Thermal-Visible Video Fusion for Moving Target Tracking and Pedestrian Classification

Alex Leykin
Indiana University
Department of Computer Science
Bloomington, IN
Email: oleykin@indiana.edu

Yang Ran
University of Maryland
Center for Automation Research
College Park, MD
Email: rany@cfar.umd.edu

Riad Hammoud
Delphi Corporation
Delphi Electronics and Safety
World Headquarters, IN, USA
Email: riad.hammoud@delphi.com

Abstract

The paper presents a fusion-tracker and pedestrian classifier for color and thermal cameras. The tracker builds a background model as a multi-modal distribution of colors and temperatures. It is constructed as a particle filter that makes a number of informed reversible transformations to sample the model probability space in order to maximize posterior probability of the scene model. Observation likelihoods of moving objects account their 3D locations with respect to the camera and occlusions by other tracked objects as well as static obstacles. After capturing the coordinates and dimensions of moving objects we apply a pedestrian classifier based on periodic gait analysis. To separate humans from other moving objects, such as cars, we detect, in human gait, a symmetrical double helical pattern, that can then be analyzed using the Frieze Group theory. The results of tracking on color and thermal sequences demonstrate that our algorithm is robust to illumination noise and performs well in the outdoor environments.

Keywords: Human Tracking, Thermal Imagery, Fusion of Color and Thermal Imagery

1. Introduction

The problem of automatic real-time pedestrian recognition has gained a lot of attention in the machine vision community and is as being identified as one of the key issues in numerous applications ranging from collision avoidance with pedestrians in the automotive world, through border surveillance, and situation awareness in autonomous vehicles and robotic systems [3, 21] through human activity recognition [13, 26]. Sensor fusion has become an increasingly important direction in computer vision and in particular human tracking systems in recent years. Human motion tracking based on the input from RGB camera already has been producing reliable results for the in-door scenes with the constant illumination and steady backgrounds. However, outdoor scenes with significant background clutter due to illumination changes, however, still appear to be challenging to handle using inputs from a conventional CCD camera. In our work we propose a method of utilizing an additional source of information - a thermal camera/sensor which produces for each pixel a gray scale mapping of the temperature at the corresponding location.

Periodic human gait analysis in surveillance videos has recently become one of the most active research areas in computer vision. The key issue is how to model the shapes resulting from individuals. In this paper, we address automatic detection of a periodic pattern generated by walk within a spatio-temporal image.

This paper aims at developing concepts from the intersection between group theory and computer animation into computer vision algorithms that can automatically analyze gait patterns in real images. Mathematically speaking, a symmetry of a subset $S$ of Euclidean space $\mathbb{R}^n$ is a rigid transformation in $\mathbb{R}^n$ that keeps $S$ set wise invariant. The set of all rigid transformations that are symmetries of a pattern has a group structure, and is called the symmetry group of the pattern. Experimentally speaking, human gait repeats itself during walking and hence belongs to periodic motion. The bipedal moving of two legs exhibits symmetry between
left and right limbs.

**Related work.** Substantial research has been accumulated in detection and tracking of people. The majority of the studies address tracking of isolated people in well controlled environment, but increasingly there is more attention to tracking specifically in *crowded environments* [4, 22, 13, 12, 14, 10]. Recently, a flurry of contributions on pedestrian localization and tracking in visible and infrared videos have appeared in the literature [11, 5, 24, 2, 32]. In [33], the P-tile method is developed to detect human head first and then human torso and legs are included by local search. Nanda [24] builds a probabilistic shape hierarchy to achieve efficient detection at different scales. In [27], a particle swarm optimization algorithm is proposed for human detection in infrared imagery. Dai et al. [5] proposed a hybrid (shape+appearance) algorithm for pedestrian detection, in which shape cue is first used to eliminate non-pedestrian moving objects and appearance cue is then used to pin down the location of pedestrians. A generalized Expectation Maximization algorithm has been employed by the authors to decompose infrared images into background and foreground layers. These approaches rely on the assumption that the person region has a much hotter appearance than the background. Davis et al. [7] proposed to fuse thermal and color sensors in a fusion-based background-subtraction framework using contour saliency map in urban settings. Information including object locations and contours from both synchronized sensors are fused together to extract the object silhouette. A higher performance is reported by fusing both sensors over visible-only and thermal-only imagery. This method is however computationally expensive as it attempts to construct a complete object contour, which does not seem a requirement in various applications like surveillance or crash-avoidance system. In [32], support vector machine and Kalman filtering are adopted for detection and tracking respectively.

We developed our generative tracking framework encouraged by recently found implementations for particle filtering. Random sampling was shown not only to successfully overcome singularities in articulated motion [8, 23], but the particle filtering approach applied to human tracking has also demonstrated potential in resolving ambiguities while dealing with crowded environments [15, 34]. Working under the Bayesian framework it has been shown that particle filters can efficiently infer both the number of objects and their parameters. Another advantage is that in dealing with distributions of mostly unknown nature, particle filters do not make Gaussian assumptions, unlike Kalman filters [16, 30].

One of the most routine actions humans perform is walk. The review by Gavrila [11] categorized human motion analysis work according to whether an explicit shape model was used and the dimensionality of the model space. Another recent work is by Liang in [31], where a hierarchical summary is given for research in related areas during 1997 to 2001. Many existing human motion analysis methods use contour based, stick figure or volumetric model in image domain that implicitly utilizes temporal information. Other methods, like the algorithm in this paper, explicitly applies a temporal model. Several solutions have been proposed for characterizing the temporal periodicity. And they could be divided into two major categories. One is to analyze periodic motion at shape or silhouette level and the other at pixel level. Little and Boyd analyzed the shape of motion and use it for real time target classification [19]. Adelson and Berger[1] presented a multiscale spatiotemporal filter bank for motion perception. In [1, 25], the motion is represented in the X-Y-t space convolving with specific impulse responses. In the second category, Yang [29] introduced video phase lock loop for perceiving the oscillations at pixel level. Liu and Picard [20] found periodicity by applying Fourier analysis along pixels’ trajectory. There is a major drawback in most of the algorithms in that they do not use knowledge on human kinematics. On the other hand methods based on complicated body models require tracking of body feature points or markers, which is unreliable in many cases. Because of the upright pose and translational global body displacement along ground surface during human walk, considering in spatio-temporal domain is more reasonable than other space. We also observe strong periodic pattern like those in texture or crystal. By bridging group theory and periodic gait pattern symmetries, we could obtain a much deeper understanding of these patterns.

**Contribution.** The goal of our tracking system is twofold: first attempt to employ all available information to achieve the noise free blob-map and second, subsequently use the blob-map to perform reliable pedestrian tracking to minimize two types of tracking errors - falsely detected people and people missed by the system. Our system segments foreground regions out of each frame by using a dynamically adapting background model. We hypothesize about the number of human bodies within such each region by using the head-candidate selection algorithm. As the next step, our system constructs a Bayesian inference model, based on the a priori knowledge of the human parameters and scene layout and geometry. Observations of the body appearances at each frame are a second driving force in our probabilistic scheme.

In this paper, we also introduce a new approach to characterize the signature generated by a walking human. To describe the computational model for this periodic helical pattern, we take the mathematical theory of symmetry groups, which is widely used in crystallographic structure research. Both observations and biometrics prove that spatio-temporal human walk patterns belong to the frieze groups because they are characterized by a repetitive motif
in the direction of walking. The structure is applied to a system capable of solving automatic detection-and-tracking, classification in both color and IR videos. Experimental results for videos acquired from both static and moving ground sensors are presented. Our algorithm demonstrates robustness to non-rigid object deformation as well as background clutter.

Despite these efforts, the challenge remains whether using stationary or moving imagery system. This is due to a number of key factors like lighting changes (shadow vs. sunny day, indoor/night vs. outdoor), cluttered backgrounds (trees, vehicles, animals), artificial appearances (clothing, portable objects), non-rigid kinematics of pedestrians, camera and object motions, depth and scale changes (child vs. adult), and low video resolution and image quality. This paper proposes a pedestrian detection and tracking approach that combines both thermal and visible information (see Figure 1) and subsequently models the motion in the scene using a Bayesian framework.

![Figure 1. Left: Thermal image of the scene Right: Color image of the same scene](image)

2. Tracking

2.1. Multi-modal Pixel Representation

Each pixel in the image is modeled as two dynamically growing vectors of codewords. For the RGB input a codeword is represented by: the average pixel $\bar{p}_{RGB}$ value and by the luminance range $I_{low}$ and $I_{hi}$ allowed for this particular codeword. If an incoming pixel is within the luminance range and the dot product of $p_{RGB}$ and $RGB$ of the codeword is less than a predefined threshold it is considered to belong to the background. For the thermal monochromatic input a codeword is represented by: intensity range $T_{low}$ and $T_{hi}$ occurring at the pixel location. Unlike for the color codewords the matching of the incoming pixels temperature $p_T \in [0, 255]$ is done by comparing the ratios of $p_T / T_{low}$ and $p_T / T_{hi}$ to the empirically set thresholds. This way we can hard limit the percentage of temperature change allowed to happen at each location. By observing several thermal sequences we have established that changes in cloud cover or shadows produced by other moving objects do not typically cause the temperature change of more than 10%. A more in-depth description of multi-modal background modelling can be found at [18, 17].

During the model acquisition stage the values are added to the background model at each new frame if there is no match found in the already existing vector. Otherwise the matching codeword is updated to account for the information from the new pixel. Empirically, we have established that there is seldom an overlap between the codewords. In the situation when this is the case, i.e. more than one match has been established for the new pixel, we merge the overlapping codewords. We assume that the background changes due to compression and illumination noise are of re-occurring nature. Therefore, at the end of training we clean up the values (“stale” codewords) that have not appeared for periods of time greater than some predefined percentage of frames in the learning stage as not belonging to the background. We keep in each codeword a so-called “maximum negative run-length (MNRL)” which is the longest interval during the period that the codeword has not occurred. One additional benefit of this modeling approach is that, given a significant learning period, it is not essential that the frames be free of moving foreground object. The background model can be learned on the fly and is helpful when tracking and model acquisition are done simultaneously.

2.2. Bayesian model: observations and states

We formulate the tracking problem as the maximization of posteriori probability of the Markov chain state. To implement Bayesian inference process efficiently we model our system as a Markov chain $M = \{x, z, x_0\}$ and employ a variant of Metropolis-Hastings particle filtering algorithm [9]. The state of the system at each frame is an aggregate of the state of each body $x_i = \{b_1, \ldots, b_n\}$. Each body, in order, is parametrically characterized as $b_i = \{x, y, h, w, c\}$, where $x, y$ are coordinates of the body on the floor map, $h, w$ its width and height measured in centimeters and $c$ is a 2D color histogram, represented as 32 by 32 bins in hue-saturation space. The body is modeled by the ellipsoid with the axes $h$ and $w$. An additional implicit variable of the model state is the number of tracked bodies $n$.

2.3. Computing Posterior Probability

The goal of our tracking system is to find the candidate state $x'$ (a set of bodies along with their parameters) which, given the last known state $x$, will best fit the current observation $z$. Therefore, at each frame we aim to maximize the posterior probability

$$P(x'|z, x) = P(z|x') \cdot P(x'|x)$$ (1)

According to Bayes rule and given (1) we formulate our goal as finding:

$$x' = \arg \max_{x'} \{P(z|x') \cdot P(x'|x)\}$$ (2)
The right hand side of equation (2) is comprised of the observation likelihood and the state prior probability. They are computed as joint likelihoods for all bodies present in the scene as described below.

2.3.1 Priors

In creating a probabilistic model of a body we considered three types of prior probabilities. The first type of priors imposes physical constraints on the body parameters. Namely, body width and height are weighted \( N(\mu_x, \sigma_x) \) and \( N(\mu_y, \sigma_y) \), with the corresponding means and variances reflecting the dimensions of a normal human body. Body coordinates \( x, y \) are weighted uniformly within the rectangular region \( R \) of the floor map. Since we track bodies which are partially out of the image boundaries \( R \) slightly exceeds the size of what corresponds to the visible part of the image to account for such cases.

The second type of priors sets the dependency between the candidate state at time \( t \) and the accepted state at time \( t-1 \). Firstly, the difference between \( w_t, h_t \) and \( w_{t-1}, h_{t-1} \) lowers the prior probability. As another factor, we use the distance between proposed body position \( (x_t, y_t) \) and \( (\hat{x}_{t-1}, \hat{y}_{t-1}) \) — the prediction from the constant velocity Kalman filter. The state of Kalman filter consists of the location of the body on the floor and its velocity. Although tracking the head seems like a first reasonable solution, we have established empirically that the perceived human body height varies as a result of walking, thus the position of the feet on the floor was chosen as a more stable reference point.

The third type of priors are physical constraints with respect to other moving and static objects in the scene. First, to avoid spatial overlap between adjacent bodies (as physically improbable) we have imposed penalties on pairs of pedestrian models located closer that their corresponding body widths would allow. Second, a similar constraint was imposed on the overlap between pedestrians and stationary obstacles, which were manually marked in the frame and converted in to 3D world coordinates.

When new body is created it does not have a correspondence, this is when we use a normally distributed prior \( N(d_0, \sigma) \), where \( d_0 \) is the location of the closest door (designated on the floor plan) and \( \sigma \) is chosen empirically to account for image noise. The same process is taking place when one of the existing bodies is being deleted.

2.3.2 Likelihoods

The second component in forming proposal probability relates the observation to the model state. First, for each existing body model the color histogram \( c \) is formed by the process of weighted accumulation, with more recent realizations of \( c \) given more weight. We then compute Bhattacharyya distance between proposed \( c_t' \) and corresponding \( c_{t-1} \) as part of the observation likelihood.

\[
P_{\text{color}} = 1 - w_{\text{color}} \ast (1 - B(c'_t, c_{t-1})),
\]

where \( w_{\text{color}} \) is an importance weight of the color matching.

To guide the tracking process by the background map at hand, we use two more components while computing model likelihood: we define the amount of blob pixels not matching any body pixels as \( P^+ \) and the amount of body pixels not matching blob pixels \( P^- \) (see eq. 4,5). Note that we use a Z-buffer \( Z \) for these as well as for computing the color histogram of the current observation in order to detect occlusions. In this buffer all the body pixels are marked according to their distance from the camera (i.e. 0=background, \( \approx \) 200=first body, \( \approx \) 200=next closest body, etc.), which we obtain during the calibration process. This way only visible pixels are considered when computing the likelihood (see Figure 2). The Z-buffer is updated after each transition to reflect the new occlusion map.

In computing the likelihood as outlined above, there is one major shortcoming overlooked in previous works [15, 34]. If the computation is done in terms of the amounts of image pixels it causes the bodies closer to the camera influence the overall configuration much more, and the bodies further away are being mostly neglected. This becomes particularly evident when the camera covers a large area, where pedestrian image presentations can vary from under 20 pixels of overall area in the back of the scene to more than 200 in front. In addition, such neglect makes the system absolutely tied to the current scene configuration and not portable to a different camera model.

To avoid these shortcomings we have utilized a so-called “distance weight plane” \( D \) which is the image of the same dimensions as the input frame and \( D_{xy} = |P_{XZ}, C_{YXZ}| \), where \( | \) — is the Euclidean distance, \( C_{XYZ} \) — camera world coordinates and \( P_{XZ} \) — world coordinates of the hypothetical point in space located at a height \( z = \frac{d}{2} \) and corresponding to the image coordinates \( (x, y) \). The map produced in this manner is a rough assessment of the actual size to image size ratio (see Figure 3).

To summarize, the implementation of z-buffer and distance weight plane allows to compute multiple-body con-
We first draw a new proposal state \( x' \) with probability \( m_t(x'|x) \) and then accept it with the probability \( \alpha(x, x') \). Notice that the proposal distribution is a time function, that is at each frame it will be formed based on the rules outlined below.

To form the proposal distribution we have implemented a number of reversible operators. There are two types of jump transitions and five types of diffuse transitions implemented in our system: Adding a body, Deleting a body, Recovering a recently deleted body, Changing body dimensions, Changing body position, Moving a body, Resizing a body. Notice that we use a set of controllable weight probabilities to add more emphasis to one or another transition type. In our application normally around 100 jump-diffuse iterations are required for each frame to reach convergence. Refer to [17] for jump-diffusion dynamics implementation specifics.

### 3. Symmetry based Pedestrian Classification

Because of the upright pose and translational global body displacement along ground surface during human walk, considering in spatio-temporal domain is more reasonable than other space. We also observe strong periodic pattern like those in texture or crystal. By bridging group theory and periodic gait pattern symmetries, we could obtain a much deeper understanding of these patterns.

#### 3.1. Symmetry Analysis

Symmetry is a fundamental concept for understanding repetitive patterns in art decoration, crystallography etc. This has been a primary motivation for developing the branch of mathematics known as Geometric Group Theory. A geometric figure is said to be symmetric if there exist isometries that permute its parts while leaving the object as a whole unchanged. An isometry of this kind is called a symmetry. The symmetries of an object form a group called the symmetry group of the object. A symmetrical group spanning in 1D is defined as a Frieze Group and is defined as a Wallpaper Group in 2D space. Because the human walking motion generates translation along planes parallel to the direction of global body translation, we are more interested in planar symmetries such as reflections and half turn. There are seven distinct subgroups (up to scaling) in the discrete Frieze group generated by translation, reflection (along the same axis or a vertical line) and a half turn (180 rotation). Therefore we use the study of symmetry as pioneered in Frieze Group theory to analyze and extract human motion based signatures regarding questions such as whether the individual is carrying a load or not.

As a human walks the swing of the limbs generate a nice and symmetrical double helical pattern that can then be analyzed using the Frieze Group theory. In order to capture and analyze the double helical gait signature of an individual, we first track a moving individual by putting a bounding box around the person being tracked. Then, we stack all the bounding boxes of all the frames to create a x-y-t volume in which the double helical signature resides. Note that the double helical signature actually resides in the x-y-t volume. Therefore, we obtain the x-t slice of this volume. Figure 4 shows an x-t slice of the volume.
double helical gait signature is clearly seen in this slice. In order to segment and extract the helix from the background we divide the helix into four quadrants and fit 1-D curve models for each of the quadrant separately. These model parameters are then refined by incorporating consistency and continuity constraints.

3.2. Double Helical Signature

In a symmetrical twin pendulum model describing hip-to-toe motion, each leg is modeled as a hand of the twin-pendulum with equal length and a uniform angular swing speed $\theta$ and period $T$. The generated gait signature are shown in the second row. The slices generated by the 'legs' of that model do contain twisted DNA-like patterns. When investigating the relationship between those twisted DNA structure, we realize that they contain different symmetries such as reflection symmetry in horizontal and vertical axis, $180^\circ$ rotation, which could be proved as follows:

Hence any gait pattern $S_z$ at height $z$ is represented as:

$$S_z(t) = \{P_1, P_2\} = \{(z/\sin \theta \cos \theta(t), t), (z/\sin \theta \cos \theta'(t), t)\}$$

(7)

The pedestrian monitoring system we design is capable of detecting and classifying humans. Most of the current approaches are based on x-y domain (frames) and depend on silhouette (hence assume static camera and high contrast edges). In this paper, we evoke the approach based on DHS (Double Helical Signature) reported in [28], which brings us a real time solution for both static and moving platforms.

4. Results and Discussion

4.1. Tracking

For testing and validation purposes we used thermal and color dataset from OTCBVS [6]. The set contains short outdoor pedestrian sequences in two locations. Each scene is filmed both with a RGB and thermal camera at the identical resolution, providing thus a pixel to pixel correspondence between two types of sensors. The operation of our color-thermal background model significantly reduces and in most cases fully eliminates two types of false foreground regions: (1) shadows as the result of a moving cloud cover (2) shadows cast by moving pedestrians.

We performed preliminary evaluation of our tracking system for the presence of three major types of inconsistencies: misses, false hits and identity switches. A miss is when the body is not detected or detected but tracked for an insignificant portion of its path ($< 30\%$). A false hit is when a new body is created where there is no actual person present. Most of the false hits are a result of more than one body in the model being assigned to a single body in the scene. An identity switch is when two or more bodies exchange their IDs once within the close proximity from each other. By counting the number of each of types of errors on a number of sequences of overall 6000 frames we have obtained results summarized in table 1.

<table>
<thead>
<tr>
<th>Seq</th>
<th>Ppl</th>
<th>$P^-$</th>
<th>$P^+$</th>
<th>$P^{+/-}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ALL</td>
<td>48</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>%</td>
<td>100</td>
<td>12.5</td>
<td>4.1</td>
<td>10.4</td>
</tr>
</tbody>
</table>

Table 1. Tracking results based on the manually observed ground truth (** - two infants, below the tracked height limit, lead tracker to some confusion; ** - 2 pedestrian covered by trees not counted). $P^-$ indicates missed people, $P^+$ indicates falsely detected people (primarily due to shading artifacts) and $P^{+/-}$ indicates two pedestrian IDs swapping
The most common mistakes made by the tracker, were false hits. We have observed that the majority of false hits (more than 50%) are short lived, i.e. typically last for only several frames. False hits are typically created when the system recognizes a blob as containing more people than the actual number present in the scene. This can be explained by the distortion in shape of the blob caused by pixel level noise. These cases can be further post-processed by temporal filtering to remove insignificantly short paths. Sometimes, however, false detections are accompanied by ID switches, when the body tracked for a long time is substituted for a false hit. This presents a more complicated case and deserves further study.

Misses typically occur due to partial or total occlusions by the scene objects or due the clothing color being too close to the background. The first type of misses is usually promptly recovered by the tracker, but if the recovery took place in a different location, a new id is assigned, that way resulting in a “switch”.

Overall performance of the tracker shows space for improvement, it produces satisfactory detection and prolonged tracking in the crowded scenes, although the ratio of pedestrian id switches has to be reduced. As it becomes apparent the complexity of the scene i.e. the number of pedestrians decreases the performance of the tracker.

4.2. Classification

Since the human motion is naturally different than other types of motions, the sequence of bounding boxes can be further analyzed to verify whether or not the tracked subject is indeed a human. We use the human motion analyzer proposed by Ran et al. [28]. The algorithm proposed tests the spatio-temporal pixels in order to prove or disprove the null hypothesis that the signal being tested is periodic. Briefly, for a given pixel location (x, y) at some slices in the bounding box, the algorithm computes the periodogram of the color value of (x, y) across some slices of the subject. A peak in the periodogram proves that the spatio-temporal signal is periodic at that pixel location. On the other hand, a flat periodogram means that the spatio-temporal is a white noise. The periodicity test is repeated for all pixel locations in the slices within the bounding box. If the histogram of periodic pixels has a peak which is higher than a certain threshold, then the subject undergoing the test is a human. Otherwise, the null hypothesis that the subject is a human is rejected. For more details on the algorithm, the reader is referred to [28]. The results are shown in Fig. 8.
References


