A Novel Video Dataset for Change Detection Benchmarking

Nil Goyette, Pierre-Marc Jodoin, *Member, IEEE*, Fatih Porikli, *Fellow, IEEE*, Janusz Konrad, *Fellow, IEEE*, and Prakash Ishwar, *Senior Member, IEEE*

***Abstract*— Change detection is one of the most commonly encountered low-level tasks in computer vision and video processing. A plethora of algorithms have been developed to date, yet no widely accepted, realistic, large-scale video data set exists for benchmarking different methods. Presented here is a unique change detection video data set consisting of nearly 90 000 frames in 31 video sequences representing six categories selected to cover a wide range of challenges in two modalities (color and thermal infrared). A distinguishing characteristic of this benchmark video data set is that each frame is meticulously annotated by hand for ground-truth foreground, background, and shadow area boundaries—an effort that goes much beyond a simple binary label denoting the presence of change. This enables objective and precise quantitative comparison and ranking of video-based change detection algorithms. This paper discusses various aspects of the new data set, quantitative performance metrics used, and comparative results for over two dozen change detection algorithms. It draws important conclusions on solved and remaining issues in change detection, and describes future challenges for the scientific community. The data set, evaluation tools, and algorithm rankings are available to the public on a website[1] and will be updated with feedback from academia and industry in the future.**

***Index Terms*— Change detection algorithms, benchmark testing, video surveillance.**

I. Introduction

DETECTION of change, and in particular motion, is a fundamental low-level task in many computer vision and video processing applications. Examples include visual surveillance (video compression, zone monitoring, people counting, anomaly detection, action recognition, etc.), smart environments (occupancy analysis, parking lot management, etc.), and content retrieval (video annotation, event detection, object tracking, forensic labeling).

Manuscript received March 16, 2013; revised December 14, 2013 and April 28, 2014; accepted June 12, 2014. Date of publication August 7, 2014; date of current version September 23, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Olivier Bernard.

N. Goyette and P.-M. Jodoin are with the Département d'informatique, Facultédes Sciences, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada (e-mail: nil.goyette@usherbrooke.ca; pierre-marc.jodoin@usherbrooke.ca).

F. Porikli is with Mitsubishi Electric Research Laboratories, Cambridge, MA 02139 USA (e-mail: fatihporikli@ieee.org).

J. Konrad and P. Ishwar are with the Department of Electrical and Computer Engineering, Boston University, Boston, MA 02215 USA (e-mail: jkonrad@bu.edu; pi@bu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2346013

[1] www.ChangeDetection.net

Change detection is closely coupled with higher level inference tasks such as detection, localization, tracking, and classification of moving objects, and is often considered to be preprocessing step. Its importance can be gauged by the large number of algorithms that have been developed to-date and the even larger number of articles that have been published on this topic. A quick search for 'motion detection' on IEEE Xplore© returns over 20,000 papers.

Among the many variants of change detection algorithms, there seems to be no single algorithm that competently addresses all of the inherent real world challenges such as sudden illumination variations, background movements, shadows, camouflage effects (photometric similarity of object and background) and ghosting artifacts (delayed detection of a moving object after it has moved away) to name a few.

Due to the tremendous effort required to build an inclusive benchmark dataset that supplies pixel-precision ground truth labels and provides a balanced coverage of the representative challenges, prior attempts to attain an objective evaluation of change detection methods have been confined to limited partial assessments.

The lack of a comprehensive dataset has a number of negative implications. Firstly, it makes it difficult to ascertain with confidence which algorithms would perform robustly when the assumptions they are built upon are violated. Many algorithms tend to overfit specific scenarios. A method tuned to be robust to shadows may not be as robust to background motion. A dataset that contains many different scenarios and applies a variety of performance measures would go a long way towards providing an objective assessment. Secondly, not all authors are willing to (or have the resources to) compare their methods against the most advanced and promising approaches. As a consequence, an overwhelming importance has been accorded to a small subset of easily implementable methods such as [1]–[3] that were developed in the late 1990's. The more recent and advanced methods have been marginalized as a result. Besides, the implementation of the same method varies significantly from one research group to another in the choice of parameters and the use of other pre- and post-processing steps. Thirdly, the fact that authors often use their own data (that are not widely available to everyone) makes a fair comparison much more problematic if not impossible.

Recognizing the importance of change detection to the computer vision and video processing communities, we have assembled a change detection benchmark dataset:

www.ChangeDetection.net (CDnet) that consists of nearly 90,000 frames in 31 video sequences representing 6 video categories (including thermal). We initially reported on this dataset at the 2012 IEEE Change Detection Workshop (within CVPR 2012) [4]. Here we expand on this initial report by providing a more comprehensive review of state-of-the-art and including results for additional 7 methods.

CDnet contains diverse motion and change detection challenges in addition to typical indoor and outdoor scenes that are encountered in most surveillance and smart environments, and video analytics applications using static cameras. A distinguishing feature of the 2012 CDnet is the fact that each image is meticulously annotated for ground-truth foreground, background, and shadow region boundaries; an effort that goes much beyond a simple binary label denoting the presence of the change. The existence of ground-truth masks permits a precise comparison and ranking of change detection algorithms. CDnet also supplies evaluation tools in *Matlab* and *Python* for quantitatively assessing the performance of different methods according to 7 distinct metrics.

The overarching objectives of this paper are:

1) To provide the research community with a rigorous and comprehensive scientific benchmarking facility, a rich dataset of videos, a set of utilities, and an access to author-approved algorithm implementations for testing and ranking of existing and new algorithms for motion and change detection. The already extensive dataset will be regularly revised and expanded with feedback from the academia and industry.
2) To establish, maintain, and update a rank list of the most accurate motion and change detection algorithms in the various categories for years to come.
3) To help identify the remaining challenges in order to provide focus for future research.

Next, we provide an overview of previous motion detection methods, surveys, and existing datasets. Details of the 2012 CDnet follow, including its categories, ground-truth annotations, and performance metrics. We then describe results and compare and contrast the performance of different methods in each category. We conclude with a discussion of solved and unsolved issues in change detection.

II. Previous Work

Instead of undertaking a detailed discussion of the huge body of existing literature on change detection, we briefly summarize the frequently encountered approaches (c.f. Table I), datasets (c.f. Table II), and surveys of change detection. An extensive survey of change detection methods can be found in [5].

A. Motion vs. Change Detection

There is a large and growing body of literature on motion and change detection the in computer vision and video processing communities. Several closely related problems that have been studied in the last two decades including *salient motion detection* [6], *background subtraction* [7]–[9], *change detection* [10]–[12], *foreground detection* [13], *foreground segmentation* [14], and *video-object segmentation* [15] share various objectives as well as core solutions. In comparison to the change detection task, motion detection may extend to camera motion, egomotion, articulated body motion, epipolar geometry, etc. to name a few.

We refer to *change detection* as the process of detecting foreground regions in a video. Such foreground regions typically correspond to moving objects, e.g., people, animals, vehicles, etc., in the scene whose movements change the photometric values of the projected image pixels. We also consider objects that become temporarily motionless and then move (e.g., a car at a traffic light) as part of the foreground. Although moving, we do not consider waves on a water surface, objects shaken by wind (trees, flags, light poles, etc.), moving shadows and object reflections as actual moving objects.

B. Change Detection

Statistical change point analysis (see [16] and references therein) refers to a large body of literature in statistics and signal processing concerning the problem of detecting changes in the statistical properties of time-series data using parametric and nonparametric statistical tests. Typical applications include neurology, bioinformatics, finance, quality control, seismology, etc. In contrast to these applications, the focus of this paper is on change detection in video, where (a) the dimensionality and size of the data can be quite high: typical video contains over one hundred thousand time-series (one time-series for each pixel), each tens of thousands of samples long, (b) there is strong spatially-localized temporal correlation among the time series, and (c) there are multiple changes across space and time, some highly correlated, some not.

Some of the commonly used change detection techniques for video application can be categorized as frame differencing, background modeling and subtraction, motion segmentation, and matrix decomposition. We briefly discuss each of these techniques below.

1) Basic Models: Frame differencing aims to detect changes in the state of a pixel, e.g., due to a moving object, by subtracting the pixel's intensity in the current frame from its intensity in the previous frame or some reference frame. Although this method is computationally very inexpensive, it is sensitive to illumination changes. Another limitation is that it cannot detect a moving object once it stops moving and when the object motion becomes small; instead it typically detects object boundaries, covered and exposed areas due to object motion.

Change detection can be achieved by building a representation of the scene, called background model, and then observing deviations from this model for each incoming frame. A sufficient change from the background model is assumed to indicate a moving object.

The simplest strategy to detect motion is to subtract the pixel's color in the current frame from the corresponding pixel's color in the background model [7]. A temporal median filter can be used to estimate a color-based background model [17]. One can also generalize to other features such

as color histograms [18], [19] and local self-similarity features [20]. In general, these temporal filtering methods are sensitive to global illumination changes and incapable to detect moving objects once they become stationary.

Early approaches use filters to predict background pixel intensities (or colors). For these models, each pixel whose observed color is far from its prediction is assumed to indicate motion. In [21] and [22], a Kalman filter is used to model background dynamics. Similarly, in [23] Wiener filtering is used to make a linear prediction at pixel level. The main advantage of these methods are their ability to cope with background changes (whether it is periodic or not) without having to assumed any parametric distribution. In case most of the pixels in a frame exhibit sudden change, the background models are assumed to be no longer valid at frame level. At this point, either a previously stored pixel-based background model is swapped in, or the model is reinitialized.

A motion history image [24], [25] is obtained by successive layering of frame differences. For each new frame, existing frame differences are scaled down in amplitude, subject to some threshold, and the new motion label field is overlaid using its full amplitude range. In consequence, image dynamics ranging from two consecutive frames to several dozen frames can be captured in a single image.

More complex background models can be used. For example, Hermade-Lopez and Rivera [26] implement a background model designed to be robust to illumination changes, cast shadows and camouflage. As opposed to other basic methods which detect motion in a maximum likelihood matter (which is sensitive to noise and isolated artifacts) they implement a quadratic Markov measure field (QMMF) which enforces spacial regularity with the help of a quadratic programming solved.

2) Parametric Background Modeling and Subtraction: In order to learn changes in time, a single Gaussian distribution was proposed to model each pixel [3]. Once the parameters of the Gaussian model have been updated over several consecutive frames, the likelihood of the current pixel color coming from this model is determined. The pixels that deviate significantly from their models are labeled as the foreground pixels. Since pixels in noisy areas are given a larger standard deviation, a larger color variation is needed in those areas to detect motion. This is a fundamental difference with the basic models for which the tolerance is fixed for every pixel. As shown by Kim *et al.* [27], a generalized Gaussian model can also be used and Morde *et al.* [25] have shown that a Chebychev inequality can also improve results. With this model, the detection criteria depends on how many standard deviations a color is from the mean.

A single Gaussian, however, is not a good model for dynamic scenes [28] as multiple colors may be observed at a pixel due to repetitive object motion, shadows or reflectance changes. A substantial improvement is achieved by using multiple statistical models to describe background color. A Gaussian Mixture Model (GMM) [29] was proposed to represent each background pixel. GMM compares each pixel in the current frame with every model in the mixture until a matching Gaussian is found. If a match is found, the mean and variance of the matching Gaussian are updated, otherwise a new Gaussian with the mean equal to the current pixel color and some initial variance is introduced into the mixture. Instead of relying on only one pixel, GMM can be trained to incorporate extended spatial information [30].

Several papers [31] improved the GMM approach to add robustness when shadows are present and to make the background models more adaptive to parasitic background motion. A recursive method with an improved update of the Gaussian parameters and an automatic selection of the number of modes was presented in [32]. Haines *et al.* [33] also propose an automatic mode selection method, but with a Dirichlet process. A splitting GMM that relies on a new initialization procedure and a mode splitting rule was proposed in [34] and [35] to avoid over-dominating modes and resolve problems due to newly static objects and moved away background objects while a multi-resolution block-based version was introduced in [36]. The GMM approach can also be expanded to include the generalized Gaussian model [37]. Let us mention that some GMM methods with an automatic mode splitting and merging procedure are sometimes considered as non-parametric methods [33].

As an alternative to mixture models, Bayesian approaches have been proposed. In [13], each pixel is modeled as a combination of layered Gaussians. Recursive Bayesian update instead of the conventional expectation maximization fitting is performed to update the background parameters and better preserve the multi-modality of the background model. A similar Bayesian decision rule with various features and a learning method that adapt to both sudden and gradual illumination changes in used in [38].

Another alternative to GMM is background clustering. In this case, each background pixel is assigned a certain number of clusters depending on the color variation observed in the training video sequence. Then, each incoming pixel whose color is close to a background cluster is considered part of the background. The clustering can be done using K-means (or a variant of it) [39], [40] or codebook [41].

3) Non-Parametric and Data-Driven Background Modeling: In contrast to parametric models, non-parametric kernel density estimation (KDE) fits a smooth probability density function to a time window with previously-observed pixel values at the same location [8]. During the change detection process, a new-frame pixel is tested against its own density function as well as those of pixels nearby. This increases the robustness against camera jitter or small movements in the background. Similar effects can be achieved by extending the support to larger blocks and using texture features that are less sensitive to inter-frame illumination variations.

Although nonparametric models are robust against small changes, they are expensive both computationally and in terms of memory use. Moreover, extending the support causes small foreground objects to disappear. As a consequence, several authors worked to improve the KDE model. For instance, a multi-level method [42] makes KDE computationally independent of the number of samples. A trend feature is used to reliably differentiate periodic background motion from illumination changes [43].

TABLE I
OVERVIEW OF 6 FAMILIES OF MOTION DETECTION METHODS

Background model	References
Basic	Running average [7], [18], [19], [20], [26] Temporal median [17] Motion history image [24], [25] Kalman filter [21], [22] Weiner filter [23]
Parametric	Single Gaussian [3] Gaussian Mixture Model (GMM) [29], [30], [31], [32], [35], [34], [36], [33] Background clustering [39], [40], [41] Generalized Gaussian Model [37], [27] Bayesian [13], [38] Chebyshev inequality [25]
Non-Parametric & Data-driven	Kernel Density Estimation (KDE) [8], [42], [43], [50] Cyclostationary [12] Stochastic K-nearest neighbors (KNN) [9], [45]. Deterministic KNN [32] Bayesian [33] Local binary patterns [47] Hidden Markov Model (HMM) [46]
Matrix Decomposition	Principal Component Analysis (PCA) [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61] Sparsity and dictionary learning [62], [63]
Motion Segmentation	Optical flow segmentation [6], [64], [65] GMM and Optical flow segmentation [40], [66]
Machine Learning	1-Class SVM [67] SVM [68], [69], [70] Neural networks [71], [72], [14], [73]

Recently, data-driven methods using random samples for background modeling have shown robustness to several types of error sources. For example, ViBe [9], [44] not only shows robustness to background motion and camera jitter but also to ghosting artifacts. Ref. [45] shows robustness on a variety of difficult scenarios due to its ability to tune its decision threshold and learning rate based on previous decisions made by the system. In both [9] and [45], a pixel is declared as foreground if it is not close to a sufficient number of background samples from the past.

A shortcoming of the above methods is that they do not account for any "temporal correlation" within video sequences, thus they are sensitive to periodic (or near-periodic) background motion. This prevents them from detecting a structured or near-periodic changes, for example alternating light signals at an intersection, motion of plants driven by wind, the appearance of rotating objects, etc. A cyclostationary background generation method based on frequency decomposition that explicitly harnesses the scene dynamics is proposed in [12]. In order to capture the cyclostationary behavior at each pixel, spectral coefficients of temporal temporal intensity profiles are computed in temporal windows and a background model that is composed of those coefficients is maintained and fused with distance maps to eliminate trail effects. An alternative approach is to use a Hidden Markov Model (HMM) with discrete states to model the intensity variations of a pixel in an image sequence. State transitions can then be used to detect changes [46]. The advantage of using HMMs is that certain events, which may not be modeled correctly by unsupervised algorithms, can be learned using the provided training samples. Some authors, such as Yao and Odobez [47], improved the Local Binary Patterns approach of Heikkila and Pietikainen [48] to be more robust to dynamic scenes.

4) Motion Segmentation: Motion segmentation refers to the assignment of groups of pixels to various classes based on the speed and direction of their movements [64]. Most approaches to motion segmentation first seek to compute optical flow from an image sequence. Discontinuities in the optical flow can help in segmenting images into regions that correspond to different objects. In [6], temporal consistency of optical flow over a narrow time window is estimated; areas with temporally-consistent optical flow are deemed to represent moving objects and those exhibiting temporal randomness are assigned to the background.

Optical flow based methods will be erroneous if brightness constancy or velocity smoothness assumptions are violated. In real imagery, such violations are quite common. Typically, optical flow methods fail in low-texture areas,around moving object boundaries, at depth discontinuities, etc. Due to the commonly imposed regularization term, most optical flow methods produce an over smooth optical flow near boundaries. This produces a halo artifact around moving objects. The resulting errors may propagate across the entire optical flow solution. As a solution, some authors [40], [65] use motion segmentation and optical flow in combination with a color-based GMM model.

5) Matrix Decomposition: Instead of modeling the variation of individual pixels, the whole image can be vectorized and used in background modeling. In [50], a holistic approach using eigenspace decomposition is proposed. For a certain number of input frames, a background matrix (called *eigen background*) is formed by arranging the vectorized representations of images in a matrix where each vectorized image is a column. An eigenvalue decomposition via Principal Component Analysis (PCA) is performed on the covariance of this matrix. The background is then represented by the most descriptive eigenvectors that encompass all possible illuminations to decrease sensitivity to illumination.

It is well known that PCA methods suffer from fundamental limitations [55], [57]. First, a video clip used to compute the eigen background should not contain large moving objects as otherwise the background will likely be corrupted. Second, most PCA methods are geared towards grayscale videos since the integration of RGB values is not trivial. Third, the PCA background model is unimodal and does not account well for videos with a dynamic background. Another limitation for most PCA methods is their sensitivity to outliers (due to a quadratic objective function) and their need to store entire video in memory in order to compute the background model and perform background subtraction.

Typical solutions to some of these problems are the so-called robust-PCA methods. These methods either replace the quadratic objective function by a robust function [73] or decompose the training video into the sum of a low-rank sparse matrix via the minimization of a L1 cost function [59], [60]. Guyon *et al.* [56] introduced a spatial term in the L1 cost function in order to add robustness to dynamic backgrounds, while Seidel *et al.* [55] showed that a weighted Lp cost function can further improve results. Some methods

[52], [54], [55], [58], [60] incorporate an online updating scheme to avoid storing the entire video in memory. Xu *et al.* [51] proposed a variation of the eigen background model which includes a recursive error compensation step for more accurate detection. Others, such as Doug *et al.* [53], proposed illumination invariant approach based on a multi-subspace PCA, each subspace representing different lighting conditions.

Instead of the conventional background and foreground definition, Porikli [61] decomposes an image into "intrinsic" background and foreground images. The multiplication of these images reconstructs the given image. Inspired by the sparseness of the intensity gradient, it applies a spatial derivative filters in the log domain to a subset of the previous video frames to obtain intensity gradient. Since the foreground gradients of natural images are Laplacian distributed and independent, the Maximum Likelihood (ML) estimate of the background gradient can be obtained by a median operator and the corresponding foreground gradient is computed. These gradient results are used to reconstruct the background and foreground intensity images using a reconstruction filter and inverse log operator. This intrinsic decomposition is shown to be robust against sudden and severe illumination changes, but it is computationally expensive.

Another background subtraction approach based on the theory of sparse representation and dictionary learning is proposed in [62]. This method makes the following two important assumptions: (1) the background of a scene has a sparse linear representation over a learned dictionary; (2) the foreground is sparse in the sense that a majority of the pixels of the frame belong to the background. These two assumptions enable handling both sudden and gradual background changes.

6) Machine Learning: Motion detection methods in this category use machine learning discriminative tools such as SVM and neural networks to decide whether or not a pixel is in motion. The parameters of these functions are learned given a training video. Lin *et al.* [67] use a probabilistic SVM to initialize the background model. They use the magnitude of optical flow and inter-frame image difference as features for classification. Han and Davis [69] model the background using a kernel density approximation with multiple features (RGB, gradient, and Haar) and employ a Kernel-SVM as a discriminative function. A somewhat similar approach has also been proposed by Hao [68]. These approaches are typical machine learning methods that need positive and negative examples for training. This is a major limitation for any practical implementation since very few videos come with manually labeled data. As a solution, Chen *et al.* [66] proposed a GPU-based 1-class SVM method called SILK. This method does not need pre-labeled training data, and permits online updating of the SVM parameters. Maddalena and Petrosino [70], [71] model the background of a video with the weights of a neural network. A very similar approach but with a post-processing MRF stage has been proposed by Schick *et al.* [14]. Results reported in the paper show great compromise between processing speed and robustness to noise and background motion. Gregorio and Giordano [72] propose a weightless neural network approach called CwisarD.

Without aiming to be exhaustive, we list below some of the most widely used datasets and describe their characteristics:

- Wallflower [23]: This is a fairly well-known dataset that continues to be used today. It contains 7 short video clips, each representing a specific challenge such as illumination change, background motion, etc. Only one frame per video has been labeled.
- PETS [74]: The Performance Evaluation of Tracking and Surveillance (PETS) program was launched with the goal of evaluating visual tracking and surveillance algorithms. The program has been collecting videos for the scientific community since the year 2000 and now contains several dozen videos. Many of these videos have been manually labeled by bounding boxes with the goal of evaluating the performance of tracking algorithms.
- CAVIAR[2]: This dataset contains more than 80 staged indoor videos representing all kinds of human behavior such as walking, browsing, shopping, fighting, etc. Like the PETS dataset, a bounding box is associated with each moving character.
- i-LIDS[3]: This dataset contains 4 scenarios (parked vehicle, abandoned object, people walking in a restricted area, doorway). Due to the size of the videos (more than 24 hours of footage) the videos are not fully labeled.
- ETISEO[4]: This dataset contains more than 80 video clips of various indoor and outdoor scenes. Since the ground truth consists mainly of high-level information such as the bounding box, object class, event type, etc., this dataset is more suitable for tracking, classification and event recognition than change detection.
- ViSOR 2009[5] [75] is a web archive whose goal is to collect, annotate, store and share surveillance videos. More than 500 videos can be downloaded, all annotated with bounding boxes. These videos are short indoor and outdoor clips (usually less than 10 seconds) often showing staged human actions and interactions. Although it has a benchmarking section, it has been left "under construction" since 2009.
- BEHAVE 2007[6]: A dataset containing 7 real videos shot by the same camera showing various human interactions such as walking in a group, chasing each other, meeting, splitting, etc. Ground truth consists of bounding boxes surrounding moving objects.
- VSSN 2006[7]: This dataset contains 9 semi-synthetic videos composed of a real background and artificially-moving objects. The videos contain animated background, illumination changes and shadows, however they do not contain any frames void of activity.
- IBM[8]: This dataset contains 15 indoor and outdoor videos taken from PETS 2001 plus additional videos. For each

[2]homepages.inf.ed.ac.uk/rbf/CAVIARDATA1
[3]hwww.homeoffice.gov.uk/science-research/hosdb/i-lids
[4]www-sop.inria.fr/orion/ETISEO
[5]www.openvisor.org/
[6]groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/
[7]mmc36.informatik.uni-augsburg.de/VSSN06_OSAC
[8]www.research.ibm.com/peoplevision/performanceevaluation.html

TABLE II
OVERVIEW OF 15 VIDEO DATASETS

Dataset	Description	Ground truth
CD.net 2012	31 videos in 6 categories : baseline, dynamic background, camera jitter, shadow, intermittent motion, and thermal.	Pixel-based labeling of 71,000 frames.
Wallflower [23]	7 short video clips, each representing a specific challenge such as illumination change, background motion, etc	Pixel-based labeling of one frame per video.
PETS [75]	Many videos aimed at evaluating the performance of tracking algorithms	Bounding boxes.
CAVIAR	80 staged indoor videos representing different human behaviors such as walking, browsing, shopping, fighting, etc.	Bounding boxes.
i-LIDS	Very long videos meant for action recognition showing parked vehicle, abandoned object, people walking in a restricted area, and doorway	Not fully labeled.
ETISEO	More than 80 videos meant to evaluate tracking and event detection methods.	High-level label such as bounding boxes, object class, event type, etc.
ViSOR 2009 [76]	Web archive with more than 500 short videos (usually less than 10 seconds)	Bounding boxes.
BEHAVE 2007	7 videos shot by the same camera showing human interactions such as walking in group, meeting, splitting, etc.	Bounding boxes.
VSSN 2006	9 semi-synthetic videos composed of a real background and artificially-moving objects. The videos contain animated background, illumination changes and shadows, however they do not contain any frames void of activity.	Pixel-based labeling of each frame.
IBM	15 videos taken from PETS 2001 plus additional videos.	Bounding box around each moving object in 1 frame out of 30.
Karlsruhe	4 grayscale videos showing traffic scenes under various weather conditions.	10 frames per video have pixel-based labeling.
Li *et al.*[38]	10 small videos with illumination changes and dynamic backgrounds.	10 frames per video have pixel-based labeling.
Karaman [77]	5 videos coming from different sources (the web, the "art live" project, etc.) with various illumination conditions and compression artifacts	Pixel-based labeling of every frame.
cVSG 2008 [15]	15 Semi-synthetic videos with various levels of textural complexity, background motion, moving object speed, size and interaction.	Pixel-based labeling obtained by filming moving objects (mostly humans) in front of a blue-screen and then pasted on top of background videos.
LIMU	8 simple indoor/outdoor videos, some borrowed from PETS2001.	pixel-based labeling for 1 frame out of 15.
USCD	18 short videos with strong background motion and/or camera motion.	pixel-based labeling.
SZTAKI	5 indoor/outdoor videos with shadows	pixel-accurate labeling of foreground and shadows for a subset of frames .
BMC 2012 [78]	29 outdoor videos, most being synthetic.	Pixel-based labeling for 10 synthetic videos and 9 real videos. Ground truth of real videos is for a small subset of images.
Brutzer *et al.* [79]	Computer-generated videos showing a 3D scene representing a street corner. The sequences include illumination changes, dynamic background, shadows, and noise.	Pixel-based labeling.

video, 1 frame out of 30 is labeled with a bounding box around each foreground moving object.

- Karlsruhe[9]: This dataset contains 4 grayscale videos from the *Institut für Algorithmen und Kognitive Systeme* [79]. These videos show traffic scenes under various weather conditions. The authors ground-truth labeled 10 frames for each video.
- Li *et al.* [38][10]: In order to validate their Bayesian motion detection method, they used a dataset made of 10 indoor/outdoor videos containing illumination changes and dynamic backgrounds. The videos are low-resolution (usually 160×120) and only 10 frames for each video sequence have been ground-truth labeled.
- Karaman *et al.* [76]: A dataset made up of 5 videos coming either from the web, from the "art live" project,[11] or from their own dataset. These videos contain various illumination conditions and compression artifacts. Videos have been manually labeled.
- cVSG 2008: In 2008, Tiburzi *et al.* [15] proposed a semi-synthetic[12] dataset consisting of 15 videos. To simplify the ground-truth labeling operation, the foreground moving objects (mostly humans) have been filmed in front of a blue-screen and then pasted on top of background videos. The resulting videos show various levels of textural complexity, background motion, moving object speed, size and interaction. Unfortunately, the web site does not allow performance evaluation.
- LIMU[13]: dataset from Kyushu University containing 8 simple indoor/outdoor videos, some being home-made while others have been borrowed from PETS2001. Pixel-accurate ground truth maps provided for 1 frame out of 15.
- USCD[14]: contains 18 short videos (less than 200 frames) showing scenes with strong background motion and/or camera motion.
- SZTAKI[15]: contains 5 videos sequences with manual groundtruth for foreground moving objects and shadows. The dataset can only be downloaded with a password provided by the owner.
- BMC 2012 [77][16]: dataset created for the Background Models Challenge (BMC) of the 2012 ACCV conference. It consists of 29 outdoor videos, some being synthetic.

[9] i21www.ira.uka.de/image_sequences/
[10] perception.i2r.a-star.edu.sg/bk_model/bk_index.html
[11] www.tele.ucl.ac.be/PROJECTS/art.live/
[12] www-vpu.ii.uam.es/CVSG/
[13] limu.ait.kyushu-u.ac.jp/dataset/en/
[14] svcl.ucsd.edu/projects/background_subtraction/
[15] web.eee.sztaki.hu/∼bcsaba/FgShBenchmark.htm
[16] bmc.univ-bpclermont.fr/

Ground truth is available for 10 synthetic videos showing two scenes: a street and a roundabout. The main challenge for these 10 videos is related to various illumination conditions. Ground truth is also available for real images although only a small subset of images have been labeled. Although they provide a software to compute Recall, Precision, F-measure, and PSNR, they provide to means of ranking methods.

- Brutzer *et al.*, 2011 [78]: Stuttgart Artificial Background Subtraction Dataset released in 2011 in conjunction with a CVPR survey paper. It contains 9 videos showing the same street corner but with different challenges (light on/off, compression artefats, etc.). Pixel-accurate ground truth is available for all sequences. As for BMC, they provide a software to compute precision, recall and F-measure values but do not explain how methods can be ranked besides the use of precision-recall curves.

Additional details regarding these datasets can be found on a web page of the European CANTATA project.[17] Many of these datasets have ground-truth information represented in terms of the bounding box for each foreground object. Furthermore, the focus of several datasets is more on tracking as well as human behavior and interaction recognition than change detection. As such, the above datasets do not contain the diversity of video categories present in the new dataset.

C. Survey Papers

To date, a number of survey papers have been written on the topic of change detection algorithms. Below, we list key survey papers that are devoted to the comparison and ranking of change and motion detection algorithms. Most of these papers use their own datasets.

- Benezeth *et al.*, 2010 [7] used a collection of 29 videos (15 camera-captured, 10 semi-synthetic, and 4 synthetic) taken from PETS 2001, the IBM dataset, and the VSSN 2006 dataset. The authors also used semi-synthetic videos composed of synthetic foreground objects (people and cars) moving over a camera-captured background.
- Bouwmans *et al.*, 2008 [80] surveyed GMM methods and used the Wallflower dataset as a benchmark.
- Bouwmans *et al.*, 2011 [5] presented one of the most complete surveys to date with more than 350 references. The paper reviewed methods spanning 6 motion detection categories and the features used by each method. The survey also listed a number of typical challenges and gave insights into memory requirements and computational complexity. The Wallflower dataset was used to compare different methods.
- Nascimento and Marques, 2006 [81] used a single PETS 2011 video sequence which they manually labeled at pixel resolution using a graphical editor.
- Hassanpour *et al.*, 2011 [82] compare 7 commonly-used methods based on time consumption, memory usage and accuracy. They only account for two 100-frame-long grayscale videos and report accuracy for only 4 frames.
- Brutzer *et al.*, 2011 [78] used a synthetic (computer-generated) dataset produced from only one 3D scene representing a street corner. The sequences included illumination changes, dynamic background, shadows and noise, while lacking frames with no activity.
- Prati *et al.*, 2001 [83] used indoor sequences containing one moving person that were manually segmented into foreground (human), shadow, and background areas. Only 112 frames were ground-truth labeled.
- Rosin and Ioannidis, 2003 [10] used a labeling program that automatically locates moving objects based on their position in space and properties such as color, size, shape, etc. These properties were not used by the change detection algorithms tested. However, the videos used were not realistic as they were limited to lab scenes with balls rolling on the floor.
- Bashir and Porikli, 2006 [84] conducted a performance evaluation of tracking algorithms using the PETS 2001 dataset by comparing the detected bounding box locations with the ground-truth.
- Parks and Fels, 2008 [85] benchmarked 7 motion detection methods and evaluated the influence of post-processing on their performance. They used 7 outdoor and 6 indoor videos containing different challenges such as dynamic backgrounds, shadows and various lighting conditions.
- Piccardi [86] reviewed 7 background subtraction methods and highlighted their strengths and weaknesses. Although no quantitative evaluation was provided, the paper included a formal investigation of computational complexity and memory requirements.
- Radke *et al.* [11] performed an extensive survey of a wide-range of algorithms devoted to the detection of all kinds of changes in images. Most of the discussion in the paper was related to background subtraction methods, pre- and post-processing, and methodologies for evaluating performance. No quantitative evaluation was included.

At a high level, the existing surveys suffer from three main limitations. First, the usual statistics reported in these papers, if any, were not computed on a well-balanced dataset composed of real (camera-captured) videos. Typically, synthetic videos, real videos with synthetic moving objects pasted in, or real videos out of which only 1 frame was manually segmented for ground truth were used. Furthermore, very few datasets contained more than 10 videos. Secondly, none of the papers was accompanied by a fully-operational website that allows users to upload their results and compare them against those of others. Thirdly, the survey papers often reported common, fairly simple motion detection methods, and did not report the performance of more complex methods.

III. The 2012 CDnet Dataset

The 2012 CDnet dataset consists of 31 videos depicting indoor and outdoor scenes with boats, cars, trucks, and pedestrians that have been captured in different scenarios and contain a range of challenges. The videos have been obtained

[17] www.hitech-projects.com/euprojects/cantata/datasets_cantata/

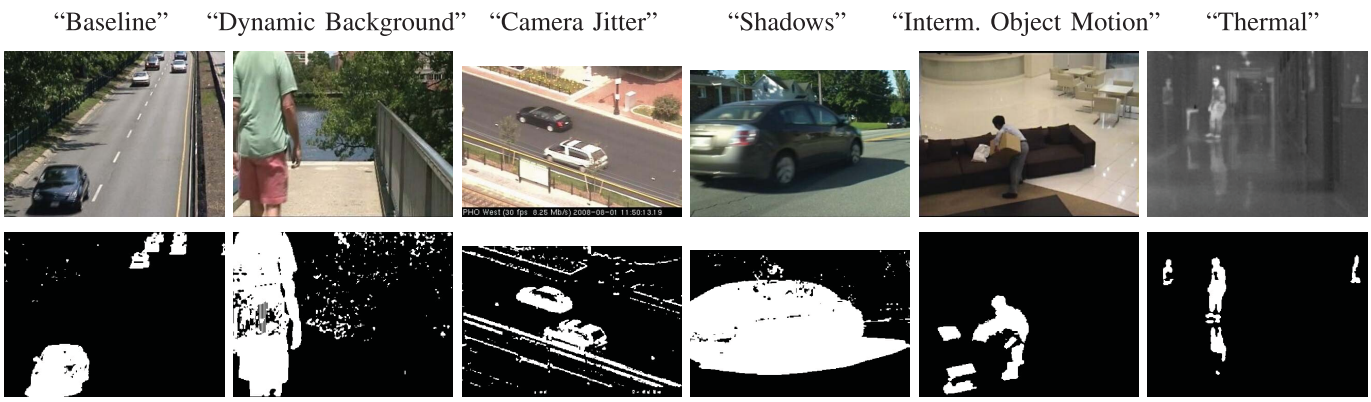

Fig. 1. Sample video frames from each of the 6 categories in the 2012 dataset available at www.changedetection.net and typical detection results obtained using basic background subtraction [7] reported in the last row of Table III.

with different cameras ranging from low-resolution IP cameras, through mid-resolution camcorders and PTZ cameras, to thermal cameras. As a consequence, spatial resolutions of the videos in CDnet vary from 320×240 to 720×576. Also, due to diverse lighting conditions present and compression parameters used, the level of noise and compression artifacts varies from one video to another. The length of the videos also varies from 1,000 to 8,000 frames and the videos shot by low-end IP cameras suffer from noticeable radial distortion. Different cameras may have different hue bias (due to different white balancing algorithms employed) and some cameras apply automatic exposure adjustment resulting in global brightness fluctuations in time. We believe that the fact that our videos have been captured under a range of settings will help prevent this dataset from favoring a certain family of change detection methods over others.

The videos are grouped into six categories according to the type of challenge each represents. We selected videos so that the challenge in one category is unique to that category. For example, only videos in the "Shadows" category contain strong shadows and only those in the "Dynamic Background" category contain strong parasitic background motion. Such a grouping is essential for a clear identification of the strengths and weaknesses of different change detection methods. With the exception of one video in the "Baseline" category, that comes from the PETS 2006 dataset, all the videos have been captured by the authors.

A. Video Categories

31 videos totaling nearly 90,000 frames are grouped into 6 categories (Fig. 1) that have been selected to cover a wide range of change detection challenges representative of typical visual data captured today in surveillance, smart environment, and video analytics applications. These 6 categories are:

1) Baseline: This category contains four videos, two indoor and two outdoor. These videos represent a mixture of mild challenges typical of the next 4 categories. Some videos have subtle background motion, others have isolated shadows, some have an abandoned object and others have pedestrians that stop for a short while and then move away. These videos are fairly easy, but not trivial, to process, and are provided mainly as reference.
2) Dynamic Background: There are six videos in this category depicting outdoor scenes with strong (parasitic) background motion. Two videos represent boats on shimmering water, two videos show cars passing next to a fountain, and the last two depict pedestrians, cars and trucks passing in front of a tree shaken by the wind (second column in Fig. 1).
3) Camera Jitter: This category contains one indoor and three outdoor videos captured by unstable (e.g., vibrating) cameras. The jitter magnitude varies from one video to another.
4) Shadows: This category consists of two indoor and four outdoor videos exhibiting strong as well as faint shadows. Some shadows are fairly narrow while others occupy most of the scene. Also, some shadows are cast by moving objects while others are cast by trees and buildings.
5) Intermittent Object Motion: This category contains six videos with scenarios known for causing "ghosting" artifacts in the detected motion, i.e., objects move, then stop for a short while, after which they start moving again. Some videos include still objects that suddenly start moving, e.g., a parked vehicle driving away, and also abandoned objects. This category is intended for testing how various algorithms adapt to background changes. One example of such a challenge is shown in the 5-th column of Fig. 1 where new objects are added to or existing objects are removed from the scene.
6) Thermal: In this category, five videos (three outdoor and two indoor) have been captured by far-infrared cameras. These videos contain typical thermal artifacts such as heat stamps (e.g., bright spots left on a seat after a person gets up and leaves), heat reflection on floors and windows (see the last column of Fig. 1), and camouflage effects, when a moving object has the same temperature as the surrounding regions.

We would like to mention that although camouflage, caused by moving objects that have very similar color/texture to the background, is among the most glaring change

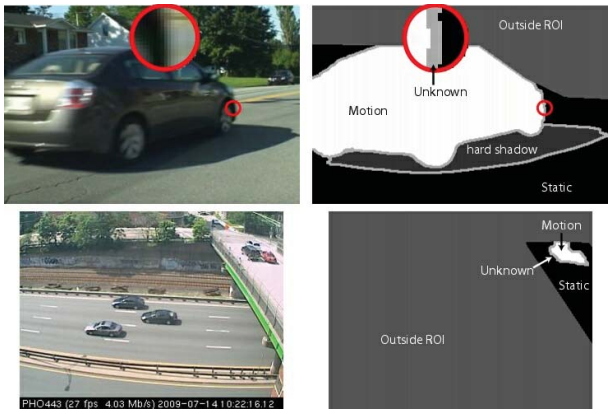

Fig. 2. Sample video frames from the *Bungalows* and *Street light* sequences and corresponding 5-class ground-truth label fields.

detection issues, we have not created a camouflage category. This is partially because almost every real video sequence contains some level of camouflage. It is difficult to create a dataset in which there is a category exclusively with camouflage challenges while other categories are void of it.

B. Ground-Truth Labels

As mentioned in Section II.B., the current online datasets have been designed mainly for testing tracking and scene understanding algorithms, and thus the ground truth is provided in the form of bounding boxes. Although this can be used to validate change detection methods, a precise validation requires ground truth at pixel resolution. Therefore, ideally, videos should be labeled a number of times by different persons and the results averaged out. This, however, is impractical due to resource and time constraints. Furthermore, it is very difficult for a person to produce uncontroversial binary ground-truth images for camera-captured videos. This is particularly difficult near moving object boundaries and in semi-transparent areas.

Due to motion blur and partially-opaque objects (e.g., sparse bushes, dirty windows, fountains), pixels in these areas may contain both the moving object and background. As a consequence, one cannot reliably classify such pixels as belonging to either *Static* or *Moving* class. Since these areas carry a certain level of uncertainty, evaluation metrics should not be computed for pixels in these areas. Therefore, we decided to produce ground-truth images with the following labels:

1) Static: assigned grayscale value of 0,
2) Shadow: assigned grayscale value of 50,
3) Non-ROI[18]: assigned grayscale value of 85,
4) Unknown: assigned grayscale value of 170,
5) Moving assigned grayscale value of 255.

The *Static* and *Moving* classes are associated with pixels for which the motion status is obvious. The *Shadow* label is associated with hard and well-defined moving shadows such as the one in Fig. 2. Hard shadows are among the most difficult artifacts to cope with and we believe that adding this extra information improves the richness and utility of the dataset (please note that evaluation metrics discussed in Section III-C consider the *Shadow* pixels as *Static* pixels). The *Unknown* label is assigned to pixels that are half-occluded or corrupted by motion blur. All pixels located close to moving-object boundaries are automatically labeled as *Unknown* (Fig. 2). This prevents evaluation metrics from being corrupted by pixels whose status is unclear.

The *Non-ROI* (not in region of interest) label serves two purposes. Firstly, since most change detection methods incur a delay before their background model stabilizes, we labeled the first few hundred frames of each video sequence as *Non-ROI*. This prevents the corruption of evaluation metrics due to errors during initialization. Secondly, the *Non-ROI* label prevents the metrics from being corrupted by activities unrelated to the category considered. An example of this situation is shown in the second row of Fig. 2, which illustrates a sequence of cars that arrive, stop at a street light and then move away. The goal of the video is to measure how well a change detection method can handle intermittent motion. However, since the scene is cluttered with unrelated activities (cars on the highway) the *Non-ROI* label puts the focus on street-light activities. Similarly, the top row in Fig. 2 illustrates the *Shadow* category; the *Non-ROI* label is used to prevent the metrics from corruption by trees moving in the background.

C. Evaluation Metrics

Finding the right metric to accurately measure the ability of a method to detect motion or change without producing excessive false positives and false negatives is not trivial. For instance, recall favors methods with a low False Negative Rate. On the contrary, specificity favors methods with a low False Positive Rate. Having the entire precision-recall tradeoff curve or the ROC curve would be ideal, but not all methods have the flexibility to sweep through the complete gamut of tradeoffs. In addition, one cannot, in general, rank-order methods based on a curve. We deal with these difficulties by reporting the average performance of each method for each video category with respect to 7 different performance metrics each of which has been well-studied in the literature. Specifically, for each method, each video category, and each metric, we report the performance (as measured by the value of the metric) of the method averaged across all the videos of the category.

Let TP = number of true positives, TN = number of true negatives, FN = number of false negatives, and FP = number of false positives. The 7 metrics that we use are:

1) Recall (Re): $TP/(TP + FN)$
2) Specificity (Sp): $TN/(TN + FP)$
3) False Positive Rate (FPR): $FP/(TP + FN)$
4) False Negative Rate (FNR): $FN/(TN + FP)$
5) Percentage of Wrong Classifications (PWC): $100(FN + FP)/(TP + FN + FP + TN)$
6) Precision (Pr): $TP/(TP + FP)$
7) F-measure: $2\frac{Pr \cdot Re}{Pr + Re}$

For the *Shadow* category, we also provide an average False Positive Rate that is confined to the hard-shadow areas (FPR-S).

[18]ROI stands for Region of Interest.

TABLE III
OVERALL RESULTS AS OF APRIL 2014 ACROSS ALL CATEGORIES (RC: AVERAGE RANKING ACROSS CATEGORIES, R: AVERAGE OVERALL RANKING)

Method	Description	RC	R	Re	Sp	FPR	FNR	PWC	F-Measure	Pr
MajVote	—	—	—	0.77	0.993	0.007	0.23	1.80	0.77	0.84
MajVote-5	—	—	—	0.83	**0.996**	**0.004**	0.17	**1.15**	0.84	**0.91**
MajVote-3	—	—	—	**0.84**	**0.996**	**0.004**	**0.16**	1.25	**0.85**	0.90
Spectral-360 [88]	Physic-based method using reflectivity	5.3	5.4	0.78	0.992	0.008	0.22	1.85	0.78	0.85
SGMM-SOD [34]	Improved version of SGMM [35]	6.3	5.6	0.77	0.994	0.006	0.23	**1.50**	0.77	0.83
PBAS [45]	Non-parametric and stochastic method	7.0	7.3	0.78	0.990	0.010	0.22	1.77	0.75	0.82
DPGMM [33]	Non-parametric Bayesian method	8.3	7.6	0.83	0.986	0.015	0.17	2.12	0.78	0.79
GPRMF [61]	Matrix factorization based on l_1 loss	8.5	12.0	0.84	0.973	0.027	0.16	3.16	**0.79**	0.81
CwisarD [73]	Weightless neural approach	8.7	11.7	0.89	0.978	0.022	0.18	2.66	0.78	0.77
ViBe+ [44]	Improved version of ViBe [9]	10.2	10.6	0.69	0.993	0.007	0.31	2.18	0.72	0.83
PSP-MRF [14]	Probabilistic super-pixels	10.8	11.6	0.80	0.983	0.017	0.20	2.39	0.74	0.75
Chebyshev probability [25]	Multistage method with Chebyshev inequality and object tracking	13.2	12.7	0.71	0.989	0.011	0.29	2.39	0.70	0.79
SC-SOBS [72]	Improved version of SOBS [71]	13.2	12.4	0.80	0.983	0.017	0.20	2.41	0.73	0.73
CDPS [26]	Probabilistic segmentation based on binary QMMF model	13.8	12.0	0.78	0.985	0.015	0.22	2.28	0.73	0.76
Multi-Layer Background Subtraction [47]	Multi-layer + local binary patterns	14.0	14.4	0.69	0.989	0.011	0.31	2.77	0.70	0.80
SOBS [71]	Neural maps	15.8	14.9	0.79	0.982	0.018	0.21	2.56	0.72	0.72
SGMM [35]	GMM + new mode initialization, updating and splitting rule	16.2	12.3	0.71	0.991	0.009	0.29	2.53	0.70	0.78
KDE Nonaka *et al.* [42]	Multi-level KDE	16.5	16.0	0.65	0.993	0.007	0.35	2.89	0.64	0.77
KNN [32]	Non-parametric KNN	16.8	15.3	0.67	0.991	0.009	0.33	2.80	0.68	0.79
GMM KaewTraKulPong [31]	Self-adapting GMM	18.2	16.0	0.51	**0.995**	**0.005**	0.49	3.11	0.59	0.82
ViBe [9]	Non-parametric and stochas- tic spatio-temporal method	18.2	20.3	0.68	0.983	0.017	0.32	3.12	0.67	0.74
KDE Elgammal [8]	Original KDE	18.2	20.3	0.74	0.976	0.024	0.26	3.46	0.67	0.68
KDE Yoshinaga *et al.* [43]	Spatio-temporal KDE	19.7	17.6	0.66	0.991	0.009	0.34	3.00	0.64	0.73
Bayesian Multi layer [13]	Bayesian layers + EM	21.2	22.7	0.60	0.983	0.017	0.40	3.39	0.63	0.74
GMM Stauffer-Grimson [29]	Original GMM	21.3	17.6	0.71	0.986	0.014	0.29	3.10	0.66	0.70
GMM Zivkovic [32]	GMM with automatic mode selection	24.3	19.7	0.70	0.985	0.015	0.30	3.15	0.66	0.71
Local-Self similarity [20]	Basic method with self- similarity measure	24.3	21.0	**0.94**	0.851	0.149	**0.07**	14.30	0.50	0.41
GMM RECTGAUSS-Tex [36]	Multiresolution GMM	24.3	22.3	0.52	0.986	0.014	0.48	3.68	0.52	0.72
Histogram over time [19]	Basic method with color histograms	25.3	23.0	0.77	0.934	0.066	0.23	6.97	0.55	0.53
Mahalanobis distance [5], [7]	Basic background subtraction	26.7	22.6	0.76	0.960	0.040	0.24	4.66	0.63	0.60
Euclidean distance [5], [7]	Basic background subtraction	28.2	23.9	0.71	0.969	0.031	0.30	4.35	0.61	0.62

For each method, the above metrics are first computed for each video in each category. For example, the recall metric for a particular video v in a category c is computed as follows:

$$\mathrm{Re}_{v,c} = TP_{v,c}/(TP_{v,c} + FN_{v,c}).$$

Then, a category-average metric for each category is computed from the values of the metric for all videos in a single category. For example, the average recall metric of category c is given by

$$\mathrm{Re}_c = \frac{1}{|N_c|}\sum_{v} \mathrm{Re}_{v,c}$$

where $|N_c|$ is the number of videos in category c. We also report an overall-average metric which is the simple average of the category-averages. For example, the overall-average recall is given by

$$\mathrm{Re} = \frac{1}{6}\sum_{c} \mathrm{Re}_c. \quad (1)$$

Similar category-average and overall-average values are also computed for the other metrics and categories accordingly. The overall-average metrics such as Re are reported in Table III while category-average metrics such as Re_c are reported on our website. Averaging metrics in this way (as opposed to pooling together all pixels across all videos and/or categories and then averaging) prevents bias that would occur should some videos be much larger in terms of frame size and/or length; summing up across videos would give overwhelming importance to larger videos.

In order to rank-order different change detection methods, we need to rationally combine the performance across different metrics (and/or categories) into a single rank that is indicative of how well a method fares *relative* to other methods in each category and across all categories. To this end, motivated by the approach followed by Young and Ferryman [74], we provide an average ranking R across all overall-average metrics, and an average ranking RC across all categories. To explain how these are computed, let $rank_i(m, c)$ denote the rank of method i for metric m in category c. The average ranking of method i in category c across all metrics is given by:

$$\mathrm{RM}_{c,i} = \frac{1}{7}\sum_{m} rank_i(m, c).$$

The overall ranking across categories RC_i of method i is then computed by taking the simple average of its average rankings

across all 6 categories:

$$\mathrm{RC}_i = \frac{1}{6}\sum_c \mathrm{RM}_{c,i}.$$

The average ranking R_i for method i across all overall-average metrics is given by

$$\mathrm{R}_i = \frac{1}{7}\sum_{m'} rank_i(m')$$

where m' is an overall-average metric such as the one computed in equation (1) and $rank_i(m')$ denotes the rank of method i according to the overall-average metric m'. We report the values of R, RC, and the 7 overall-average metrics for different methods in Table III.[19] The category-wide overall rankings and category-average metrics are available on the 2012 section of *www.changedetection.net* website.

IV. Methods Tested

A total of 28 change detection methods were evaluated. 6 methods are relatively simple as they rely on plain background subtraction, of which 2 use color features (Euclidean and Mahalanobis distance methods described in [5] and [7]), one uses RGB histograms over time [19], one uses local self-similarity features [20], and one use a more complex model which implements a quadratic Markov measure field [26]. Two fairly old, but frequently-cited methods have also been tested: KDE-based estimation by Elgammal *et al.* [8] and GMM by Stauffer and Grimson [29].

We also have results for 6 improved GMM methods. The self-adapting GMM by KaewTraKulPong [31], the improved GMM method by Zivkovic and Heijden [32], the multiresolution block-based GMM (RECTGAUSS-Tex) by Dora *et al.* [36], the SGMM and SGMM-SOD methods by Evangelio *et al.* [34], [35] which rely on a new initialization procedure and novel mode splitting rule, and an automatic mode selection method with a Dirichlet process (DPGMM) by Haines *et al.* [33].

We also report results on novel KDE methods namely the multi-level KDE by Nonaka *et al.* [42], and spatio-temporal KDE by Yoshinaga *et al.* [43] as well as results for 3 machine learning methods based on neural maps: CwisarD [72], SOBS, and SC-SOBS [70], [71]. We also report results for several non parametric methods such as a simple K-nearest neighbor method [32] and 3 stochastic methods based on background sample selection namely ViBe [9], ViBe+ [44], and Hofmann's self-adaptive method (PBAS) [45]. We also included a recursive per-pixel Bayesian approach by Porikli and Tuzel [13], a post-processing method based on probabilistic super-pixels (PSP-MRF) [14], a probabilistic method for matrix factorization based on the l_1 loss [60] and a recent method using local binary patterns, a multi-layer background subtraction [47].

[19]All evaluation metrics are computed as empirical averages across the test-set pixels whose number is on the order of 10^8–10^9 for each video sequence. The confidence intervals of these empirical averages are therefore on the order of 10^{-4}–10^{-5}.

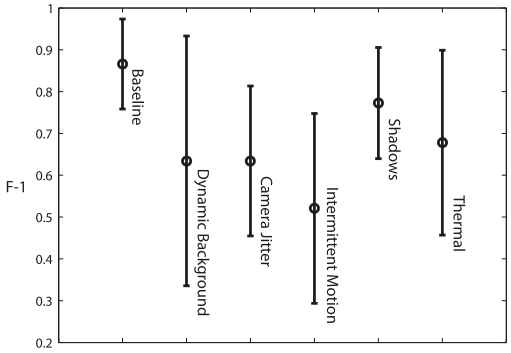

Fig. 3. Mean and standard deviation F-Measure for all methods over each category.

And last, we report results for 2 fairly complex commercial methods. One that does pixel-level detection using the Chebyshev inequality and peripheral and recurrent motion detectors by Morde *et. al.* [25] and Spectral-360 [87], a patented but non-published method.

For each method, only one set of parameters was used for all the videos. These parameters were selected according to the authors' recommendations or, when not available, were adjusted to enhance the overall results. All parameters are available on the *changedetection.net* website.

V. Experimental Results

The overall results as of April 2014 are shown in Table III where the methods have been sorted according to their average ranking across categories (RC). A more comprehensive tabulation of performance can be found on the website, where a visitor can re-sort the methods by the average overall ranking R as well as individual average metrics.

We also added to Table III results from three pixel-based *majority vote* methods. *MajVote* is a majority vote over the 28 methods whereas *MajVote-3* and *5* are the majority vote of the best combination of 3 and 5 methods (here CDPS, Chebyshev probability, GPRMF, SGMM-SOD and Spectral360). These methods have been obtained by testing every possible combination of 3 and 5 methods. Interestingly, the best combination of 3 and 5 methods does not include Spectral-360 and PBAS, two of the top performing methods.

It should come as no surprise that the simplistic methods based on plain background subtraction [7], [19], [20] are at the bottom of the table, whereas more recent methods [14], [34], [44], [45], [71], [87] are at the top.

According to RC, the top performing method is Spectral-360 [87] which is a patented but non-published method. Although the patent is hard to read, we understand that instead of using plain RGB or texture information as is usually the case, they model the physical surface of the objects with a spectral reflectance descriptor model. They mention that their physic-based model together with a sophisticated correlation technique between foreground and background spectral reflectance to detect motion is the key of their success.

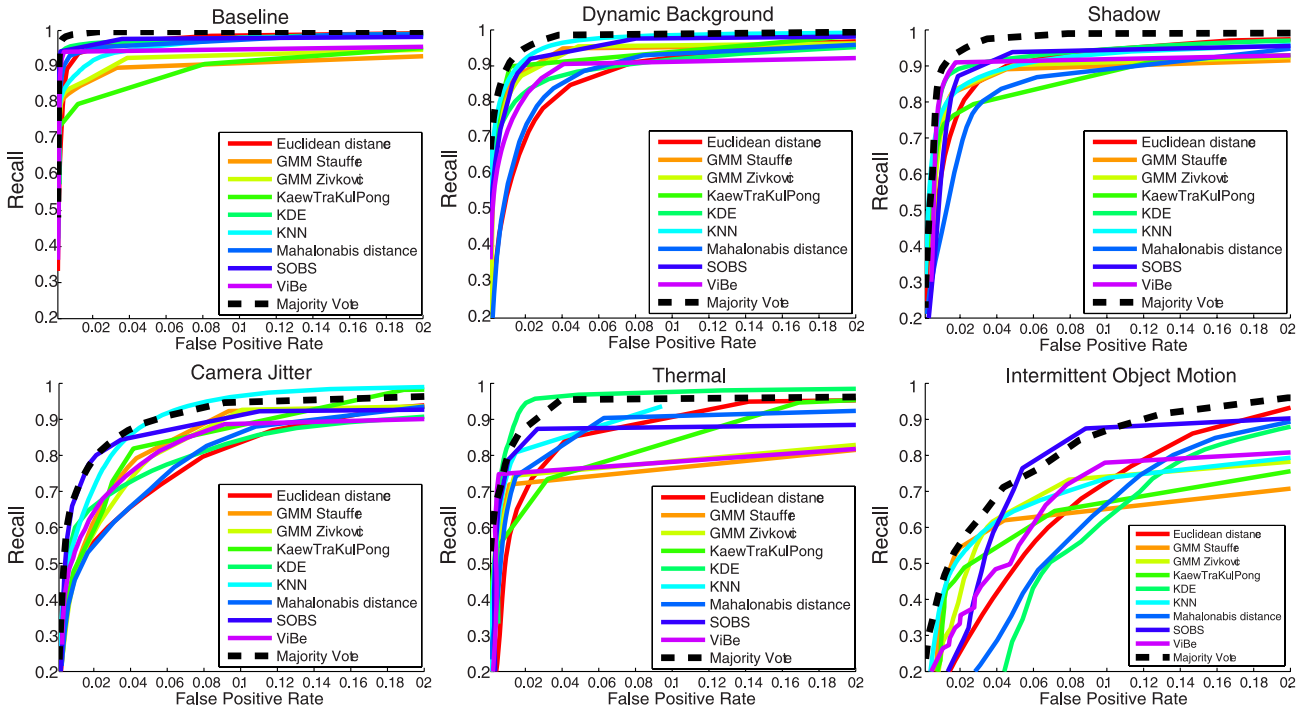

Fig. 4. ROC curves obtained for each category using 9 methods and the majority vote among these methods.

The second top method SGMM-SOD [34] is a modified version of Stauffer and Grimson's GMM method [29]. The success of SGMM-SOD can be attributed to three innovations: i) a different initialization procedure of newly created modes, ii) a background updating scheme that adapts to changing conditions, and iii) a new splitting rule which avoids over-dominating modes. PBAS, the method ranked number 3 [45], uses a non-parametric probabilistic model for the background at each spatial location based on a random subset of pixel values from the recent past. Such a stochastic non-parametric model makes these methods robust to instabilities (background motion and camera jitter) and intermittent motion artifacts.

A bit surprising are the majority vote results. MajVote is in the top 3 methods when considering PWC, F-measure and Precision while MajVote-5 and MajVote-3 outperform all 28 methods. This calls for an important conclusion: as of today, there is no single best method that can serve as a "silver bullet" for each scenario. It seems that the methods are complementary by nature and combining them improves results. We found this to be true even when combining low-rank methods. This suggests modular change detection methods that can adapt to the content of the video are likely to perform very well.

A similar conclusion can be drawn from Table IV and Fig. 3. Table IV shows the highest ranked methods for each video category. As one can see, the 3 best overall methods (Spectral-360, SGMM-SOD and PBAS) are not necessarily the best methods over each category. Fig. 3 shows the mean and std-dev of the F-measure for each category across all the methods. This figure visually captures the difficulty of each category: "Baseline" is the easiest whereas "Intermittent Motion" is the most challenging category. Fig. 3 also shows how different the methods can be.

TABLE IV
THREE HIGHEST RANKED METHOD FOR EACH CATEGORY

Category	1st	2nd	3rd
Baseline	SC-SOBS [72]	SOBS [71]	GPRMF [61]
Dynamic Background	CwisarD [73]	DPGMM [33]	Chebyshev [25]
Shadows	GPRMF [61]	Spectral-360 [88]	SGMM-SOD [34]
Camera Jitter	GPRMF [60]	CwisarD [72]	PSP-MRF [14]
thermal	SGMM-SOD [34]	Chebyshev [25]	Spectral-360
Interm. Motion	SGMM-SOD [34]	CDPS [26]	KDE Nonaka[42]

A. Metrics

An interesting observation from Table III is that recall, specificity, FNR and FPR are not good indicators of the overall ranking. According to recall and FNR, the best method is the Local-Self similarity which is actually ranked 25th. According to specificity and FPR, the best method is GMM by Kaewtrakulpong and Bowden which is at the 18th position. On the other hand, PWC, F-Measure and precision correlate better with the average rankings as they give the best score to SGMM-SOD, GPRMF, and Spectral360, some of the highest ranked methods.

Between F-measure, PWC and precision, PWC and F-measure correlate slightly better with the average ranking. When methods are re-ordered according to PWC or F-measure, methods do not shift in ranking by more than 5 positions (except for GPRMF [60] and CwisarD [72]). As for precision, re-ranking often leads to a shift of up to 9 positions. So, if only one metric were to be used to rank methods, PWC or F-measure would probably be the best choices.

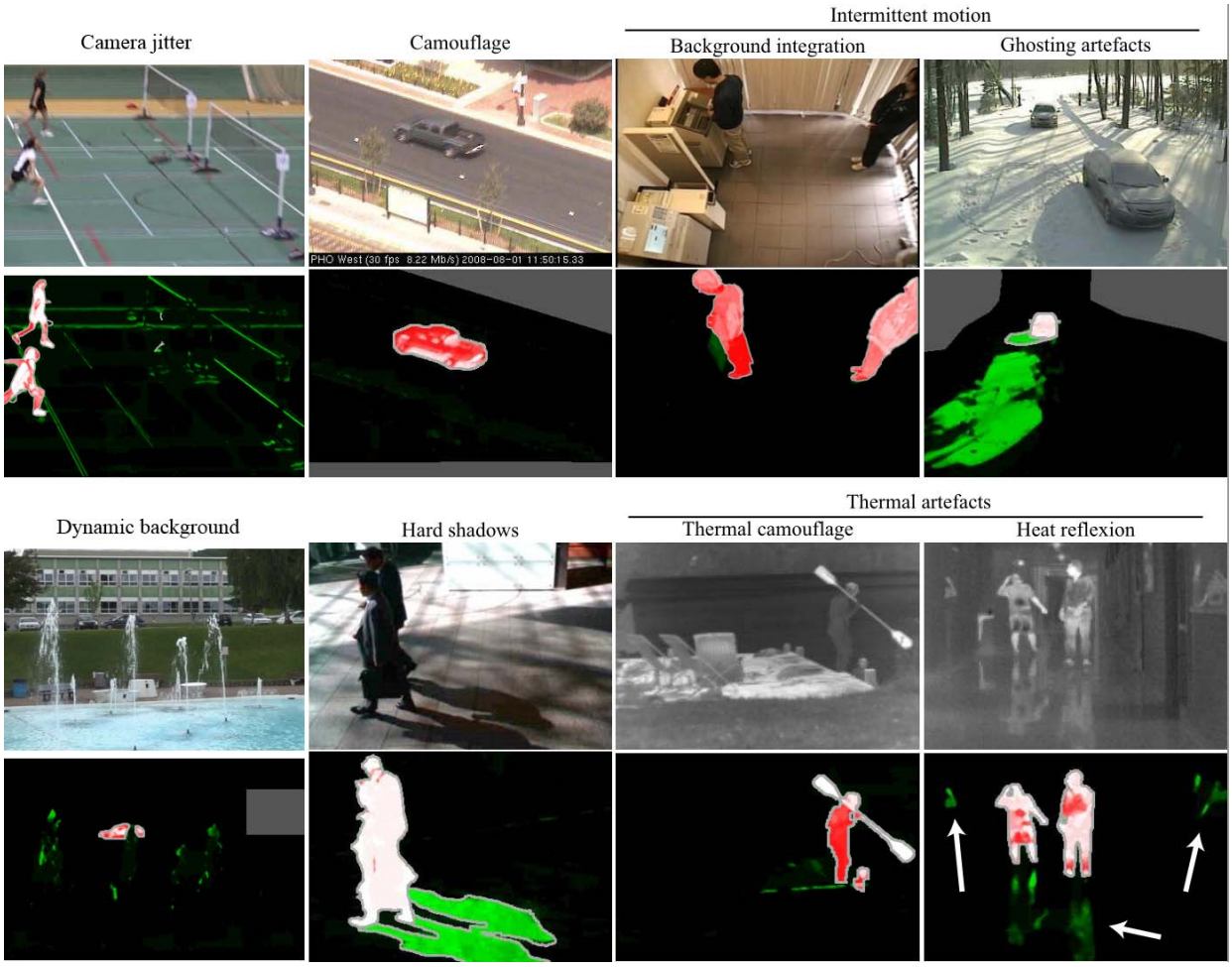

Fig. 5. Total error maps: failure areas for a large proportion of methods. Red shows locations with many false positives, green - false negatives, white - true positives and black - true negatives.

B. Solved and Remaining Issues

In order to assess the challenge that each video category poses for the tested methods, we ranked the categories according to score obtained by all methods in a given category. As can be seen in Table V, videos with intermittent motion pose the largest challenge in terms of the F-Measure (0.51), FPR (0.033) and PWC (6.0%). On the other hand, videos exhibiting steady background motion seem to be less challenging. Many methods had difficulty with thermal videos as most of the time they suffered from camouflage problems, resulting in large FNR scores.

TABLE V
MEDIAN F-MEASURE, FPR, FNR AND PWC OBTAINED BY ALL 29 METHODS IN EACH CATEGORY

Category	F-Measure	FPR	FNR	PWC
Interm. Motion	0.51	0.033	0.47	6.0
Camera Jitter	0.69	0.018	0.27	2.9
Dynamic Back.	0.65	0.009	0.20	1.2
Thermal	0.68	0.004	0.40	2.8
Shadows	0.78	0.011	0.17	2.1
Baseline	**0.87**	**0.003**	**0.13**	**0.9**

Also it is not clear from Table V how much of a challenge do hard shadows pose for the tested methods. In order to verify this, we computed FPR within shadow areas (FPR-S) for all videos in the "Shadows" category. As can be seen in Table VI, the tested methods attained FPR in shadow areas between 0.2 and 0.64. This large FPR indicates that none of the methods deals with shadows effectively.

ROC curves obtained for 9 methods and the majority vote computed from them are shown in Fig. 4. First, except for the thermal category, the majority vote outperforms every method. This result correlates well with Table III and the performance of majority voting. These curves also show that intermittent motion and thermal videos are very challenging. These two categories display the greatest disparity between methods tested.

In order to identify where the methods fail, we computed an error map for each frame of every video. Then, for each frame of each video, we integrated all error maps into a total error map by calculating the proportion of methods with a true positive, a false positive, a true negative, or a false negative at each pixel (Fig. 5). The red and green areas show where methods suffer from false positives and false negatives, respectively, whereas white and black areas show true positives and true negatives, respectively. A careful inspection of the

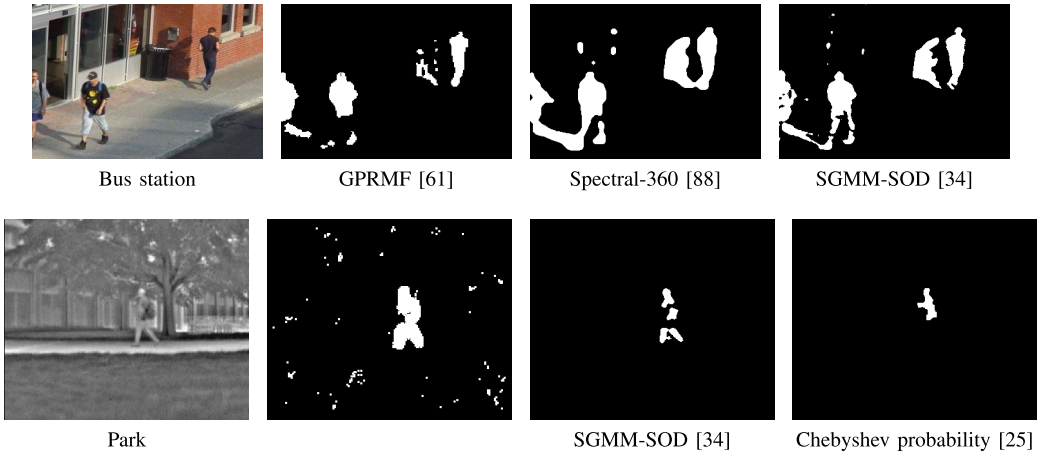

Fig. 6. Comparison of the three best methods' results of the left image's category.

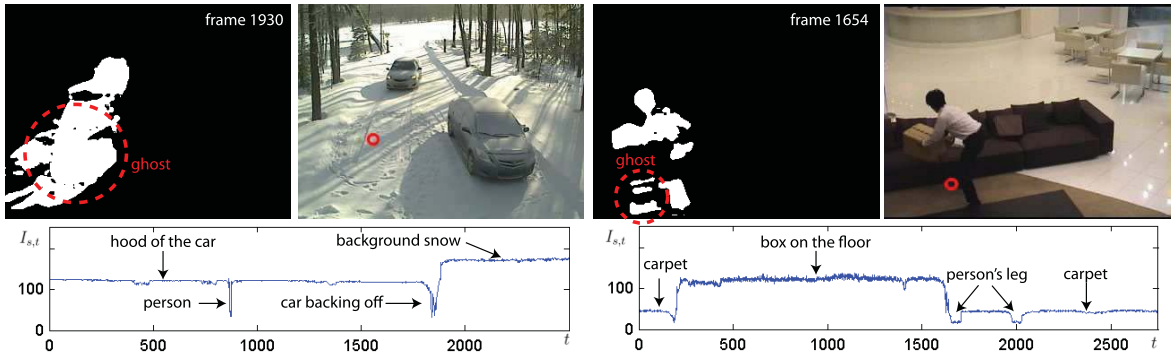

Fig. 7. Illustration of ghosting artifacts, i.e., a transition from object color to background color (car to snow on left and box to carpet on right).

total error maps leads us to the following observations:

1) camouflage is a problem that spans all categories and for which none of the methods tested in this paper has a decisive solution,
2) shadow-robust methods are only effective on soft shadows; dealing with hard shadows is still an open issue,
3) methods have a hard time dealing with thermal videos due to reflection and camouflage issues,
4) intermittent motion is another challenge that none of the tested methods handles well.

In order to illustrate these observations, we put in Fig. 6 the results for 3 of the top performing methods on 2 different videos. As can be seen, all of them generate false positives and false negatives due to camouflage, shadows and intermittent motion. This shows that there is still room for improvement, even for the top performing methods.

Intermittent motion causes methods to suffer from false positives and false negatives. For example, a car pulling off a driveway leads to false positives (ghosting artifact) while people waiting in line cause false negatives. The ghosting issue is illustrated in Fig. 7. When we look at a pixel's history, a ghosting artifact appears after a strong transition which reveals the background color. Unfortunately, transition from the object to the background color can barely be distinguished from a background to object color transition (at least using only pixel history). This is shown in the right plot of Fig. 7 where the color history goes from carpet to box and then from box to carpet.

TABLE VI
RANKING OF METHODS ACCORDING TO FPR IN SHADOW AREAS FOR VIDEOS IN THE "SHADOWS" CATEGORY

Method	FPR-S
Bayesian Multi-Layer [13]	0.33
KDE Nonaka *et al.* [42]	0.39
KDE Yoshinaga *et al.* [43]	0.40
KNN [32]	0.40
GMM KaewTraKulPong [31]	0.41
DPGMM [33]	0.42
Chebyshev probability [25]	0.42
RECTGAUSS-Tex [36]	0.48
SGMM [35]	0.49
GPRMF [61]	0.49
ViBe+ [44]	0.53
GMM Stauffer-Grimson [29]	0.54
GMM Zivkovic [32]	0.54
ViBe [9]	0.55
CwisarD [73]	0.56
SOBS [71]	0.57
Euclidean distance [5], [7]	0.58
PBAS [45]	0.58
PSP-MRF [14]	0.59
Histogram over time [19]	0.59
Mahalanobis distance [5], [7]	0.59
CDPS [26]	0.59
SC-SOBS [72]	0.60
Spectral-360 [88]	0.62
KDE Elgammal [8]	0.62
SGMM-SOD [34]	0.63
Local-Self similarity [20]	0.64

Ghosts and false negatives are related to how fast the background is updated: too quick of an update leads to false negatives whereas too slow of an update leads to ghosts. These

problems are difficult to handle because solving one issue (say preventing the method from merging people waiting in line with the background) exacerbates the other issue (say, ghosting artifacts). Some methods such as ViBe [9] provide tentative solutions to ghosting artifacts by pooling information spatially. But such solutions are not sufficient to remove large ghosting artifacts.

VI. Conclusions

We would like to summarize the salient findings of this study through five conclusions which, we hope, would inspire and guide future work on this topic.

1) Among the 7 metrics, **PWC** and the **F-measure** correlate best with the average ranking across categories.
2) Videos with small **recurrent background motion** (ripples on the water, trees shaken by the wind) do not pose a heavy challenge for the top performing methods. The same conclusion applies to **baseline videos**.
3) None of the above methods tested is robust to **hard shadows, intermittent motion**, and **camouflage**. These are open issues that are yet to be solved.
4) Contrary to common belief, detecting humans and moving objects in **thermal videos** is not trivial. It is often accompanied by reflections and camouflage effects that no method handles well.
5) As of today, methods are complementary in nature and combining them with a majority vote helps improving results. Future research should consider modular change detection methods that would incorporate complementary approaches.

VII. Summary and Outlook

Change detection task plays a pivotal role in many computer vision application as an early preprocessing step. Despite the multitude of algorithms developed to date, there is no clear way to establish which method responds well to certain challenges, and thus selecting the "optimal" algorithm for a given task is difficult.

In order to address this problem, we have prepared a comprehensive dataset, called CDnet. In the 2012 version of CDnet, each video sequence has been very carefully hand-labeled to allow a accurate, objective and quantitative ranking of change detection algorithms on 7 fidelity metrics.

The CDnet undertaking aims to provide the research community with a rigorous scientific benchmarking facility with a rich dataset of videos for testing and ranking of existing methods. It offers utility code, documentation, and consolidated algorithms for change detection. It also provides an access to author-approved algorithm implementations.

This dataset will be regularly revised and expanded with feedback from the academia and industry. We hope to maintain and update a rank list of the most accurate change detection algorithms in the various categories for years to come.

References

[1] C. Stauffer and E. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 1999.

[2] A. M. Elgammal, D. Harwood, and L. S. Davis, "Non-parametric model for background subtraction," in *Proc. 6th Eur. Conf. Comput. Vis.*, 2000, pp. 751–767.

[3] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, Jun. 1997.

[4] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection.net: A new change detection benchmark dataset," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 1–8.

[5] T. Bouwmans, "Recent advanced statistical background modeling for foreground detection: A systematic survey," *Recent Patents Comput. Sci.*, vol. 4, no. 3, pp. 147–176, 2011.

[6] L. Wixson, "Detecting salient motion by accumulating directionally-consistent flow," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 774–780, Aug. 2000.

[7] Y. Benezeth, P.-M. Jodoin, B. Emile, H. Laurent, and C. Rosenberger, "Comparative study of background subtraction algorithms," *J. Electron. Imag.*, vol. 19, no. 3, pp. 1–12, 2010.

[8] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proc. IEEE*, vol. 90, no. 7, pp. 1151–1163, Jul. 2002.

[9] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1709–1724, Jun. 2011.

[10] P. Rosin and E. Ioannidis, "Evaluation of global image thresholding for change detection," *Pattern Recognit. Lett.*, vol. 24, no. 14, pp. 2345–2356, 2003.

[11] R. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: A systematic survey," *IEEE Trans. Image Process.*, vol. 14, no. 3, pp. 294–307, Mar. 2005.

[12] F. Porikli and C. Wren, "Change detection by frequency decomposition: Wave-back," in *Proc. Workshop IAMIS*, 2005.

[13] F. Porikli and O. Tuzel, "Bayesian background modeling for foreground detection," in *Proc. ACM Vis. Surveill. Sensor Netw.*, 2005.

[14] A. Schick, M. Bäuml, and R. Stiefelhagen, "Improving foreground segmentations with probabilistic superpixel Markov random fields," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 27–31.

[15] F. Tiburzi, M. Escudero, J. Bescos, and J. M. Martinez, "A ground truth for motion-based video-object segmentation," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 17–20.

[16] D. S. Matteson and N. A. Jame, "A nonparametric approach for multiple change point analysis of multivariate data," *J. Amer. Statist. Assoc.*, vol. 109, no. 505, pp. 334–345, 2013.

[17] N. J. B. McFarlane and C. P. Schofield, "Segmentation and tracking of piglets in images," *Mach. Vis. Appl.*, vol. 8, no. 3, pp. 187–193, 1995.

[18] D. Kit, B. Sullivan, and D. Ballard, "Novelty detection using growing neural gas for visuo-spatial memory," in *Proc. IEEE IROS*, Sep. 2011, pp. 1194–1200.

[19] J. Zheng, Y. Wang, N. L. Nihan, and M. E. Hallenbeck, "Extracting roadway background image: A mode based approach," *J. Transp. Res. Rep.*, vol. 1944, no. 1, pp. 82–88, 2006.

[20] J.-P. Jodoin, G. Bilodeau, and N. Saunier, "Background subtraction based on local shape," École Polytechnique, Montréal, QC, Canada, Tech. Rep., 2012.

[21] K. Karman and A. von Brandt, "Moving object recognition using an adaptive background memory," in *Time-Varying Image Processing and Moving Object Recognition*, vol. 2, V. Capellini, Ed. Amsterdam, The Netherlands: Elsevier, 1990, pp. 297–307.

[22] J. Zhong and S. Sclaroff, "Segmenting foreground objects from a dynamic textured background via a robust Kalman filter," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 44–50.

[23] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 1. Sep. 1999, pp. 255–261.

[24] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[25] A. Morde, X. Ma, and S. Guler, "Learning a background model for change detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 15–20.

[26] F. J. Hernandez-Lopez and M. Rivera, "Change detection by probabilistic segmentation from monocular view," *Mach. Vis. Appl.*, vol. 25, no. 5, pp. 1175–1195, 2014.

[27] H. Kim, R. Sakamoto, I. Kitahara, T. Toriyama, and K. Kogure, "Robust foreground extraction technique using Gaussian family model and multiple thresholds," in *Proc. ACCV*, 2007, pp. 758–768.

[28] X. Gao, T. E. Boult, F. Coetzee, and V. Ramesh, "Error analysis of background adaption," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2000, pp. 503–510.

[29] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, Aug. 2000.

[30] P.-M. Jodoin, M. Mignotte, and J. Konrad, "Statistical background subtraction using spatial cues," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 12, pp. 1758–1763, Dec. 2007.

[31] P. KaewTraKulPong and R. Bowden, "An improved adaptive background mixture model for realtime tracking with shadow detection," in *European Workshop on Advanced Video Based Surveillance Systems*. New York, NY, USA: Springer-Verlag, 2001.

[32] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognit. Lett.*, vol. 27, no. 7, pp. 773–780, 2006.

[33] T. S. F. Haines and T. Xiang, "Background subtraction with dirichlet processes," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 97–111.

[34] R. H. Evangelio, M. Pätzold, and T. Sikora, "Splitting Gaussians in mixture models," in *Proc. IEEE Int. Conf. AVSS*, Sep. 2012, pp. 300–305.

[35] R. H. Evangelio and T. Sikora, "Complementary background models for the detection of static and moving objects in crowded environments," in *Proc. 8th IEEE Int. Conf. AVSS*, Sep. 2011, pp. 71–76.

[36] D. Riahi, P.-L. St-Onge, and G. Bilodeau, "RECTGAUSS-tex: Block-based background subtraction," Dept. génie informatique et génie logiciel, École Polytechn. de Montreal, Montreal, QC, Canada, Tech. Rep. EPM-RT-2012-03, 2012.

[37] M. S. Allili, N. Bouguila, and D. Ziou, "Finite general Gaussian mixture modeling and application to image and video foreground segmentation," *J. Electron. Imag.*, vol. 17, no. 1, pp. 1–23, 2008.

[38] L. Li, W. Huang, I. Y.-H. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Trans. Image Process.*, vol. 13, no. 11, pp. 1459–1472, Nov. 2004.

[39] D. Butler, S. Sridharan, and M. Bove, "Real-time adaptive background segmentation," in *Proc. ICME*, Jul. 2003, pp. 341–344.

[40] P. Jaikumar, A. Singh, and S. K. Mitra, "Background subtraction in videos using Bayesian learning with motion information," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 1–10.

[41] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground-background segmentation using codebook model," *Real-Time Imag.*, vol. 11, no. 3, pp. 172–185, 2005.

[42] Y. Nonaka, A. Shimada, H. Nagahara, and R. Taniguchi, "Evaluation report of integrated background modeling based on spatio-temporal features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 9–14.

[43] S. Yoshinaga, A. Shimada, H. Nagahara, and R. Taniguchi, "Background Model Based on Intensity Change Similarity Among Pixels," in *Proc. Frontiers Comput. Vis.*, Jan. 2013, pp. 276–280.

[44] M. Van Droogenbroeck and O. Paquot, "Background subtraction: Experiments and improvements for ViBe," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 32–37.

[45] M. Hofmann, P. Tiefenbacher, and G. Rigoll, "Background segmentation with feedback: The pixel-based adaptive segmenter," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 38–43.

[46] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, and J. M. Buhmann, "Topology free hidden Markov models: Application to background modeling," in *Proc. 8th IEEE Int. Conf. Comput. Vis.*, Jul. 2001, pp. 294–301.

[47] J. Yao and J. Odobez, "Multi-layer background subtraction based on color and texture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.

[48] M. Heikkila and M. Pietikainen, "A texture-based method for modeling the background and detecting moving objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 657–662, Apr. 2006.

[49] A. Mittal and N. Paragios, "Motion-based background subtraction using adaptive kernel density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2004, pp. 302–309.

[50] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 831–843, Aug. 2000.

[51] Z. Xu, P. Shi, and I. Y.-H. Gu, "An eigenbackground subtraction method using recursive error compensation," in *Proc. PCM*, 2006, pp. 779–787.

[52] Y. Li, "On incremental and robust subspace learning," *Pattern Recognit.*, vol. 37, no. 7, pp. 1509–1518, 2004.

[53] Y. Dong, T. X. Han, and G. N. Desouza, "Illumination invariant foreground detection using multi-subspace learning," *Int. J. Knowl.-Based Intell. Eng. Syst.*, vol. 14, no. 1, pp. 31–41, 2010.

[54] J. Rymel, J. Renno, D. Greenhill, J. Orwell, and G. Jones, "Adaptive eigen-backgrounds for object detection," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2004, pp. 1847–1850.

[55] F. Seidel, C. Hage, and M. Kleinsteuber, "pROST: A smoothed ℓp-norm robust online subspace tracking method for background subtraction in video," *Mach. Vis. Appl.*, vol. 25, no. 5, pp. 1227–1240, 2013.

[56] C. Guyon, T. Bouwmans, and E. Zahzah, "Foreground detection via robust low rank matrix factorization including spatial constraint with Iterative reweighted regression," in *Proc. 21st ICPR*, Nov. 2012, pp. 2805–2808.

[57] C. Guyon, T. Bouwmans, and E. Zahzah, "Robust principal component analysis for background subtraction: Systematic evaluation and comparative analysis," in *Principal Component Analysis*, P. Sanguansat, Ed. Rijeka, Croatia: InTech, 2012, ch. 12, pp. 223–238.

[58] J. He, L. Balzano, and A. Szlam, "Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1568–1575.

[59] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, 2011.

[60] N. Wang, T. Yao, J. Wang, and D.-Y. Yeung, "A probabilistic approach to robust matrix factorization," in *Computer Vision*. Berlin, Germany: Springer-Verlag, 2012, pp. 126–139.

[61] F. Porikli, "Multiplicative background-foreground estimation under uncontrolled illumination using intrinsic images," in *Proc. 7th IEEE Workshops Appl. Comput. Vis.*, Breckenridge, CO, USA, Jan. 2005, pp. 20–27.

[62] C. Zhao, X. Wang, and W.-K. Cham, "Background subtraction via robust dictionary learning," *EURASIP J. Image Video Process.*, vol. 2011, pp. 1–12, Jan. 2011.

[63] E. Memin and P. Perez, "Dense estimation and object-based segmentation of the optical flow with robust techniques," *IEEE Trans. Image Process.*, vol. 7, no. 5, pp. 703–719, May 1998.

[64] X. Lu and R. Manduchi, "Fast image motion segmentation for surveillance applications," *Image Vis. Comput.*, vol. 29, nos. 2–3, pp. 104–116, 2011.

[65] D. Zhou and H. Zhang, "Modified GMM background modeling and optical flow for detection of moving objects," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Oct. 2005, pp. 2224–2229.

[66] L. Cheng, M. Gong, D. Schuurmans, and T. Caelli, "Real-time discriminative background subtraction," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1401–1414, May 2011.

[67] H.-H. Lin, T.-L. Liu, and J.-H. Chuang, "A probabilistic SVM approach for background scene initialization," in *Proc. IEEE Int. Conf. Image Process.*, Jun. 2002, pp. 893–896.

[68] Z. Hao, W. Wen, Z. Liu, and X. Yang, "Real-time foreground-background segmentation using adaptive support vector machine algorithm," in *Proc. 17th ICANN*, 2007, pp. 603–610.

[69] B. Han and L. S. Davis, "Density-based multifeature background subtraction with support vector machine," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 1017–1023, May 2012.

[70] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1168–1177, Jul. 2008.

[71] L. Maddalena and A. Petrosino, "The SOBS algorithm: What are the limits?" in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 21–26.

[72] M. D. Gregorio and M. Giordano, "A wisard-based approach to CDNet," in *Proc. 1st BRICS Countries Congr. (BRICS-CCI)/11th Brazilian Congr. (CBIC) Comput. Intell.*, 2013.

[73] F. De La Torre and M. J. Black, "A framework for robust subspace learning," *Int. J. Comput. Vis.*, vol. 54, nos. 1–3, pp. 117–142, 2003.

[74] D. P. Young and J. M. Ferryman, "PETS metrics: Online performance evaluation service," in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Syst.*, Oct. 2005, pp. 317–324.

[75] R. Vezzani and R. Cucchiara, "Video surveillance online repository (ViSOR): An integrated framework," *Multimedia Tools Appl.*, vol. 50, no. 2, pp. 359–380, 2010.

[76] M. Karaman, L. Goldmann, D. Yu, and T. Sikora, "Comparison of static background segmentation methods," *Proc. SPIE*, vol. 5960, pp. 2140–2151, Jun. 2005.

[77] A. Vacavant, T. Chateau, A. Wilhelm, and L. Lequievre, "A benchmark dataset for outdoor foreground/background extraction," in *Proc. 11th ACCV Workshops*, 2012, pp. 291–300.
[78] S. Brutzer, B. Heidemann, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1937–1944.
[79] S.-C. S. Cheung and C. Kamath, "Robust background subtraction with foreground validation for urban traffic video," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 2330–2340, Jan. 2005.
[80] T. Bouwmans, F. El Baf, and B. Vachon, "Background modeling using mixture of Gaussians for foreground detection: A survey," *Recent Patents Comput. Sci.*, vol. 1, no. 3, pp. 219–237, 2008.
[81] J. Nascimento and J. Marques, "Performance evaluation of object detection algorithms for video surveillance," *IEEE Trans. Multimedia*, vol. 8, no. 8, pp. 761–774, Aug. 2006.
[82] H. Hassanpour, M. Sedighi, and A. R. Manashty, "Video frame's background modeling: Reviewing the techniques," *J. Signal Inf. Process.*, vol. 2, no. 2, pp. 72–78, 2011.
[83] A. Prati, R. Cucchiara, I. Mikic, and M. M. Trivedi, "Analysis and detection of shadows in video streams: A comparative evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2001, pp. 571–577.
[84] F. Bashir and F. Porikli, "Performance evaluation of object detection and tracking systems," in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Syst.*, Jun. 2006.
[85] D. Parks and S. Fels, "Evaluation of background subtraction algorithms with post-processing," in *Proc. IEEE 5th Int. Conf. AVSS*, Sep. 2008, pp. 192–199.
[86] M. Piccardi, "Background subtraction techniques: A review," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Oct. 2004, pp. 3099–3104.
[87] M. Sedky, C. Chibelushi, and M. Moniri, "Object segmentation using full-spectrum matching of albedo derived from colour images," U.S. Patent 20 120 008 858 A1, Jan. 12, 2012.

Nil Goyette received the B.Sc. and M.Sc. degrees in computer science from the University of Sherbrooke, Sherbrooke, QC, Canada, in 2010 and 2013, respectively, where he studied video surveillance and computer vision. He created, along with the co-authors, the changedetection.net initiative, which encapsulates a rigorous and comprehensive academic benchmarking effort for testing and ranking existing and new algorithms for change and motion detection. He is currently working on medical imaging in Quebec.

Pierre-Marc Jodoin (M'07) is currently a Canadian Computer Engineer and an Associate Professor with the Department of Computer Science, University of Sherbrooke, Sherbrooke, QC, Canada. He received the Ph.D. (Hons.) degree in computer vision and video analytics from the Université de Montréal, Montreal, QC, Canada, in 2007. His research interests include video analytics and surveillance, image processing, medical imaging, and 3D reconstruction. He currently serves as an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, and an Invited Editor of the *Pattern Recognition* (Elsevier) and *Signal Processing* (Elsevier) journals. He is also the Director of the Sherbrooke Research Center on Smart Environments, which he co-founded in 2012. He also co-founded the Sherbrooke Medical Image Processing Service in 2010, and in 2011, Imeka.ca, a company specialized in medical imaging and brain tractography, and started the changedetection.net initiative in 2011, one of the significant benchmarking effort in the field of video analytics.

Fatih Porikli (F'14) is a Professor with the Research School of Engineering, Australian National University, Canberra, ACT, Australia. He is also acting as the leader of the Computer Vision Group with National ICT Australia Ltd., Sydney, NSW, Australia. He received the Ph.D. degree from New York University, New York, NY, USA, in 2002. Previously, he served as a Distinguished Research Scientist at Mitsubishi Electric Research Laboratories, Cambridge, MA, USA. He has contributed broadly to object detection, motion estimation, tracking, image-based representations, and video analytics. He is the Co-Editor of two books entitled *Video Analytics for Business Intelligence* and *Handbook on Background Modeling and Foreground Detection for Video Surveillance*. He is an Associate Editor of five journals, including the IEEE SIGNAL PROCESSING MAGAZINE, *SIAM Imaging Sciences*, *EURASIP Journal of Image & Video Processing*, *Machine Vision Applications* (Springer), and *Real-time Image & Video Processing* (Springer). His publications received three best paper awards, and he was a recipient of the R&D100 Award in the Scientist of the Year category in 2006. He served as the General and Program Chair of several IEEE conferences.

Janusz Konrad (M'93–SM'98–F'08) received the M.Eng. degree from the Technical University of Szczecin, Szczecin, Poland, in 1980, and the Ph.D. degree from McGill University, Montreal, QC, Canada, in 1989. From 1989 to 2000, he was with INRS-Télécommunications, Montreal. Since 2000, he has been with Boston University, Boston, MA, USA. He is currently an Area Editor of the *EURASIP Signal Processing: Image Communications* journal and an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING. He was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE COMMUNICATIONS MAGAZINE, the IEEE SIGNAL PROCESSING LETTERS, and the *EURASIP International Journal on Image and Video Processing*. He was a member of the IMDSP Technical Committee of the IEEE Signal Processing Society, the Technical Program Co-Chair of ICIP-2000, the Tutorials Co-Chair of ICASSP-2004, the Technical Program Co-Chair of AVSS-2010, and the General Chair of AVSS-2013. He was a co-recipient of the 2001 IEEE SIGNAL PROCESSING MAGAZINE Award for a paper co-authored with Dr. C. Stiller, and the 2004–2005 *EURASIP Image Communications* Best Paper Award for a paper co-authored with Dr. N. Bozinovic, the AVSS-2010 Best Paper Award, and the 2010 ICPR Aerial View Activity Classification Challenge Award. His research interests include image and video processing, stereoscopic and 3D imaging and displays, visual sensor networks, and human–computer interfaces.

Prakash Ishwar (SM'07) received the B.Tech. degree in electrical engineering from IIT Bombay, Mumbai, India, in 1996, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 1998 and 2002, respectively. After two years as a Post-Doctoral Researcher with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA, USA, he joined the faculty of Boston University, Boston, MA, USA, where he is currently an Associate Professor of Electrical and Computer Engineering. His research interests are in statistical signal processing, machine learning, visual information analysis and processing, information theory, and security.

Dr. Ishwar was a recipient of the 2005 United States National Science Foundation CAREER Award, a co-recipient of the Best Paper Award at the 2010 IEEE International Conference on Advanced Video and Signal-Based Surveillance and the 2010 Aerial View Activity Classification Challenge Award at the IEEE International Conference on Pattern Recognition. He is currently an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING and an elected member of the IEEE Signal Processing Theory and Methods Technical Committee, and was a member of the IEEE Image, Video, and Multidimensional Signal Processing Technical Committee.