YORK
UNIVERSITÉ
UNIVERSITY

redefine THE POSSIBLE.

Background Image Modelling for Change Detection

Hang Gao and Richard P. Wildes

Technical Report EECS-2016-01

January 7 2016

Department of Electrical Engineering and Computer Science
4700 Keele Street, Toronto, Ontario M3J 1P3 Canada

Background Image modelling for change detection

Hang Gao and Richard P. Wildes

Abstract

Change detection relative to a background image model is a potentially enabling capability for a variety of image analysis tasks, including automated video surveillance and monitoring. In this report, various combinations of chromatic, spatial and dynamic features are proposed and evaluated with reference to background modelling for image change detection. The features are embedded in a popular background modelling framework and empirically evaluated relative to each other as well as the state-of-the-art on a standard, publicly available change detection dataset.

Contents

1 Introduction **4**

1.1 Motivation 4

1.2 Related work 5

1.2.1 Overview 5

1.2.2 PDF models 6

1.2.2.1 Gaussian and Gaussian Mixtures 6

1.2.2.2 Student's t or Dirichlet mixture models 7

1.2.2.3 Non-parametric estimation models 7

1.2.2.4 KDE-GMM hybrid models 8

1.2.3 Subspace models 8

1.2.3.1 Reconstructive subspace models 8

1.2.3.2 Robust subspace models 9

1.2.3.3 Subspace tracking models 9

1.2.4 Sample-based models 10

1.2.5 Neural network models 10

1.2.6 Cluster models 12

1.2.7 Other related approaches 12

1.2.8 Evaluations 13

1.3 Contributions 14

1.4 Outline 16

2 Technical approach **17**

2.1 Overview 17

2.2 Primitive feature representation 17

2.2.1 Dynamic features 18

2.2.2 Spatial orientation features . . . 21
2.2.3 Chromatic features . . . 22
2.3 Background image modelling . . . 22
2.3.1 ViBe framework . . . 22
2.3.1.1 Model and classification process . . . 23
2.3.1.2 Model initialization . . . 23
2.3.1.3 Model update . . . 25
2.3.2 Multiple features in ViBe framework . . . 25
2.3.2.1 Pixel model and classification process . . . 25
2.3.2.2 Model initialization and update . . . 28
2.3.2.3 Feature combination . . . 29
2.4 Recapitulation . . . 30

3 Empirical evaluation **32**
3.1 Datasets . . . 32
3.2 Results . . . 32
3.2.1 Comparison among different feature combinations . . . 32
3.2.2 Comparison with state of the art . . . 46
3.3 Discussion . . . 46
3.3.1 Comparison among different feature combinations . . . 46
3.3.2 Comparison with state of the art . . . 48

4 Conclusion **49**
4.1 Summary . . . 49
4.2 Future work . . . 49

Appendix **51**

References **59**

1 Introduction

1.1 Motivation

The majority of current real-world surveillance and monitoring systems are based primarily on the performance of human operators that are expected to watch, often simultaneously, a large number of screens that show video streams captured by different cameras. The task assigned to the human is to detect events of interest (e.g., nefarious or otherwise unusual activities) and decide on appropriate actions to take (e.g., alert further security personnel). A disadvantage of this paradigm is the likelihood that the operators will miss a significant event owing to the monotonous, yet difficult nature of their task. Monotony results from extended periods of time when nothing interesting is happening; difficulty arises as subtle changes should be monitored simultaneously across multiple displays.

Automated video surveillance and monitoring systems have potential to alleviate the human workload by assuming some of the processing demands. Along these lines, change detection relative to a background model is of potentially enabling importance. The background model itself is an image that captures the range of pixelwise brightness measurements expected under typical conditions as a camera views a particular scene, e.g., as exemplified during a training period [78]. Subsequent analyses are performed relative to this reference background image. For example, new objects entering the scene can be detected and motion of moving targets can be tracked relative to the background. The information that is gleaned from such processing can serve to alert human operators or serve as a component of a larger automated system.

The accuracy and precision of such automated analyses is directly related to the ability of the background model to adequately capture typical background scene variations (e.g., variable lighting and camera parameters, cluttered and dynamic backgrounds); however, no extant approach to background modelling is robust to the wide range of conditions present

in real world scenes [33, 78]. Ability to automatically model such background scene variation from input video will vastly extend the operational scenarios in which video analytics and surveillance can be deployed with success.

1.2 Related work

1.2.1 Overview

Much research has considered the challenge of background modelling and foreground detection, with particular growth in this area during the last decade. Several surveys can be found in the literature; in chronological order, the following are notable. Mc Ivor [69] surveyed a representative sample of the published techniques for background subtraction and analyzed them with respect to foreground detection, background maintenance and post-processing. Piccardi [78] provided a review of the main methods and an original categorization based on speed, memory requirements and accuracy. Radke et al. [80] presented a systematic survey of the common processing steps and core decision rules in modern change detection algorithms, including significance and hypothesis testing, predictive models, techniques based on the shading model [77] and background modelling. Elhabian et al. [26] surveyed many existing schemes in the literature of background removal, including common pre-processing algorithms used in different situations, different background models, commonly used model updating techniques and model initialization methods. Cristani et al. [13] proposed a comprehensive review of the background subtraction methods that consider frequency bands other than just the visible portion of the optical spectum (e.g., also including audio and infrared). Bouwmans et al. [7] provided a comprehensive survey of statistical background modelling methods that they subsequently extended and revised [5, 6, 8]. Brutzer et al. [9] identified the main challenges of background subtraction in the field of video surveillance and then compared the performance of nine background subtraction methods with post-processing according to their ability to meet those challenges.

In the following discussion, an overview of this vast literature will be given that is organized according to five underlying mathematical modelling approaches: Probability Density Function (PDF) models, subspace models, sample-based models, neural network models and clustering models.

1.2.2 PDF models

Many approaches have been developed based on the assumption that the history of a pixel's intensity value can be modeled by one or more PDFs. Such approaches can be categorized by the particular density function that they employ as well as their method of estimation.

1.2.2.1 Gaussian and Gaussian Mixtures

The most commonly employed density model is the Gaussian. What appears to be the earliest application of a PDF to background modelling made use of a single Gaussian [101]. This work subsequently was extended via an updating procedure that relied on pixel change classification [32] as well as through use of a "generalized" Gaussian function [45].

Use of a single Gaussian limits background modelling to be applicable to only unimodal phenomena. Thus, even a simple two state process (e.g., a light that could be either of two colours) are beyond the scope of such models. A straightforward extension that does capture such phenomena comes via the Gaussian Mixture Model (GMM), which correspondingly has been applied to background modelling [90]. Various extensions also have been developed, including improved update procedures [43, 52], ability to automatically select the number of components [28, 111, 112] and the ability to estimate not only the mean and variance of each model, but also the probability distribution of these parameters [79].

Various other efforts have considered ways to enhance further the descriptive richness of GMMs so as to meet more challenging problems in image modelling. To robustly detect foreground in the presence of sudden illumination changes and shadows, a mixture model of generalized Gaussians was used [1]. To deal with the dual challenges of fast adaptation to

scene changes and achieving a representation of an empty scene, research has considered dual GMMs: one with fast update and the other with slow update [27]. Other research has concentrated on responding to the challenges of dynamic backgrounds, via use of Fuzzy GMMs [24], coarse, block-based detection [81] and larger regional support [94] for GMM parameter estimation. Yet other research has attempted to handle multiple difficult challenges in a single system (shadows, illumination variation, dynamic background, stopped and removed objects), e.g., an approach was proposed based on split Gaussian models for foreground and background that is further augmented with motion-based change detection [98].

1.2.2.2 Student's t or Dirichlet mixture models

Other parametric distributions than Gaussians have been used for mixture modelling of backgrounds. Student's t-mixture model (STMM) has been shown to be very robust against noise encountered in practice due to its more heavy-tailed nature [71]. Other reseach used Dirichlet mixture models to provide more robust estimation in the presence of a dynamic background; e.g., a Dirichlet process Gaussian mixture model followed by probabilistic regularization [35, 36] and a Dirichlet mixture model updated by an online variational Bayes approach [30].

1.2.2.3 Non-parametric estimation models

A background having fast variations cannot be accurately modeled with a small number of PDFs. Non-parametric techniques have been used to address this challenge. One such approach estimated the probability of observing pixel intensity values based on a sample of recent values over time using Kernel Density Estimation (KDE) for each pixel [25]. This KDE framework subsequently was extended to be used over a joint domain-range representation of image pixels, in order to directly model the multi-modal spatial uncertainties and complex dependencies between the domain and range [86]. Yet another approach employed projection pursuit density estimation (PPDE) [76], which had faster run-time and was able to deal with

a higher dimensional feature spaces compared to KDE.

1.2.2.4 KDE-GMM hybrid models

Some research has considered combining non-parametric and parametric estimation techniques to overcome some shortcomings of both. In order to deal with highly dynamic scenes, a KDE-GMM hybrid model was constructed to model the spatial dependencies of neighboring pixel colors [58]. To approximate the background color distribution precisely and also be robust to various spatial noise sources, a mixture of a non-parametric regional model (KDE) and a parametric pixel-wise model (GMM) was proposed [23].

1.2.3 Subspace models

Assuming foreground objects appeared rarely compared to dominant background scenes, backgrounds can be modelled as the primary subspace of the input video space, and foregrounds can be detected as the outliers or sparse parts.

1.2.3.1 Reconstructive subspace models

Early application of subspace learning to background/foreground estimation computed the difference between the PCA reconstruction of an image and the original input [73]. Several limitations of this model and corresponding improvements can be found in the survey made by Bouwmans [4]. As examples, to deal with PCA's limitation that foreground objects must be small and can't be static for an extended time in training sequences, an error compensation process is introduced to reduce the foreground objects' influence [103] and an iterative optimal projection method is used to exclude foreground regions while establishing the background model [44]. Another limitation is that PCA is a batch model so that it's computationally intensive to update. To solve this problem, several methods have been proposed to keep the background model updated incrementally [54, 82, 87, 97]. Besides, it is not straightforward to integrate multi-channel data in such approaches, which has been

addressed via weighted 2-dimensional PCA [37].

Beyond PCA, other reconstructive subspace models also have been used, e.g., Independent Component Analysis (ICA) [104], Incremental $Rank - (R_1, R_2, R_3)$ Tensor-based Subspace learning Algorithm (IRTSA) [53], Locality Preserving Projections (LPP) [49] and Tensor Locality Preserving Projections (Ten-LoPP) [50]. Compared to PCA, these approaches are very efficient in the case of illumination changes.

1.2.3.2 Robust subspace models

In robust subspace models, the background and the foreground are separated based on a low-rank and sparse decomposition. With respect to foreground/background image estimation, this decomposition has been done in a Robust Principal Components Analysis (RPCA) [10] as well as Robust Non-negative Matrix Factorization (RNMF) [34, 51]. Robust subspace models avoid using free parameters, e.g. a distance threshold, and directly extract foreground objects as the sparse outliers by simple optimization algorithms. A limitation of these models is that they make strong assumptions regarding strict enforcement of low-rank structure and sparsity and therefore have trouble dealing with the highly noise corrupted nature of real-world video surveillance data.

1.2.3.3 Subspace tracking models

In order to estimate and track non-stationary subspaces when streaming data vectors are corrupted with outliers and have missing data values, GRASTA (Grassmannian Robust Adaptive Subspace Tracking Algorithm) was presented, which also was able to separate background and foreground on-line [38]. To be able to detect foreground robustly in camera jitter, low-rank subspaces were maintained by Transformed GRASTA (t-GRASTA) through an image alignment process [39]. Different from GRASTA, which used the l_1-norm function as a convex relaxation of the ideal sparsifying function, pROST (l_p-quasi-norm Robust subspace tracking) approached the problem with a smoothed l_p-quasi-norm function and it

outperforms GRASTA in the case of multi-modal backgrounds [85].

1.2.4 Sample-based models

In sample-based models, for each pixel, a set of values taken in the past at the same location or in the neighbourhood are stored as background samples. Subsequently, this set is compared to the current pixel value in order to determine whether that pixel belongs to the background.

Visual Background Extractor (ViBe) was the first proposed sample-based algorithm, where a random selection policy was used to ensure a smooth exponentially decaying lifespan for the sample values that constitute the pixel models [2]. A series of modifications that alter the operation of ViBe have been made [93], e.g., the inhibition of propagation around internal borders, a color distortion metric and an adaptive threshold. Based on the ViBe framework, a Pixel-based Adaptive Segmenter (PBAS) method was proposed, which used pixel-level feedback loops to dynamically adjust the internal parameters of ViBe without user intervention [41]. Further improvements were made to PBAS, such as, combining spatiotemporal binary features with color information and measuring local segmentation noise levels based on the detection of blinking pixels [88, 89]. Additionally, instead of the random selection policy in ViBe, the efficacy of stored background samples was estimated based on occurrence statistics so that it is capable of removing the least useful background samples [95]. To overcome the drawback that single models cannot support clear judgment of whether a current pixel comes from the neighbourhood or not, two interrelated sample models have been used: a self-model, which consists of history values at the same position, and a neighbourhood-model described by neighbourhood pixel values [106].

1.2.5 Neural network models

In this category, the background is represented by means of a neural network suitably trained on representative frames, then the network further learns how to classify each pixel as background or foreground.

To achieve real-time object segmentation of high-resolution images, several parallelized neural network architectures were proposed, e.g. a neural network (NN) that formed an unsupervised Bayesian classifier [14, 15], an unsupervised competitive neural network (CNN) [60] and a dipolar CNN that improved over CNN by using a dipolar representation to capture the intrinsic directionality of data [61].

Another class of approaches were proposed based on self organization through artificial neural networks named Self-Organizing Background Subtraction (SOBS), which was robust to moving backgrounds, gradual illumination changes, cast shadows and bootstrapping [64]. Subsequently, several improvements were made to SOBS. A SOBS-CF approach was proposed that added a spatial coherence variant to SOBS to enhance robustness against false detections and also formulated a fuzzy model to deal with decision problems typically arising when crisp settings (binary-valued rule based) are involved [65]. The algorithm SOBS-SC provided further robustness against false detections by introducing spatial coherence into the background update procedure in SOBS [66]. An enhanced version of SOBS called 3dSOBS+ introduced initial background estimation, shadow detection and removal and spatial coherence into SOBS [67]. Other self-organizing neural network (SONN) methods include the Growing Neural Gas (GNG), which learns the distribution of visual features [48]; a hierarchical architecture composed of SONN, which models inherent hierarchical relations in the input data and is more flexible in adaptation than the original SONN [74]; and an ART-type network (adaptive resonant theory) approach, which not only possesses the stable self-organized structure and learning ability of an ART-based network, but also uses a neural merging process to adapt to the variability of the input data [63].

The above methods were based on weighted neuron models, approaches relying on a weightless neural network architecture were also proposed such as WiSAR [16] and CwisarDH [17], which have yielded good results in terms of recent overall ranking on the standard CDnet change detection dataset [33].

Finally, instead of using a neural network alone, a multivalued discrete neural network

was used to detect and correct some of the deficiencies and errors of an underlying Mixture of Gaussians algorithm [62].

1.2.6 Cluster models

Cluster models suppose that each pixel can be temporally represented by clusters. They are particularly useful in representing multi-modal backgrounds. In order to represent a compressed form of a background model for a long image sequence, a codebook algorithm was proposed, which quantized background values at each pixel into codebooks [46]. To improve the basic codebook algorithm, two features have been added: layered modelling and adaptive codebook updating [47]. To economize computation time and save space, an online clustering background reconstruction algorithm was proposed, where cluster centers and appearance probabilities of each cluster were calculated and clusters with the appearance probability greater than threshold are selected as the background pixel intensity value [102]. Instead of clustering based on intensity or color, other features were also used in cluster models, such as similarity of intensity changes [107], local self-similarity descriptors [42], local binary patterns (LBP) and photometric invariant color measurements [105].

1.2.7 Other related approaches

Some background subtraction models are built based on simple and basic methods such as median models [68], histogram analysis over time [110], as well as Mahalanobis distance and Euclidean distance [3]. Some methods estimate the expected background using filters such as the Self-Adaptive Kalman filter [29] or Chebyshev filter [12]. Other algorithms achieve probabilistic segmentation of foreground/background based on the Quadratic Markov Measure Field models [40], Markov Random Field build on probabilistic superpixels [83] or Chebyshev probability inequality [70]. Learning algorithms were also used to classify imagery into foreground and background, e.g. probabilistic support vector machine (SVM) [57] and on-line updated Support Vector Regression (SVR) [96]. In order to fill in homogeneous re-

gions of foreground objects and detect sudden global changes of brightness, an integrated background model was presented, which consists of three complementary approaches: pixel-level, region-level and frame-level modellings [72, 91]. To deal with incessantly passing or temporally static foreground, a multilayer background modelling algorithm was presented, where object-wise regions were clustered using spatio-temporal cohesion together with spectral similarity by comparing inputs with background layers [75]. To be robust in the cases of sudden illumination fluctuation as well as burst moving background, the background has been modelled based on co-occurrence probability-based pixel pairs (CP3) [55, 56]. To detect foreground objects from dynamic and shadow backgrounds in real time, a multiscale background model was presented, which can propagate motion measurements across different scales [59]. A physics-based change detection technique called Spectral-360 was also used in change detection, which is based on the dichromatic color reflectance model [84].

1.2.8 Evaluations

Several datasets have been built to evaluate and compare background subtraction algorithms, e.g., Wallflower [91], PETS [108], BMC [92], CDnet dataset [33]. A comprehensive introduction to those datasets can be found in Bouwmans's survey [6]. Among all the datasets available, the CDnet and BMC datasets are the most recent and large-scale. The BMC dataset consists of 10 synthetic and 9 real videos and focuses on outdoor situations with weather variations such as wind, sun and rain. The 2012 CDnet dataset consists of 31 videos representing 6 categories (Baseline, Dynamic Background, Camera Jitter, Intermittent Object Motion, Shadows, Thermal) selected to cover a wide range of challenges in surveillance. Moreover, CDnet provides a website, *www.changedetection.net*, that allows users to upload results and compare them against those of others. So, the comparison of many recent and more complex methods is available under the benchmark of CDnet dataset.

Currently, there are 40 methods ranked on the 2012 CDnet dataset and 35 provide results for all categories, see Table 1. The ranking in Table 1 is calculated via averaging

ranking values across all the categories; so, the smaller the ranking value is, the better the corresponding method performs. According to the ranking value, except for unpublished and anonymous methods, sample-based models [41, 88, 89] perform the best. The original GMMs [90] and GMMs with mild improvement [43, 111] place around the bottom ten. But dual GMMs with fast and slow updates [27] and Dirichlet process Gaussian mixture model [35] are within the top ten. The only physics-based method [84] is within the top five. Neural network models rank in the middle. There are few subspace and cluster models appearing in the ranking. The simple and basic methods such as histogram [110], Mahalanobis distance and Euclidean distance [3] are at the bottom.

1.3 Contributions

In the light of previous research in background image modelling and foreground detection, the following are the main contributions of the current work.

- An analysis is developed that relates marginalized spatiotemporal oriented energy (MSOE), spatial orientation (SO) and chromatic measurements (CM) features to background modelling.
- The three features are algorithmically embedded in a selected framework called ViBe and the resulting algorithm is implemented in software for background image modelling and foreground detection.
- The performance of different feature combinations (MSOE, SO and CM) is compared based on the CDnet dataset. The developed change detection system is evaluated both qualitatively and quantitatively with state-of-the-art algorithms based on the same dataset.

Index	Method Name on 2012 CDnet dataset results	Model Category	Ranking
1	CDet	anonymous	3.50
2	PAWCS [89]	sample-based	3.50
3	SuBSENSE [88]	sample-based	5.17
4	Spectral-360 [84]	other	7.67
5	SGMM-SOD [27]	PDF	8.50
6	PBAS-PID	unpublished	9.50
7	PBAS [41]	sample-based	11.17
8	DPGMM [35]	PDF	11.67
9	GPRMF	anonymous	11.67
10	Pixel Based Adaptive Foreground Extractor (PBAFE)	anonymous	12.00
11	CwisarD [16]	neural network	12.50
12	PSP-MRF [83]	other	14.67
13	SOBS_CF [65]	neural network	16.17
14	CDPS [40]	other	17.00
15	Chebyshev prob. with Static Object detection [70]	other	17.17
16	SC-SOBS [66]	neural network	17.67
17	Multi-Layer Background Subtraction [105]	cluster	17.67
18	RMoG (Region-based Mixture of Gaussians) [94]	PDF	18.50
19	SGMM [28]	PDF	19.50
20	KNN [112]	PDF	20.17
21	SOBS [64]	neural network	20.33
22	KDE-Integrated Spatio-temporal Features [72]	other	21.83
23	GMM - KaewTraKulPong [43]	PDF	22.00
24	KDE - ElGammal [25]	PDF	23.33
25	KDE - Spatio-temporal change detection [107]	cluster	23.83
26	Bayesian Background [79]	PDF	25.83
27	GMM - Stauffer & Grimson [90]	PDF	26.67
28	TUBITAK UZAY 1 [52]	PDF	28.17
29	GMM — Zivkovic [111]	PDF	29.00
30	Local-Self similarity [42]	cluster	29.33
31	GMM - RECTGAUSS-Tex [81]	PDF	29.33
32	pROST [85]	subspace	29.50
33	Histogram [110]	other	30.83
34	Mahalanobis distance [3]	other	32.83
35	Euclidean distance [3]	other	34.00

Table 1: Average ranking across all the categories in 2012 CDnet dataset, 2015-05-13 3:00 p.m.

1.4 Outline

This first section of the report has motivated research in background image modelling and foreground detection and thereby placed the present work in context. Section 2 describes the proposed technical approach, including a novel feature set for representing background images as well as specification of how these features can be embedded in a sample-based modelling framework. Section 3 provides extensive empirical evaluation of the proposed approach. Finally, Section 4 provides a summary and conclusions.

2 Technical approach

2.1 Overview

In this section, the developed approach to background modelling and change detection will be developed in detail. Change detection is realized by the comparison between new observations and a background model. The presentation will be expanded based on how the background model is represented and organized. First, the primitive feature representation will be presented. Each pixel in a real-world digital video sequence is usually stored in the form of RGB color or gray-level intensity; however, this representation is not enough for understanding the content of videos. In the proposed approach, dynamic features, spatial orientation features and chromatic features are extracted as the primitive representation for background model and new observations. Second, the process of organizing a background model and classifying new observations based on the proposed primitive features will be presented. A wide variety of approaches previously have been proposed for background model definition, as introduced in the related work section of this report. For the current work, a sample-based framework is adopted, which derives from an algorithm called ViBe [2], as it previously has been shown to be a strong general performer and the proposed features map well onto this model. After briefly introducing ViBe, details regarding mapping multiple features into the framework will be presented.

2.2 Primitive feature representation

The employed primitive feature descriptors for background modelling include dynamic features, spatial orientation features and chromatic features. They separately gather different aspects of information from input imagery. For example, when a person walks down a street, his motion is captured by dynamic features, the texture or pattern on his clothes or face is captured by spatial orientation features and the color or intensity information is captured

by chromatic features. Judicious combination of these different features can support background modelling that can capture a wide range of real-world scenarios as well as distinguish subsequent change relative to an acquired model.

2.2.1 Dynamic features

Video sequences induce a variety of patterns in visual spacetime [100]. For example, static objects have a very different spacetime orientation pattern compared to moving objects. Therefore, a spatiotemporal oriented decomposition of an input video is an efficient approach to represent the information that is implicit in videos. In particular, the input video is subjected to an oriented, bandpass decomposition in both space and time via application of a spatiotemporal filter bank [18].

Previous work based on this approach has yielded state-of-the-art performance in dynamic texture recognition [20] and scene recognition [21] as well as allied areas of target tracking [11] and human action recognition [22]. Such success owes to the approach's ability to represent a wide range of imaging variations (owing to lighting, motion, etc.). Further, it is amenable to real-time computation on GPUs [22, 109], which is crucial for application to time critical tasks, e.g., surveillance.

The filtering is realized in terms of second derivative of $3D$ Gaussian filters,

$$G_{2\theta}(x, y, t) = \frac{\partial^2 \kappa e^{-(x^2+y^2+t^2)}}{\partial \theta^2} \tag{1}$$

and their Hilbert transforms, $H_{2\theta}(x, y, t)$, where vector θ represents the direction of the filter's $3D$ axis of symmetry and κ is a normalization factor. The Hilbert transform is defined such that it keeps the amplitude of a signal the same while it shifts its phase by $\frac{\pi}{2}$. Both $G_{2\theta}(x, y, t)$ and $H_{2\theta}(x, y, t)$ are steerable, i.e. can be written as a linear sum of rotated versions of itself [31],

$$G_{2\theta}(x, y, t) = \sum_{i=1}^{M_g} k_i(\theta) G_{2\theta_i}(x, y, t), \tag{2}$$

$$H_{2\theta}(x, y, t) = \sum_{j=1}^{M_h} k_j(\theta) H_{2\theta_j}(x, y, t), \tag{3}$$

where $G_{2\theta_i}(x, y, t)$ is a basis filter for steering $G_{2\theta}(x, y, t)$ and $k_i(\theta)$ is the corresponding interpolation function, and similarly for $H_{2\theta}(x, y, t)$. There must be at least $M_g = 6$ basis filters for steering $G_{2\theta}(x, y, t)$ and $M_h = 10$ for steering $H_{2\theta}(x, y, t)$ [18]. So, at least 10 directions are needed to steer the pair, which are conveniently taken as the verticies of a dodecahedron with antipodal directions identified, as the dodecahedron evenly samples the sphere. The filters are taken in quadrature, to yield the following local oriented energy measure,

$$E_\theta(x, y, t) = (G_{2\theta} * I)^2 + (H_{2\theta} * I)^2, \tag{4}$$

where $I \equiv I(x, y, t)$ denotes the input imagery and $*$ symbolizes convolution. If a pair of filters has the same frequency response but differs in phase by $\frac{\pi}{2}$, they are in quadrature. The G_2 and H_2 are employed because they can measure local orientation direction and strength [31]. Additionally, the steerable and separable formulation of these filters leads to efficient computations.

For the case of capturing pure pattern dynamics [20], each spatiotemporal oriented energy measurement, (4), is confounded with spatial orientation. To remove this difficulty, the spatial orientation component is discounted by marginalizing this attribute. A pattern exhibiting a single spacetime orientation (e.g., motion) manifests itself as a plane through the origin in the frequency domain [99]. Correspondingly, summation of a set of energy measurements (4) spanning such a frequency domain plane removes the influence of purely spatial oriented energy. Let each plane be parameterized in terms of its unit normal, $\hat{\mathbf{n}}$, which is also the orientation of the pure dynamic pattern in visual spacetime. The energy measure, (4), can now be refined to become marginalized spatiotemporal oriented energy (MSOE),

$$\tilde{E}_{\hat{\mathbf{n}}}(x, y, t) = \sum_{i=0}^{N} E_{\theta_i}(x, y, t), \tag{5}$$

where θ_i represents one of the $N+1$ equal spaced orientations distributed within the plane of normal $\hat{\mathbf{n}}$ and $N=2$ is the order of the Gaussian derivative filter (4); for details see [19];

The resulting oriented energies, (5), are confounded with local contrast. This makes it impossible to determine whether a high response from a particular filter is indicative of a close match with the underlying structure or is instead a low match that yields a high response due to significant contrast in the signal. To obtain a purer measure of spatiotemporal dynamic energy, the energy measures (5) are normalized by the sum of the oriented responses at each point,

$$\hat{E}_{\hat{\mathbf{n}}_i}(x,y,t) = \frac{\tilde{E}_{\hat{\mathbf{n}}_i}(x,y,t)}{\sum_{\hat{\mathbf{n}}_j \in S} \tilde{E}_{\hat{\mathbf{n}}_j}(x,y,t) + \epsilon}, \tag{6}$$

where S denotes the set of spatiotemporal orientations, with $\hat{\mathbf{n}}_i$ a particular sample and ϵ a small constant that serves as a noise floor and to avoid numerical sensitivity when the overall energy at a point is small. In addition, to the set of normalized oriented energies at a point, (6), a normalized ϵ is computed, as

$$\hat{E}_{\epsilon}(x,y,t) = \frac{\epsilon}{\sum_{\hat{\mathbf{n}}_j \in S} \tilde{E}_{\hat{\mathbf{n}}_j}(x,y,t) + \epsilon}, \tag{7}$$

to explicitly capture lack of texture within the region. (Note that regions where texture is less apparent, e.g., a region of clear sky, the summation in the denominator approaches zero; hence, the normalized ϵ, (7), approaches one and thereby indicates lack of structure.) In the current implementation, the spatiotemporal orientation set, S, consists of six different spacetime orientations, corresponding to, leftward, rightward, upward and downward motion, static and flicker (vertically and horizontally). This MSOE distribution at each spacetime point (x,y,t) is maintained as a histogram with $k=6$ bins,

$$F_{MSOE}(x,y,t) = \left[h_1(x,y,t), \ldots, h_k(x,y,t)\right]. \tag{8}$$

where

$$h_1(x, y, t) = \hat{E}_{static}(x, y, t) + \hat{E}_{\epsilon}(x, y, t), \tag{9}$$

denotes the non-moving energy by combining static energy and non-texture energy, while $h_2(x, y, t), \ldots, h_k(x, y, t)$ respectively corresponds to leftward, rightward, upward, downward motion and flicker energy.

2.2.2 Spatial orientation features

Spatial oriented information is obtained by measuring spatial orientations (SO), which are defined in terms of a set of evenly distributed first-order spatial derivatives. In practice, the measuremts are calculated via application of differently oriented first-order Gaussian derivatives,

$$G_{1\theta}(x, y) = \frac{\partial \kappa e^{-(x^2+y^2)}}{\partial \theta}, \tag{10}$$

along (2D) directions θ. In particular, the measurements are defined as

$$g_i(x, y, t) = G_{1\theta_i} * I(x, y, t) \tag{11}$$

with $\theta_i \in \{0°, 30°, 60°, 90°, 120°, 150°\}$ one of $k = 6$ evenly distributed directions in the image plane. Notice that, unlike the dynamic feature measurements, (6), the spatial measurements, (11), are neither converted to (unsigned) energies nor normalized. Preliminary investigation revealed that maintaining the sign and magnitude of contrast is important to distinguish spatial structure in background modelling, while it is not necessary for purely dynamic characterization.

Analogously to the dynamic feature measurements, (8), the spatial measurements at the spacetime point (x, y, t) are combined by concatenating the $k = 6$ Gaussian derivatives to yield,

$$F_{SO}(x, y, t) = [g_1(x, y, t), \ldots, g_k(x, y, t)]\,. \tag{12}$$

2.2.3 Chromatic features

State-of-the-art background subtraction systems [6, 33] often make use of chromatic information in their modelling. Correspondingly, chromatic measurements (CM) will be incorporated in the proposed primitive feature descriptor. For color input imagery, three measurements at each point taken as RGB color space observations will be used as the chromatic feature representation

$$F_{CM}(x, y, t) = [R, G, B]\,. \tag{13}$$

For grayscale input, e.g. thermal imagery, only intensity of each point is used.

2.3 Background image modelling

The framework for the proposed background image modelling algorithm is based on an algorithm called Visual Background Extractor (ViBe) [2]. ViBe is a sample-based algorithm and it previously has served as the underlying basic framework for many state-of-the-art algorithms, e.g., PBAS [41] and SuBSENSE [88], which are among the top algorithms by CDnet database metrics [33]. Moreover, the ViBe framework is suitable for integrating multiple features ranging from grey scale and color through various derived measures. In the following: First, the basic ViBe framework is briefly introduced; second, the developed approach to mapping multiple features into the framework is described.

2.3.1 ViBe framework

In this section, the general ViBe model and its application to foreground detection will be introduced first. With the model in hand, initialization and update mechanisms will be presented.

2.3.1.1 Model and classification process

ViBe is a sample-based background model. Each pixel, (x, y), in the background image is modeled by a collection of N background sample values

$$B(x, y) = \{b_1, \ldots, b_n, \ldots, b_N\} \tag{14}$$

where b_n is one of the background samples, whose selection is explained in section 2.3.1.2. To classify a current input pixel $p(x, y, t)$ according to its corresponding model $B(x, y)$, it is compared with the closest samples within the background sample set. Here, both the pixel value, p, as well as the sample value, b_i, are given in terms of some primitive feature representation, e.g., as presented in Section 2.2. Denoting by d_n the distance between input $p(x, y, t)$ and b_n, the closest samples are defined as

$$\{b_n \mid d_n < R, n \in [1, N]\} \tag{15}$$

where R is a fixed threshold [1]. That is, those background samples within the intersection of the sphere of radius R centered on $p(x, y, t)$ and the collection of model samples $B(x, y)$ in Figure 1, e.g. b_3. If the number of closest samples is larger than or equal to a given threshold, then the pixel $p(x, y, t)$ is classified as background.

2.3.1.2 Model initialization

The background model is initialized from the first frame. Under the assumption that neighboring pixels share a similar distribution at a given time, the sample set of $B(x, y)$ is filled with values randomly taken from the neighborhood of (x, y) in the first frame, i.e. at time t_0,

[1] The exact nature of the distance metric d_n depends on the feature representation. As explained later, we use the L_1 metric based on preliminary experimentation.

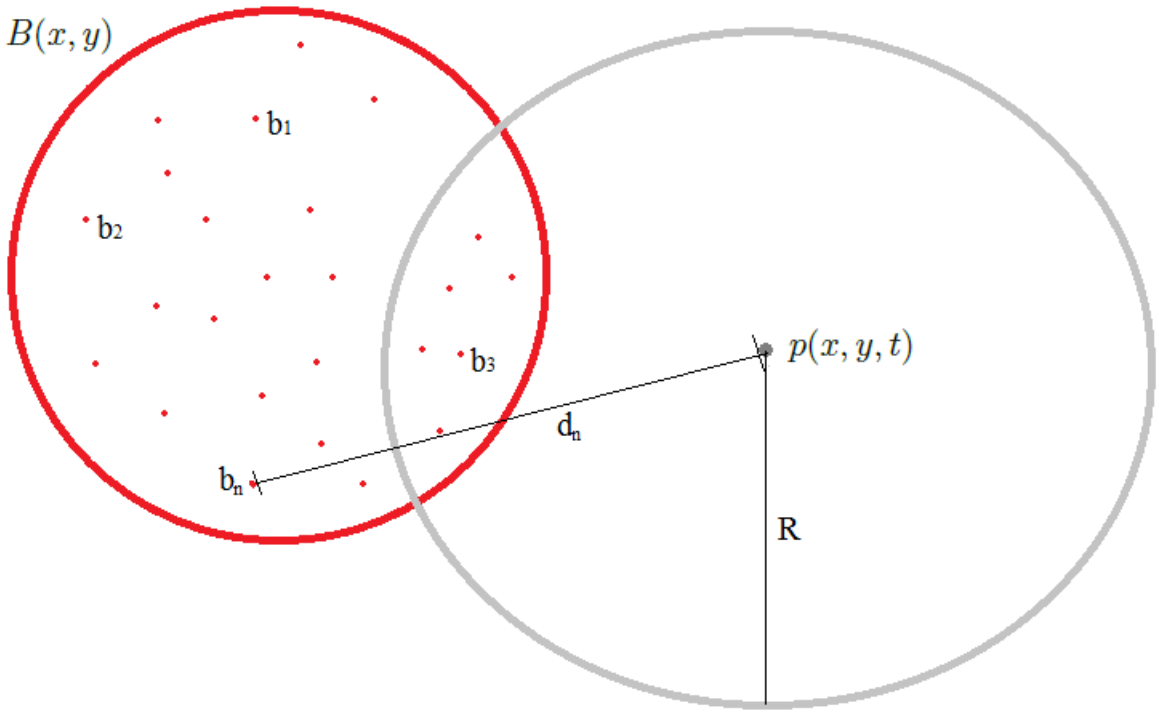

Figure 1: Sample-Based Background Modelling and Foreground Detection. Comparison of a currently observed pixel value, $p(x, y, t)$, with a set of background sample values, b_i, yield classification as background vs. foreground. For the pixel to be classified as background, there must be at least a certain number of samples, b_i, whose distance, d_i, is less than a threshold, R.

$$B(x,y) = \{p(\bar{x},\bar{y},t_0) \mid (\bar{x},\bar{y}) \in \mathcal{N}(x,y)\} \tag{16}$$

where $\mathcal{N}(x,y)$ is the neighborhood of position (x,y). The probability of chosing $(\bar{x},\bar{y})$ is decided by a Gaussian distribution centred on (x,y), so that pixels closer to (x,y) have higher probabilities appearing in $B(x,y)$.

2.3.1.3 Model update

When $p(x,y,t)$ is classified as background, whether it will be used to update its corresponding background model $B(x,y)$ is determined probabilistically. If $p(x,y,t)$ is decided to update the background model, a sample b_n randomly chosen from its corresponding background sample sets (14) with an uniform probability will be substituted by $p(x,y,t)$. When $p(x,y,t)$ is classified as background, it also is used to update the background sample set $B(\bar{x},\bar{y})$, where $(\bar{x},\bar{y})$ is one pixel chosen randomly from its neighborhood $\mathcal{N}(x,y)$ again with a uniform probability distribution. The procedure of updating $B(\bar{x},\bar{y})$ is the same as updating $B(x,y)$. The update process has been theoretically proven to ensure a smooth exponentially decaying lifespan for the background samples [2].

2.3.2 Multiple features in ViBe framework

The original ViBe algorithm [2] only used chromatic features. To map multiple features into ViBe, the framework is extended as follows.

2.3.2.1 Pixel model and classification process

Formally, each input pixel $p(x,y,t)$ consists of M feature descriptors,

$$p(x,y,t) = \{p^{f_1} \dots p^{f_m} \dots p^{f_M}\} \tag{17}$$

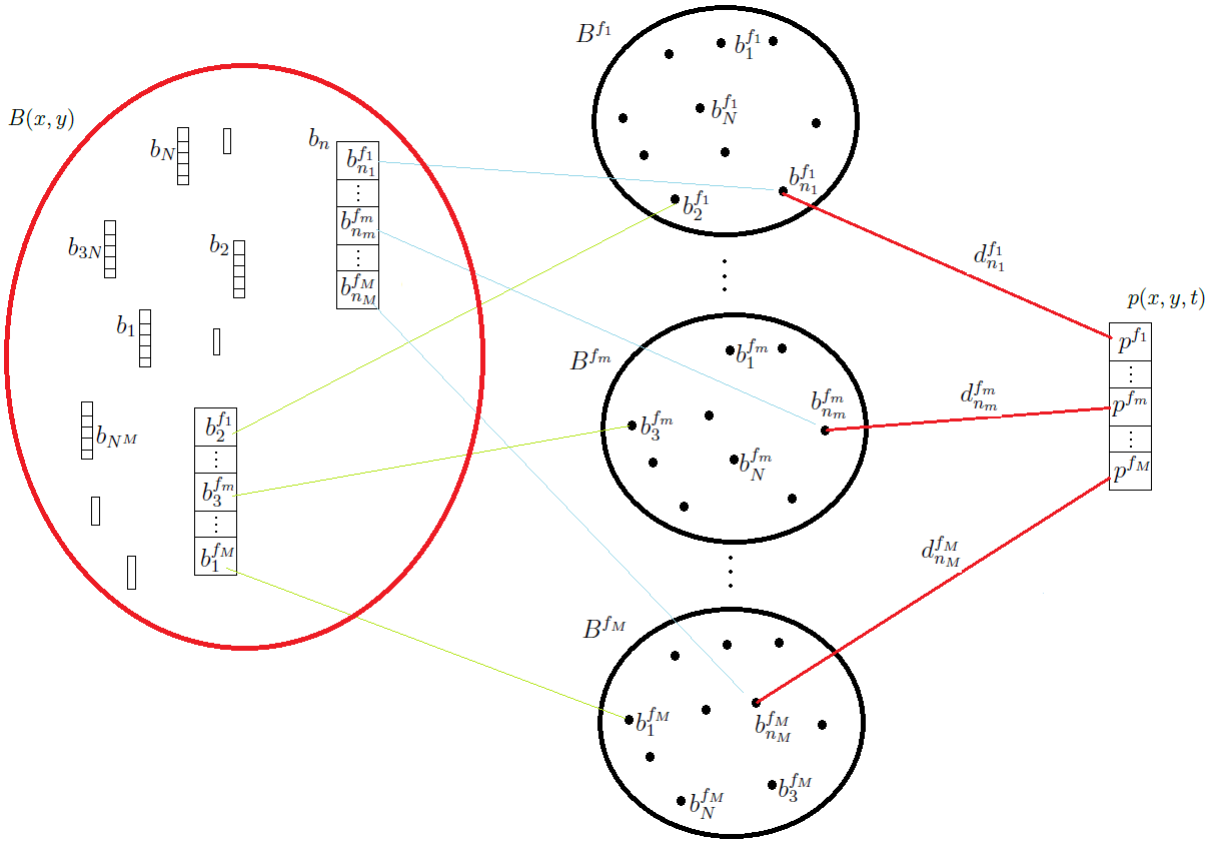

Figure 2: Extension of Sample-Based Background Modelling and Foreground Detection to Multiple Features. See text for explanation.

where $m \in [1, M]$ is the index for each feature, see Figure 2. Correspondingly, there are M sets of background feature samples for the background model, $\{B^{f_1} \dots B^{f_m} \dots B^{f_M}\}$, as shown in Figure 2. In each feature set, there are N samples; and one sample with index n_m in B^{f_m} is denoted by $b^{f_m}_{n_m}$, that is

$$B^{f_m} = \{b^{f_m}_{n_m} \mid n_m \in [1, N]\} \tag{18}$$

Similarly, there is a distance between the input pixel's feature value, p^{f_m} and one of the corresponding background feature samples, $b^{f_m}_{n_m}$,

$$d^{f_m}\left(b^{f_m}_{n_m}, p^{f_m}\right). \tag{19}$$

As in the case of single features, Section 2.3.1, the distance function must be defined appropriately for the particular feature types considered. In the present work, there could be at most three different distance functions separately for MSOE, SO and CM feature comparisons. For the current implementation, L_1 distance function is used for all features.

By choosing one sample from each background feature sample set, we can get one background sample with all features,

$$b_n(x, y) = \{b^{f_m}_{n_m} \mid m \in [1, M]\} \tag{20}$$

where $n_m \in [1, N]$ could be the same or different values for different features. Thus, the background sample set $B(x, y)$ for multiple features, see in Figure 2 is

$$B(x, y) = \{b_n(x, y) \mid n \in [1, N^M]\} \tag{21}$$

The distance between the input pixel (17) and one background sample (20) in the back-

ground set (21) is

$$d_n(x,y) = \sum_{m=1}^{M} \omega^{f_m}(x,y,t) \cdot d^{f_m}\left(b_{n_m}^{f_m}, p^{f_m}\right), \tag{22}$$

where $\omega^{f_m}(x,y,t)$ is the weight for feature f_m in position (x,y) and time t, as defined in Section 2.3.2.3. Hereafter, $\omega^{f_m}(x,y,t)$ will be simplified to ω^{f_m}.

To classify the current input pixel (17) according to its corresponding pixel model, ViBe compares it to the closest samples within the background sample set (21). The closest samples are defined the same as (15), which are those whose distance (22) is less than a fixed distance threshold R. If the number of closest samples is larger than or equal to a given threshold $\#min$, the current pixel $p(x,y,t)$ is classified as background.

Finally, following initial pixelwise background detection, the results are postprocessed with a 11×11 spatial median filter to remove isolated false positive and false negative responses. In preliminary experiments, median filter processing provided superior results to alternative morphological operations.

2.3.2.2 Model initialization and update

The background model is initialized using features from the first frame $t = t_0$. For pixel (x,y), each background feature sample set (18) is initially filled with the corresponding feature values selected from the pixel's neighborhood according to a 2D Gaussian probability distribution centered at that pixel,

$$B^{f_m}(x,y) = \{p^{f_m}(\bar{x},\bar{y},t_0) \mid (\bar{x},\bar{y}) \in \mathcal{N}^{f_m}(x,y)\} \tag{23}$$

where $\mathcal{N}^{f_m}(x,y)$ is the neighborhood of the pixel (x,y) in the feature m image.

The update procedure for the ViBe framework containing multiple features also follows the same procedures mentioned in section 2.3.1.3. If $p(x,y,t)$ is classified as background, whether it will be used to update the background model is decided probabilistically. If

$p(x, y, t)$ is decided to update background feature sample sets (18), then

$$b_{n_m}^{f_m}(x, y) = p^{f_m}(x, y, t), \forall m \in [1, M] \tag{24}$$

where n_m is randomly chosen with an uniform probability from $[1, N]$ for each m.

If $p(x, y, t)$ is also chosen by a probability to update one of the neighboring pixels of $(\bar{x}, \bar{y})$, then

$$b_{n_m}^{f_m}(\bar{x}, \bar{y}) = p^{f_m}(x, y, t), \forall m \in [1, M] \tag{25}$$

where $n_m \in [1, N]$ is randomly picked for each m and $(\bar{x}, \bar{y})$ is a randomly selected pixel from the neighborhood of (x, y), both with an uniform probability.

2.3.2.3 Feature combination

There are seven different feature combinations from our three primitive feature representations – MSOE (8), SO (12) and CM (13). Different selected feature combinations lead to different weight specifications, ω in (22).

When only one feature representation is extracted from input imagery, i.e. $M = 1$, then there are three possibilities,

$$p(x, y, t) \equiv \{F_{CM}\} \| \{F_{SO}\} \| \{F_{MSOE}\}, \tag{26}$$

with $\|$ standing for "or". In this case, there is no need to use a weight, so implicitly $\omega^{f_1} = 1$.

When there are two or three features in the system, the weight is used to adjust the local (pixelwise) emphasis placed on different features according to how static vs. dynamic a pixel is measured to be. In particular, recall that h_1 captures the static energy of a pixel, (9). Similarly, since the h_i are normalized (8), $1 - h_1$ captures the proportion of the local structure that is accounted for by non-static, i.e., dynamic energy. Correspondingly, define

$$\alpha = h_1(x, y, t), \tag{27}$$

$$\beta = 1 - h_1(x, y, t), \tag{28}$$

where α represents the pixel's static status value and β represents its dynamic status value.

Now, additional weights for various feature combinations are defined as follows. If MSOE features, (8), are employed by the system, then the weight assigned to MSOE feature is the pixel's dynamic value (28), which means the system relies more heavily on MSOE features when the input pixel presents as dynamic. In contrast, both CM (13) and SO (12) features capture static attributes (color and spatial orientation, resp.) and are more heavily relied upon when the input pixel presents as static via weighing with (27). More precisely, let $f_1 \equiv F_{MSOE}$, then

$$M = 2 : \omega^{f_1} = \beta, \omega^{f_2} = \alpha \tag{29}$$

$$M = 3 : \omega^{f_1} = \beta, \omega^{f_2} = \omega^{f_3} = \alpha \tag{30}$$

with f_2 and f_3 standing for either F_{SO} or F_{CM}. Alternatively, if only CM, (13), and SO, (12), features are used in the system, then $\omega^{f_1} = \omega^{f_2} = \alpha, (f_1 \equiv F_{CM}, f_2 \equiv F_{SO})$, which can be simplified to $\omega^{f_1} = \omega^{f_2} = 1$.

2.4 Recapitulation

By way of summary, Figure 3 provides a flow diagram that captures the entire proposed approach.

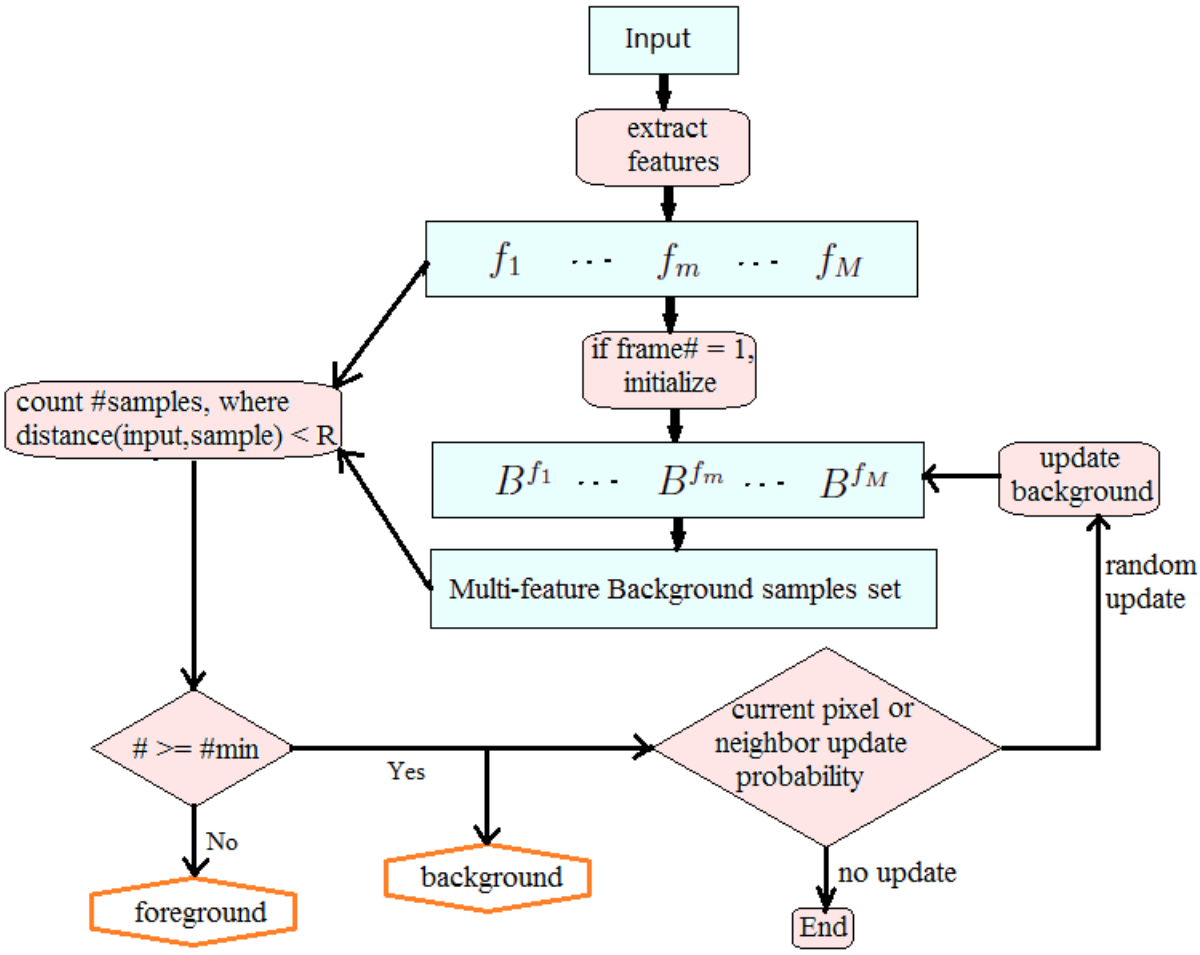

Figure 3: Flow diagram of proposed background subtraction system.

3 Empirical evaluation

3.1 Datasets

To evaluate the performance of the proposed system and compare it with state-of-the-art algorithms, a standard, publicly available dataset, the ChangeDetection.net Video Database, is considered [33]; see Figure 4 as well as the corresponding website `www.changedetection.net`. This dataset contains six video categories with four to six video sequences in each category. The categories encompass a baseline set, dynamic background, camera jitter, intermittent object motion, shadows and thermal imagery. The dataset thereby includes a wide variety of challenging, real-world scenarios. All sequences are available with manually constructed groundtruth that identifies change relative to a training portion of the video. The groundtruth images are produced with 5 labels: static (grayscale value is 0), shadow (50), non-ROI (85), unknown (170), moving (255); see Figure 5. Static and shadow labels are background, moving labels are foreground and other labels are not included into the statistics computation.

3.2 Results

3.2.1 Comparison among different feature combinations

Letter 'C', 'S' and 'D' separately represents chromatic features (CM), spatial orientation features (SO) and dynamic features (MSOE). 'CS' represents the combination of chromatic features and spatial orientation features. 'CD' represents the combination of chromatic features and dynamic features. 'SD' represents the combination of spatial orientation features and dynamic features. 'CSD' represents the combination of chromatic features, spatial orientation features and dynamic features. These letter representations are used hereafter.

The CDnet evaluation protocol is followed to rank order the different proposed feature combinations [33]. Seven individual performance metrics are considered: recall(Re), Speci-

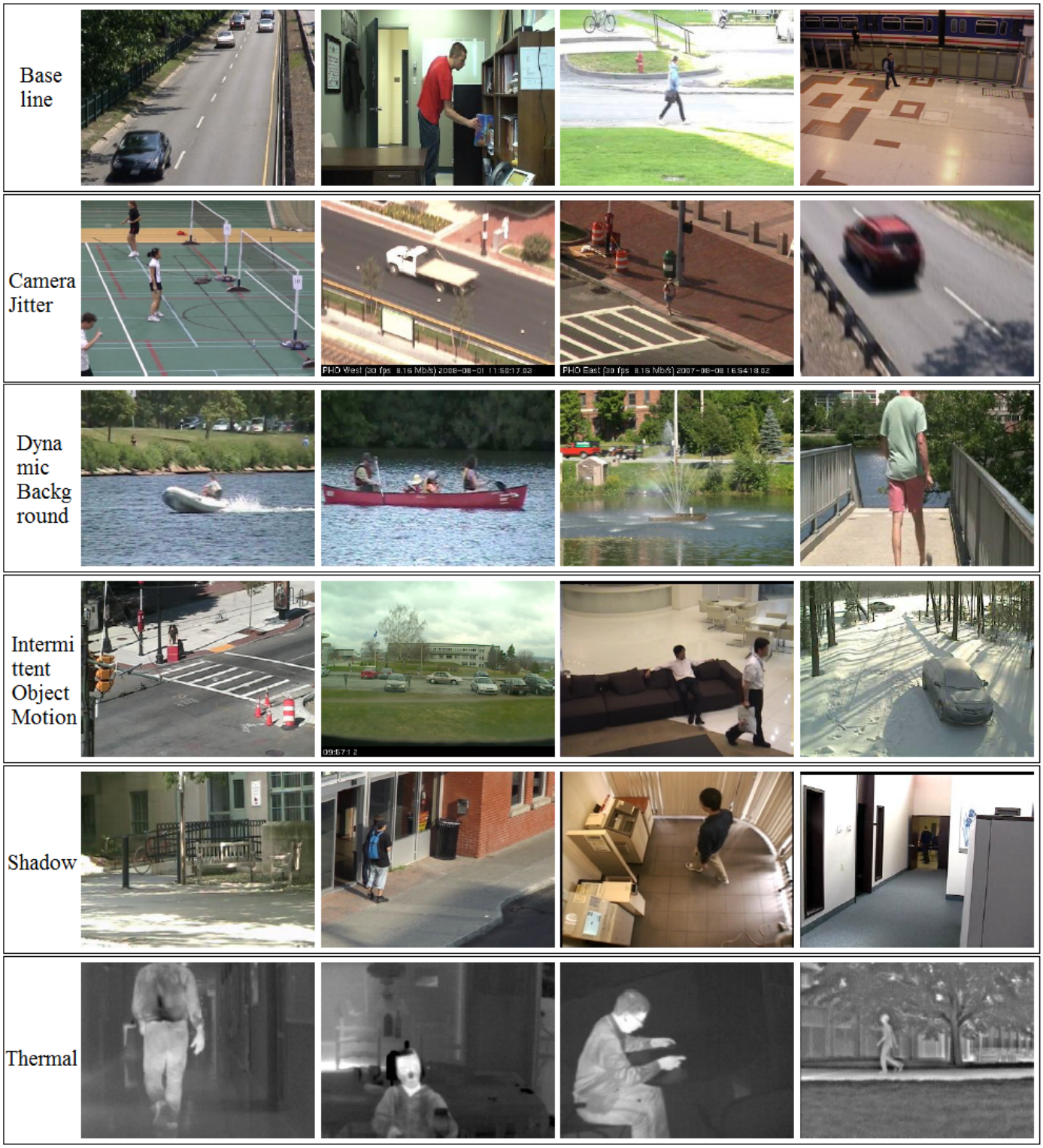

Figure 4: Sample Images from CDnet Dataset.

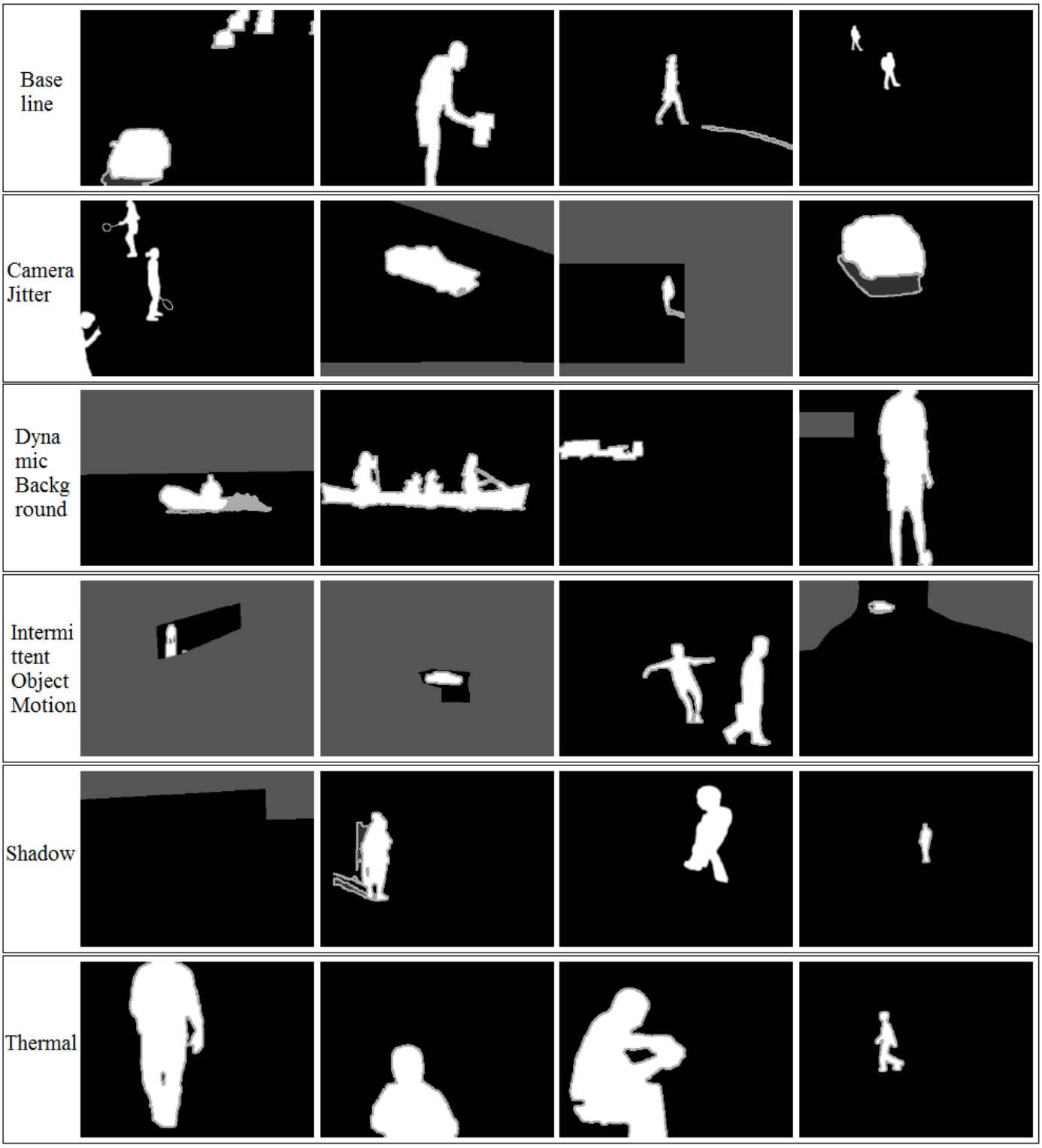

Figure 5: Groundtruth of Sample Images from CDnet Dataset.

ficity(Sp), False Positive Rate(FPR), False Negative Rate(FNR), Percentage of Wrong Classifications (PWC), Precision(Pr) and F-measure. Average performance across all videos in each background category is reported for each metric. Also reported is the average ranking across all metrics, which is calculated according to the average of the seven metrics' ranking numbers.

The rankings of the seven feature combinations for each background category as well as over all are presented in Tables 2 - 8. Based on average ranking, it is seen that: 'C' performs best in the baseline, dynamic background and thermal categories; 'CS' performs best in the camera jitter and shadow categories; 'SD' performs best in the intermittent object motion category. Moreover, 'CS' performs best overall. Although 'SD' is worse than 'S' or 'D' in most categories based on average ranking, the recall and FNR of 'SD' from all tables are better than 'S' or 'D'.

The binary detection results of sample images for each feature combination are presented in Figure 6 - 12. A number of general observations can be made by study of the image-based detection results. Feature combination 'CS' detects results with the highest accuracy and the fewest false positive regions, see Figure 6. Colour alone, 'C', (Figure 7) generally provides the least amount of blur in its detected regions, but detects more false positive regions compared with features containing both color and spatial or spatiotemporal features. Feature combinations 'CD' and 'CSD' (Figures 8 and 9) detect most of the foreground regions in the dynamic background and camera jitter categories, with little false positive detection except in the vicinity of the detected foreground objects' boundaries. The spatial and dynamic features alone, 'S' or 'D', provide the lowest recall, as shown in Figures 11 and 10. Feature combination 'SD', Figure 12, detects more true foreground regions than either features 'S' or 'D' alone.

As a complementary approach to evaluating the various feature combinations, ROCs are presented in the Appendix to this report.

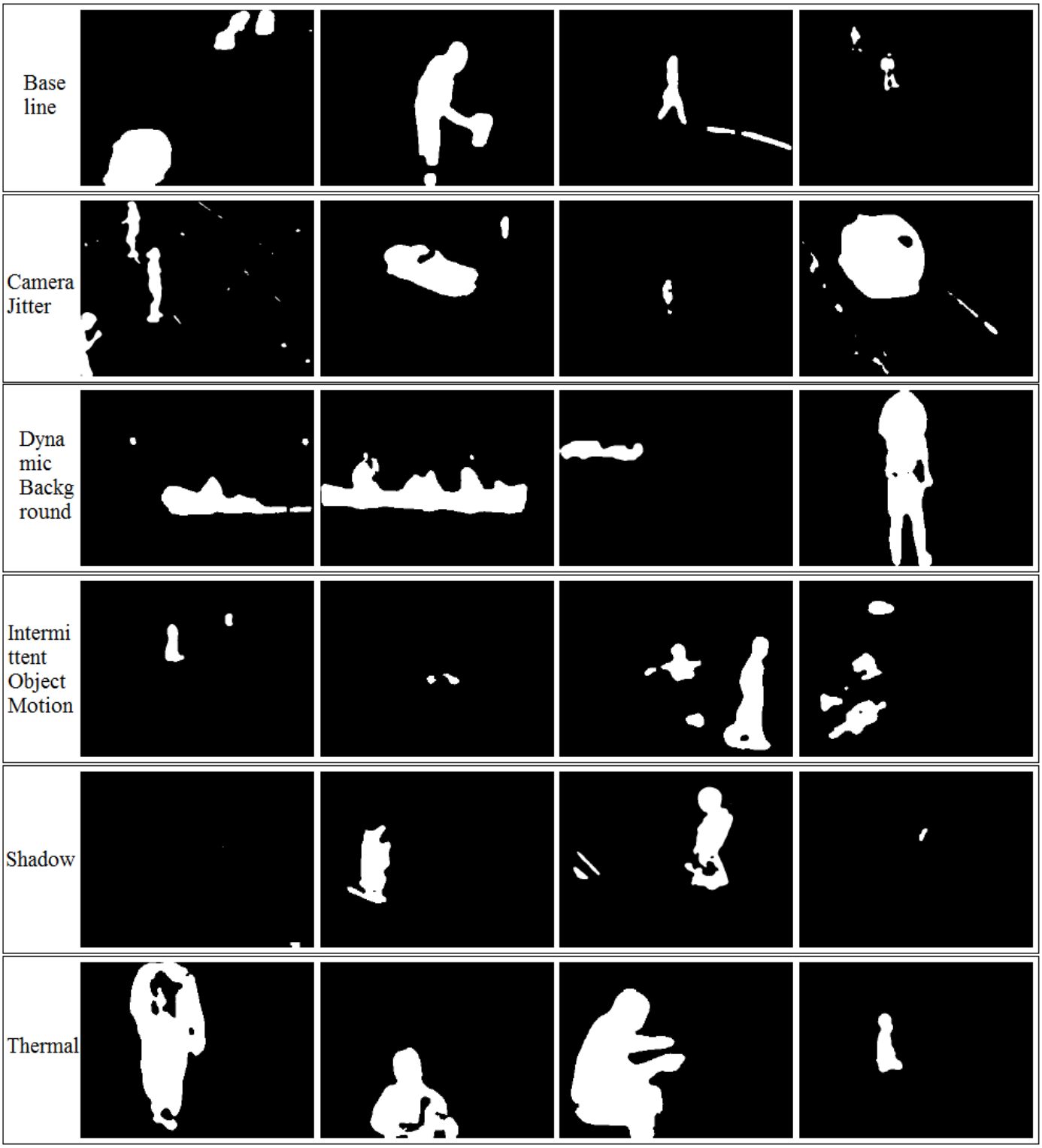

Figure 6: Sample Images' binary results for feature combination 'CS'.

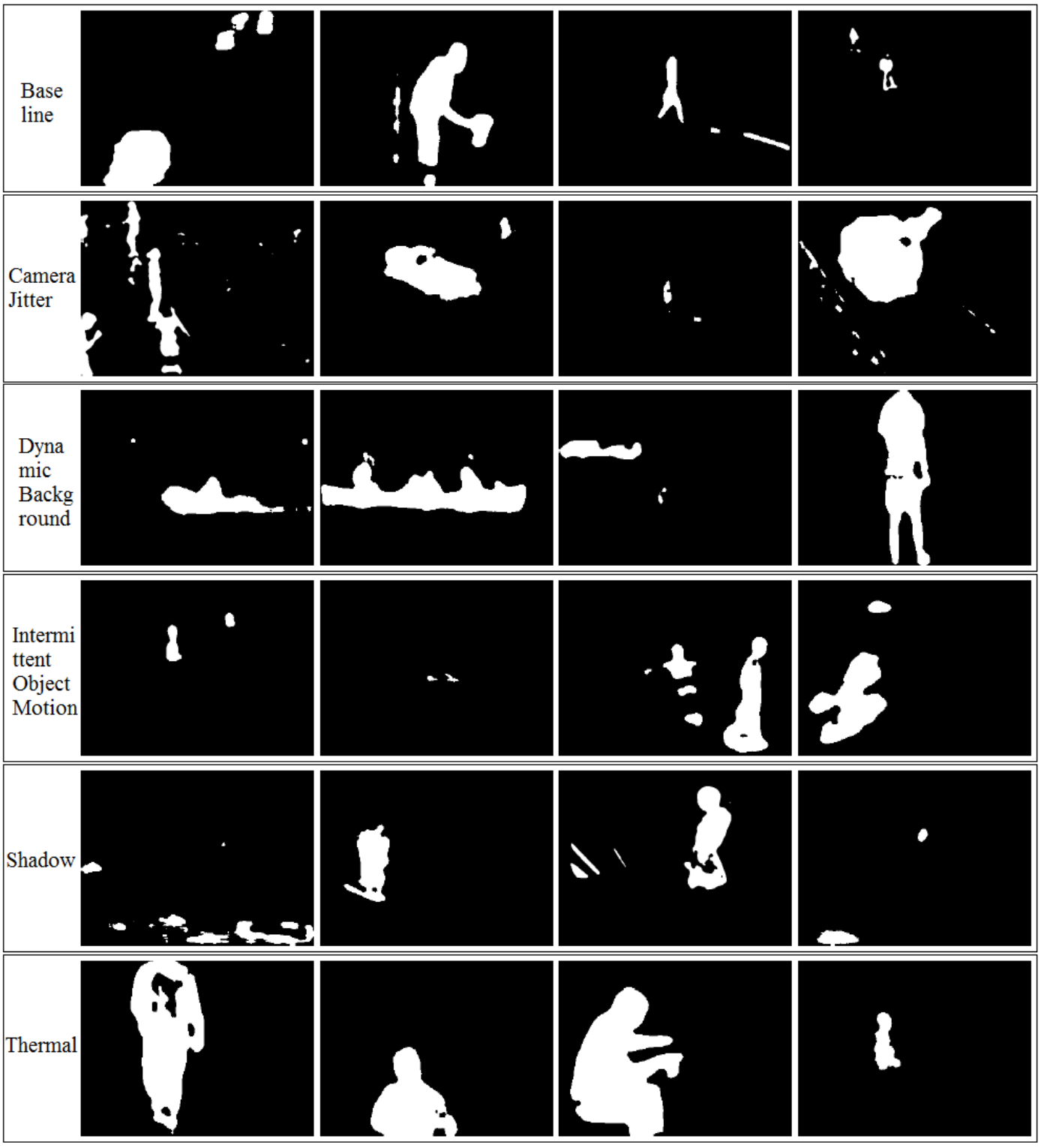

Figure 7: Sample Images' binary results for feature combination 'C'.

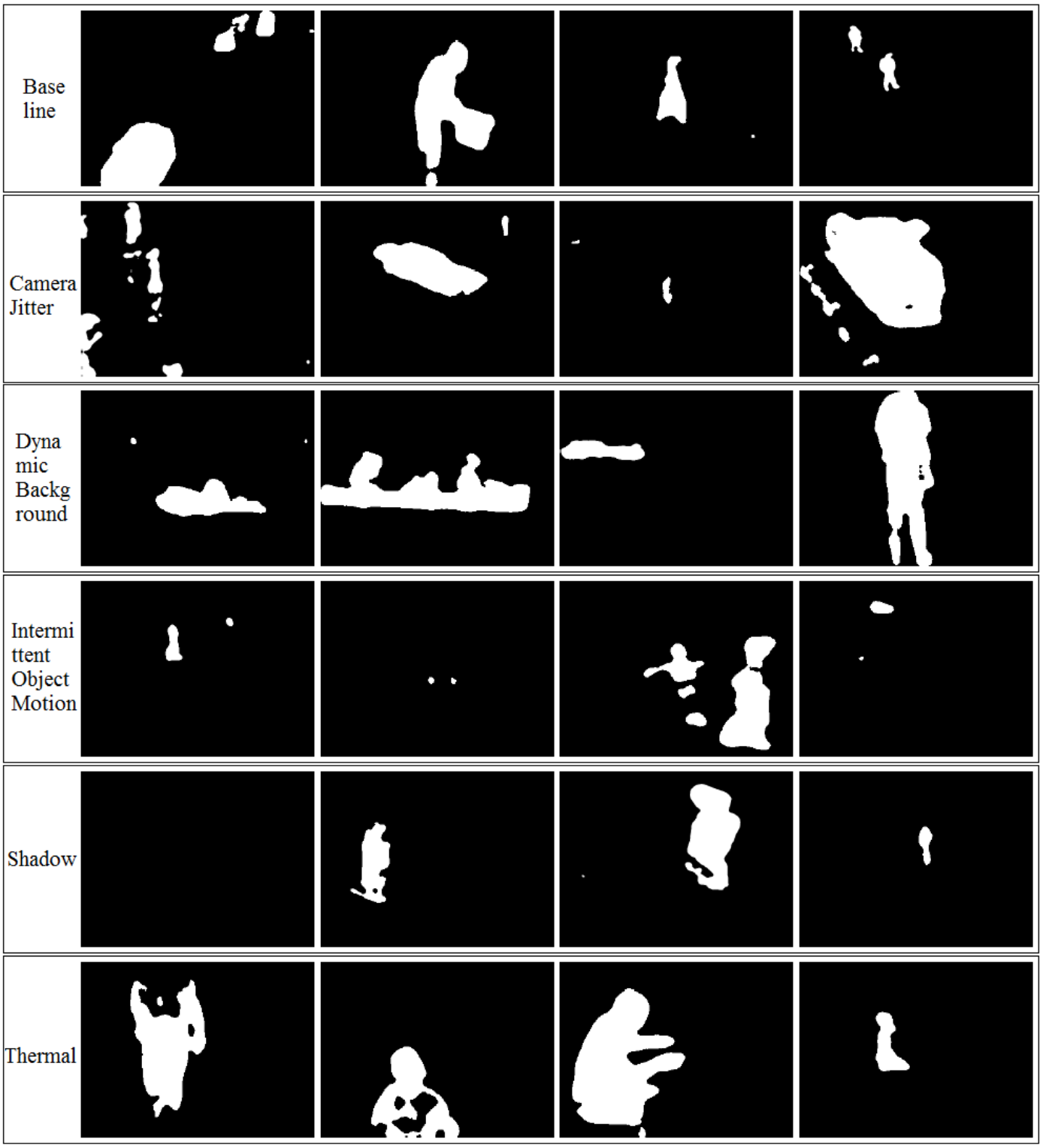

Figure 8: Sample Images' binary results for feature combination 'CD'.

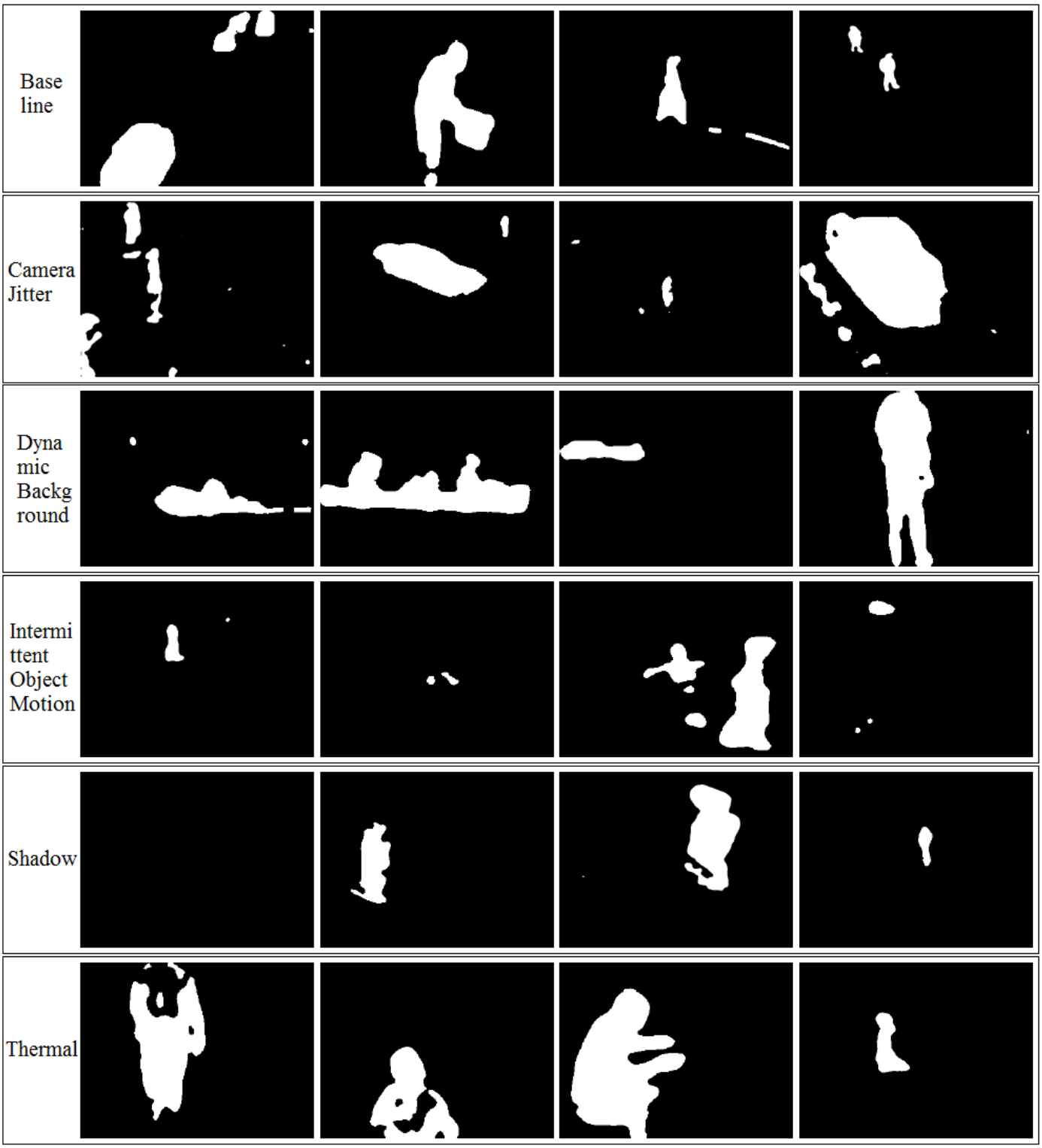

Figure 9: Sample Images' binary results for feature combination 'CSD'.

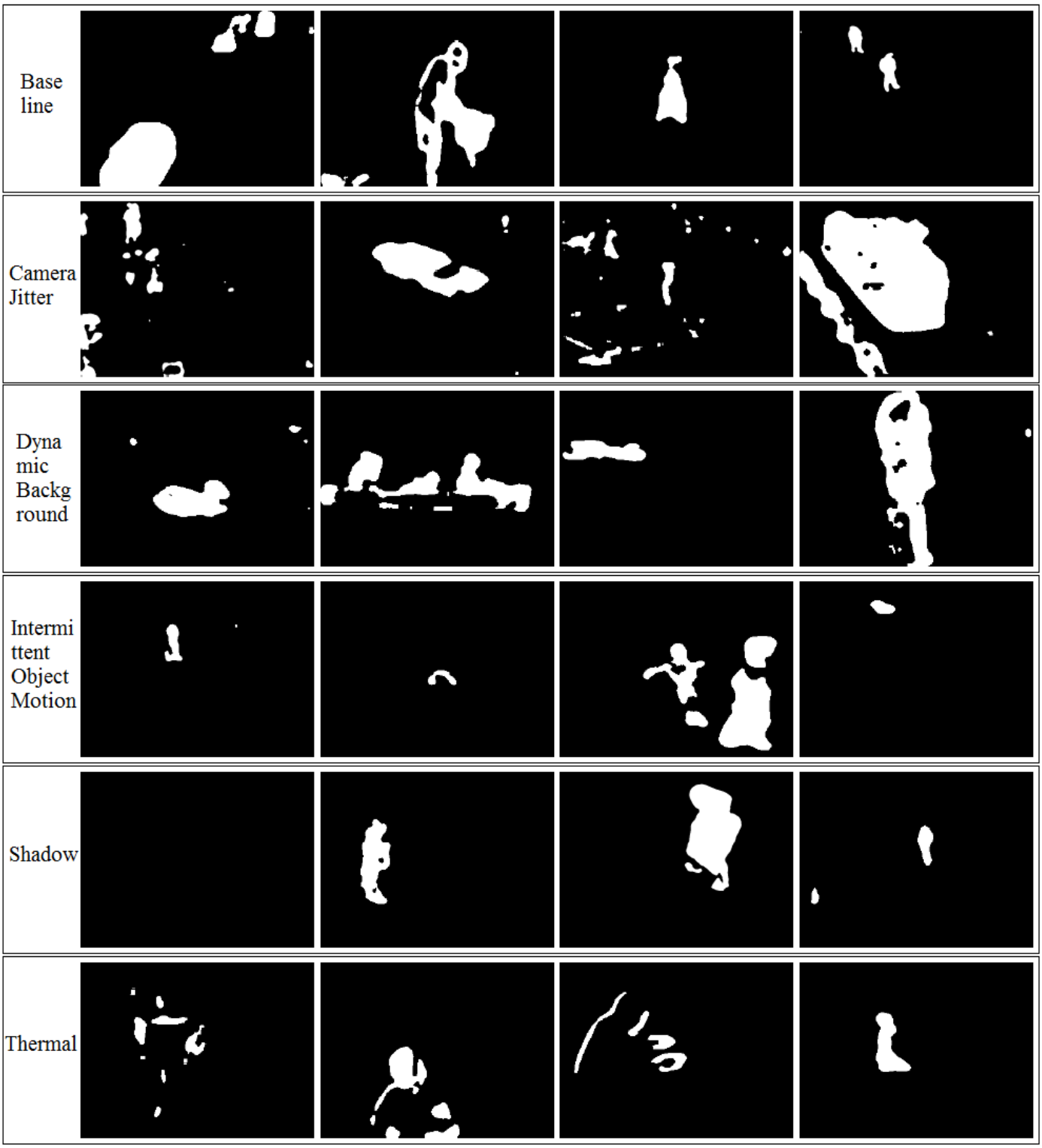

Figure 10: Sample Images' binary results for feature combination 'D'.

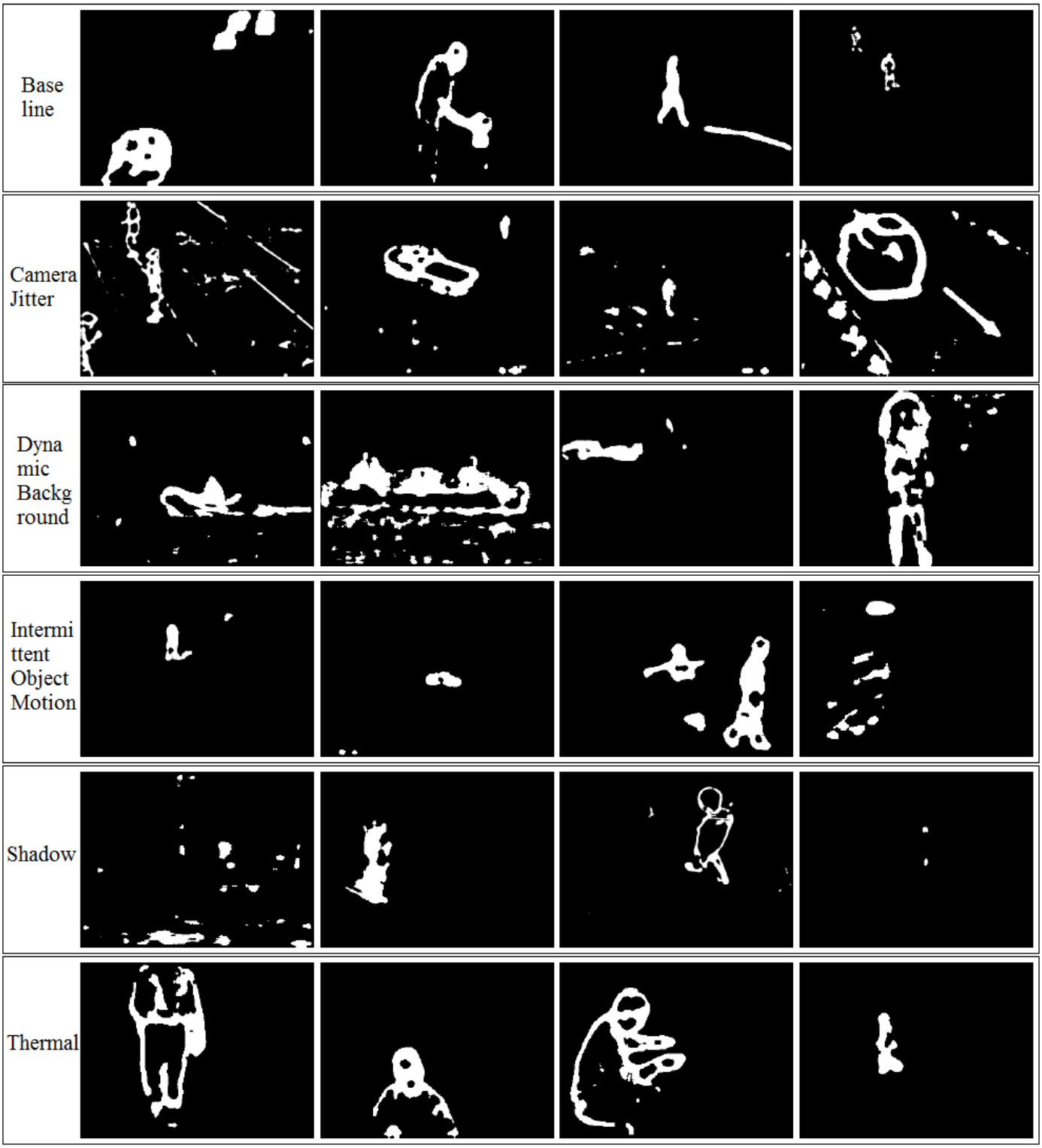

Figure 11: Sample Images' binary results for feature combination 'S'.

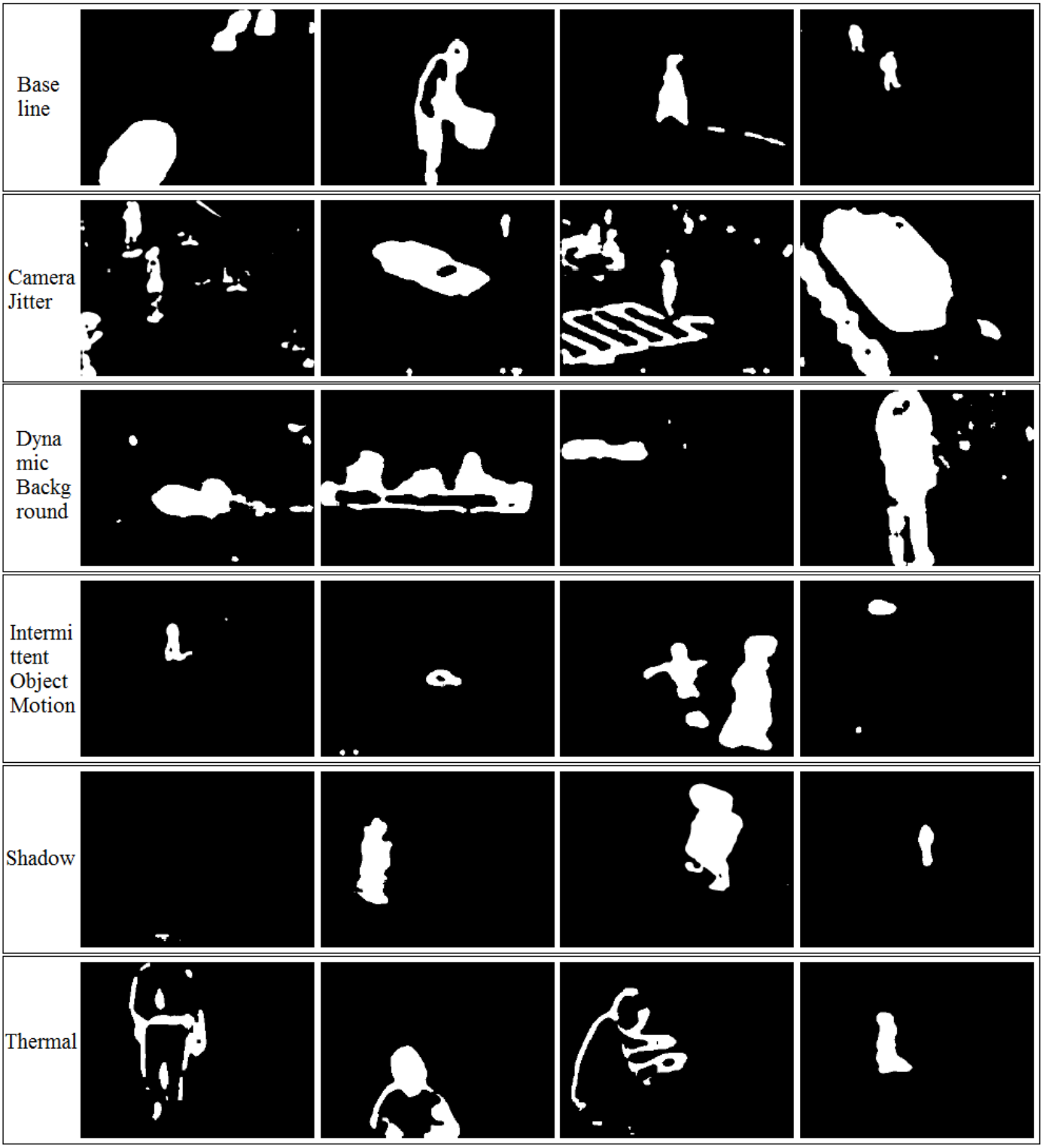

Figure 12: Sample Images' binary results for feature combination 'SD'.

Method Overall	Average ranking	Average Re	Average Sp	Average FPR	Average FNR	Average PWC	Average F-Measure	Average Precision
CS	1.29	0.7894	0.9830	0.0170	0.2106	2.7052	0.7544	0.8015
C	2.86	0.8203	0.9745	0.0255	0.1797	3.0948	0.7361	0.7417
CD	3.14	0.7517	0.9825	0.0175	0.2483	3.1626	0.6816	0.7283
CSD	3.29	0.7862	0.9787	0.0213	0.2138	3.3408	0.6946	0.7163
D	5.29	0.6395	0.9784	0.0216	0.3605	4.2213	0.5562	0.6408
S	6.00	0.5496	0.9782	0.0218	0.4504	4.6617	0.5184	0.6565
SD	6.14	0.7516	0.9633	0.0367	0.2484	5.2341	0.5755	0.5793

Table 2: Overall 7 combinations comparison

Method Baseline	Average ranking	Average Re	Average Sp	Average FPR	Average FNR	Average PWC	Average F-Measure	Average Precision
C	2.00	0.9263	0.9972	0.0028	0.0737	0.5352	0.9142	0.9033
CS	2.14	0.8917	0.9975	0.0025	0.1083	0.7289	0.8982	0.9114
CD	2.86	0.9305	0.9918	0.0082	0.0695	1.1693	0.8237	0.7547
CSD	3.86	0.9303	0.9911	0.0089	0.0697	1.2708	0.8144	0.7430
S	5.00	0.6329	0.9968	0.0032	0.3671	1.8510	0.6946	0.8713
D	6.00	0.8230	0.9880	0.0120	0.1770	2.1440	0.7102	0.6681
SD	6.14	0.8797	0.9855	0.0145	0.1203	2.1625	0.7171	0.6437

Table 3: Baseline 7 combinations comparison

Method camera-Jitter	Average ranking	Average Re	Average Sp	Average FPR	Average FNR	Average PWC	Average F-Measure	Average Precision
CS	1.29	0.8429	0.9841	0.0159	0.1571	2.1080	0.7826	0.7433
CSD	2.29	0.8409	0.9730	0.0270	0.1591	3.1274	0.7395	0.6882
C	3.43	0.8462	0.9562	0.0438	0.1538	4.7178	0.6369	0.5599
CD	3.86	0.7821	0.9687	0.0313	0.2179	3.7214	0.6910	0.6685
S	5.43	0.5314	0.9693	0.0307	0.4686	5.0343	0.4752	0.4714
D	5.71	0.6682	0.9420	0.0580	0.3318	6.7117	0.5228	0.4951
SD	6.00	0.8341	0.9043	0.0957	0.1659	9.8055	0.4982	0.4206

Table 4: Camera Jitter 7 combinations comparison

Method Dynamic Background	Average ranking	Average Re	Average Sp	Average FPR	Average FNR	Average PWC	Average F-Measure	Average Precision
C	2.00	0.8661	0.9930	0.0071	0.1339	0.7927	0.7018	0.6662
CD	2.14	0.8825	0.9901	0.0099	0.1175	1.0552	0.6812	0.6612
CS	2.71	0.8665	0.9860	0.0140	0.1335	1.4714	0.6736	0.6737
CSD	3.43	0.9041	0.9806	0.0194	0.0959	1.9764	0.6313	0.6040
D	5.00	0.7513	0.9824	0.0176	0.2487	2.1503	0.5246	0.5311
SD	5.71	0.8596	0.9484	0.0516	0.1404	5.3579	0.4107	0.3193
S	7.00	0.6766	0.9325	0.0675	0.3234	7.2199	0.2873	0.2160

Table 5: Dynamic Background 7 combinations comparison

Method Intermittent Motion	Average ranking	Average Re	Average Sp	Average FPR	Average FNR	Average PWC	Average F-Measure	Average Precision
SD	3.29	0.6020	0.9738	0.0262	0.3980	5.5949	0.5972	0.6216
S	3.57	0.5837	0.9755	0.0245	0.4163	5.6038	0.5654	0.6339
D	3.57	0.3909	0.9876	0.0124	0.6091	5.5519	0.4636	0.6810
CS	3.57	0.6651	0.9399	0.0601	0.3349	7.7990	0.6066	0.6741
C	4.43	0.6922	0.9161	0.0839	0.3078	9.3183	0.5841	0.6405
CD	4.71	0.4965	0.9695	0.0305	0.5035	6.4589	0.5059	0.6725
CSD	4.86	0.5804	0.9533	0.0467	0.4196	7.3340	0.5603	0.6533

Table 6: Intermittent Motion 7 combinations comparison

Method Shadow	Average ranking	Average Re	Average Sp	Average FPR	Average FNR	Average PWC	Average F-Measure	Average Precision
CS	2.71	0.8741	0.9921	0.0079	0.1259	1.3477	0.8624	0.8611
CD	2.86	0.9426	0.9783	0.0217	0.0574	2.3301	0.8076	0.7220
C	3.29	0.9019	0.9876	0.0124	0.0981	1.5845	0.8153	0.7541
CSD	3.86	0.9391	0.9781	0.0219	0.0609	2.4037	0.8022	0.7160
S	4.43	0.5020	0.9967	0.0033	0.4980	2.8660	0.5926	0.8469
D	5.29	0.9162	0.9741	0.0259	0.0838	2.8226	0.7586	0.6648
SD	5.57	0.9381	0.9720	0.0280	0.0619	2.9684	0.7590	0.6496

Table 7: Shadow 7 combinations comparison

Method Thermal	Average ranking	Average Re	Average Sp	Average FPR	Average FNR	Average PWC	Average F-Measure	Average Precision
C	1.71	0.6894	0.9971	0.0029	0.3106	1.6202	0.7645	0.9260
CS	1.86	0.5962	0.9982	0.0018	0.4038	2.7766	0.7031	0.9452
S	3.86	0.3710	0.9982	0.0018	0.6290	5.3955	0.4956	0.8992
CSD	4.00	0.5223	0.9962	0.0038	0.4777	3.9322	0.6198	0.8934
CD	4.43	0.4762	0.9963	0.0037	0.5238	4.2405	0.5799	0.8907
SD	6.00	0.3960	0.9956	0.0044	0.6040	5.5156	0.4707	0.8208
D	6.14	0.2874	0.9964	0.0036	0.7126	5.9473	0.3571	0.8047

Table 8: Thermal 7 combinations comparison

3.2.2 Comparison with state of the art

Among the above seven combinations, 'CS', 'C' and 'CD' have ranked top three for overall videos, see Table 2. So, 'CS', 'C' and 'CD' are chosen to compete with all the methods published on the 2012 CDnet results of the website `www.changedetection.net`, the final rankings are shown in Figure 13. Overall, it is seen that the best performing of the proposed feature combinations, 'CS', is the tenth best approach based on average ranking, when compared to the state-of-the-art, see Figure 13. Among methods better than 'CS', PAWCS, SuBSENSE, PBAS and PBAS-PID are sample-based background models, DPGMM and SGMM-SOD are PDF background models, Spectral-360 is a physical model.

3.3 Discussion

3.3.1 Comparison among different feature combinations

The result that 'CS' is the best in Table 2 shows that by adding spatial orientation features (SO) to chromatic features (CM), the overall performance is improved. 'C' is the best for videos in the baseline category, see Table 3. This result is due to the fact that localization is very important for the current dataset evaluation and spatial or spatiotemporal features can blur detection results by involving surrounding information in their recovery, e.g. the obvious difference of the third image in baseline category between Figure 7 and Figure 8. 'CS' and 'CSD' are better than 'C' in Table 4. This result indicates that spatial and spatiotemporal features improve foreground detection for videos with heavy camera jitter, as colour alone does not supply adequate information. In Table 5, the result that 'CD' is better than 'CS' shows that spatiotemporal features can help more than spatial orientation features in dynamic background scenarios, as they explicitly capture the temporal as well as spatial variations. 'SD', 'S' and 'D' are the top three combinations for videos in intermittent object motion category, see Table 6. This result arises due to foreground and background objects having low color contrast in the videos with intermittent object motion, as can be

Method	Average ranking	Average Re	Average Sp	Average FPR	Average FNR	Average PWC	Average F-Measure	Average Precision
PAWCS	2.00	0.8547	0.9949	0.0051	0.1453	1.1402	0.8579	0.8746
CDet	3.86	0.9034	0.9917	0.0083	0.0966	1.1574	0.8608	0.8397
SuBSENSE	3.86	0.8281	0.9938	0.0062	0.1719	1.5447	0.8260	0.8576
Spectral-360	9.43	0.7770	0.9920	0.0080	0.2230	1.8516	0.7770	0.8461
SGMM-SOD	10.00	0.7697	0.9938	0.0062	0.2303	1.4960	0.7661	0.8339
PBAS-PID	10.14	0.7967	0.9902	0.0098	0.2033	1.6904	0.7720	0.8162
DPGMM	10.86	0.8275	0.9855	0.0145	0.1725	2.1159	0.7763	0.7928
PBAS	12.14	0.7840	0.9898	0.0102	0.2160	1.7693	0.7532	0.8160
PSP-MRF	15.86	0.8037	0.9830	0.0170	0.1963	2.3937	0.7372	0.7512
CS	16.14	0.7894	0.9830	0.0170	0.2106	2.7052	0.7544	0.8015
CwisarD	16.43	0.8178	0.9781	0.0219	0.1822	2.6607	0.7780	0.7739
GPRMF	16.86	0.8372	0.9734	0.0266	0.1628	3.1583	0.7944	0.8144
SC-SOBS	16.86	0.8017	0.9831	0.0169	0.1983	2.4081	0.7283	0.7315
SGMM	17.29	0.7073	0.9910	0.0090	0.2927	2.5311	0.7008	0.7812
CDPS	17.29	0.7769	0.9848	0.0152	0.2231	2.2747	0.7281	0.7610
RMoG (Region-based Mixture of Gaussians)	17.57	0.6042	0.9950	0.0050	0.3958	2.7798	0.6607	0.8247
Pixel Based Adaptive Foreground Extractor (PBAFE)	17.71	0.8109	0.9758	0.0242	0.1891	2.8904	0.7711	0.7913
Chebyshev prob. with Static Object detection	17.71	0.7133	0.9888	0.0112	0.2867	2.3856	0.7001	0.7856
SOBS_CF	18.71	0.8211	0.9788	0.0212	0.1789	2.6466	0.7273	0.7139
Multi-Layer Background Subtraction	20.14	0.6936	0.9888	0.0112	0.3064	2.7658	0.6993	0.7960
SOBS	20.29	0.7882	0.9818	0.0182	0.2118	2.5642	0.7159	0.7179
C	20.43	0.8203	0.9745	0.0255	0.1797	3.0948	0.7361	0.7417
KNN	20.57	0.6707	0.9907	0.0093	0.3293	2.7954	0.6785	0.7882
GMM | KaewTraKulPong	21.43	0.5072	0.9947	0.0053	0.4928	3.1051	0.5904	0.8228
KDE - Integrated Spatio-temporal Features	21.57	0.6507	0.9932	0.0068	0.3493	2.8905	0.6418	0.7663
KDE - Spatio-temporal change detection	23.14	0.6576	0.9910	0.0090	0.3424	3.0022	0.6437	0.7341
GMM | Stauffer & Grimson	23.86	0.7108	0.9860	0.0140	0.2892	3.1037	0.6624	0.7012
CD	24.71	0.7517	0.9825	0.0175	0.2483	3.1626	0.6816	0.7283
GMM | Zivkovic	26.00	0.6964	0.9845	0.0155	0.3036	3.1504	0.6596	0.7079
Local-Self similarity	27.43	0.9354	0.8512	0.1488	0.0646	14.2954	0.5016	0.4139
TUBITAK UZAY 1	27.71	0.7794	0.9756	0.0244	0.2206	3.7014	0.6475	0.6237
KDE - ElGammal	28.29	0.7442	0.9757	0.0243	0.2558	3.4602	0.6719	0.6843
GMM | RECTGAUSS-Tex	29.00	0.5156	0.9862	0.0138	0.4844	3.6842	0.5221	0.7190
Bayesian Background	29.29	0.6018	0.9826	0.0174	0.3982	3.3879	0.6272	0.7435
pROST	30.00	0.6735	0.9790	0.0210	0.3265	3.2534	0.6350	0.6734
Mahalanobis distance	31.57	0.7607	0.9599	0.0401	0.2393	4.6631	0.6259	0.6040
Histogram	32.00	0.7698	0.9343	0.0657	0.2302	6.9682	0.5485	0.5251
Euclidean distance	32.86	0.7048	0.9692	0.0308	0.2952	4.3465	0.6111	0.6223

Figure 13: Overall CDnet Rank

observed in the sample images in Figure 4; thus, chromatic features only make this result worse than purely spatial or spatiotemporal features. The rank in Table 7 shows that spatial or spatiotemporal features improve the detection results over using only chromatic features for videos containing shadows. This result is due to the ability of the such features to model surface texture and thereby avoid falsely labelling regions that go into and out of shadow as change because the consistency of surface texture is maintained. This can be clearly observed in the first image of the shadow category in Figures 7 and 8. In Figure 7, the shadow cast onto the ground surface is wrongly detected as foreground, but feature combination 'CD' avoids detecting purely brightness change, as shown in Figure 8. Since objects in thermal videos are lacking in textures, 'C' is better than 'CS'. As noted above, 'SD' is worse than 'S' or 'D' in most categories based on average ranking, but the recall and FNR of 'SD' from all tables are better than 'S' or 'D'. This result is likely due to the fact that combining the spatial and dynamic texture features results in more foreground being detected in general.

3.3.2 Comparison with state of the art

As expected, the rankings of the three proposed algorithms ('CS', 'C' and 'CD') that have been selected for comparison to the state-of-the-art maintain their relative orderings in this large comparison. Four of the state-of-the-art algorithms (PAWCS, SubSENSE, PBAS and PBAS-PPID, with PAWCS the current top performer) that rank above the top performing of the proposed algorithms ('CS') share interesting similarities with 'CS'. First, they all make use of both chromatic and spatial texture features. Second, they all are sample-based approaches. Here, it also is interesting to note that adding spatial features to realize 'CS' further improves performance of 'C' in the rank. The distinguishing attribute of the higher performing sample-based approaches is that they also incorporate a feedback mechanism. These observations suggest that further refinement of 'CS' to include feedback could boost its performance to be competitive with the very best current change detection algorithms.

4 Conclusion

4.1 Summary

In this report, we presented a background modelling algorithm based on dynamic, spatial orientation and chromatic features. Amongst the various proposed feature combinations, the combination of chromatic plus spatial orientation features performs best overall. In comparison to the state-of-the-art, this combination ranks tenth best. More generally: The relationship between the three features and background modelling has been explicitly analyzed and presented. Moreover, a state-of-the-art background model (ViBe) has been introduced and different feature combinations have been embedded into the framework to form a foreground detection algorithm. Finally, the resulting algorithm has been implemented in software and empirically evaluated both qualitatively and quantitatively based on a standard publicly available change detection dataset.

4.2 Future work

In the light of the work that has been described in this report, several directions for future work can be considered, as follows.

- The top amongst the state-of-the-art algorithms (PAWCS) is a sample-based approach, as are three more of the approaches that rank above the proposed approach (SubSENSE, PBAS and PBAS-PID). Significantly, the proposed approach also is sample-based; however, those that rank above also incorporate a feedback mechanism that the proposed approach does not. Therefore, a promising direction for future research is to embed the best of the proposed feature combinations in a sample-based model with feedback loops.

- The proposed approach is based on purely local feature measurements. Preliminary experiments suggested that regional aggregations of local measurements can be effective

in isolating false negatives as well as eliminating false positives. Thus, an interesting direction for future research would be to use aggregations of the current features in sample-based background modelling.

- The current system was developed for ease of experimentation with various algorithmic variations and therefore was not optimized for execution rate. Since many envisioned applications involve real-time demands, follow-on research that targets efficient implementation would be of great interest.

Appendix

In complement to the standard CDnet evaluation results presented in the main body of this report, this appendix presents ROCs for the proposed features and their combinations. The plots are shown in Figures 14 - 20. It is seen that the ROC rankings are mostly consistent with those presented in Tables 2 - 8; however, a few points of discrepancy can be noted, as follows. First, combination 'SD' is better than 'D' in Figures 14 and 16, which is a different ranking than presented in Tables 2 and 4. Addition of the spatial orientation features causes an increase in Recall, with little change to the FPR, which improves the 'SD' ROC results. However, the PWC and Specificity of 'SD' become worse, which pulls its rank down. Second, feature 'S' is the worst in all ROC curves, but it only ranks the worst in Table 5. These results can be explained by the fact that 'S' alone can only detect regions with texture; therefore, its Recall cannot be increased usefully beyond this inherent limitation via lower thresholds. Finally, while not necessarily inconsistent with the results shown in the Tables 2 - 8, it is interesting to note that the curves for 'C' show a tendency to cross those of 'CD' and 'CSD', indicating threshold sensitivity in their relative performances.

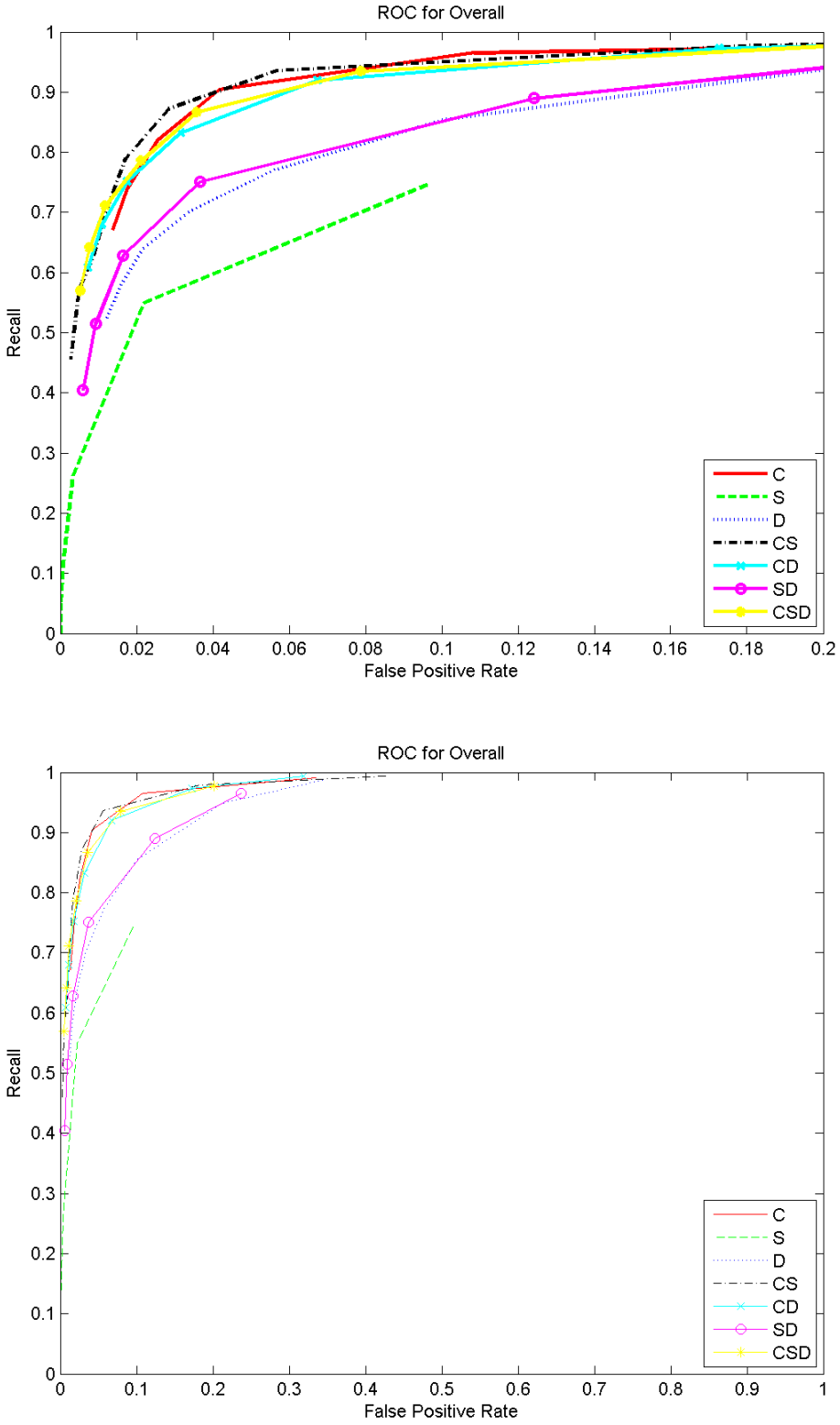

Figure 14: ROCs calculated overall categories. The bottom plot shows the full operating range; the top provides zoomed in view.

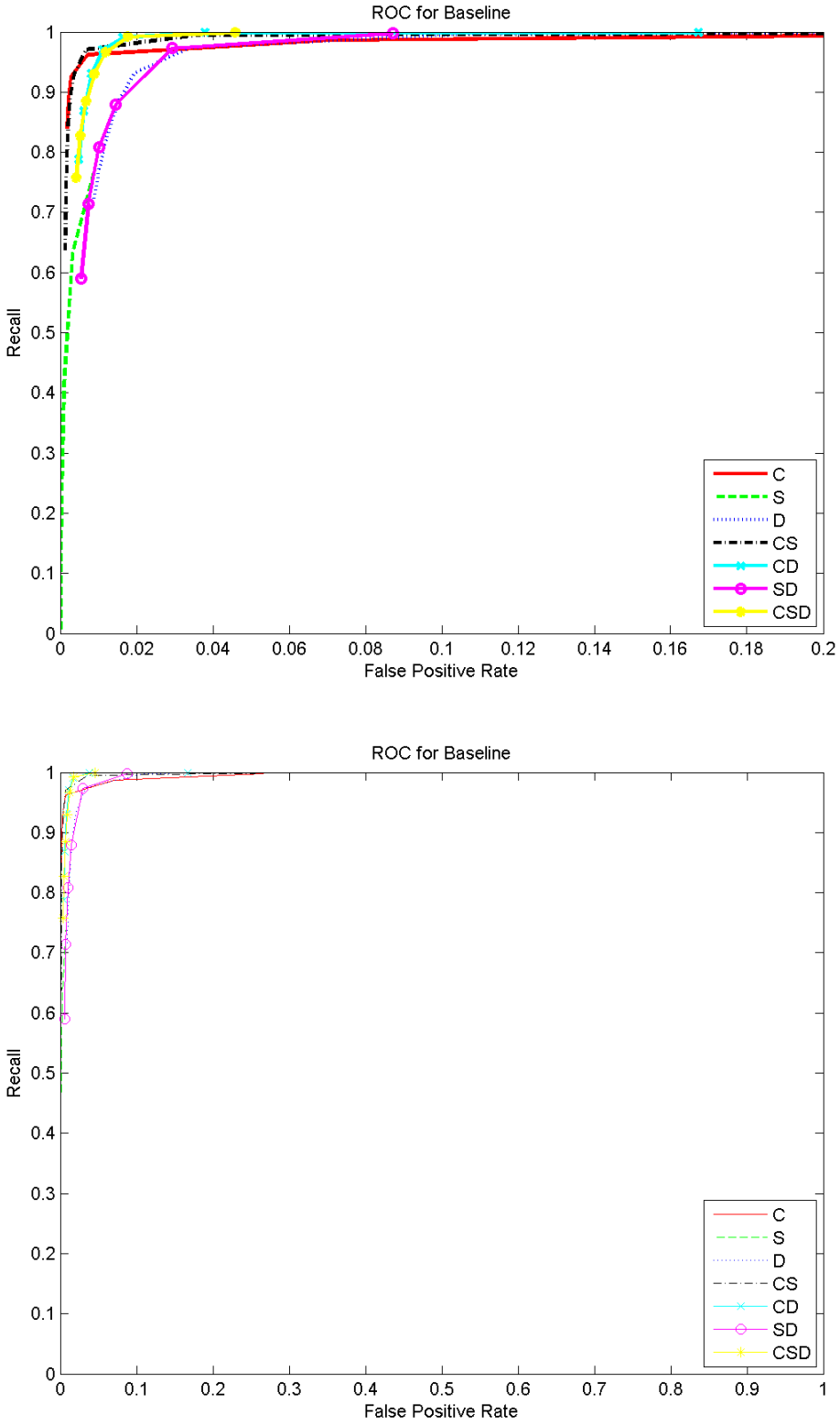

Figure 15: ROCs for the Baseline category. The bottom plot shows the full operating range; the top provides zoomed in view.

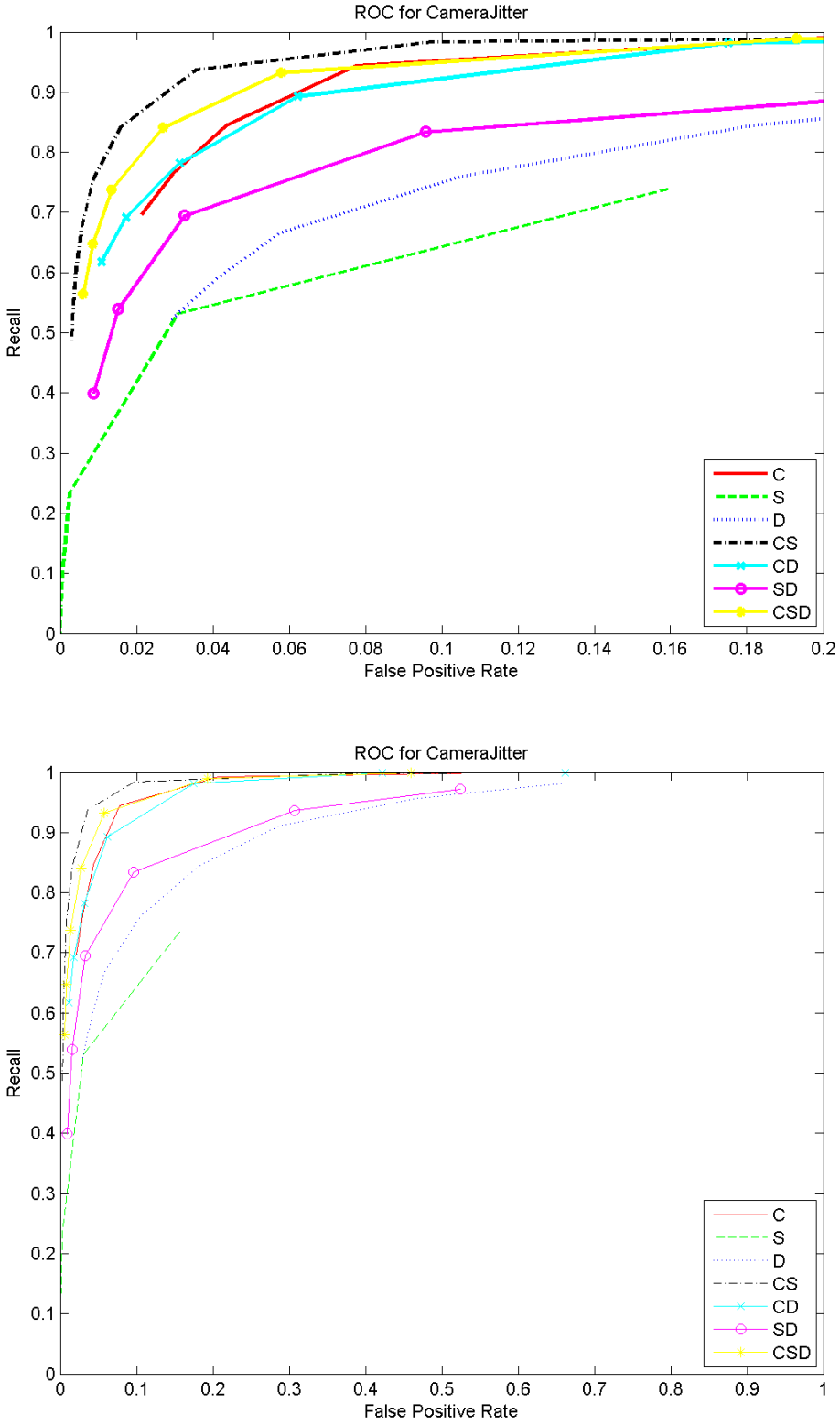

Figure 16: ROCs for the Camera Jitter category. The bottom plot shows the full operating range; the top provides zoomed in view.

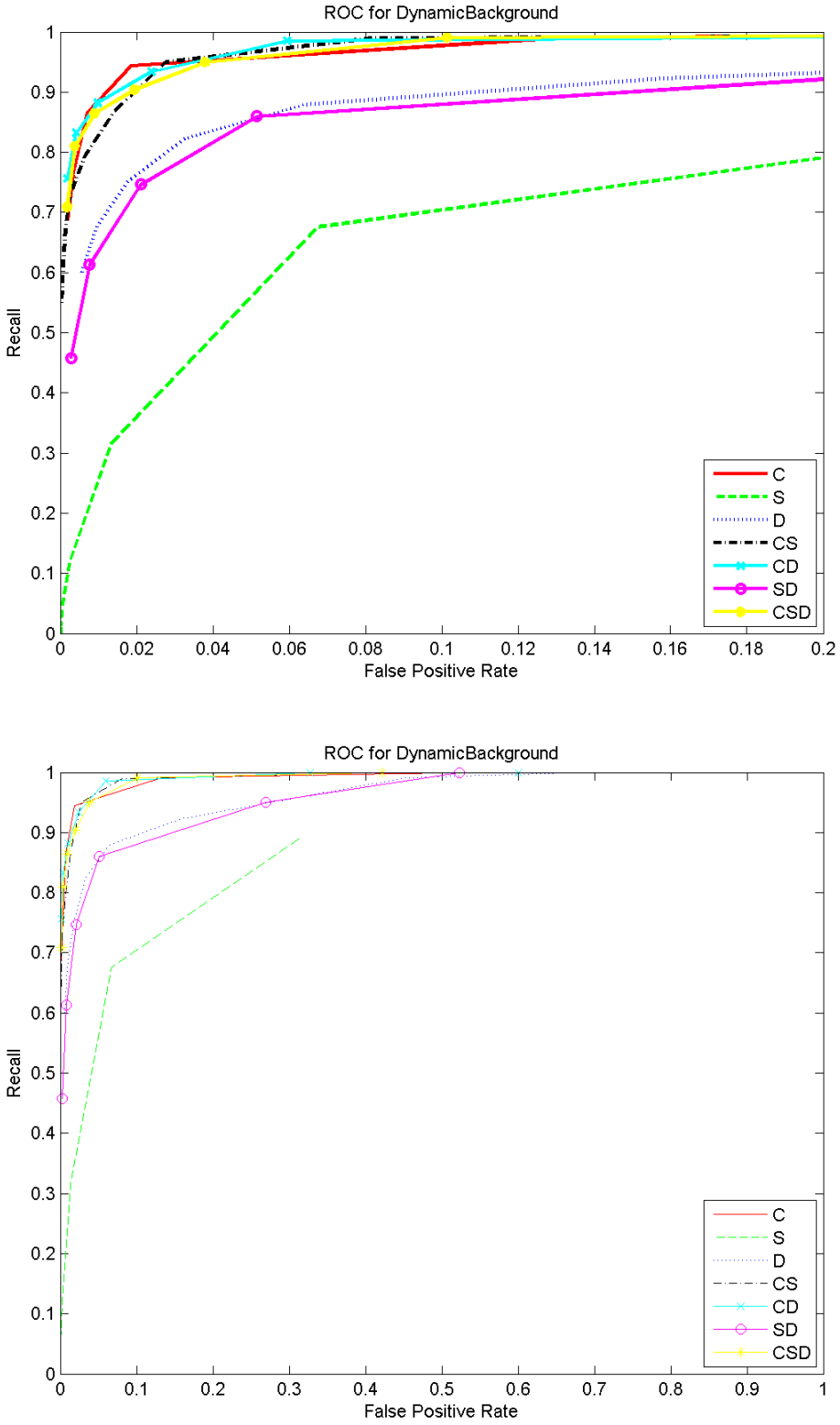

Figure 17: ROCs for the Dynamic Background category. The bottom plot shows the full operating range; the top provides zoomed in view.

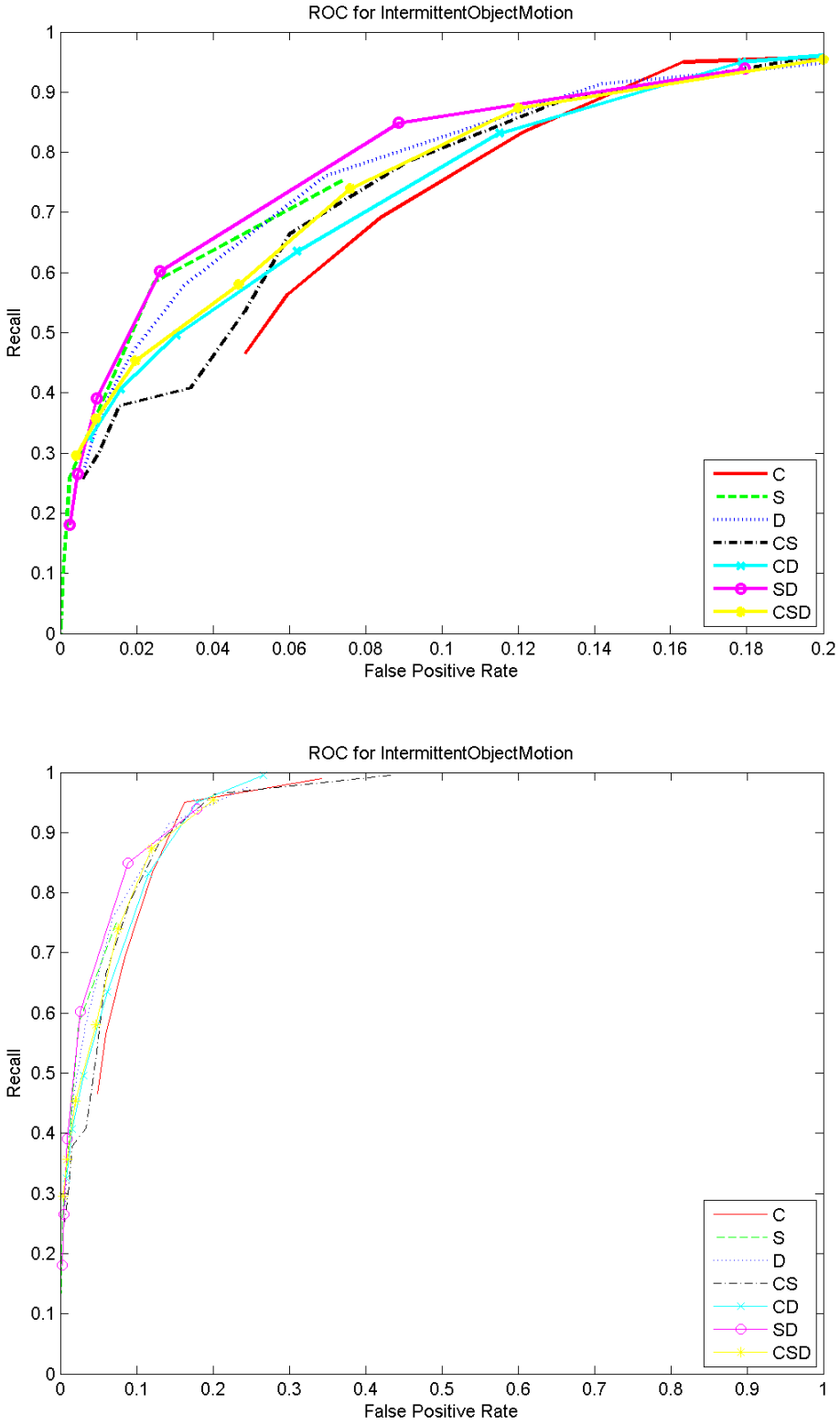

Figure 18: ROCs for the Intermittent Object Motion category. The bottom plot shows the full operating range; the top provides zoomed in view.

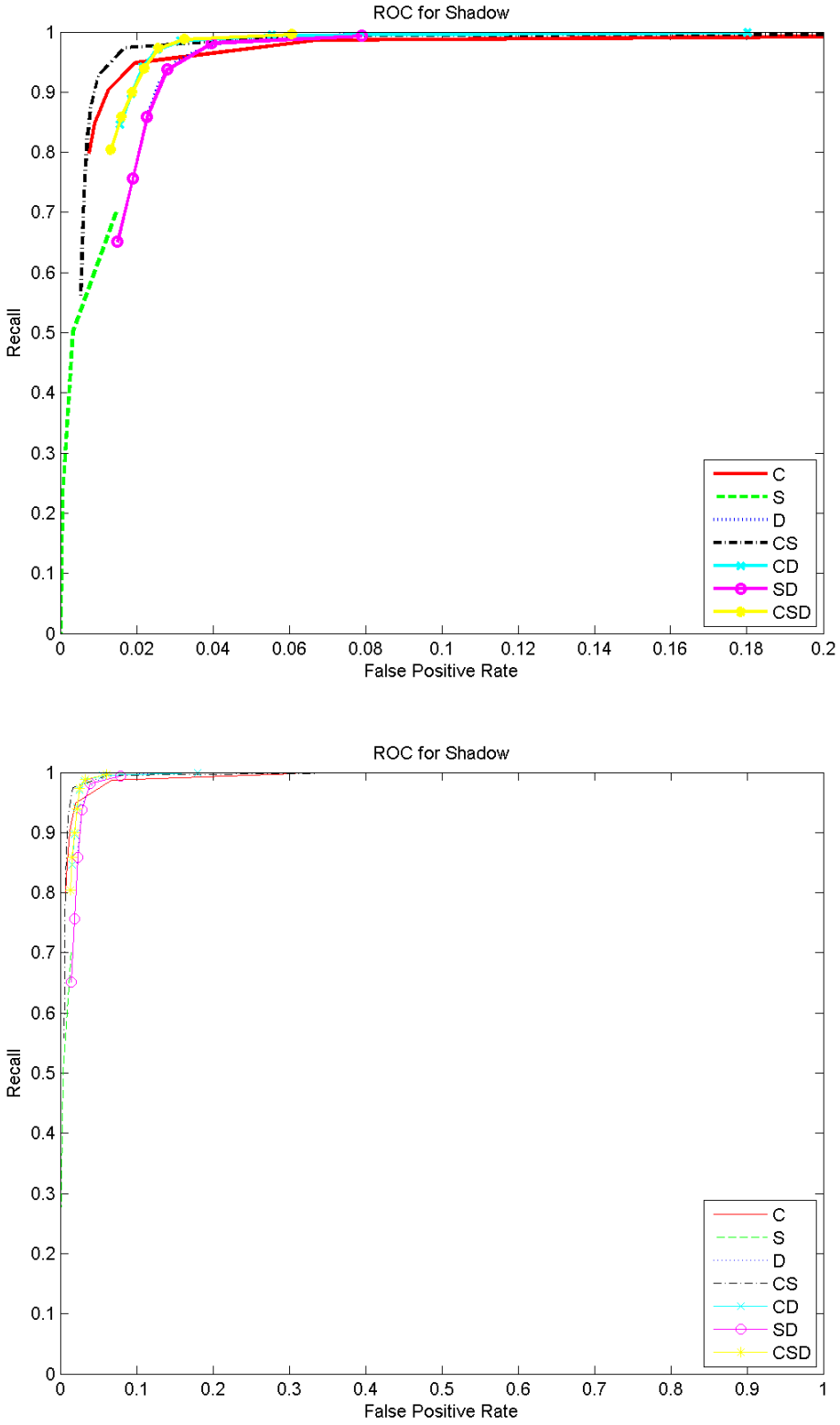

Figure 19: ROCs for the Shadow category. The bottom plot shows the full operating range; the top provides zoomed in view.

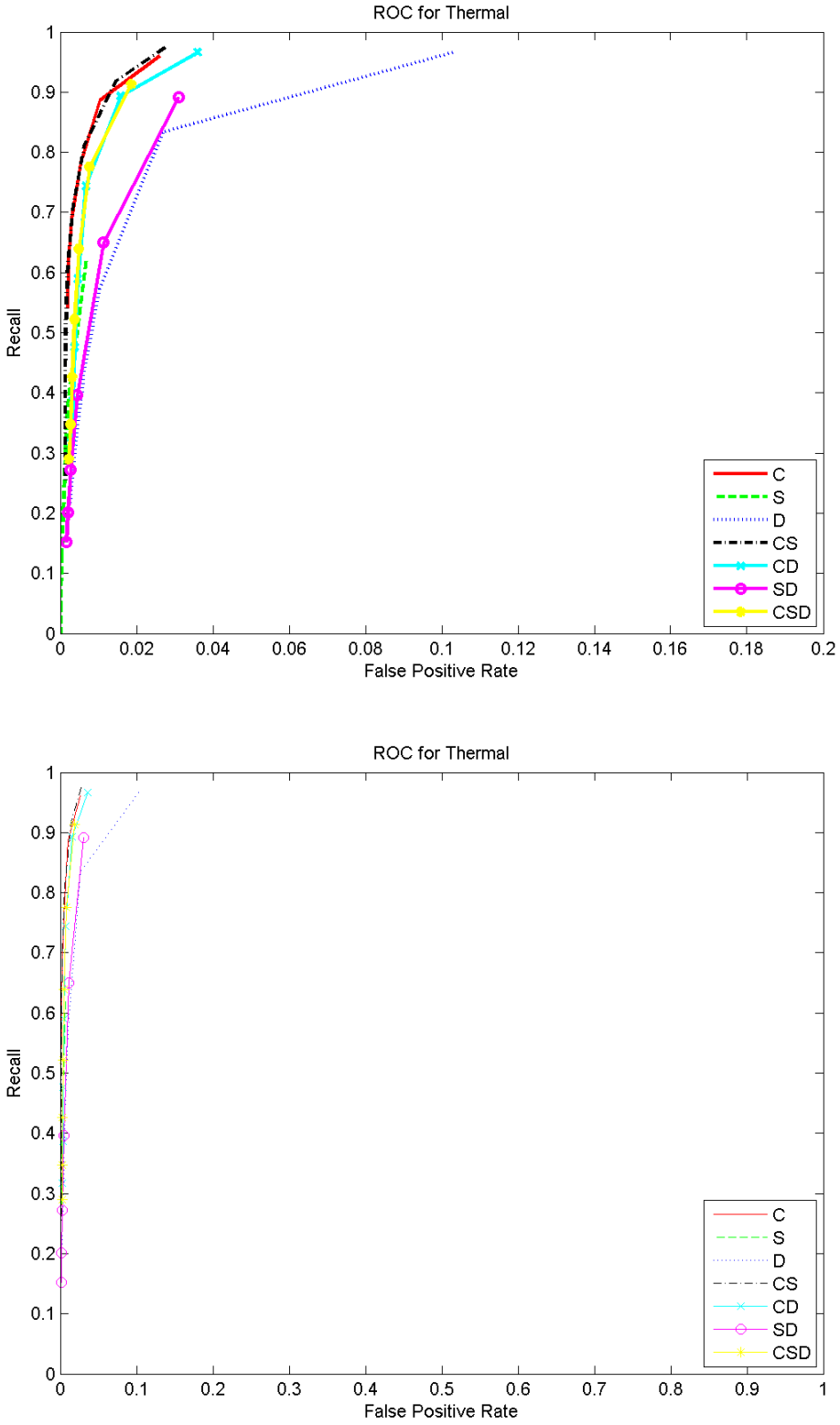

Figure 20: ROCs for the Thermal category. The bottom plot shows the full operating range; the top provides zoomed in view.

References

[1] M. S. Allili and N. Bouguila. A robust video foreground segmentation by using generalized gaussian mixture modeling. In *Proceedings of Canadian Conference on Computer and Robot Vision*, pages 503–509, 2007.

[2] O. Barnich and M. Van Droogenbroeck. Vibe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image Processing*, 20(6):1709–1724, 2011.

[3] Y. Benezeth, P. Jodoin, B. E. H. Laurent, and C. Rosenberger. Comparative study of background subtraction algorithms. *Journal of Electronic Imaging*, 19(3):033003, 2010.

[4] T. Bouwmans. Subspace learning for background modeling: A survey. *Recent Patents on Computer Science*, 2(3):223–234, 2009.

[5] T. Bouwmans. Recent advanced statistical background modeling for foreground detection - A systematic survey. *Recent Patents on Computer Science*, 4(3):147–176, 2011.

[6] T. Bouwmans. Traditional and recent approaches in background modeling for foreground detection: An overview. *Computer Science Review*, 11-12:31–66, 2014.

[7] T. Bouwmans, F. El Baf, and B. Vachon. Statistical background modeling for foreground detection: A survey. In C.H.Chen, editor, *Handbook of Pattern Recognition and Computer Vision*. World Scientific Publishing, 4th edition, 2010.

[8] T. Bouwmans, J. Gonzalez, C. Shan, M. Piccardi, and L. Davis. Special issue on background modeling for foreground detection in real-world dynamic scenes. *Machine Vision and Applications*, 25(5):1101–1103, 2014.

[9] S. Brutzer, B. H. ferlin, and G. Heidemann. Evaluation of background subtraction techniques for video surveillance. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1937–1944, 2011.

[10] E. J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3), 2011. Article 11.

[11] K. J. Cannons, J. M. Gryn, and R. P. Wildes. Visual tracking using a pixelwise spatiotemporal oriented energy representation. In *Proceedings of European Conference on Computer Vision*, pages 511–524, 2010.

[12] R. Chang, T. Gandhi, and M. M. Trivedi. Vision modules for a multi-sensory bridge monitoring approach. In *Proceedings of IEEE Intelligent Transportation Systems Conference*, pages 971–976, 2004.

[13] M. Cristani, M. Farenzena, D. Bloisi, and V. Murino. Background subtraction for automated multisensor surveillance: A comprehensive review. *EURASIP Journal on Advances in Signal Processing*, 2010:343057, 2010.

[14] D. Culibrk, O. Marques, D. Socek, H. Kalva, and B. Furht. A neural network approach to bayesian background modeling for video object segmentation. In *Proceedings of International Conference on Computer Vision Theory and Applications*, pages 25–28, 2006.

[15] D. Culibrk, O. Marques, D. Socek, H. Kalva, and B. Furht. Neural network approach to background modeling for video object segmentation. *IEEE Transactions on Neural Networks*, 18(6):1614–1627, 2007.

[16] M. De Gregorio and M. Giordano. A WiSARD-based approach to CDnet. In *Proceedings of BRICS Congress on Computational Intelligence and Brazilian Congress on Computational Intelligence*, pages 172–177, 2013.

[17] M. De Gregorio and M. Giordano. Change detection with weightless neural networks. In *Proceedings of IEEE Workshop on Change Detection*, pages 409–413, 2014.

[18] K. G. Derpanis and J. M. Gryn. Three-dimensional nth derivative of Gaussian separable steerable filters. In *Proceedings of International Conference on Image Processing*, pages III – 553–6, 2005.

[19] K. G. Derpanis and R. P. Wildes. Early spatiotemporal grouping with a distributed oriented energy representation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 232–239, 2009.

[20] K. G. Derpanis and R. P. Wildes. Spacetime texture representation and recognition based on a spatiotemporal orientation analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1193–1205, 2012.

[21] K. G. Derpanis, M. Lecce, K. Daniilidis, and R. P. Wildes. Dynamic scene understanding: The role of orientation features in space and time in scene classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1306 – 1313, 2012.

[22] K. G. Derpanis, M. Sizintsev, K. J. Cannons, and R. P. Wildes. Action spotting and recognition based on a spatiotemporal orientation analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):527–540, 2013.

[23] J. Ding, M. Li, K. Huang, and T. Tan. Modeling complex scenes for accurate moving objects segmentation. In *Proceedings of Asian Conference on Computer Vision*, pages 82–94, 2010.

[24] F. El Baf, T. Bouwmans, and B. Vachon. Type-2 fuzzy mixture of gaussians model: Application to background modeling. In *Proceedings of International Symposium on Visual Computing*, pages 772–781, 2008.

[25] A. Elgammal, D. Harwood, and L. Davis. Non-parametric model for background subtraction. In *Proceedings of European Conference on Computer Vision*, pages 772–781, 2000.

[26] S. Y. Elhabian, K. M. El-Sayed, and S. H. Ahmed. Moving object detection in spatial domain using background removal techniques-state-of-art. *Recent Patents on Computer Science*, 1(1):32–54, 2008.

[27] R. H. Evangelio and T. Sikora. Complementary background models for the detection of static and moving objects in crowded environments. In *Proceedings of IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 71–76, 2011.

[28] R. H. Evangelio, M. Patzold, and T. Sikora. Splitting gaussians in mixture models. In *Proceedings of IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 300–305, 2012.

[29] D. Fan, M. Cao, and C. Lv. An updating method of self-adaptive background for moving objects detection in video. In *Proceedings of International Conference on Audio, Language and Image Processing*, pages 1497–1501, 2008.

[30] W. Fan and N. Bouguila. Online variational learning of finite dirichlet mixture models. *Evolving Systems*, 3(3):153–165, 2012.

[31] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[32] S. Fu, G. Jiang, and M. Yu. An effective background subtraction method based on pixel change classification. In *Proceedings of International Conference on Electrical and Control Engineering*, pages 4634–4637, 2010.

[33] N. Goyette, P. Jodoin, F. Porikli, J. Konrad, and P. Ishwar. changedetection.net: A new change detection benchmark dataset. In *Proceedings of IEEE Workshop on Change Detection*, pages 1–8, 2012.

[34] N. Guan, D. Tao, Z. Luo, and J. Shawe-Taylor. MahNMF: manhattan non-negative matrix factorization. *arXiv*, 1207.3438, 2012.

[35] T. S. Haines and T. Xiang. Background subtraction with dirichlet processes. In *Proceedings of European Conference on Computer Visions*, pages 97–111, 2012.

[36] T. S. Haines and T. Xiang. Background subtraction with dirichlet process mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):670–683, 2014.

[37] B. Han and R. Jain. Real-time subspace-based background modeling using multi-channel data. In *Proceedings of International Symposium on Visual Computing*, pages 162–172, 2007.

[38] J. He, L. Balzano, and J. C. Lui. Online robust subspace tracking from partial information. *arXiv*, 1109.3827, 2011.

[39] J. He, D. Zhang, L. Balzano, and T. Tao. Iterative Grassmannian optimization for robust image alignment. *Image and Vision Computing*, 32(10):800–813, 2014.

[40] F. J. Hernandez-Lopez and M. Rivera. Change detection by probabilistic segmentation from monocular view. *Machine Vision and Applications*, 25(5):1175–1195, 2014.

[41] M. Hofmann, P. Tiefenbacher, and G. Rigoll. Background segmentation with feedback: The pixel-based adaptive segmenter. In *Proceedings of IEEE Workshop on Change Detection*, pages 38–43, 2012.

[42] J. Jodoin, G. Bilodeau, and N. Saunier. Background subtraction based on local shape. *arXiv*, 1204.6326v2, 2012.

[43] P. KaewTraKulPong and R. Bowden. An improved adaptive background mixture model for realtime tracking with shadow detection. In *Proceedings of European Workshop on Advanced Video Based Surveillance Systems*, pages 135–144, 2001.

[44] S. Kawabata, S. Hiura, and K. Sato. Real-time detection of anomalous objects in dynamic scene. In *Proceedings of International Conference on Pattern Recognition*, pages 1171 – 1174, 2006.

[45] H. Kim, R. Sakamoto, I. Kitahara, T. Toriyama, and K. Kogure. Robust foreground extraction technique using Gaussian family model and multiple thresholds. In *Proceedings of Asian Conference on Computer Vision*, pages 758–768, 2007.

[46] K. Kim, T. H. Chalidabhongse, D. Hanuood, and L. Davis. Background modeling and subtraction by codebook construction. In *Proceedings of International Conference on Image Processing*, pages 3061–3064, 2004.

[47] K. Kim, T. H. Chalidabhongse, D. Hanuood, and L. Davis. Real-time foreground-background segmentation using codebook model. *Real-Time Imaging*, 11(3):172185, 2005.

[48] D. Kit, B. Sullivan, and D. Ballard. Novelty detection using Growing Neural Gas for visuo-spatial memory. In *Proceedings of IEEE International Conference on Intelligent Robots and Systems*, pages 1194–1200, 2011.

[49] M. G. Krishna, V. M. Aradhya, M. Ravishankar, and D. R. Babu. LoPP: Locality preserving projections for moving object detection. In *Proceedings of International Conference on Computer, Communication, Control and Information Technology (C3IT)*, pages 624–628, 2012.

[50] M. G. Krishna, M. Ravishankar, and D. R. Babu. Ten-LoPP: Tensor locality preserving projections approach for moving object detection and tracking. In *Proceedings of*

International Conference on Computing and Information Technology (IC2IT), pages 291–300, 2013.

[51] A. Kumar and V. Sindhwani. Near-separable non-negative matrix factorization with l_1 and Bregman loss functions. *arXiv*, 1312.7167, 2013.

[52] D. Lee. Effective Gaussian mixture learning for video background subtraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):827–832, 2005.

[53] X. Li, W. Hu, Z. Zhang, and X. Zhang. Robust foreground segmentation based on two effective background models. In *Proceedings of ACM international conference on Multimedia information retrieval*, pages 223–228, 2008.

[54] Y. Li. On incremental and robust subspace learning. *Pattern Recognition*, 37(7): 1509–1518, 2004.

[55] D. Liang and S. Kaneko. Improvements and experiments of a compact statistical background model. *arXiv*, 1405.6275v1, 2014.

[56] D. Liang, S. Kaneko, and M. Hashimoto. Co-occurrence-based adaptive background model for robust object detection. In *Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 401–406, 2013.

[57] H. Lin, T. Liu, and J. Chuang. A probabilistic SVM approach for background scene initialization. In *Proceedings of International Conference on Image Processing*, pages 893–896, 2002.

[58] Z. Liu, W. Chen, K. Huang, and T. Tan. A probabilistic framework based on KDE-GMM hybrid model for moving object segmentation in dynamic scenes. In *Proceedings of International Workshop on Visual Surveillance*, 2008.

[59] X. Lu. A multiscale spatio-temporal background model for motion detection. In *Proceedings of International Conference on Image Processing*, pages 3268–3271, 2014.

[60] R. M. Luque, E. Dominguez, E. J. Palomo, and J. Munoz. A neural network approach for video object segmentation in traffic surveillance. In *Proceedings of International Conference on Image Analysis and Recognition*, pages 151–158, 2008.

[61] R. M. Luque, D. Lopez-Rodriguez, E. Dominguez, and E. J. Palomo. A dipolar competitive neural network for video segmentation. In *Proceedings of Advances in Artificial Intelligence IBERAMIA*, pages 103–112, 2008.

[62] R. M. Luque, D. Lopez-Rodriguez, E. Merida-Casermeiro, and E. J. Palomo. Video object segmentation with multivalued neural networks. In *Proceedings of International Conference on Hybrid Intelligent Systems*, pages 613–618, 2008.

[63] R. M. Luque, E. Dominguez, E. J. Palomo, and J. Munoz. An ART-type network approach for video object detection. In *Proceedings of European Symposium on Artificial Neural Networks-Computational Intelligence and Machine Learning*, pages 423–428, 2010.

[64] L. Maddalena and A. Petrosino. A self-organizing approach to background subtraction for visual surveillance applications. *IEEE Transactions on Image Processing*, 17(7): 1168–1177, 2008.

[65] L. Maddalena and A. Petrosino. A fuzzy spatial coherence-based approach to background/foreground separation for moving object detection. *Neural Computing and Applications*, 19(2):179–186, 2010.

[66] L. Maddalena and A. Petrosino. The SOBS algorithm: what are the limits? In *Proceedings of IEEE Workshop on Change Detection*, pages 21–26, 2012.

[67] L. Maddalena and A. Petrosino. The 3dSOBS+ algorithm for moving object detection. *Computer Vision and Image Understanding*, 122:65–73, 2014.

[68] N. McFarlane and C. Schofield. Segmentation and tracking of piglets in images. *Machine Vision and Applications*, 8(3):187–193, 1995.

[69] A. M. McIvor. Background subtraction techniques. In *Proceedings of Image and Vision Computing New Zealand*, 2000.

[70] A. Morde, X. Ma, and S. Guler. Learning a background model for change detection. In *Proceedings of IEEE Workshop on Change Detection*, pages 15–20, 2012.

[71] D. Mukherjee and Q. M. J. Wu. Real-time video segmentation using Student's t mixture model. In *Proceedings of International Conference on Ambient Systems, Networks and Technologies*, pages 153–160, 2012.

[72] Y. Nonaka, A. Shimada, H. Nagahara, and R. Taniguchi. Evaluation report of integrated background modeling based on spatio-temporal features. In *Proceedings of IEEE Workshop on Change Detection*, pages 9–14, 2012.

[73] N. M. Oliver, B. Rosario, and A. P. Pentland. A Bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):831–843, 2000.

[74] E. J. Palomo, E. Dominguez, R. M. Luque, and J. Munoz. Image hierarchical segmentation based on a GHSOM. In *Proceedings of International Conference on Neural Information Processing*, pages 743–750, 2009.

[75] D. Park and H. Byun. Object-wise multi-layer background ordering for public area surveillance. In *Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 484–489, 2009.

[76] T. V. Pham and A. W. M. Smeulders. Efficient projection pursuit density estimation for background subtraction. In *Proceedings of IEEE International Workshop on Visual Surveillance*, 2006.

[77] B. T. Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975.

[78] M. Piccardi. Background subtraction techniques: a review. In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, pages 3099–3104, 2004.

[79] F. Porikli and O. Tuzel. Bayesian background modeling for foreground detection. In *Proceedings of ACM International Workshop on Video Surveillance and Sensor Networks*, pages 55–58, 2005.

[80] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image change detection algorithms: A systematic survey. *IEEE Transactions on Image Processing*, 14(3):294–307, 2005.

[81] D. Riahi, P. St-Onge, and G. Bilodeau. RECTGAUSS-Tex: Block-based background subtraction. Technical report, Department of Computer Engineering and Software Engineering, Ecole Polytechnique de Montreal, 2012.

[82] J. Rymel, J. Rennno, D. Greenhill, J. Orwell, and G. Jones. Adaptive eigen-backgrounds for object detection. In *Proceedings of International Conference on Image Processing*, pages 1847–1850, 2004.

[83] A. Schick, M. Bauml, and R. Stiefelhagen. Improving foreground segmentations with probabilistic superpixel Markov Random Fields. In *Proceedings of IEEE Workshop on Change Detection*, pages 27–31, 2012.

[84] M. Sedky, M. Moniri, and C. C. Chibelushi. Spectral-360: A physics-based technique for change detection. In *Proceedings of IEEE Workshop on Change Detection*, pages 405–408, 2014.

[85] F. Seidel, C. Hage, and M. Kleinsteuber. pROST: A smoothed l_p-norm robust online

subspace tracking method for realtime background subtraction in video. *Machine Vision and Applications*, 25:1227–1240, 2014.

[86] Y. Sheikh and M. Shah. Bayesian modeling of dynamic scenes for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1778–1792, 2005.

[87] D. Skocaj and A. Leonardis. Incremental and robust learning of subspace representations. *Image and Vision Computing*, 26(1):2738, 2008.

[88] P. St-Charles, G. Bilodeau, and R. Bergevin. SuBSENSE: A universal change detection method with local adaptive sensitivity. *IEEE Transactions on Image Processing*, 24 (1):359–373, 2015.

[89] P. St-Charles, G. Bilodeau, and R. Bergevin. A self-adjusting approach to change detection based on background word consensus. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, 2015.

[90] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 246–252, 1999.

[91] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *Proceedings of International Conference on Computer Vision*, pages 255–261, 1999.

[92] A. Vacavant, T. Chateau, A. Wilhelm, and L. Lequievre. A benchmark dataset for outdoor foreground/background extraction. In *Proceedings of Asian Conference on Computer Vision*, pages 291–300, 2012.

[93] M. Van Droogenbroeck and O. Paquot. Background subtraction: Experiments and

improvements for ViBe. In *Proceedings of IEEE Workshop on Change Detection*, pages 32–37, 2012.

[94] S. Varadarajan, P. Miller, and H. Zhou. Spatial mixture of gaussians for dynamic background modelling. In *Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 63–68, 2013.

[95] B. Wang and P. Dudek. A fast self-tuning background subtraction algorithm. In *Proceedings of IEEE Workshop on Change Detection*, pages 401–404, 2014.

[96] J. Wang, G. Bebis, and R. Miller. Robust video-based surveillance by integrating target detection with tracking. In *Proceedings of IEEE International Workshop on Object Tracking and Classification in and beyond the Visible Spectrum*, 2006.

[97] L. Wang, L. Wang, M. Wen, Q. Zhuo, and W. Wang. Background subtraction using incremental subspace learning. In *Proceedings of International Conference on Image Processing*, pages V–45 – V–48, 2007.

[98] R. Wang, F. Bunyak, G. Seetharaman, and K. Palaniappan. Static and moving object detection using flux tensor with split Gaussian models. In *Proceedings of IEEE Workshop on Change Detection*, pages 420–424, 2014.

[99] A. B. Watson and A. J. Ahumada. Model of human visual-motion sensing. *Journal of the Optical Society of America A*, 2(2):322–342, 1985.

[100] R. P. Wildes and J. R. Bergen. Qualitative spatiotemporal analysis using an oriented energy representation. In *Proceedings of European Conference on Computer Vision*, pages 768–784, 2000.

[101] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[102] M. Xiao, C. Han, and X. Kang. A background reconstruction for dynamic scenes. In *Proceedings of International Conference on Information Fusion*, pages 1–7, 2006.

[103] Z. Xu, P. Shi, and I. Y. Gu. An eigenbackground subtraction method using recursive error compensation. In *Proceedings of Pacific Rim Conference on Multimedia*, pages 779–787, 2006.

[104] M. Yamazaki, G. Xu, and Y. Chen. Detection of moving objects by independent component analysis. In *Proceedings of Asian Conference on Computer Vision*, pages 467–478, 2006.

[105] J. Yao and J. Odobez. Multi-layer background subtraction based on color and texture. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[106] B. Yin, J. Zhang, and Z. Wang. Background segmentation of dynamic scenes based on dual model. *IET Computer Vision*, 8(6):545–555, 2014.

[107] S. Yoshinaga, A. Shimada, H. Nagahara, and R. ichiro Taniguchi. Background model based on intensity change similarity among pixels. In *Proceedings of Korea-Japan Joint Workshop on Frontiers of Computer Vision*, pages 276–280, 2013.

[108] D. P. Young and J. M. Ferryman. Pets metrics: On-line performance evaluation service. In *Proceedings of International Conference on Computer Communications and Networks*, pages 317–324, 2005.

[109] A. Zaharescu and R. P. Wildes. Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing. In *Proceedings of European Conference on Computer Vision*, pages 563–576, 2010.

[110] J. Zheng, Y. Wang, N. L. Nihan, and M. E. Hallenbeck. Extracting roadway back-

ground image: a mode-based approach. *Transportation Research Record: Journal of the Transportation Research Board*, 1944:82–88, 2006.

[111] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of International Conference on Pattern Recognition*, volume 2, pages 28–31, 2004.

[112] Z. Zivkovic and F. Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, 27(7):773–780, 2006.