A One-Threshold Algorithm for Detecting Abandoned Packages Under Severe Occlusions Using a Single Camera

Ser-Nam Lim Larry S. Davis
CS Dept., University of Maryland, College Park, USA

Abstract

We describe a single-camera system capable of detecting abandoned packages under severe occlusions, which leads to complications on several levels. The first arises when frames containing only background pixels are unavailable for initializing the background model - a problem for which we apply a novel discriminative measure. The proposed measure is essentially the probability of observing a particular pixel value, conditioned on the probability that no motion is detected, with the pdf on which the latter is based being estimated as a zero-mean and unimodal Gaussian distribution from observing the difference values between successive frames. We will show that such a measure is a powerful discriminant even under severe occlusions, and can deal robustly with the foreground aperture effect - a problem inherently caused by differencing successive frames. The detection of abandoned packages then follows at both the pixel and region level. At the pixel-level, an "abandoned pixel" is detected as a foreground pixel, at which no motion is observed. At the region-level, abandoned pixels are ascertained in a Markov Random Field (MRF), after which they are clustered. These clusters are only finally classified as abandoned packages, if they display temporal persistency in their size, shape, position and color properties, which is determined using conditional probabilities of these attributes. The algorithm is also carefully designed to avoid any thresholding, which is the pitfall of many vision systems, and which significantly improves the robustness of our system. Experimental results from real-life train station sequences demonstrate the robustness and applicability of our algorithm.

1. Introduction

The problem of detecting abandoned packages in *crowded places* is increasingly becoming an important surveillance issue. One can easily realize that crowded places are desirable and vulnerable targets of terror attacks, due to their high human densities. Detecting abandoned packages in places with high human densities introduced several challenging problems caused by severe occlusions, including the difficulties faced in modeling the background and identifying static objects, and if only a single camera is considered, only limited glimpses of the abandoned package are available over a short period of time. We consider only a single camera in this paper, with the motivation that a robust single-camera algorithm operating under severe occlusions, when extended to multiple cameras, will be extremely useful.

Given a single camera, a typical approach to the problem would be to perform change detection, followed by a threshold-based approach to detect static objects, before classifying them as possible packages based on appearance. Several researchers have thus focused on first building a background model, with the assumption that frames containing only background pixels are available (e.g., [3, 21]) in the initial phase. In this aspect, our system eliminates such a requirement by building the background model incrementally based on a novel discriminative measure, and estimating its density using kernel density estimation ([6]). While background modeling typically utilizes the history of the pixel values, our proposed measure do so conditioned on the fact that no motion should be observed at the given pixel. The intuition is simple; given frames containing moving foreground objects, the only pixels that we are interested in during the background modeling phase are those that lie in static region. While several researchers have looked into similar problem, such as [10] where the dominant mode at a pixel is used as the background, or [5] who suggested building the background model by searching for input frame in an image sequence that has background visible at a particular pixel, they depend on the fact that the particular background pixel is seem more frequently than foreground pixels over a short period of time - an assumption that quickly becomes invalid under severe occlusions.

In order to use our proposed measure, we need an efficient and effective way to detect motion, for which several approaches exist. One such approach which has been worked on extensively in the past is optical flow (e.g., [8, 22, 13]), albeit that not only is it hard to compute optical flow on the basis of image measurements only, it is also relatively slow. We address these concerns by differencing successive frames (e.g., [7, 15, 18, 20, 19]) instead, which is extremely fast and its pdf can be easily derived as a zero-mean and unimodal Gaussian distribution by observing difference values between successive frames over time. Such an approach, however, suffers from missed detections caused by homogeneous moving objects - a problem commonly known as the foreground aperture problem - unless more elaborate image processing scheme such as the ones described in [15, 20] are employed. Herein thus lies another advantage of our proposed measure; it is able to deal effectively with the problem, since each homogeneous moving region occludes the true background pixel for a short period of time, and different such regions are more than likely to exhibit different color properties. In other words, using our proposed measure, the true background pixel is much more likely to exhibit higher frequency than these “homogeneous pixels” over a sufficient period of time, given that their frequency increases whenever they become visible.

Using such a robust approach for background modeling, coupled with the superior performance of using successive frames differencing, allows the system to efficiently detect pixels belonging to potential abandoned packages as those foreground pixels detected by the background model but not the frame differencing phase. We again face problems caused by homogeneous moving regions when homogeneous pixels occlude an “abandoned pixel”, since they are equally likely to be classified (wrongly) as abandoned, being foreground and mistaken as static. We give a two-step approach to overcome this problem, which begins with deriving the pdf of observing a pixel value, given that it is detected as a foreground pixel and that no motion has been observed, with the values coming from homogeneous pixels similarly filtered out as before. Such a pdf reflects the presence of abandoned pixel or lack thereof, which can be determined by thresholding the corresponding variance - an approach that we want to avoid. To do so, we give a Markov Random Field (MRF) formulation for ascertaining abandoned pixels that considers the influence of a pixel’s neighborhood, with the optimal configuration derived as the one with the Maximum A Posteriori (MAP) probability, thereby avoiding any form of thresholding.

Abandoned pixels can then be clustered, and each cluster can be identified as abandoned packages after discarding those that are improbable abandoned packages due to their sizes. Such an approach can potentially fail when confusion arises in differentiating between real abandoned packages and other static objects, such as a person standing still. Our system avoids doing so by evaluating conditional probabilities based on the color histogram, shape, size and position of a given cluster, and identifying from the corresponding pdfs true abandoned packages as those with small variances, whereby we commit the first threshold of our system. The avoidance of thresholding, that has been the pitfall of many vision systems, makes our system extremely robust, and is perhaps the most important contribution of this paper.

2. Motion Detection

Given the speeds, u and v, of a pixel with intensity I, in the x and y directions respectively, the well-known optical flow constraint equation is given as:

$$-\frac{\partial I}{\partial t} = \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v. \tag{1}$$

The goal, then, is to solve for u and v using the smoothness constraint, whereby it is expected that adjacent pixels have largely similar motion except near the motion boundaries. In practice, optical flow computation is often prone to errors, and performs slowly due to the optimization process that steps through different combinations of u and v.

On the other hand, a further look at Eqn. 1 shows that if we just want to detect motion (i.e., $|u| > 0$ and/or $|v| > 0$) and is uninterested in the value of u and v specifically, then we can just check that $|\frac{\partial I}{\partial t}| > 0$. This would advantageously retain video rate performance, bearing in mind that doing so still suffers from the foreground aperture problem commonly associated with optical flow computation, when a homogeneous moving region causes $\frac{\partial I}{\partial x} \approx 0$ and $\frac{\partial I}{\partial y} \approx 0$, producing little change in intensity over time.

Checking that $|\frac{\partial I}{\partial t}| > 0$ is simple; we can just compute the differences in pixel values between the current frame and previous frames. For this purpose, the current frame at time t is differenced from n preceding frames, between time $t-n$ to $t-1$, giving us the difference value of a pixel, D_t, as:

$$D_t = \sum_{\tau=t-n}^{t-1} w_\tau * |C_\tau - C_t|, \tag{2}$$

where w_τ represents a normalized weighting scheme for preceding frames, such that earlier frames are given smaller weights, and C is the notation for the pixel value. The distribution of the difference values of a “static pixel” (i.e., pixel where no motion is observed) over time is zero-mean and unimodal. To see this, consider that additive Gaussian noise, n_t, is added to the true pixel value, λ, causing it to be observed as C_t, i.e.,:

$$C_t = \lambda + n_t, \tag{3}$$

where $n_t \sim N(\mu_n, \sigma_n^2)$ is a Gaussian distribution with mean μ_n and variance σ_n^2. Differencing of C_t and C_{t+1} removes the λ term, leaving the difference of n_t and n_{t+1}, of which the distribution is exactly that of D_t given as:

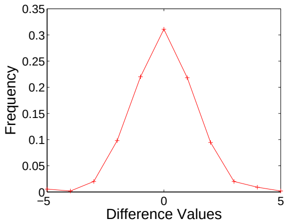


Figure 1. The difference values of an unobstructed static pixel between successive frames over 550 frames is measured against the frequencies. It shows a zero-mean, unimodal Gaussian distribution.

$$P(D_t) = \frac{e^{-\frac{[-D_t-(\mu_n-\mu_n)]^2}{2(\sigma_n^2+\sigma_n^2)}}}{\sqrt{2\pi(\sigma_n^2+\sigma_n^2)}}, \quad = \frac{e^{-\frac{D_t^2}{4\sigma^2}}}{\sqrt{4\pi\sigma_n^2}}, \tag{4}$$

with mean zero and variance $2\sigma_n^2$. An example is shown in Fig. 1, where the difference values of a static pixel (a pixel on the ceiling was chosen so that it remains unobstructed throughout) over 550 frames are measured against the frequencies with which they occur. $P(D_t)$ measures the probability of *not* observing motion, and under severe occlusions, its pdf can be estimated from the history of the difference values between successive frames, based on the understanding that it would be unimodal and zero-mean. We do this by first finding the mode that has center closest to zero, followed by delimiting it by the immediate left and right neighboring mode. Let the frequency of the first be f_0, the left neighbor be f_ℓ and the right neighbor be f_r. Then a histogram at the three values becomes respectively, $\frac{f_0}{f_0+f_\ell+f_r}$, $\frac{f_\ell}{f_0+f_\ell+f_r}$ and $\frac{f_r}{f_0+f_\ell+f_r}$. The variance of the pdf is then estimated as $\frac{1}{2\pi(\frac{f_0}{f_0+f_\ell+f_r})^2}$. We show such a plot and the estimated pdf in Fig. 2(a) and (b) respectively, whereby multiple modes caused by motion can be clearly seen.

It is important to know that such a pdf does not differentiate between homogeneous pixels and static pixels. While we can increase the value of n in Eqn. 2 to alleviate the problem, too large a value of n will result in wrongly classifying a static pixel as a non-static one. Empirically, we have found that n should not exceed 5 under video rate processing.

3. Background Modeling

The foreground aperture problem, together with severe occlusions that allow only limited glimpses of the true background pixel, make building the background extremely challenging. In view of these problems, we propose a novel discriminative measure to identify background pixels, that is based on the conditional probability of observing a pixel value when no motion is detected. To build the background, we first obtain the history of pixel values and difference values from time $t-\Delta t$ to $t-1$, given as $\{C_{t-\Delta t}, ..., C_{t-1}\}$ and $\{D_{t-\Delta t}, ..., D_{t-1}\}$ respectively, and evaluate the frequency with which each of the possible unique pixel values (we use grayscale here, so the pixel values range from 0 to 255) are seen. Given an unique pixel value, C_i, we compute its frequency, f_i, in the histogram as:

$$f_i = \sum_{\tau=t-\Delta t}^{t-1} \frac{P(|C_\tau - C_i|) * P(D_\tau)}{P(D_\tau)}, \quad = P(\Delta C_i|D), \tag{5}$$

where $P(\Delta C_i|D)$ is the conditional probability that C_i is observed based on $P(C_\tau - C_i)$, when no motion has been detected. $|C_\tau - C_i|$ and D_τ are assumed to be independent of each other, so that their joint probability can be easily computed from the respective pdf. Additionally, $P(|C_\tau - C_i|)$ behaves similarly as $P(D_\tau)$, so that both values can be derived using Eqn 4. This is illustrated in Fig. 2(c) and (d).

One can easily realize that f_i is very effective in identifying background pixels, even under severe occlusions and the presence of homogeneous moving regions. Intuitively, such a frequency measure for homogeneous pixels will be low, since they are only detected as static for a short period of time and different homogeneous moving regions are expected to exhibit vastly different color properties, whereas the same frequency measure when used for a background pixel is expected to be high since its frequency increases whenever it becomes visible. This is illustrated in Fig. 3. The frequency of a pixel on a specular surface, under severe occlusions, was recorded over 100, 200 and 300 frames in (a), (b) and (c) respectively. Each time the modes were correctly identified (manually verified). The choice of a specular pixel allows us to illustrate clearly the effectiveness of such a frequency measure, where the pixel values fluctuate greatly between the two main peaks in the plots, as moving objects cast reflections on the surface.

Given n unique pixel values, the density of the resulting background model is estimated with Gaussian kernel density estimation, so that the probability of observing a new pixel value, C_t, is given as:

$$P(C_t) = \frac{1}{\sum_{i=1}^{n} f_i} \sum_{i=1}^{n} \sum_{j=1}^{f_i} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(C_t-C_i)^2}{2\sigma^2}}, \tag{6}$$

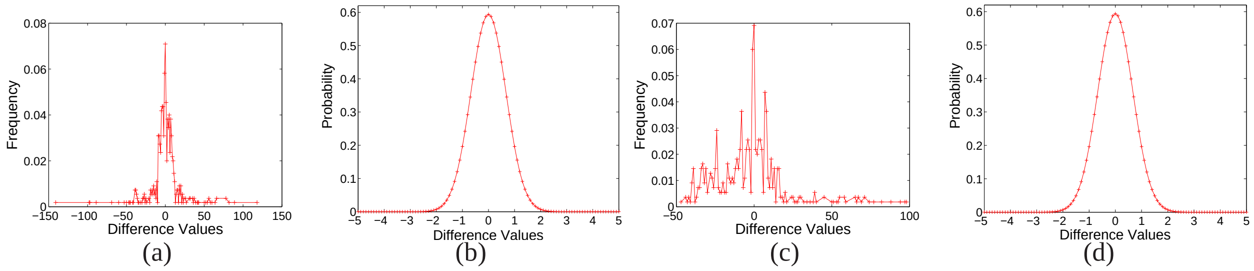


Figure 2. (a) Under severe occlusions, we can see many modes besides the one centered at zero, with the former indicating presence of motion and the latter indicating static or homogeneous pixels. (b) The system finds the mode with center closest to zero, delimiting it by the left and right neighboring mode, and computing the variance of the pdf over the same range of data. (c) In this plot, the difference values are computed as the difference between the current pixel value and the true background pixel value, at the same pixel location as (a). (d) The pdf associated with (c), estimated in the same manner as (b), is very similar to (b). All the plots are measured over 550 frames.

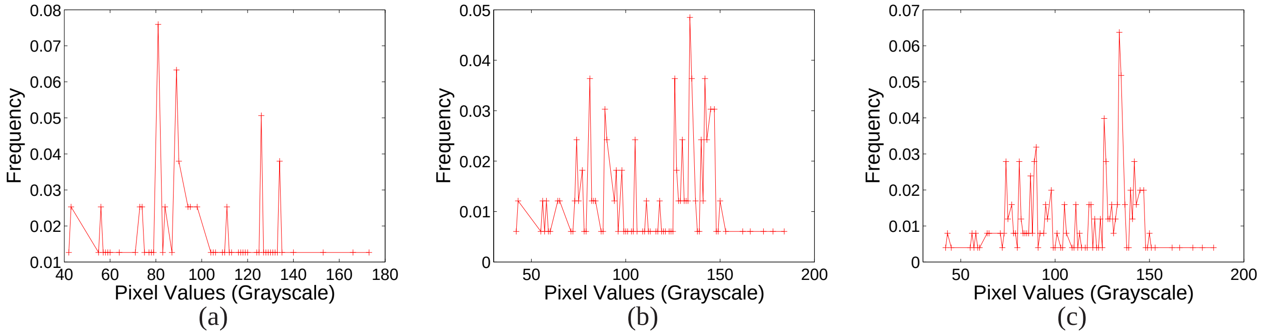


Figure 3. We show here the modes of a background pixel, identified using the frequency measure in Eqn. 5, over 100, 200 and 300 frames in (a), (b) and (c) respectively. The modes are correctly identified each time. The background pixel belongs to a specular floor surface and the two main peaks are caused by moving objects casting reflections on the surface from time to time.

where σ is the chosen bandwidth. Several methods currently exist for automatic selection of the bandwidth, notably plug-in and cross-validation methods (e.g., [12, 17]), but the simple method suggested in [6] works reasonably well and is used in our system. Here, a closer look at Eqn. 6 will reveal that we have effectively "transformed" the observed frequency to the frequency measure given by Eqn. 5 in performing a non-parametric summation of the probabilities with respect to each unique pixel value.

4. Abandoned Package Detection

4.1. Pixel-level Detection

In addition to being highly effective in identifying background pixels, the frequency measure proposed in Eqn. 5 also allows the background model to be initialized as soon as enough glimpses of the background pixel can be collected. Intuitively, once such initialization is realized for a pixel, foreground pixels belonging to abandoned packages that occlude the background pixel can be identified as pixels that are static (Eqn. 4), but yet are classified as foreground by the background model. It should, however, be clear by now that under severe occlusions, simply doing so will not be able to deal effectively with the foreground aperture problem. Instead, we propose an approach as follow.

For each pixel, we consider the histogram for the set of unique pixel values, C_i, seen at the pixel during time interval between $t - \Delta t$ and t, with the frequency of each unique pixel value, f_i, computed as:

$$\begin{aligned} f_i &= \sum_{\tau=t-\Delta t}^{t} \frac{P(|C_\tau - C_i|) * P(D_\tau) * P(\bar{C}_i)}{P(D_\tau) * P(\bar{C}_i)}, \\ &= P(\Delta C_i | D, F), \end{aligned} \qquad (7)$$

where $P(\Delta C_i|D,F)$ represents the conditional probability that C_i is observed (measured with $P(C_\tau - C_i)$), given that it is static and foreground. The joint probability is again computed on the assumption that the random variables are independent of each other. $P(\bar{C}_i)$ is the probability of seeing C_i as a foreground pixel and is the complement of Eqn. 6. We show such a distribution in Fig. 4, collected over 400 frames of a severely occluded scene. A package

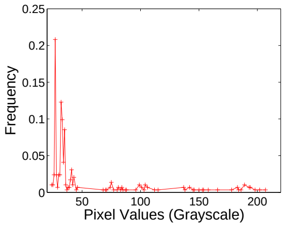


Figure 4. The plot of the distribution of a pixel over 400 frames is shown here. The frequency is measured according to Eqn. 7, and a package was abandoned midway through the frames, causing the main peak seen here.

was abandoned midway, and the distribution at a pixel location occupied by the package, is shown here before and after the package was left. It shows that the system was able to pick up correctly the abandoned pixel corresponding to the highest peak in the plot, using the frequency measure in Eqn. 7. As a result, the system can now obtain such a distribution for every pixel, look for the dominant mode (highest peak), obtain its sample probability, $P(\mu)$ (in fig. 4, its $\approx$ 0.21), and compute the corresponding sample variance as $\sigma^2 = \frac{1}{2\pi P(\mu)^2}$, with the assumption that the underlying distribution is normal. Such a sample variance reflects the likelihood that the pixel is an abandoned pixel and is used in the following section for ascertaining and clustering abandoned pixels. For convenience, we will call these variance values as “abandoned variances”.

4.2. Region-level Detection

4.2.1 Clustering Abandoned Pixels

After obtaining the abandoned variances, further processing is performed at the region-level to group the abandoned pixels into clusters. Two choices exist - we can either classify a pixel as abandoned by thresholding its abandoned variance and performing morphological operation on the abandoned pixels, or we can utilize more elaborate processing to avoid thresholding at this stage. For the latter, we propose a formulation that utilizes a MAP-MRF (Maximum A Posteriori-Markov Random Field) labeling technique ([1, 9]), which has the additional advantage of considering the obvious importance of the relationship of a pixel to its neighbors in ascertaining that it is indeed an abandoned pixel.

Consider a configuration, $f = \{f_1, ..., f_m\}$, where m is the number of pixels in the image, and the set of abandoned variances, $d = \{\sigma_1^2, ..., \sigma_m^2\}$. Each $f_i \in f$ is assigned label 1 or 0, to indicate whether the corresponding pixel is an abandoned pixel or not respectively. Clearly, such a configuration is Markovian, i.e., the label of a pixel interacts only with the neighboring labels. Due to the Hammersley-Clifford theorem ([4, 11]) that establishes the equivalence between the properties of MRFs and Gibbs distribution, the probability of a configuration, $P(f)$, can be written as:

$$P(f) = \frac{1}{Z} e^{-\frac{1}{T}U(f)}, \tag{8}$$

where T is called the temperature and usually assumed to be 1, and given that F is the set containing all possible configurations, Z, called the partition function, can be written as $Z = \sum_{f \in F} e^{-\frac{1}{T}U(f)}$. For the purpose of performing MAP, Z is fortunately inconsequential, since $P(f) \propto e^{-\frac{1}{T}U(f)}$. We then compute $U(f)$, consisting of only pair-site cliques, to encourage smoothness in the clustering:

$$U(f) = \sum_{i}^{m} \sum_{i' \in N_i} \frac{1}{2}(f_i - f_{i'})^2, \tag{9}$$

where N_i is the 8-neighborhood system of i. To obtain the configuration f_{max} with the maximum a posteriori probability, the set d gives us the abandoned variance of each pixel, with a smaller variance implying a higher likelihood of being abandoned. A weighting scheme, W_σ, is used for modeling the variance that comprises two separate exponential functions for label 0 and 1 respectively. Although they are not probability density functions, we will see that they suffice for the purpose of maximizing the posterior probability. They are given as:

$$W_\sigma(\sigma_i|f_i) \quad = \quad \begin{cases} e^{-\frac{\sigma_i}{\theta_1}} & f_i = 1, \\ e^{-\theta_0 e^{-\frac{\sigma_i}{\theta_0}}} & f_i = 0. \end{cases} \tag{10}$$

The above exponential functions are designed to satisfy several conditions. Firstly, for label 1, the function should be monotonically decreasing as the variance increases, and the opposite should be true for that of label 0. Secondly, we want to be able to perform MAP without needing to compute Z, for obvious performance reason, and this is achieved by using exponential functions. Lastly, the probability given by one function at a particular value of variance should complement as much as possible that of the other function, i.e., if ρ is the probability of being label 0, then the probability of being label 1 should be as close as possible to $1 - \rho$. We achieve this (approximately) by setting $\theta_0 = 4$ and $\theta_1 = 12$, noting that these values are set once and is inconsequential to the robustness of the algorithm. A plot of both functions with these θ-values is shown in Fig. 5.

Based on Eqn. 10, the likelihood density can be written as:

$$P(d|f) = \prod_{i=1}^{m} W_\sigma(\sigma_i|f_i), \tag{11}$$

Since the posterior probability is:

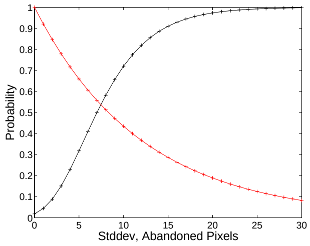


Figure 5. θ_0 and θ_1 are set to 4 and 12 respectively. The functions complement each other approximately as shown.

$$P(f|d) \propto e^{-U(f|d)}, \tag{12}$$

taking the log of $P(f|d) \propto P(d|f)P(f)$ gives:

$$U(f|d) = U(d|f) + U(f), \tag{13}$$

where $U(d|f)$ can be written as:

$$U(d|f) = \sum_{i=1}^{m}(1 - f_i)\theta_0 e^{-\frac{\sigma_i}{\theta_0}} + f_i\frac{\sigma_i}{\theta_1}. \tag{14}$$

The MAP estimate of f_{max} then becomes:

$$f_{max} =^{\arg\min}_{f} U(f|d). \tag{15}$$

Unfortunately, optimization of the above cost function is an exponential problem, since there would be 2^m different combinations of f. Our problem is however simpler, given that abandoned package is expected to be a rare event, so that it is very unlikely that there would be a lot of (if any) true abandoned pixels at any given time. Our proposal, is then, to minimize the number of different combinations of f to be considered in each of a number of iterations. The pixels are split into different sets, so that the size of each set is sufficiently small to avoid any significant deterioration in performance. Optimization is then performed on an initial set, after which optimization of subsequent sets is conditioned on the state of the sets that have been optimized. Such an optimization procedure is in essence what is known as the Iterated Conditional Modes (ICM) approach ([2]). While the ICM approach might converge to some local instead of global maxima of the a posteriori probability, it performs exceptionally well for our problem, without causing significant slowdown.

With the pixels properly labeled, we can then cluster pixels that have been positively labeled. The clustering process is "loosely" performed, whereby any pixel within a 8-neighborhood system of a pixel is cluster together, for the same reason that abandoned package is rare, and it is very unlikely that multiple packages would be abandoned at the same time.

4.2.2 Region-level Semantics

The set of candidate abandoned packages that have been identified allows further ascertainment at the region-level, of which there are several advantages. In addition to reascertaining abandoned pixels, it can also help in situations where the system needs to distinguish between true abandoned packages and other static objects, such as a person standing still, that greatly confuses the system. This is particularly important in our application of interest - detecting abandoned packages in train stations - where severe occlusions arise from high human traffic, and where people very often stood in place. Although a coarse filtering step, during which oversized or undersized clusters are discarded, proves to be very effective, noise in the foreground detection phase can sometimes break the object into multiple smaller foreground regions or the object might only be partially visible due to occlusions. To overcome this problem, we adopt an approach that is based on the observation that true abandoned package remains absolutely stationary (as compared to, say, a person standing in place). Then, it can be expected that the shape and color of a true abandoned package, at the initially detected location, would remain relatively constant over time.

Given a candidate abandoned package initially detected at time t, the system first notes its size (ς_t), shape (δ_t), position (ρ_t) and color histogram (C_t). Because we expect the abandoned package to be stationary, we look for it at ρ_t in subsequent frames; this simplifies the task, which would otherwise require tracking. Within the same image region, given by ς_t at ρ_t, in a subsequent frame, we then evaluate whether the package is still there, both in terms of shape and color histogram. For the former, the Hausdorff distance is used ([14, 21]). Edges are first detected in the initial and subsequent frame within the boundaries given by ς_t, yielding two sets, A_t and B_t respectively, containing points lying on detected edges. The Hausdorff distance, $H(A_t, B_t)$, between them is then given as:

$$H(A_t, B_t) = \max(h(A_t, B_t), h(B_t, A_t)), \tag{16}$$

where

$$h(A_t, B_t) = \max_{a \in A_t} \min_{b \in B_t} P(D_b) * |a - b|. \tag{17}$$

The Hausdorff distance, $H(A_t, B_t)$, measures the distance of the point of A_t that is farthest from any point of B_t, and is particularly useful in our context for comparing shapes, since we clearly do not have to worry about scaling and transformations (the package is expected to be stationary). By adding the term, $P(D_b)$, that represents the probability of observing no motion at b, we also impose the requirement that the pixel used in the calculation should be static.

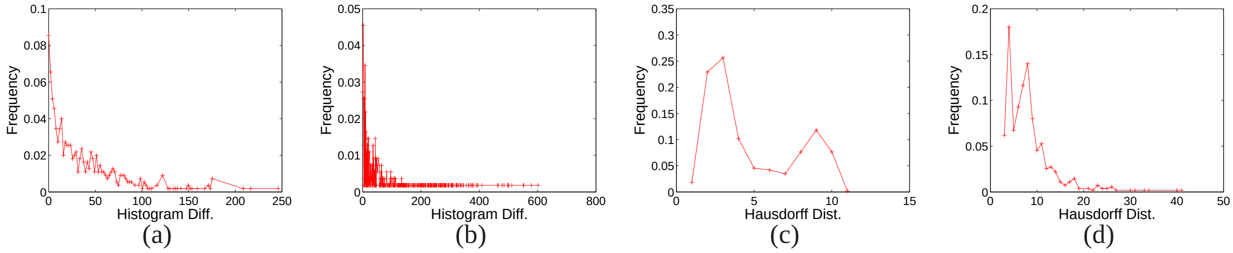


Figure 6. (a) The color histogram differences over 550 frames are computed for an unobstructed region, and the dominant mode is predictably centered at zero. (b) Here, the color histogram differences are computed for a different region that experienced severe occlusions. (c)(d) The Hausdorff distances over the same frames are computed for an unobstructed region and one that has severe occlusions respectively. We look for modes with centers closest to zero for both the color histogram difference and Hausdorff distance, and use them to estimate the corresponding pdfs needed for computing the probability in Eqn. 19.

Following shape comparison using the Hausdorff distance, differences in color properties are then evaluated. The main concerns here are the efficiency and effectiveness of any such comparison algorithms. We adopt a simple approach as follow. We first convert the initial and subsequent frame to grayscale, and split the resulting histograms, X_t and Y_t respectively, into 16 bins each. The grayscale histogram distance measure, $d_{hist}(X_t, Y_t)$, is then computed by the following quadratic form ([16]):

$$d_{hist}(X_t, Y_t) = (X_t - Y_t)^T P_{D_Y} W P_{D_Y} (X_t - Y_t), \quad (18)$$

where W is a 16×16 weight matrix, that gives the similarity between different bins, and contain ones on the diagonal, and P_{D_Y} is the matrix containing the probability of observing motion for each pixel used in the computation. Each element of W is computed as $1 - \frac{(|row_{diag} - row_{elem}|)}{16}$, where row_{diag} and row_{elem} are respectively the row index of the diagonal element and the row index of the element in the same column. Using these measures, observations made from time $t + 1$ to $t + \Delta t$, $\{H(A_t, B_{t+1})), ..., H(A_t, B_{t+\Delta t})\}$ and $\{d_{hist}(X_t, Y_{t+1}), ..., d_{hist}(X_t, Y_{t+\Delta t})\}$, allow the system to finally classify a cluster as abandoned package when the following probability exceeds T:

$$P(D_{hausdorff}|D_{color}) = \frac{P(D_{hausdorff}) * P(D_{color})}{P(D_{color})}, \quad (19)$$

where $P(D_{hausdorff})$ and $P(D_{color})$ are respectively the probability of observing differences in the Hausdorff distance and color histogram. The rationale in such a measure is to condition the detection of abandoned packages on the consistency in the color properties of candidate packages, thereby achieving robustness even under severe occlusions and the presence of foreground aperture effect. Since we expect $P(D_{hausdorff})$ and $P_(D_{color})$ to be unimodal and zero-mean, we estimate their densities from the observations by again looking for the mode with center closest to zero, and delimiting it with the neighboring modes. We show in Fig. 6(a) and (b), the color histogram differences measured over 550 frames, for an unobstructed region and one with severe occlusions respectively. Each of them clearly shows mode centered at zero value, that is extracted as the pdf for use in Eqn. 19. Fig. 6(c) and (d) show the corresponding plots for the Hausdorff distances measured over the same frames. Finally, Fig. 7 shows the plot for the values computed by Eqn. 19 for a real abandoned package, the dominant mode of which exhibits a small variance.

5. Experimental Results

We have applied our algorithm to several challenging video sequences that have been collected from extremely crowded train stations, two of which are shown here in Fig. 8 and Fig. 9. In Fig. 8(a), we show in the leftmost image the result of performing successive frames differencing, which expectedly revealed missed detections in homogeneous moving regions. In the following image to the right, we show abandoned pixels in green, which were detected as foreground in static regions, in a short amount of time after the package was left by the perpetrator. Then, clustering of the abandoned pixels is demonstrated in the third image from the left, where a blue box was used to bound the detected cluster as candidate abandoned package. Finally, we show in the rightmost image the edge map of the scene, used in computing the Hausdorff distance. Evidently, as we proceed from (a) to (d), the edge maps within the static region of the candidate abandoned package gave Hausdorff distances that were expectedly small. That, coupled with similar color properties, causes a discernable mode shown in Fig. 7, so that the package was finally classified as abandoned (bounded by a red box in the third image from the left in (d)), demonstrating the effectiveness of our algorithm in picking up the presence of abandoned packages even under

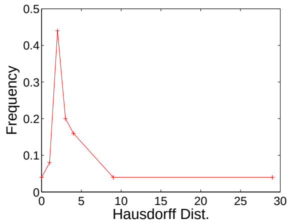


Figure 7. Using the measure given by Eqn. 19, we are able to detect a significant mode for a real abandoned package.

severe occlusions.

We then show in fig. 9 results that allow the readers to appreciate the merits of region-level detection, using Hausdorff distance and color histogram similarity. In (a), in the third image from the left, the lady circled by a red ellipse was standing in place and causing abandoned pixels to be wrongly detected (green pixels in the second image from left). However, as we have pointed out, such objects in the scene can be distinguished from real abandoned packages because they are seldom absolutely stationary. So, in (b), we see that the edge map in the corresponding region was very different from that in (a), thus avoiding false detection of the lady as an abandoned package. The idea is further illustrated in (c) and (d), where a falsely detected candidate package was bounded in blue box in the third image from the left in (c). This was correctly removed in (d) when the same region exhibited different shape and color. The real abandoned package in this sequence was eventually detected in (e) and (f). Note that due to the large amount of specularities and lighting changes in this sequence, there were constant detections on the ceilings and signboards. A closer look, however, allows one to realize that they did not cause any significant problem since both the background model and motion detection phase have positively detected them.

Results shown here, together with additional results, are also provided in the accompanying video sequences.

6. Conclusions

We have described a system capable of detecting abandoned packages under severe occlusions. The most important contribution of this paper is the statistical framework used to propagate probability associated with the decision made in each step to the next, requiring the system to threshold only at the last stage. Several novelties can be claimed by the statistical framework. These include, firstly, the proposal of a strong discriminative measure to identify background pixels, even under severe occlusions and the presence of homogeneous moving regions. Moreover, the statistical evaluation of the shape and color properties of abandoned packages using Hausdorff distance and a simple quadratic histogram similarity measure, coupled with an MRF formulation for clustering abandoned pixels, allow the system to robustly identify true abandoned packages. By avoiding thresholding in our algorithm, it becomes extremely robust for use in real world surveillance applications. Further work that can be extended from here includes the use of our algorithm to first detect abandoned packages, after which the system can backtrack in time to determine the perpetrator who is expected to be near the location of the package at the instance it was first left behind. It would also be a good idea to extend our algorithm to multiple cameras, so that problems caused by severe occlusions can be more effectively dealt with.

References

[1] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, B(36):192–236, 1974. 5

[2] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, B(48):259–302, 1986. 6

[3] M. D. Beynon, D. J. V. Hook, M. Seibert, and A. Peacock. Detecting abandoned packages in a multi-camera video surveillance system. In *IEEE Conference on Advanced Video and Signal Based Surveillance, Miami, Florida*, Jul 2003. 1

[4] P. Clifford. Markov random fields in statistics. *Geoffrey Grimmett and Domnic Welsh (Eds.), Disorder in Physical Systems: A Volume in Honour of John M. Hammersley*, pages 19–32, 1990. 5

[5] S. Cohen. Background estimation as a labeling problem. In *ICCV, Beijing, China*, Oct 2005. 1

[6] A. Elgammal, D. Harwood, and L. S. Davis. Non-parametric model for background subtraction. In *ECCV, Dublin, Ireland*, Jun 2000. 1, 4

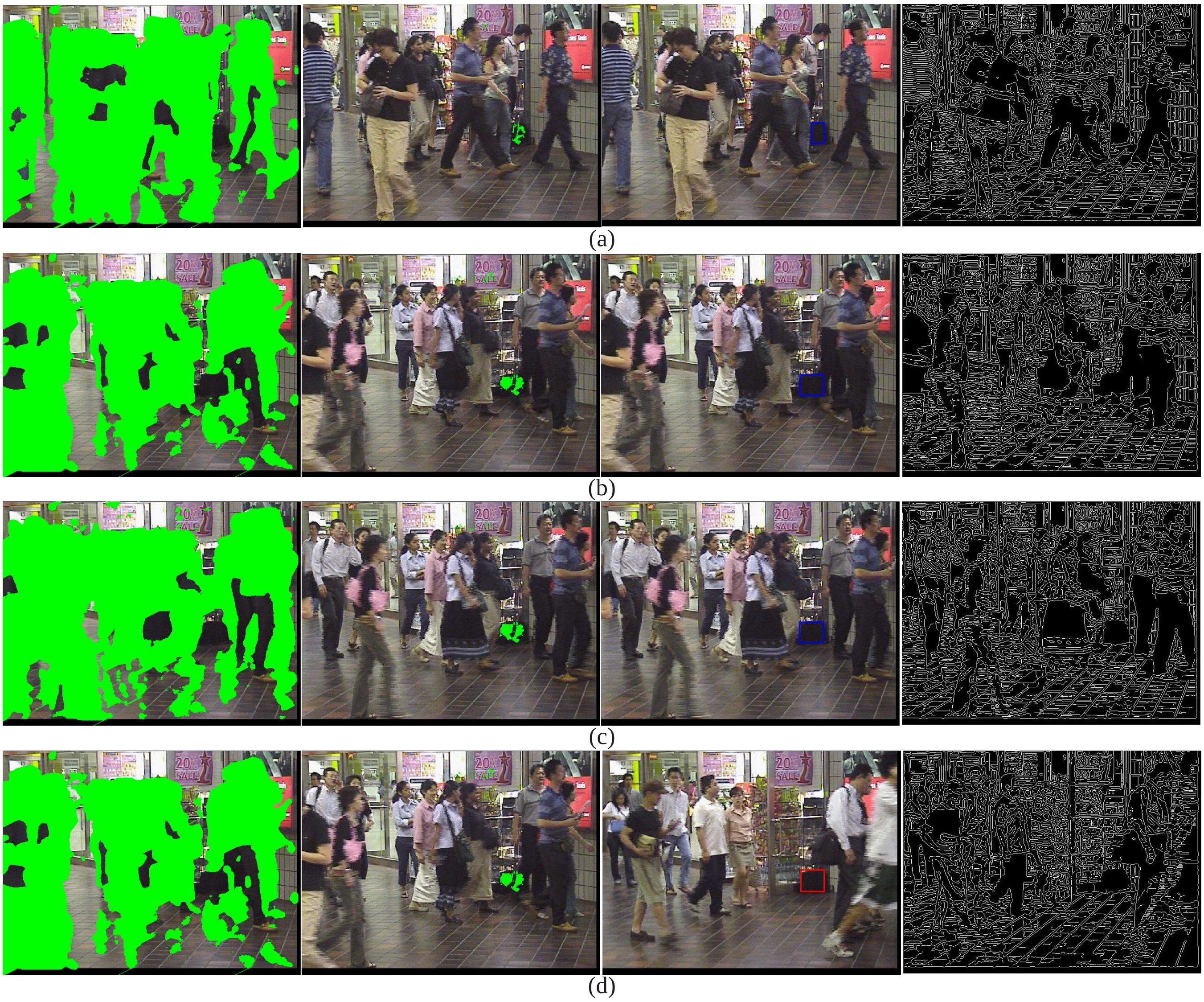
[7] D. Farin, P. H. N. de With, and W. Effelsberg. Robust background estimation for complex video sequences. In *ICIP, Barcelona, Spain*, Sep 2003. 2

[8] L. Gaucher and G. Medioni. Accurate motion flow estimation with discontinuities. In *ICCV, Kerkyra, Greece*, Sep 1999. 1

[9] S. Geman and D. Geman. Stochastic relaxation, gibbs distribution and the bayesian restoration of images. *IEEE PAMI*, 6(6):721–741, 1984. 5

[10] D. Gibbins, G. Newsam, and M. Brooks. Detecting suspicious background changes in video surveillance of busy scenes. In *3rd IEEE Workshop on Applications of Computer Vision, Sarasota, Florida*, Dec 1996. 1

[11] J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. *Unpublished*, 1971. 5


(a)

(b)

(c)

(d)

Figure 8. Please refer to the first paragraph of Sec. 5 for detailed explanation. From left to right in each row: successive frames differencing, foreground detections in static regions, abandoned package detections and edge map.

[12] W. Hardle. Smoothing techniques, with implementations in s. *Springer, New York*, 1991. 4

[13] B. K. P. Horn and B. G. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185 – 203, 1981. 1

[14] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using hausdorff distance. *IEEE PAMI*, 15(9):850–863, 1993. 6

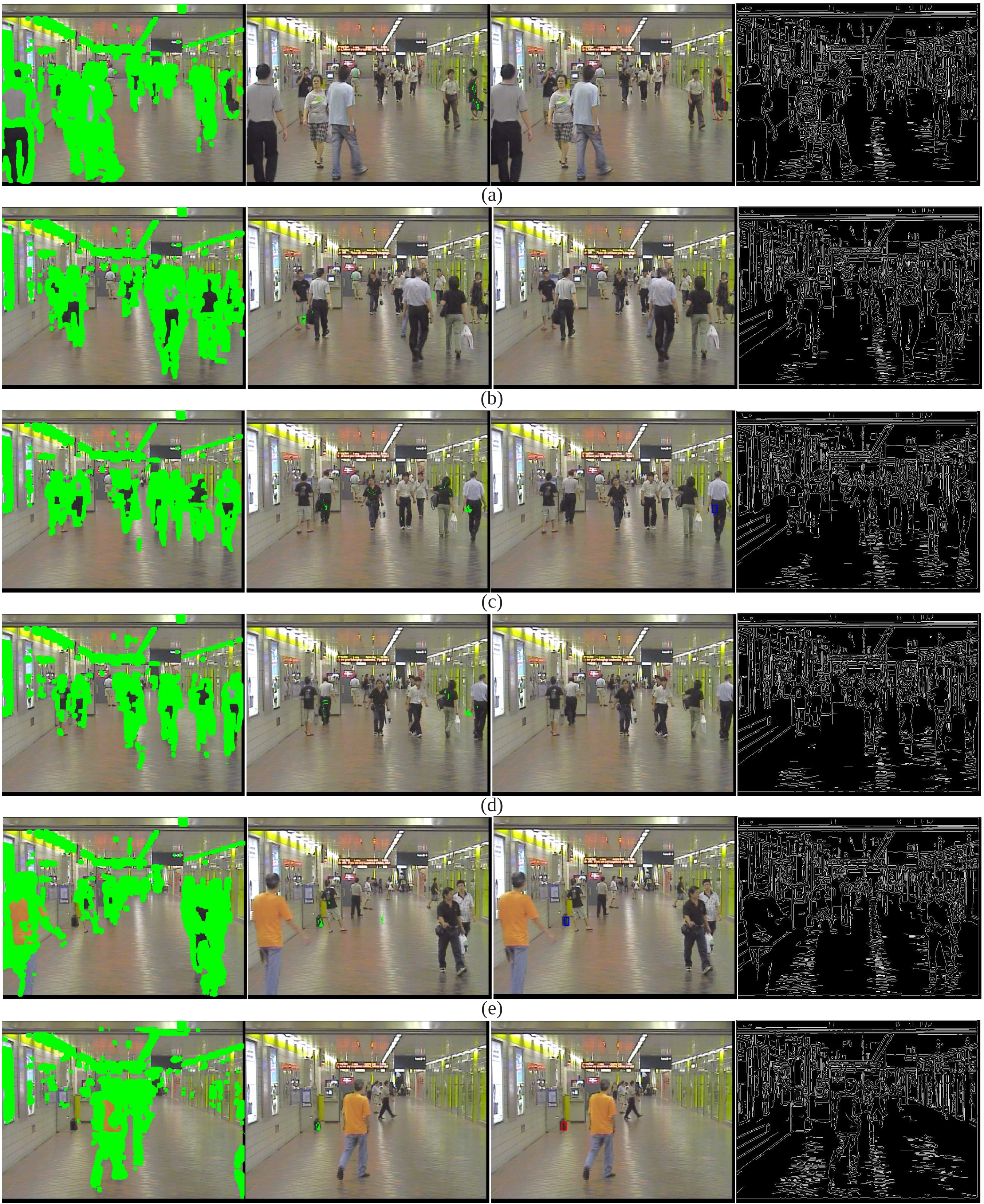
[15] M.-S. Lee. Detecting people in cluttered indoor scenes. In *CVPR, Hilton Head Island, South Carolina*, Jun 2000. 2

[16] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, and C. Faloutsos. The qbic project: Querying images by content using color, texture, and shape. *Storage and Retrieval for Image and Video Databases, volume SPIE*, 1908, 1993. 7

[17] B. U. Park and B. A. Turlach. Practical performance of several data driven bandwidth selectors. *Computational Statistics*, 7:251–270, 1992. 4

[18] J. Ramesh and N. H. H. On the analysis of accumulative difference pictures from image sequences of real world scenes. *IEEE PAMI*, 1(2):206–214, 1979. 2

[19] J. Ramesh, M. W. N., and A. J. K. Segmentation through the detection of changes due to motion. *Computer Graphics and Image Processing*, 1(11):13–34, 1979. 2



(a)

(b)

(c)

(d)

(e)

(f)

Figure 9. Please refer to the second paragraph of Sec. 5 for detailed explanation.

[20] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *ICCV, Kerkyra, Greece*, Sep 1999. 2

[21] J. Wang and W.-T. Ooi. Detecting static objects in busy scenes. *Technical Report TR99-1730, Department of Computer Science, Cornell University*, Feb. 1, 6

[22] Y. Weiss. Smoothness in layers: Motion segmentation using nonparametric mixture estimation. In *CVPR, San Juan, Puerto Rico*, Jun 1997. 1