Background Modeling and Object Tracking Using Multi-Spectral Sensors

Cheng-Yao Chen
Department of Electrical Engineering
Princeton University
Princeton, New Jersey, USA
chengc@princeton.edu

Wayne Wolf*
Department of Electrical Engineering
Princeton University
Princeton, New Jersey, USA
wolf@princeton.edu

ABSTRACT

In this paper, we present a multi-spectral video surveillance system. Improved background modeling and appearance-based object tracking are proposed with both signal-level and decision-level multi-spectral information fusion. In addition to modeling observations in each spectral channel by a typical pixel-level mixture-of-Gaussian-based model, we also model high level factors such as confidence of each modality, motion, object area, and lighting with a hierarchical probabilistic model feedback. This hierarchical model can enhance the performance of different challenging environment conditions, such as global illumination changes, and random parameter failures of background subtraction. Moreover, real-world vision problems include occlusion and merge/split are managed by our non-parametric tracking methodology and appearance-distance histogram. Our experiment in object tracking shows that under normal conditions, our system extends the capability of single spectral sensor, and under severe environment conditions, the overall system performance outperforms traditional direct fusion techniques in tracking reliability. This promising performance also encourages us to further extend our techniques to general multi-spectral and multi-modal surveillance.

Categories and Subject Descriptors

I.4.6 [**Image Processing and Computer Vision**]: Segmentation—*pixel classification*; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*sensor fusion, tracking*

General Terms

Algorithms, Design

Keywords

Background modeling, object appearance tracking, multi-spectral sensor, object occlusion, object merge and split, sensor fusion, video surveillance

*This work is supported by ARO grant # W911NF-05-1-0480.

VSSN'06, October 27, 2006, Santa Barbara, California, USA.

1. INTRODUCTION

Background modeling and appearance-based object tracking are important techniques which are applied commonly in video surveillance. Even though with great progresses in the computer vision area, their performance and robustness are still greatly affected and limited by the environment conditions especially with typical visible spectrum sensor only. Those environmental factors include lighting, shadow, weather, occlusion, etc. Solutions for those performance limitations are using different spectral modalities and/or using multi-sensory setups. Adding another spectral modality within the same sensor can not only improve visual signal reliability in normal conditions but also enhance the signal robustness in challenging conditions. As the costs of infrared cameras keep decreasing, it will be more and more appealing to deploying multi-spectral sensors. On the other hand, multi-sensory setup will help enhance fault tolerance at system level. It is our belief that applying multi-spectral and multi-sensory system setups will be the final solution for reliable real-world video surveillance. In this article, we, however, will first only focus on applying multi-spectral sensor information processing to enhance the robustness of background modeling and appearance-based object tracking.

A great deal of research has studied methodologies for using multi-modal or multi-spectral separately in background subtraction and appearance-based object tracking. Here we will review some of them in order to have a better insight of the advantages of using our proposed approach. Stauffer and Grimson [9] proposed a statistical framework to model the scene background. They characterized each image pixel with a mixture-of-Gaussian model and adaptively updated the model with new observations. This method has been shown as one of the most effective approach to describe the background, and it serves as the framework of our pixel-level background model. However, with another sensor modality, using higher-level object information to update the background model will be beneficial.

Javed et al. [7] presented a bottom-to-top hierarchical background subtraction using both color and gradient information. Although they did not use different spectral modalities, the usage of an extra modality from gradient information is illustrative of how to fuse another modality to typical visible spectrum background modeling in the signal level and decision level. Their approach separates foreground objects and updates the background model using three different level information from pixel, region, to frame. We consider this work as an important framework for single multi-modal sensor setup, and we extend the idea of signal fusion and the high level information feedback to multi-spectral sensors specifically while the assumptions and observations for each channel are significantly different in our work.

Conaire at el. [1] proposed a background information fusion technique in long-wave infrared (LWIR) and visible spectrum video which is the closest to out proposed work. However, they made several assumptions in background modeling based on human tracking which may not be applicable for general object tracking. Moreover, there is no framework dealing with inconsistency between two channels except rule-based decisions. In their following work, Conaire at el. [2] presented an idea of using "Transferable Belief Model" to adaptively update the appearance model by combining foreground segments information in different spectral channels. Their usage of "belief fusion" is crucial to solve the decision inconsistency between different channels, but their algorithm uses only the current observation and does not have sufficient information about the camera belief of each individual object along the time when updating the appearance. Additionally, they did not discuss the framework for object merge, split, and occlusion, lighting correction, by taking the advantage of different spectral modalities.

Torresan et al. [10] also presented a model-level infrared and visible spectrum information fusion system for pedestrian detection and tracking. They tried to apply this bi-modal system to solve several problems mentioned previously including, object merge, and occlusion. They reported successful improvements as compared to traditional visible spectrum only solutions. However, they only used rule-based decisions in the model level which will again have problems when there is an inconsistency between the two spectral observations.

The methodology that we propose in this article is a hierarchical information fusion approach from pixel level to object level. It applies multi-spectral sensors to enhance the background modeling and appearance object tracking in real-world surveillance environment considerations. Additionally, high-level camera confidence model for each individual detected object is proposed to resolve the problem of signal fusion among multiple channels. Improved fault tolerance for challenging vision problems such as lighting variations, object merge, split, and occlusion are going to be discussed as well.

This paper is organized as follows. In Section 2, we present the proposed method for background modeling of a multi-spectral sensor with hierarchical information feedback. In Section 3, we propose a environment invariant pixel-level image fusion for appearance tracking. We use a probabilistic similarity model to deal with object occlusion in single camera view in Section 4. Object merge and split decisions by appearance-distance information are discussed in Section 5. In Section 6, we discuss the performance of the proposed method with public and customized testing sequences. Finally, we briefly conclude our work with some future directions in Section 7.

2. BACKGROUND MODELING BY MULTI-SPECTRAL INFORMATION

As mentioned before, we apply a mixture-of-Gaussian model to describe the background scenes. We first perform this background modeling which updates the background model by pixel-level observations. Then after grouping the objects by the connected component analysis in each channel, we feedback observations from previous object-level information to each background pixel in each channel. We also assume that the pixel error probability of each channel at the feature level is independent, so each channel can be updated separately. Moreover, in our system, image registration is done properly so that the correspondence of each pixel in each channel is already calibrated. As a result, we do not combine independent observations until final foreground regions are needed. The probability density function of kth Gaussian at pixel(x, y) at time t,

$$p(b(x,y)|t) = \frac{1}{\sqrt{2\pi\sigma(x,y)_k^2}}\, e^{\left(\frac{-(b(x,y)_t - m(x,y)_k)^2}{2\sigma(x,y)_k^2}\right)}$$

Where $b(x, y)$ is the current observation, and $m(x, y)_k$ and $\sigma(x, y)_k^2$ are the mean and variance for kth Gaussian at location (x, y) respectively.

Shadow detection is performed first to prevent putting shadows, which are caused by moving objects, into the background modeling. For visible spectrum, most shadows happen during the daytime and less commonly in the nighttime. However, if the full moon or the street light is present, the observation in visible spectrum is still interfered strongly by the shadow pixels. On the other hand, shadows for moving objects are less critical in infrared channel due to small sensitivity of temperature or reflectivity beyond visible spectrum. As a result, removing shadows from the visible channel by infrared channel observation is informative. Our shadow detection process is motivated by Cucchiara et al. [3] and is enhanced by the infrared channel observations. First, we perform image color space conversion. We convert the color image from RGB format to HSV in possible foreground regions. The criteria for a shadow region is defined as if its HSV observations and corresponding infrared observations follow,

$$h \approx H_{adj},\ s \approx S_{adj},\ v \leq \xi,\ i \neq I_f$$

$$R_{adj} = |D_{obj}| \times |M_{obj}| + \gamma$$

Where h, s, and v are the HSV values at the given pixel location. H_{adj} and S_{adj} are the average hue and saturation values at the corresponding adjacent region. Where i, the observation at the same location in infrared channel, is not categorized as a foreground pixel I_f. ξ is a small threshold for low luminance property of the shadow. R_{adj} defines adjacent region radius which is estimated by the previously identified object size,D_{obj}, its projected motion magnitude,$|M_{adj}|$, and a tolerance threshold,γ. If a pixel is categorized as a shadow pixel, we do not update the observation to the background model.

If a shadow is caused by background objects or moving clouds, there will not be an obvious moving foreground object pertaining to it. If the shadow is caused by a moving foreground object, the angle between the object and the shadow is then determined by the angle between the two largest group motion vectors which are normalized by the total pixels within each region. This can alleviate the side effects result from non-rigid object motion. Afterwards, we construct and update a lighting map which describe largest global illumination source locations by recursively minimizing the mean square error of angles between all the objects and their corresponding shadows. In this article, we assume only one global illumination source is present. This light source location information will be used later when we approximate the illumination correction in the next section.

There are two parameters describing the updating properties of adaptive background modeling. η is related to prior weighting and α is related to the updating speed. If pixel observations of the two rectangular blobs are not consistent in each channel, it will cause a mismatch in foreground region area and pixel observations are relatively "unstable" as compared to pixels in other regions. So within the non-overlapping or inconsistent areas, we decrease the weighting η of the maximum Gaussian to be less than the normal setting for the current observation, and re-normalize the weighting

of other distribution correspondingly by the following equation,

$$\eta_{mod}^{max} = \eta_{org}^{max} - (1 - (A_{vis} \bigcap A_{inf})) \times k_A$$

$$\eta_{mod}^{others} = \eta_{org}^{others} + (\eta_{org}^{max} - \eta_{mod}^{max})/(\#\ of\ distributions - 1)$$

Where k_A is an adaptive coefficient changing with the overall image contrast. It is learned from sample training sequences by minimizing the detection error of manually indexed foreground objects. This operation relates the confidence of background pixel updates to the degree of observation overlap. As our methods shows, in unstable regions, the weighting of predominant errors could be decreased and the real background or foreground pixels could be revealed in the following frames.

The object motion magnitude would imply the updating speed of the background. For example, when the foreground object movement is fast, the background pixel near the object region should be updated slower and vice versa due to the imaging capture defects. By using this observation, random motion noises caused by lighting or sensor sensitivity will not be included into the background. We calculate a group motion vector and project it back to possible motion trace regions. This group motion vector is calculated by correlating all the pixel's motion vectors within a foreground object blob. As a result, the updating speed α in the trace regions is decreased by the following equation,

$$\alpha_{mod} = \alpha_{org} \times k_M / |M_{group}|$$

Where k_M again is an adaptive coefficient learned from the sample sequences by estimation maximization of manually indexed foreground objects. $|M_{group}|$ is the magnitude of the motion vector normalized to the known camera coordinate for alleviating three dimension distortion. This operation shows that if the observations are not consistent, then a decrease in the current reliability for that observation is applied. Moreover, a exponential decaying time modification factor is attached to each pixel. If within a region there is no foreground object detected for a certain amount of time, the background updating speed and weighting pertaining to that region will go back to the normal settings gradually.

Assuming the size of the object and the motion direction will not change significantly during each observation period, we can model the object confidence factor from each sensor by a two-dimension Gaussian distribution defined as follows,

$$C_{obj}(s) = \frac{1}{\sqrt{2\pi\sigma_{m,a}^2(s)}}\, e^{\left(\frac{-(O_{m,a}(s)-\mu_{m,a}(s))^2}{2\sigma_{m,a}^2(s)}\right)} \quad s = v\ \ or\ \ i$$

Where s is either visible or infrared. $O_{m,a}(s)$ are the current observations of object motion and area size. $\mu_{m,a}$ and $\sigma_{m,a}$ are means and covariance of the tracked motion and area respectively in each modality. This Gaussian confidence model indicates that if the object size of a new observation of a given object is closer to the mean size of its previous observations, it is more reliable. For example, if an object enters the scene, its observed size will be increasing gradually and then reach a steady state. It will decrease when it leaves the scene. The same activity can be observed when there is a occlusion event. For object motion, it has a similar observation activity as the object area size. As a result, this confidence model can describe the object observation reliability within the view. This model can greatly alleviate the problems of sudden global illumination changes of video cameras and sudden polar reversals of thermal sensors.

Finally, $P(x, y)$ is then defined as a foreground pixel if it satisfies the following condition,

$$C_{obj}(v) \times P_v(x,y) + C_{obj}(i) \times P_i(x,y) > I_v \times \mu_v + I_i \times \mu_i$$

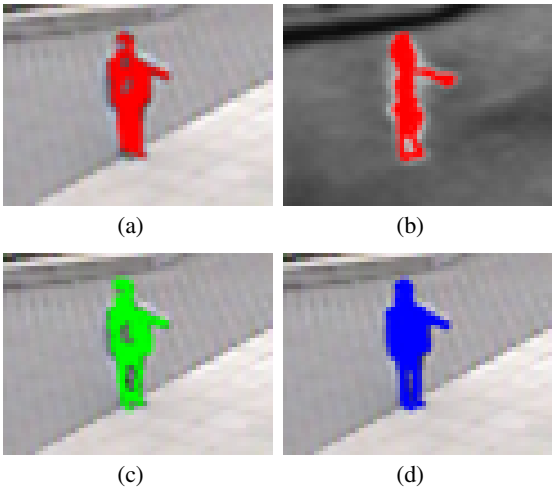

Figure 1: Foreground pixels in different modalities. ((a): Foreground pixels in the visible spectrum. (b): Foreground pixels in the infrared spectrum. (c): Foreground pixels in the direct signal fusion. (d): Foreground pixels in the proposed weighted fusion.)

Where I_v and I_i are the overall image contrast which are measured previously by the dynamic range of each pixel in visible and infrared spectrum respectively. μ_v and μ_i are thresholds of detecting foreground pixels in the visible spectrum and the infrared spectrum respectively. This adaptive threshold based on the current global contrast ratio further provides robust foreground observations. Sample results of updating the background modeling with the discussed bottom-up approach is illustrated in Figure 1. From Figure 1 (a) to (d) are visible only, infrared only, direct fusion, and our approach respectively. Direct combination refers to treating infrared signal as the fourth channel besides three channels in visible spectrum with the same weighting and performing the background modeling directly. For direct combination and our approach, the foreground pixels are drawn on the visible channel image for comparison and visualization. As we can see, the foreground regions are better extracted by our proposed approach especially for those foreground pixels whose color are closer to the background near the object outlier.

3. OBJECT TRACKING WITH WEIGHTED MULTI-SPECTRAL FUSION

As soon as we determine foreground regions, we can match those rectangular blobs with previously recognized appearance models. Signal-level fusion appearance based object modeling has several advantages compared to object-level or decision-level fusion. Those advantages include possible smaller number of feature data dimensions for object modeling, and better robustness when one or more modalities fail constantly. The latter one is especially important since there is usually no prior knowledge of the signal robustness of each channel along the time. However, the challenge of signal-level fusion is how well the representation fused from selected signals covers the original feature space with physical observation supports if possible.

Infrared images help us remove shadows and some occlusions. However, for appearance based tracking, the infrared channel does not provide enough resolution as compared to visible spectrum.

On the other hand, the visible spectrum contains more information about the object appearance but is more noisy when in a complex background. As a result, combining two channels can improve the performance and robustness of appearance model. Lighting correction is always a challenging issue in visible spectrum image processing. Moreover, the appearance in visible spectrum changes significantly when the lighting condition changes. Here we propose a signal-level image fusion appearance model invariant to global lighting variations.

If the object size is small, the shape of the object is less important when reflecting the environment lighting. Then we assume the directional cylinder reflectance is applicable to the object reflectance. As a result, the reflected intensity in visible spectrum can be characterized by the angle between the reflectance normal vector and the light source location. For each spectral channel, the response is defined as follows,

$$f(T) = \int_{\lambda_1}^{\lambda_2} \varepsilon(\lambda)^2 \frac{\pi c}{\lambda^4} \left(\frac{1}{e^{hc/\lambda kT} - 1}\right) d\lambda$$

In HSV representation, the corresponding λ for maximum $f(T)$ determines the color which is related to the hue and saturation value. The intensity of $f(T)$ which is the V value corresponds to the total reflected phon flux. In the LWIR, the intensity of the response corresponds to the total amount energy emitted by the object. Scribner et al. [8] showed that there is a mild anti-correlation between LWIR and visible spectrum, but they are still compliment enough to provide useful information. Based on our coverage of the spectral, we can represent a "virtual eye" which can perceive the total received energy by defining a new V channel with the reflected and emitted energy at the same time. If both the lighting and temperature reach a equilibrium and independent, then the appearance in the new modified V channel should also be consistent. If we simply add V channel and the intensity channel together, when one of the observation fails to work the appearance would change significantly. Thus, we combine the intensity of the infrared channel and the V channel in color image with our embedded sensor confidence model.

With sensor confidence model, if the observations are not stable, we can still have a stable appearance model. As mentioned above, the intensity of v in the visible spectrum and the intensity in the LWIR is not fully compliment. Instead, they are non-linearly mild anti-correlated when their intensities are below certain values. As a result, we calculate a calibration curve to fix this problem so that we can add two channels together at any value. This calibration curve is learned from fitting a non-linear equation so that the correlation coefficient of the intensity in the visible spectrum and LWIR is close to zero. In our experiment, we apply a fourth-order polynomial to adjust the LWIR intensity. After normalizing the range of intensity in infrared image from 0 to 1 to be the same as V in visible spectrum, the new fused V is calculated by the following equation,

$$V_{mod} = \frac{C_{obj}(v)}{C_{obj}(v) + C_{obj}(i)} \times V_{vis} \times (1 - \cos(\theta))$$

$$+ \frac{C_{obj}(i)}{C_{obj}(v) + C_{obj}(i)} \times f(V_{inf})$$

Where $f(.)$ is the calibration function. θ is the angle between light source and the normal vector of camera viewing plane. This angle is estimated by the collective moving object shadows with the weak perspective camera model. The lighting correction based on assumptions include that there is only one global diffuse light source, the object can be described as an Lambertian reflector, and the polar angle ϕ is ignored. As a result, the observed reflected intensity is only proportional to the $\cos(\theta)$. The object surface reflectance is linearly interpolated and centered. θ is estimated by the maximum length vector of foreground object and its shadow if it exists. The initialization of the location of global light source is described as follows. First, we select a small number of training frames to provide initial information about the light source. Second, we select foreground objects manually and convert the reflectance of each foreground object by spherical harmonic transform. Third, we select the first six spherical harmonic transform coefficients of each foreground object as the reflectance vector. Finally, we select a global illumination source by minimizing the square root sum of euclidian distance of coefficients of each object reflectance vector. We then apply this initial light source information for following lighting compensation with a confined searching range for the new light source location. Figure 2 demonstrates the relationship among illumination source location, object reflectance surface, and the light incident path. As also can be seen in the figure, static shadows are not selected because of the absence of moving object in their neighborhood. This new v representation covers properties of intrinsic object reflectivity, extrinsic incident illumination, and object emissivity. Our channel confidence model also provide better weighting among them. This fused v channel is more robust to sudden illumination changes and blinking effects in some lighting conditions based on the combination of two spectral signal and the confidence model.

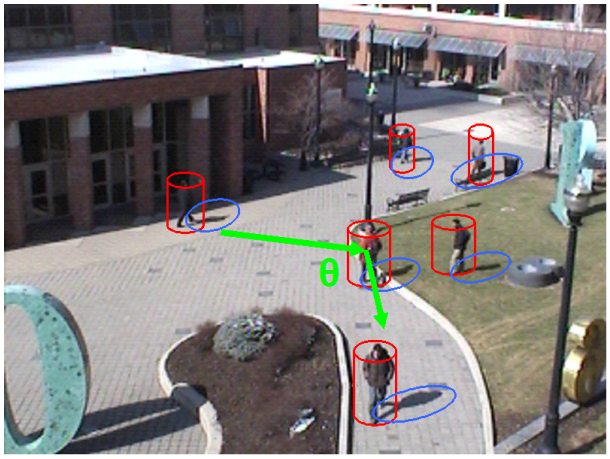

Figure 2: Illumination relationships among the light source, objects, and the camera (Red cylinders: object reflectance surfaces, Blue ellipses: moving shadows, Green lines: estimated incident light paths)

Isard and MacCormick [6] presented a particle filter with mixture-of-Gaussian distributions for color-based appearance tracking. It is suitable for multiple object tracking, and we use the same representation for our appearance tracking and updating mechanism. However, we add our new invariant V channel in the appearance model instead of traditional H and S channels only in HSV representation or other color representations in visible spectrum. There are several supports for using this method in this article, such as the same mixture-of-Gaussian background modeling, Gaussian distribution based confidence model, and thermal error patterns are usually modeled as Gaussian distributions as well. Here we use a 10-component mixture of Gaussian model in our fused signal domains for each response based on our preliminary k-means clustering ex-

periment. For comparison, we found that in order to represent the same foreground objects in the visible spectrum requires at least a 18-component mixture of Gaussian model in the normal condition. This again confirms that a signal-level of fusion is efficient.

Adaptive weighting is also updated by the relationship of confidence factors. Additionally, within the object blobs, we also increase the weighting at the edges of infrared channel, since for infrared channel, more information is preserved at the boundary positions. The non uniformed weighted is applied again with a 5 x 5 Gaussian mask so that pixels with larger distance from the center will have larger weighting coefficients.

4. OBJECT TRACKING UNDER OCCLUSION

If there is an obvious mismatch between two observations in two spectral channels, it possibly results from object occlusion. Typical occlusion detection works on detecting sudden or gradually disappear objects. A probabilistic framework of occlusion detection is given in Elgammal and Davis [5]. We applied the same estimation maximization approach as the main building block but we further extended it with our multi-spectral system setup. Since we do not presume consistent observations between two spectral channels, the occlusion detection can be more efficient and effective by our weighted fusion which embeds camera confidence models. First of all, we have to assume that before the object is occluded, it has been detected before so that we have registered its isolated appearance model.

There are three possible types of occlusions in our multi-spectral system, including, visible spectrum occlusion, infrared spectrum occlusion, and complete occlusion. For the first two types of occlusions, they are automatically resolved by our method since embedded camera confidence model can manage those partial signal occlusions. For the third type of occlusion, it can also be managed by maximizing a non-parametric model of our fused appearance with extra spatial and time information. We then define a likelihood function and apply an estimation maximization method to find the maximum likelihood of the appearance based on our weighted fused image histogram, spatial distribution, and object motion history. Since we use our fused histogram with lighting compensation, we are able to assume that within the blob if the same appearance would appear as the same regardless of its relative locations. This is based on assumptions that the global illumination and the thermal equilibrium do not change significantly during the occluded period. Whenever there is a new foreground object appearing in the view, we calculate the following expectation maximization similarity criterion for the reported possible lost objects to address the possible occlusion problem,

$$S_{obj}^{k} = arg_k \ max \sum_{i=1}^{n} logPr(O(i,t)|H(k,t))$$

Where n represents the total number of the histogram dimension, and k is the total number of object trackers which were observed before but lost. S_{obj}^{k} is the similarity between the newly registered foreground object i and the reported lost object k. $O(i,t)$ is the current observation of the newly found object i. $H(k,t)$ represents the kth tracker appearance histogram with time distribution adjustment at time t. $H(k,t)$ include two types of temporal related information, object tracked time and motion direction history. The object tracked time is a ratio calculated by dividing the tracked length of a lost object by its length of lost length. If the ratio is larger, we are more confident in linking that object to close foreground regions. For motion direction history, we record the object movement direction along with a exponential decaying weighting. This exponential decaying weighting will not only have larger weighting on latest object movement direction but also preserve consistent previous movement trait. This weighted history can provide the probability of the object motion direction if its movement is consistent, and also it can alleviate the problem for burst movement. These two temporal related criteria provide a lost tracker with possibilities to pick up the reappearing object even if the new foreground object is not spatially close enough or is with insignificant changes in appearance. However, if the object is occluded for a long time, the existing occluded tracker information may be disappear because of this exponential time decaying and lost time ratio. In this case, a new tracker will be created. Besides, if S_{obj} is below a small predefined threshold, the new foreground object is also considered as a new tracker without matching to any of the lost trackers.

5. MERGE AND SPLIT BY APPEARANCE-DISTANCE HISTOGRAM

The performance of multiple object merge and split can also be greatly improved by the correlation with two spectral channel observations. We would like to calculate the relationship between appearance and proximity by color and infrared intensity information at the same time. Typically, people used proximity of the object to define the merge and split. Here since we have more compliment observations, we should be able to make better decisions about the object merge and split. We characterized the distance between different appearance set by a appearance-distance joint histogram. At each distance, we have a corresponding histogram to describe the object appearances. At euclidian distance k, the histogram of appearance set i and j is defined as follows,

$$A^{i,j}(k) = \sum_{x} \sum_{y} P_{x,y}^{i,j}(k)$$

where $A^{i,j}(k)$ is the sum of color-thermal appearance pairs of set i and set j at pixel distance k. $P_{x,y}^{i,j}(k)$ is one when the set of the hue and the modified v value at (x,y) is the color-thermal set i and within euclidian distance k the set of the hue and the modified v value is the color-thermal set j. It is zero otherwise.

It is reasonable to assume that the appearance and the distance are independent between each object if there is no prior knowledge of the object. This appearance-distance histogram is a three dimensional histogram with three independent axes, hue, modified v, and euclidian pixel distance. If the distance is 1, the diagonal entries will be connected clusters with same appearances. We quantized the object hue color to twelve true color values and modified v value to ten levels since it is more efficient in storage size and less error sensitive compared to full hue and modified v scales. The number of distance is calculated based on the threshold used in morphological operations in connected components. The threshold used in connected component analysis determines the connectivity of objects in the object blob formation process so if we decide to split or merge objects we should use that factor in our distance calculation. We assume that when the object first observed, the appearance of the whole object is registered. This assumption means that only the previously observed object pixels would be taken into account. As a result, the new coming pixels for a gradually appearing object will only be put into the appearance histogram for the next frame. Then the appearance-distance histogram can be used to determine the degree of object merge and split.

When objects cross each other, only one dominant object can be observed. As a result, when they split to different direction, the appearance-distance histogram will have an increasing amount in

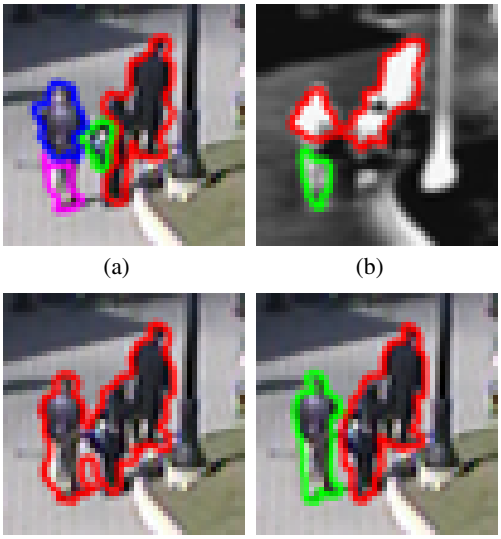

Figure 3: Merge/Split results in different spectrums. ((a): Object segments in the visible channel - incorrectly fractional. (b): Object segments in the infrared channel - incorrectly merged. (c): Object segments in the direct fusion - incorrectly merged. (d): Object segments in our proposed fusion - correctly split.)

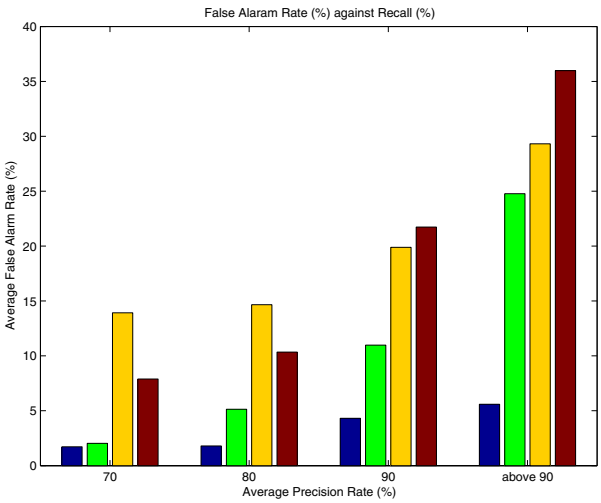

Figure 4: Average object recall and false alarm rate for different system requirements (From left to right of each bin cluster: Blue: proposed weighted fusion, Green: direct fusion, Yellow: infrared only, Red: visible only)

the off-diagonal entries at distance k and smaller than k, where k is the proximity threshold for connected component, since off-diagonal terms refer to different color-thermal pairs. The values of off-diagonal entries can help us determine which appearance is moving away or moving closer to other color-thermal appearance groups. As a result, we define the merge and split by,

$$\int_{k'<=k} \frac{\partial A^{i,j}(k')}{\partial t} dk' \quad > \theta \ \ or < \theta$$

If we assume only one part is split from the group blob or merge into the group blob each time, we can determine the split and the merge component by separate the largest component with mixed appearance entries from the original component at k. This regulation can be translated to a maximum volume cut with largest distance from the diagonal axis in the appearance-distance histogram which is also similar to maximizing the Bhattacharyya distance between two sets of appearance histograms, but our histograms are multi-dimensional. Every time, we split or merge one part of object blobs and recursively check the threshold for each newly generated blobs until there is no further cuts or combinations. Additionally, this cut or combination is non-rectangular and generates results comparable to Davis and Shama [4] with a much lower computation complexity but higher storage penalties. However, although objects are non-rectangular, we still use rectangular blobs for simplicity. As a result, possible overlapped regions will be observed. Among those overlapped regions, pixels are updated based on the larger corresponding blobs.

The color appearance can be different in different parts of the object in the visible spectrum, so they may not satisfy the appearance-distance criteria. However, if we assume that the object has a uniform temperature distribution if it emits, we can still observe a more uniform moving blob. With only infrared channel, it is almost impossible to distinguish human body temperature, so it needs other appearance descriptions. In either situation, we would still satisfy the overall criteria when both channels are available. This kind of extra object information can not be observed if only visible spectrum signal or infrared spectrum is used. By using this criteria with multi-spectral data, we can perform better in split and merge decisions as compared to traditional morphological operations in visible spectrum only. Figure 3 shows the effects of merge and split. As we can observe from the figure, objects are incorrectly split into pieces in visible spectrum due to different colors within the object. They are also incorrectly merged in both infrared only and direct combination method. For our appearance-distance criteria, those three parts can be separated into two parts even from the closely combined object in the first iteration.

6. EXPERIMENT RESULTS

We used IEEE OTCBVS WS Series Bench [4] to test our single multi-spectral sensor background modeling and appearance tracking. However, the benchmark does not have enough challenging scenes except for shadows, so we added several artificial errors into it. Linearly adjusted contrast infrared images and low intensity color images were created from the original sequences to simulate the effects of low temperature contrast and nighttime scenes during a 24-hour surveillance. We measured the dynamic range of image contrast from sample images which were taken during different lighting and temperature conditions. We used this dynamic range to modify the test bench. As a result, ratio 1 refers to the best contrast ratio, ratio 0.5 refers to contrast ratio linearly interpolated to one half of the full dynamic range, 0 refers to the minimum ratio, and so on.

We would like to first demonstrate that under normal conditions, using our proposed multi-spectral system can enhance the surveillance performance. False alarm is one of the most undesirable properties in video surveillance, so we use false alarm rate under the same recall rate as the performance measurement. In Figure 4, we compare the false alarm rate while changing different recall rates for each method, including, the visible spectrum only, infrared only, direct fusion, and our proposed fusion. Every time, we modify the thresholds to achieve the desired recall rate or above

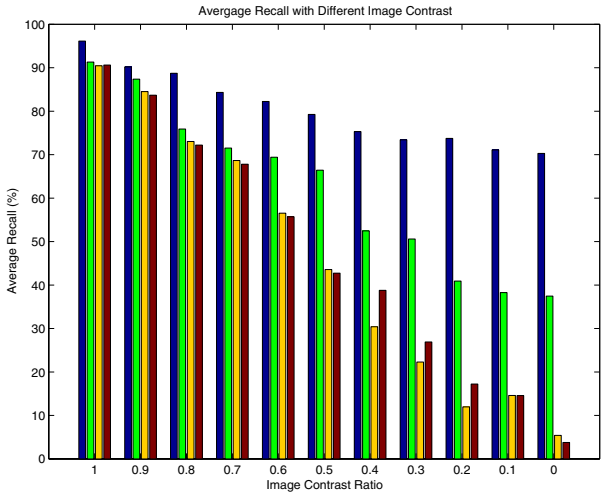

Figure 5: Average object recall of different image contrast ratios (From left to right of each bin cluster: Blue: proposed weighted fusion, Green: direct fusion, Yellow: infrared only, Red: visible only)

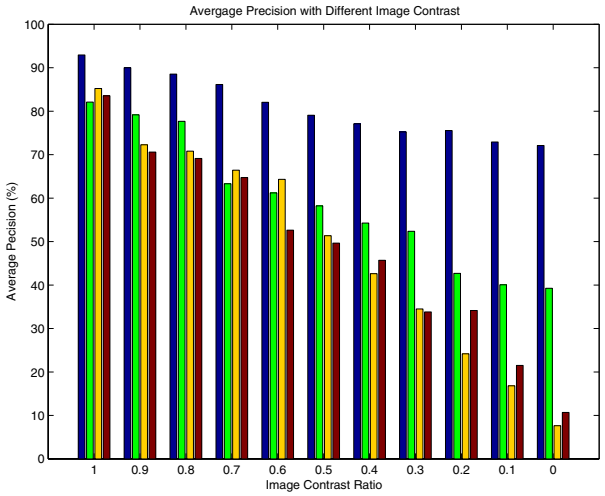

Figure 6: Average object precision of different image contrast ratios (From left to right of each bin cluster: Blue: proposed weighted fusion, Green: direct fusion, Yellow: infrared only, Red: visible only)

depending on the system characteristics. We can observe that while maintaining the same object recall rate, our system greatly outperformed other systems with greatly reduced false alarm rates. Moreover, as the figure shows, direct fusion of the visible spectrum and infrared spectrum improves the performance as compared to single spectrum, but it still generates high false alarms when the required precision rate is high. This observation clearly indicates that even under normal conditions, the observations in each spectral channel are not consistent and in turn limit the performance. As a result, a direct fusion technique is not applicable.

Secondly, we would like to show the advantages of multi-spectral system when the environmental conditions are poor. We used our linearly adjusted sequences to test this situation. For every experiment in the direct fusion and our weighted fusion, we only changed one signal channel while the other signal channel stays in 0.8 contrast and calculated the average of two. This is based on our observation that two channels will not have the best contrast case at the same time, nor does the worst contrast. In this experiment, we used *recall* and *precision* as our performance metrics since they were more illustrative in demonstrating different spectrum properties in different image contrasts. From Figure 5 and 6, we can observe that when the image contrast ratio close to the best condition, all three methods perform comparably in recall, but the visible only and infrared only already have fairly low precision rates. When it is slightly worse than the premium conditions, the performance of direct fusion and proposed weighted fusion start to perform much better than single spectral tracking. However, when the image contrast ratio keeps decreasing to lower than 0.5, our proposed fusion system shows a significant performance advantage while direct fusion performs even worse than as compared to the default 0.8 in the other spectrum. This indicates that when one of the spectrum fails to work, direct fusion will degrade the overall performance even if the other spectrum signal is still quite robust. Moreover, the performance of our proposed system stays relatively stable in most conditions. Especially when one of the signal channel has extremely low contrast, our system performs the same as one strong signal from the default and is not affected by the partial failure. This observation indicates the effectiveness of our camera signal confidence modeling.

Thirdly, we would like to show that using our proposed fusion method, the histogram matching stays consistently in different environment variations. This also indicates that under poor conditions, our system has a better chance to match the same objects without modifying the threshold. The average error of the histogram appearance matching is summarized in Figure 7. The x coordinate represents the contrast ratio for each spectrum. The y coordinate is the average sum of the histogram error. Here we only record matching errors with the closest matches. As we can observe from the figure, without any additional information from other spectral observations, the average matching error increases exponentially when the image quality gets worse. Also, the visible channel starts relatively well in good conditions as compared to the infrared channel but the performance worsens greatly faster than the infrared when the image contrast goes low. On the other hand, matching errors increase only linearly when two spectral channels are utilized in our proposed weighted fusion, and also the slope is flat. Moreover, direct fusion does not perform better than single spectrum only, and it also has an exponential increasing slope. This also shows that under different environment conditions, unweighted spectral fusion is not applicable.

Finally, we also inserted artificial occlusions in the sequences. We randomly selected five regions within the scene and placed a "black box" which blocks out the observations in that region artificially. The performance of object merge and split is calculated manually by three different subjects. The average performance of occlusion, merge and split are calculated based on 0.8 and 0.4 image contrast condition and is shown as Figure 8. From Figure 8, we can observe that with only single spectral sensor, performance for object occlusion, merge and split is much worse than direct signal combination and our confidence weighted method. The visible spectrum performs poorly in both merge and split because of the variations of object color appearances. The infrared spectrum performs relatively better in merge but fairly poor in split since there is not enough resolution for the split operations. Additionally, di-

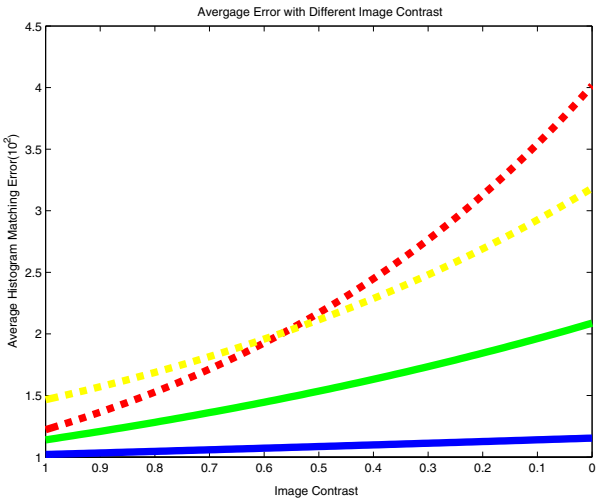

Figure 7: Average object histogram matching error (From top to bottom: Red-Dash: visible only, Yellow-Dash: infrared only, Green-Solid: direct fusion, Blue-Solid: proposed weighted fusion)

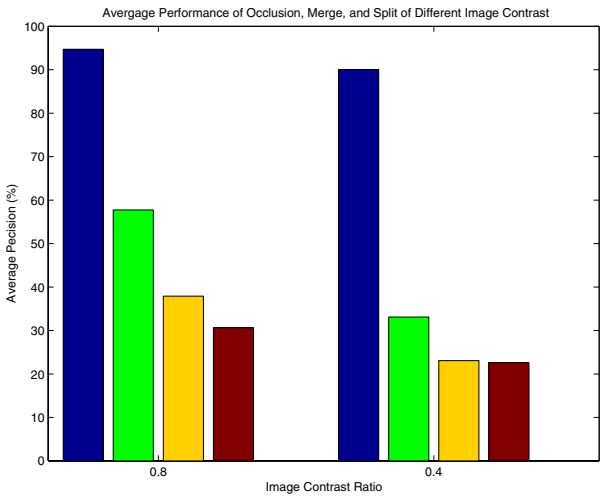

Figure 8: Object occlusion, merge/split performance (From left to right of each bin cluster: Blue: proposed weighted fusion, Green: direct fusion, Yellow: infrared only, Red: visible only)

rect fusion does not noticeably outperform single spectrum systems when the image contrast is low. On the other hand, our method significantly outperforms all the other methods when the image contrast ratio is low.

7. CONCLUSION

We present a robust background modeling and a self-adaptive appearance based tracking algorithm which efficiently and effectively take the advantage of multi-spectral system setups. We use a hierarchical probabilistic model to fuse and update the image information in different spectrums and in different processing levels. Moreover, object tracking with merge and split are shown to be better managed by our multi-spectral method. Object occlusion can also be better resolved and recovered later when occluded objects appear again. Quantitative analysis also shows that our proposed method outperforms traditional systems especially in the low contrast image environment which is inevitable in real-world surveillance.

In this work, Although we successfully improve the performance of background modeling and appearance-based tracking by applying two spectral signals, there are still limitations for single sensor setup in the fully unconstrained environment, such as permanent occlusions and appearance changes. Currently, we are working on collaborating multiple overlapped multi-spectral sensors. We hope to conquer more challenging real-world surveillance problems with a multi-spectral multi-sensory system. In that system setup, we are going to fuse the signal one level up from the frame level to the network level. Furthermore, LWIR is mainly advantageous for human tracking. For many other non thermal-emitting objects, a multi-band infrared fusion is more desirable. This will further increase the difficulties of signal fusion and need further research efforts.

8. REFERENCES

[1] C. Ó. Conarie, E. Cooke, N. O'Connor, N. Murphy, and A. Smeaton. Background modeling in infrared and visible spectrum video for people tracking. In *Proceedings of IEEE International Workshop on Object Tracking and Classification in and beyond Visible Spectrum*, 2005.

[2] C. Ó. Conarie, E. Cooke, N. O'Connor, N. Murphy, and A. Smeaton. Multispectral object segmentation and retrieval in surveillance video. In *Proceedings of IEEE International Conference on Image Processing*, 2006.

[3] R. Cucchiara, C. Grana, M. Pocardi, and A. Prati. Detecting moving objects, ghosts, and shadows in video streams. In *IEEE Transactions on Patern Analysis and Machine Intelligence*, 25(10):1337-1342, October, 2003.

[4] J. Davis and V. Sharma. Fusion-based background-subtraction using contour saliency. In *Proceedings of IEEE International Workshop on Object Tracking and Classification in and Beyond the Visible Spectrum*, 2005.

[5] A. M. Elgammal and L. S. Davis. Probabilistic framework for segmenting people under occlusion. In *Proceedings of IEEE International Conference on Computer Vision*, pages 145-152, 2001.

[6] M. Isard and J. MacCormick. Bramble: A bayesian multiple-blob tracker. In *Proceedings of IEEE International Conference on Computer Vision*, pages 34-41, 2001.

[7] O. Javed, K. Shaphique, and M. Shah. A hierarchical approach to robust background subtraction using color and gradient information. In *Proceedings of IEEE Internaional Workshop on Motion and Video Computing*, pages. 22-27, 2002.

[8] D. Scribner, P. Warren and J. Schuler. Extending color vision methods to bands beyond the visible. In *Journal of Machine Vision Application*, 11(6):306-312, 2000.

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

[10] H. Torresan, B. Turgeon, C. Ibarra-Castanedo, P. Hebert, and X. Maldague. Advanced surveillance systems: Combining video and thermal imagery for pedestrian detection. In *Proceedings of SPIE - The International Society for Optical Engineering , Thermosense XXVI*, pages 506-515, 2004.