away from each other. In the M2 tracker, a densely located multiple object can be tracked by segmenting and calculating its position on a 3-D ground plane. The M2 tracker is fully automatic and does not require any manual input or initializations. Furthermore, it is able to handle occlusions and partial occlusions caused by the dense location of multiple objects.

Kang *et al*. [93] proposed a continuous tracking algorithm using a combination of stationary and moving cameras. There are two models to address the tracking problem; the motion model and the appearance model. The motion model is obtained using a Kalman filtering process, which predicts the position of the moving object, while the appearance model is obtained using multiple color distribution components to describe the object. Tracking is performed by maximizing the joint probability of two models. The moving camera and the stationary camera are registered using a homography transform. Occlusion handling, deriving accurate motion measurements, and camera handoff are performed through fusion of these cameras. Javed *et al*. [94] proposed a multi-camera tracking algorithm which can track even when observations of objects are not available for relatively large time periods due to non-overlapping camera views without camera calibration. In this paper, the object tracking of non-overlapping view is performed by learning camera topology and path probability of an object using the Parzen window [95]. During the training phase, inter-camera time intervals, location of exit/entrances, and velocities of objects are jointly modeled to constrain correspondences in the Bayesian framework. Once learning is complete, the object correspondences are assigned using the maximum a posteriori (MAP) estimation framework and learned parameters are updated with trajectory changes.

4. CONCLUSION

Early visual surveillance systems have been highly dependent on human operators when monitoring a visual event and searching for a features in a video database. Those passive surveillance systems have suffered the cost and the efficiency of surveillance inevitably. Therefore, recent advances in visual surveillance have been focused on intelligence techniques including automatic image interpretation and wide area surveillance systems to reduce maintenance cost and dependency on human operators.

To automatically analyze images and extract high-level information, image enhancement, motion detection, object tracking and behavior understanding researches have been actively studied together or separately.

In the research of motion detection on IVS back-ground subtraction has been studied in deep depth re-search because of computational effectiveness and high accuracy. In particular, many researchers have been pay-ing attention to make a model for the background against multi-modal property and background changes by change of weather condition, deformation of object and moving background structure robustly and effectively. In

studies to date, probabilistic approach, statistical approach and pattern recognition approach are major approaches for modeling background. In the probabilistic approach, distribution of background has been estimated with probability theory such as Gaussian mixture model and kernel density estimation. In the statistical approach, statistical properties have used as constraint features for modeling the background such as median, mean, variance of the background. Moreover, pattern recognition theory like spectral analysis of image has been focused as a method for finding background pattern. Eigenbackground is the representative pattern recognition approach by finding the fundamental pattern between image sequences. Furthermore, in recent studies, motion detection algorithm for moving platform has been actively studied to relax limitations of conventional motion detection algorithm such as assumption of using fixed cameras.

The research on object tracking can be classified as point tracking, kernel tracking and contour tracking according to the representation method of a target object. In point tracking approach, statistical filtering method has been used to estimating the state of target object. Kalman filter and particle filter are the most popular filtering method. In kernel tracking approach, various estimating methods are used to find corresponding region to target object. Mean-shift tracking and particle filter tracking are the most famous kernel tracking research recently. Contour tracking can be divided into state-space method and energy function minimization method according to the way of evolving contour. Contour tracking has been applied to track the object with a complex shape and various changes of shape due to a good representation ability of contour. Especially, Condensation algorithm has caused big impact due to its good performance to non-rigid object. The recent issues of object tracking are to find the way of handling the occlusion, tracking non-rigid shape object during changing object shape and illumination of the scene. Particularly, particle filter has been often referenced regardless of representation method of object due to the ability of robust estimation by handling a non-Gaussian distribution.

In the behavior understanding, researches for human body modeling, estimation of body part location, and human activity recognition have been organically related for automatically recognizing the human action. On human body modeling research, stick figure, 2d contour, and 3d volumetric figure are commonly used to representing the human body as a whole unit or separated body part unit. Localizing the human body part is performed with body part tracking method. In the research on estimation of body part location, there are two main classes of estimation, top-down approach and bottom-up approach. Top-down approach matches a projection of the human body with the input image. On the other hand, bottom-up approach tracks human body part by assembling individual body part into the human body model with specific constraints. Moreover, combining top-down and bottom-up approach are widely used in the recent research to compensate for the disadvantages of each approach combining. For human activity recognition, researches can be divided into general feature sequence matching and prior knowledge based approach. In the general feature sequence matching approach, human activity recognition is simply regarded as a classification problem of the feature sequence usually treating time varying feature sequences. Therefore, HMM, DTW and DBN are widely used due to ability for matching time varying. Instead, several studies recognize human activity by using prior knowledge such as physical law, contextual information.

Along with image interpretation task, wide area surveillance task using multiple sensors to widen the scope of the surveillance area has been also under active study, especially synchronizing sensor, effective installation of multiple sensors and integrating numerous data to extract high-level information.

Increasing number of cameras makes the research of camera calibration and the installation method of multiple sensors more important. In a wide area surveillance environment, camera calibration cost is very high because of large number of cameras. Thus a self-calibration of camera has been important research area. Moreover, it is also important to find an optimal deployment of sensors so that the configuration of sensor covers the entire area with the minimum number of sensor.

In the same context, researches for integrating numerous data increased by adding sensors have been important. To solve difficulties of occlusion situation and error correction in motion detection, it is important to attempt of combining contextual information between different types of sensor. Cooperative camera system with PTZ and fixed camera and referencing stereo modules are the representative examples. In particular, seamless tracking via sharing the local network of camera has been intensively studied by integrating data from neighbor cameras. It is also mainly studied to find corresponding object among multiple sensors, estimating camera topology and moving path of object on cooperative camera system research.

We have represented researches related to IVS including image interpretation and wide area surveillance techniques. There have been a lot of researches as reflecting the growing demand and the importance for safety and security. According to our survey, automation of surveillance and reduction of cost are main subject in IVS. In order to make surveillance system automatic, there are a lot of attempts combining pattern recognition and data mining researches based on computer vision techniques. Especially, it will be contribute significantly in improving accuracy and effectiveness of surveillance, if algorithm can handle an uncertainty of the scene such as illumination change, non-rigid object, occlusion between object and an undefined human activity. To reduce the cost of surveillance, distributed system and data communication techniques are well combined based on multiple sensors network. Also, when installation of sensor and integrating data from sensor network can be performed with minimal manual reconfiguration, this will contribute in expanding surveillance area enough.

REFERENCES

[1] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. on Systems, Man, and Cybernetics - Part C*, vol. 34, no. 3, pp. 334-352, August 2004.

[2] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, and L. Wixson, "A system for video surveillance and monitoring," *Carnegie Mellon University Technical Report*, CMU-RI-TR-00-12, 2000.

[3] N. T. Siebel and S. Maybank, "The advisor visual surveillance system." *Proc. of the ECCV Workshop on Applications of Computer Vision*, pp. 103-111, 2004.

[4] C. F. Shu, A. Hampapur, M. Lu, L. Brown, J. Connell, A. Senior, and Y. Tian, "IBM smart surveillance system (S3): an open and extensible framework for event based surveillance," *Proc. of IEEE Conf. Advanced Video and Signal Based Surveillance*, pp. 318-323, 2005.

[5] C. Regazzoni, V. Ramesh, and G. L. Foresti, "Special issue on video communications, processing, and understanding for third generation surveillance systems," *Proc. of the IEEE*, vol. 89, no. 10, pp. 1355-1367, 2001.

[6] V. Kober, "Robust and efficient algorithm of image enhancement," *IEEE Trans. on Consumer Electronics*, vol. 52, no. 2, pp. 655-659, 2006.

[7] M. J. Seow and V. K. Asrai, "Homomorphic processing system and ratio rule for color image enhancement," *Proc. of the IEEE International Conf. Neural Network*, vol. 4, pp. 2507-2511, 2004.

[8] T. C. Aysal and K. E. Barner, "Quadratic weighted median filters for edge enhancement of noisy images," *IEEE Trans. on Image Processing*, vol. 15, no. 11, pp. 3294-3310, November 2006.

[9] S. Thurnhofer and S. K. Mitra, "A general framework for quadratic Voltera filters for edge enhancement," *IEEE Trans. on Image Processing*, vol. 5, no. 6, pp. 950-963, June 1996.

[10] K. E. Barner and T. C. Aysal, "Polynomial weighted median filtering," *Proc. of IEEE International Conf. Acoustics Speech and Signal Processing*, vol. 4, no. 4, pp. 153-156, 2006.

[11] J. Duan and G. Qiu, "Novel histogram processing for color image enhancement," *Proc. of the 3rd International Conf. Image and Graphics*, pp. 55-58, 2004.

[12] Y. T. Kim, "Contrast enhancement using brightness preserving bi-histogram equalization," *IEEE Trans. on Consumer Electronics*, vol. 43, no. 1, pp. 1-8, February 1997.

[13] Y. Wang, Q. Chen, and B. M. Zhang, "Image Enhancement based on equal area dualistic sub-image histogram equalization method," *IEEE Trans. on Consumer Electronincs*, vol. 45, no. 1, pp. 68-75, February 1999.

[14] S. D. Chen and A. R. Ramli, "Minimum mean brightness error bi-histogram equalization in contrast enhancement," *IEEE Trans. on Consumer Electronics*, vol. 49, no. 4, pp. 1310-1319, November 2003.

[15] C. Wang and Z. Ye, "Brightness preserving histogram equalization with maximum entropy: a variational perspective," *IEEE Trans. on Consumer. Electronics,* vol. 51, no. 4, pp. 1326-1334, November 2005.

[16] D. Y. Tsai, Y. B. Lee, M. Sekiya, S. Sakaguchi, and I. Yamada, "A method of medical image enhancement using wavelet analysis," *Proc. of International Conf. Signal Processing,* vol. 1, pp. 723-726, 2002.

[17] S. S. Agaian, K. Panetta, and A. M. Grigoryan, "Transform based image enhancement algorithms," *IEEE Trans. on Image Processing*, vol. 10, no. 3, pp. 367-382, 2001.

[18] S. Aghagolzadeh and O. K. Ersoy, "Transform image enhancement," *Optical Engineering*, vol. 31, no. 3, pp. 614-626, 1992.

[19] F. T. Arslan and A. M. Grigoryan, "Image enhancement by the tensor transform," *Proc. of IEEE International Symp. on Biomedical Imaging: Macro to Nano,* vol. 1, pp. 816-819, April 2004.

[20] F. T. Arslan and A. M. Grigoryan, "Fast splitting alpha-rooting method of image enhancement: tensor representation" *IEEE Trans. on Image Processing*, vol. 15, no. 11, pp. 3375-3384, November 2006.

[21] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780-785, July 1997.

[22] C. Stauffer and E. Grimson, "Learning patterns of activity using real time tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747-757, August 2000.

[23] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: real-time surveillance of people and their activities," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809-830, August 2000.

[24] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831-843, August 2000.

[25] T. Horprasert, D. Harwood, and L. S. Davies, "A robust background subtraction and shadow detection," *Proc. of Asian Conf. Computer Vision*, pp. 8-11, 2000.

[26] A. J. Lipton, H. Fujiyoshi, and R. S. Patil, "Moving target classification and tracking from real-time video," *Proc. of the IEEE Workshop Applications of Computer Vision*, pp. 8-14, 1998.

[27] J. Barron, D. Fleet, and S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12,

no. 1, pp. 42-77, February1994.
[28] A. Bevilacqua and P. Azzari, "High-quality real time motion detection using PTZ cameras," *Proc. of IEEE Conf. Advanced Video and Signal Based Surveillance*, pp. 23-23, 2006.
[29] R. Cucchiara, A. Prati, and R. Vezzani , "Advanced video surveillance with pan tilt zoom cameras," *Proc. of the 6th IEEE International Workshop on Visual Surveillance*, 2006.
[30] C. Micheloni and G. L. Foresti, "A robust feature tracker for active surveillance of outdoor scenes," *Electronic Letters on Computer Vision and Image Analysis*, pp. 21-34, 2003.
[31] D. Murray and A. Basu, "Motion tracking with an active camera," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 449-459, May 1994.
[32] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Proc. of the Imaging understanding workshop*, pp. 121-130, 1981.
[33] J. Shi and C. Tomasi," Good features to track," *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pp. 593-600, 1994.
[34] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: a survey," *ACM Computing Surveys*, vol. 38, no. 4, 2006.
[35] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615-1630, October 2005.
[36] V. Salari and I. Sethi, "Feature point correspondence in the presence of occlusion," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 12, no. 1, pp. 87-91, January 1990.
[37] T. Broida and R. Chellappa, "Estimation of object motion parameters from noisy images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 8, no. 1, pp. 90-99, January 1986.
[38] H. Tanizaki, "Non-gaussian state-space modeling of nonstationary time series," *Journal of the American Statistical Association*, vol. 82, pp. 1032-1063, December 1987.
[39] G. Hager, M. Dewan, and C. Stewart, "Multiple kernel tracking with SSD," *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pp. 790-797, 2004.
[40] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking" *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564-575, May 2003.
[41] R. Collins, "Mean-shift blob tracking through scale space," *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 234-240, June 2003.
[42] T. Lindeberg, "Scale-space theory: A basic tool for analyzing structures at different scales," *Journal of Applied Statistics*, vol. 21, no. 2, pp. 224-270, 1994.
[43] S. Birchfield and S. Rangarajan, "Spatiograms versus histograms for region based tracking," *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1158-1163, 2005.
[44] M. Black and A. Jepson, "Eigentracking: robust matching and tracking of articulated objects using a view-based representation," *International Journal of Computer Vision*, vol. 26, no. 1, pp. 63-84, 1998.
[45] J. Lim, D. Ross, R. Lin, and M. H. Yang, "Incremental learning for visual tracking," *Advances in Neural Information Processing Systems*, pp. 793-800, 2005.
[46] S. Avidan "Support vector tracking," *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pp. 184-191, 2001.
[47] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, 2nd Edition, pp. 318-350, 1998.
[48] N. Paragios and R. Deriche "Geodesic active regions and level set methods for supervised texture segmentation," *International Journal of Computer Vision*, vol. 46, no. 3, pp. 223-247, 2002.
[49] N. Peter Freund, "Robust tracking of position and velocity with Kalman snakes," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 6, pp. 564-569, June 1999.
[50] M. Isard and A. Blake, "Condensation - conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5-28, 1998.
[51] A. Yilmaz, X. Li, and M. Shah, "Contour based object tracking with occlusion handling in video acquired using mobile cameras," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1531-1536, November 2004.
[52] J. Sethian, *Level Set Methods: Evolving Interfaces in Geometry, Fluid Mechanics Computer Vision and Material Sciences*, Cambridge University. Press, 1999.
[53] J. K. Aggarwal and P. Sangho, "Human motion: modeling and recognition of actions and interactions," *Proc. of the 2nd International Symposium* on *3D Data Processing, Visualization and Transmission*, pp. 640-647, September 2004.
[54] K. Sato and J. K. Aggarwal, "Temporal spatio-velocity transform and its applications to tracking and interaction," *Computer Vision and Image Understanding*, vol. 96, no. 2, pp. 100-128, November 2004.
[55] I. Haritaoglu, D. Harwood, and L. Davis, "Hydra: Multiple people detection and tracking using silhouettes," *Proc. of the 2nd IEEE Workshop on Visual Surveillance*, pp. 6-13, 1999.
[56] M. Pantic, A. Pentland, A. Nijholt, and T. Huang, "Human computing and machine understanding of human behavior: a survey," *Proc. of the International Conf. Multimodal interfaces*, pp. 239-248, 2006.
[57] H. Fujiyoshi and A. Lipton, "Real-time human motion analysis by image skeletonization," *Proc. of the Workshop on Applications of Computer Vision*, pp. 15-21, 1998.
[58] S. Ju, M. Black, and Y. Yaccob, "Cardboard

people: a parameterized model of articulated image motion," *Proc. of the IEEE International Conf. Automatic Face and gesture Recognition*, pp. 38-44, 1996.

[59] M. Leung and Y. Yang, "First sight: a human body outline labeling system," *IEEE Trans. on Pattern Analysis Machine Intelligence*, vol. 17, no. 4, pp. 359-377, April 1995.

[60] J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata, "Nonrigid motion analysis: articulated and elastic motion," *Computer Vision and Image Understanding*, vol. 70, no. 2, pp. 142-156, 1997.

[61] J. Park, S. Park, and J. K. Aggarwal, "Model-based human motion capture from monocular video sequences," *Lecture Notes in Computer Science: Computer and Information Sciences*, vol. 2869, pp. 405-412, 2003.

[62] D. Gavrila, *Vision-Based 3-D Tracking of Humans in Action*, Ph.D. thesis, Department of Computer Science, University of Maryland, 1996.

[63] J. O'Rourke and N. Badler, "Model-based image analysis of human motion using constraint propagation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 2, no. 6, pp. 522-536, November 1980.

[64] C. Myers, L. Rabinier, and A. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 28, no. 6, pp. 623-635, December 1980.

[65] L. R Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

[66] K. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*, Ph.D. Thesis, University of California at Berkeley, 2002.

[67] A. Bobick and A. Wilson, "A state-based technique for the summarization and recognition of gesture," *Proc. of the International Conf. Computer Vision*, pp. 382-388, June 1995.

[68] J. Yang, Y. Xu, and C. S. Chen, "Human action learning via hidden markov model," *IEEE Trans. on System, Man and Cybernetics*, vol. 27, no. 1, pp. 34-44, January 1997.

[69] T. Starner, J. Weaver, and A. Pentland, "Real-time American sign language recognition using desk and wearable computer-based video," *IEEE Trans. on Pattern Anal. Mach. Intell*, vol. 20, no. 12, pp. 1371-1375, December 1998.

[70] N. M. Oliver, B. Rosario, and A. P. Pentland, "A bayesian computer vision system for modeling human interactions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831-843, August 2000.

[71] S. Park and J. K. Aggarwal, "Recognition of two-person interactions using a hierarchical Bayesian network," *Proc. of ACM SIGMM International Workshop on Video Surveillance*, pp. 65-76, 2003.

[72] A. F. Bobick and J. Davis, "Real-time recognition of activity using temporal templates," *Proc. of the IEEE CS Workshop on Applications of Computer Vision*, pp. 39-42, 1996.

[73] A. F. Bobick and J. Davis, "Real-time recognition of activity using temporal templates," *Proc. of the IEEE Computer Society Workshop on Computer vision*, pp. 39-42, 1996.

[74] A. D. Wilson, A. F. Bobick, and J. Cassell, "Temporal classification of natural gesture and application to video coding," *Proc. of the IEEE Conf. Computer Vision and Pattern Recognition*, pp. 948-954, 1997.

[75] F. Bremond and G. Medioni, "Scenario recognition in airborne video imagery," *Proc. of the International Workshop Interpretation of Visual Motion*, pp. 57-64, 1998.

[76] S. Intille and A. Bobick, "Representation and visual recognition of complex, multi-agent actions using belief networks," *MIT Technical Report,* no. 454, 1998.

[77] R. Mann, A. Jepson, and J. Siskind, "Computational perception of scene dynamics," *Computer Vision and Image Understanding*, vol. 65, no. 2, pp. 113-128, February 1997.

[78] M. Brand and I. Essa, "Causal analysis for visual gesture understanding," *MIT Technical Report*, 1995.

[79] Y. Ivanov and A. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 852-872, August 2000.

[80] D. Ayers and M. Shah, "Recognizing human action in a static room," *Proc. of the IEEE Computer Society Workshop on Interpretation of Visual Motion*, pp. 42-46, 1998.

[81] R. T. Collins and Y. Tsin, "Calibration of an outdoor active camera system," *Proc. of the Conf. Computer Vision and Pattern Recognition*, pp. 528-534, June 1999.

[82] J. Davis and X. Chen, "Calibrating pan-tilt cameras in wide-area surveillance networks," *Proc. of IEEE International Conf. Computer Vision*, vol. 1, pp. 144-149, 2003.

[83] E. Hemayed, "A survey of camera self-calibration," *Proc. of IEEE Conf. Advanced Video and Signal Based Surveillance*, pp. 351-357, 2003.

[84] I. Pavlidis, V. Morellas, P. Tsiamyrtzis, and S. Harp, "Urban surveillance systems: from the laboratory to the commercial world," *Proc. of the IEEE*, vol. 89, no. 10, pp. 1478-1495, 2001.

[85] I. Paulidis and V. Morellas, "Two examples of indoor and outdoor surveillance systems," in P. Remagnino, G. A. Jones, N. Paragios, and C. S. Regazzoni Eds., *Video-based Surveillance Sys-tems*, Kluwer Academic Publishers, Boston, pp. 39-51, 2002.

[86] C. Micheloni, G. L. Foresti, and L. Snidaro, "A co-operative multicamera system for video-surveillance of parking lots," *Proc. of IEE Symp. on Intelligent Distributed Surveillance Systems*, pp.

21-24, 2003.
[87] J. Krumm, S. Harris, B. Meyers, B. Brumit, M. Hale, and S. Shafer, "Multi-camera multi-person tracking for easy living," *Proc. of the 3rd IEEE International Workshop on Visual Surveillance*, pp. 8-11, 2000.
[88] N. T. Nguyen, S. Venkatesh, G. West, and H. H. Bui, "Multiple camera coordination in a surveillance system," *ACTA Automatica Sinica*, vol. 29, no. 3, pp. 408-421, 2003.
[89] D. MaKris, T. Ellis, and J. Black, "Bridging the gaps between cameras," *Proc. of International Conf. Multimedia and Expo*, June 2004.
[90] R. T. Collins, A. J. Lipton, H. Fujiyoshi, and T. Kanade, "Algorithms for cooperative multisensor surveillance," *Proc. of the IEEE*, vol. 89, no. 10, pp. 1456-1477, 2001.
[91] M. Xu, L. Lowey, and J. Orwell, "Architecture and algorithms for tracking football players with multiple cameras," *Proc. of the IEE Workshop on Intelligent Distributed Surveillance Systems*, pp. 51-56, 2004.
[92] A. Mittal and L. S. Davis, "M2 tracker: a multi-view approach to segmenting and tracking people in a cluttered scene," *International Journal of Computer Vision*, vol. 51, no. 3, pp. 189-203, 2003.
[93] J. Kang, I. Cohen, and G. Medioni, "Continuous tracking within and across camera streams," *Proc. of IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 267-272, 2003.
[94] O. Javed, Z. Rasheed, K. Shafique, and M. Shah, "Tracking across multiple cameras with disjoint views," *Proc. of IEEE International Conf. Computer Vision*, vol. 1, pp.952-957, 2003.
[95] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, 2nd edition, pp. 164-173, 2000.
[96] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. on Systems, Man and Cybernetics - Part C*, vol. 34, no. 3, August 2004.
[97] S. Hong, H. Lee, K.-A. Toh, and E. Kim, "Gait recognition using multi-bipolarized contour vector," *International Journal of Control, Automation, and Systems*, vol. 7, no. 5, pp. 799-808, 2009.

In Su Kim received his B.S in the School of Electrical Engineering and Computer Science from Hanyang University, Seoul, Korea in 2003. He is currently a candidate for Ph.D degree in the School of Electrical Engineering and Computer Science at Seoul National University. His research interests include computer vision, pattern recognition, motion detection and object classification for embedded visual surveillance system.

Hong Seok Choi receives his B.S. and M.S. degrees in the School of Electrical Engineering from Seoul National University, Seoul, Korea, in 2000 and 2002, respectively. He is working toward the Ph.D. degree in the School of Electrical Engineering and Computer Science at Seoul National University. His research interests are in the area of computer vision, pattern recognition, and embedded surveillance system.

Kwang Moo Yi received his B.S. Degree in the department of Electrical Engineering and Computer Science from Seoul National University, Seoul, Korea, in 2007. Currently he is a Ph.D. candidate student in the department of Electrical Engineering and Computer Science from Seoul National University, Seoul, Korea. His research interests include computer vision, visual tracking, motion segmentation, motion detection, unsupervised Bayesian learning, and so on.

Jin Young Choi received his B.S., M.S., and Ph.D. degrees in Electrical Engineering and Computer Science from Seoul National University in 1982, 1984, 1993, respectively. He was a researcher at the Electronics and Telecommunications Research Institute (ETRI) from 1984 to 1994. He was a visiting professor at the University of California, Riverside from 1998 to 1999. He is currently a professor at the School of Electrical Engineering and Computer science at Seoul National University, Korea. He is a director of the Perception and Intelligence Research Center. His research interests include visual surveillance, intelligent systems, and adaptive control.

Seong G. Kong received his B.S. and M.S. degrees in Electrical Engineer-ing from Seoul National University, Seoul, Korea, in 1982 and 1987, respect-ively, and his Ph.D. degree in Electrical Engineering from the University of Southern California, Los Angeles, in 1991. From 1992 to 2000, he was an Associate Professor of Electrical Engineering at Soongsil University, Seoul, Korea. He was Chair of the Department from 1998 to 2000. During 2000-2001, he was with School of Electrical and Computer Engineering at Purdue University, West Lafayette, IN, as a Visiting Scholar. From 2002 to 2007, he was an Associate Professor at the Department of Electrical and Computer Engineering, University of Tennessee, Knoxville. Currently, he is an Associate Professor at the Department of Electrical and Computer Engineering, Temple University, Philadelphia, PA. He published more than 70 refereed journal articles, conference papers, and book chapters in the areas of image processing, pattern recognition, and intelligent systems.Dr. Kong was the Editor-in-Chief of Journal of Fuzzy Logic and Intelligent Systems from 1996 to 1999. He is an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS. He received the Award for Academic Excellence from Korea Fuzzy Logic and Intelligent Systems Society in 2000 and the Most Cited Paper Award from the Journal of Computer Vision and Image Understanding in 2007 and 2008, two years in a row. He is a Technical Committee member of the IEEE Computational Intelligence Society.