Pedestrian Recognition in Aerial Video Using Saliency and Multi-Features

Chunping Liu, Xu Fang, Xingbao Wang, and Shengrong Gong

***Abstract*—Pedestrian recognition in aerial video is a challenge problem for the problem of low resolution, camera movement and target's blurred detail in aerial video. This paper proposes weighted region matching algorithm with Kalman filter, Multi-features fusion model and saliency segmentation (KMFS-WRM) to detect and recognize pedestrian. The KMFS-WRM algorithm first uses Kalman filter algorithm to mark candidate's region, which can avoid the problem of selecting candidates under supervision. Then we proposed the fusion algorithm of multi-feature, including HOG, LBP and SIFT features, namely HLS model to detect the pedestrian in aerial video. Our proposed detection method is robust for whether the camera is moving. And instructing human percept and concept, we segment the pedestrians in marked region using Context-Aware saliency detection algorithm that proposed by Goferman et al. and revised the segmentation results by HST model (Head Shoulder and Torso) and AAM model (Active Appearance Model) to obtain the candidates set. Last the matching of voter and candidates set using weighted region matching algorithm. Experimental results in complex aerial video demonstrated that our KMFS-WRM algorithm not only cuts down calculated complexity, but also improves adaptive and real-time ability. Moreover proposed method outperforms recent state-of-the-art methods.**

***Index Terms*—Pedestrian detection, saliency detection, Kalman filter algorithm, weight region matching.**

I. Introduction

Pedestrian detection and recognition is an essential and significant task in any intelligent video surveillance system, since it provides the fundamental semantic information of video understanding. Along with the urgent demand for video surveillance systems in many security-sensitive occasions, e.g., some large squares, stadiums, super-markets *et al*., the research on non-contact pedestrian detection and recognition draws more attention. Most existing methods of pedestrian detection and recognition mainly depend on the extraction of features based on image details [1]-[4]. As one of main information source for non-contact pedestrian detection and recognition, aerial video has seen widely used in both military and commercial world application where its advantages over traditional video outweigh its disadvantages, such as poorer spatial resolution, moving video camera, targets with much smaller and more obscure details, complex background and high density pedestrian flow. Therefore, interpretation of aerial video for scene understanding and object tracking become an important problem domain.

Existing much works in aerial video analysis have been done on identifying single objects, e.g., detecting and tracking moving objects and specific objects detecting, such as rooftops [5], cars [6], [7] or roads [8], building [9], [10], airport, bridge etc. However, these models are often designed for rigid objects where appearance constraints are fairly constant between instances of the object. There are difficulties for pedestrian recognition in aerial video. In [11], weighted region matching (WRM) algorithm is proposed to solve the problem of human identity recognition over a set of unordered low quality aerial images. But the algorithm has some defect: 1) the pedestrian's segmentation with a lot of noise: because to identify pedestrian is marked in a rectangular box, if the target contains numerous noises, resulting in a large number of noise regions in matching stage, the accuracy of recognition will be reduced; 2) a fixed number of candidate: selecting a fixed number of candidates by supervised learning in [11], the WRM algorithm can only be used in static images. However, in the actual video surveillance, such as square, subway, train stations and other places with crowd pedestrian flow, the situation of more or less than the fixed number of candidate often occur, obviously, this hypothesis will greatly reduce self-adaptability and robustness. Namely the number of candidates has been not automatically confirmed in terms of scene content.

To solve the above problem, this paper presents weighted region matching algorithm based on Kalman filter, Multi-features fusion model and saliency detection (KMFS-WRM). In the pedestrian segmentation stage, generally, the traditional image segmentation algorithms only rely on color and appearance, which are changing with environment, resulting in the poor precision of pedestrian segmentation. Psychophysical and physiological evidence indicates that human have the ability to fixate on visual salient area of an image while scanning it with a very quick speed [12]. We wish to find out the visual salient regions in the aerial video. So this paper adds human visual perception and concept to segment target using context-aware saliency detection [13]. Most images' foreground is more eyes catching, which can treat as salient points. We convert gray-scale image to binary image which consists of salient points, and get good target. For the poor adaptability and computational complexity, we use Kalman filter algorithm [14] to mark candidate region. Based on the current location of the target region, it can predict the location of the target

Manuscript received December 19, 2013; revised May 5, 2014. This work was supported in part National Science Foundation of China (NSFC Grant No. 61272258, 61170124, 61170020, 61301299).

Chunping Liu, Xu Fang, and Shengrong Gong are with School of Computer Science and Technology, Soochow University, Suzhou 215006, China (e-mail: cpliu@suda.edu.cn, fangxu_8040595@126.com, shrgong@suda.edu.cn).

Xingbao Wang was with School of Computer science and technology, Soochow University, Suzhou 215006, China. He is now with Anhui USTCiflytek Co., LTD (e-mail: wangxingbao@163.com).

region in the next frame and mark the region by rectangular box. Then we use proposed HLS model to detect pedestrian, and segment the pedestrian by saliency detection. And last used WRM algorithm [11] to select target in the marked region. Fig. 1 shows the system framework. Through a lot of the complex background and large pedestrian flow aerial videos, such as grassland, football field, square *et al*., it can demonstrate that the algorithm outperforms other methods in accuracy, timeliness and adaptability.

The rest of this paper is organized as follows. Section II briefly introduces the original WRM mehtod. Section III represents our proposed system framework of pedestrian recognition in aerial video using KMFS-WRM. Experimental results are provided in Section IV. Finally, Section V concludes the paper.

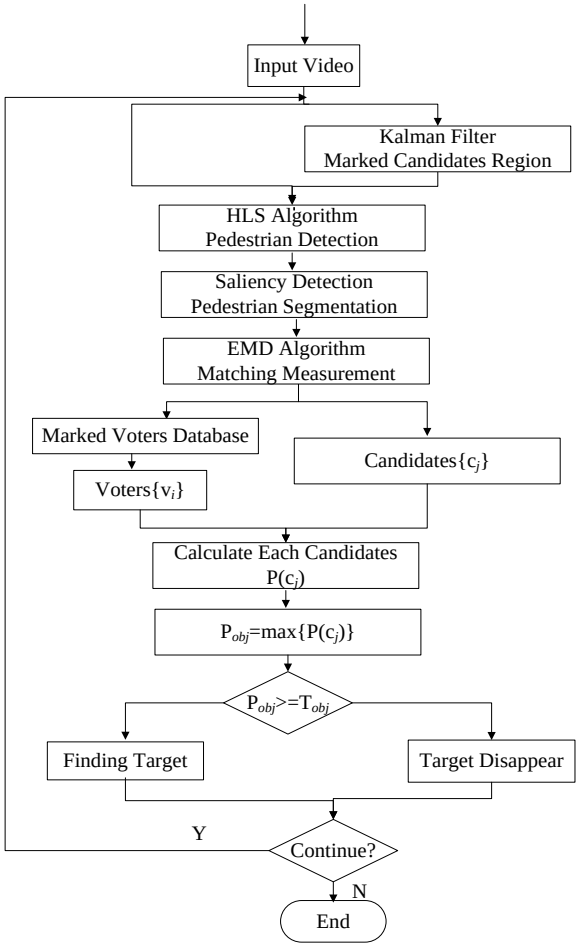


Fig. 1. The proposed system framework.

II. Background

In the subsection, we briefly introduce related WRM algorithm.

WRM algorithm is the novel solution of Pedestrian detection and recognition over a set of unordered low quality aerial images [11]. The main idea of WRM algorithm is the voter-candidate race. The proposed matching operation expects as an input a few images where the target has been recognized. When the input images and each query image undergo the following preprocessing steps: 1) human detection; 2) blob extraction; and 3) alignment, we can have two sets of blobs, the voters and the candidates. The training pedestrians' set $\{v_i\}$ is described as voters, and pedestrians' set $\{c_j\}$ appearing to be detected in video is described as candidates. Eventually, the voters vote for each candidate, and then the candidate with the most votes is the target.

Given the set of the voters $V=\{v_i, i=1,\cdots,n\}$ and the set of the candidates $C=\{c_j, j=1,\cdots,m\}$, then the probability of blob c_j corresponding to the target is denoted by:

$$P_T(c_j)=\sum_{i=1}^{n} P(c_j|v_i)P(v_i) \tag{1}$$

where $P(v_i)$ is the prior of voter. $P(c_j|v_i)$ can transform matching measurement into $D(c_j,v_i)$ between voters and candidates. If w_i is a weight assigned to voter i, $P(v_i)$ is equal to w_i and $D \in [0,1]$ is the normalized distance between c_j and v_i, then formulas (1) can be rewritten in a similar to the form of mixture Gaussians model:

$$P_T(c_j) \propto \sum_{i=1}^{n} (\exp(-D(c_j,v_i)/\tau)) \times w_i \tag{2}$$

where τ is a constant parameter. In order to solve the recognition problem efficiently, we need to provide a robust representation of the distance between every voter-candidate pair. Moreover, we need to specify the weight of every voter according to its importance in representing the target's specific information.

For the robust representation of the distance, the WRM method use Earth Mover Distance algorithm (EMD) [15] describes the problem of matching measurement. So we can convert matching problem into two sets mapping problem. For each region, we apply comentropy *JD* and Euclidean distance (*ED*) between two centroids of regions to describe EMD. At last, we combine *JD* and *ED* to obtain final EMD value by the formulas (3)-(5).

$$JD(V,C)=\sum_i [p_i \log_2(\frac{p_i}{(p_i+q_i)/2}) + q_i \log_2(\frac{q_i}{(p_i+q_i)/2})] \tag{3}$$

$$ED(R_V,R_C)=\sqrt{(R_{V,x}-R_{C,x})^2+(R_{V,y}-R_{C,y})^2} \tag{4}$$

$$D(V,C)=JD(V,C)+\zeta ED(R_V,R_C) \tag{5}$$

The weight of each voter in WRM algorithm is assigned by Page Rank algorithm (PR) [16]. The process of determining the voter's weight is as follows:

1) Calculating distances between each region and other regions in undirected graph G=(R,E) by equation (4), where the regions' set R represents the nodes of graph, and *E* is the set of edges connecting each pair of regions.
2) Finding nearest *K* regions between v_i and v_j($i \neq j$) by K-nearest neighbor algorithm. *K*=4 in our experiment.
3) Computing the weight of each region w_k^{pr} by PR algorithm.
4) Computing the normalized weight w_i of each voter's v_i is calculated.

$$w_i = \frac{\sum_{k=1}^{s} w_k}{\sum_{j=1}^{n} w_j} \tag{6}$$

where final region's weight $w_k = w_k^{pr} \times w_k^s$, w_k^s is region's area. For further refinement and scene learning, matched candidates with high confidence can be added to the voters set V in order to capture more information of the target. The details referred to [11].

III. Proposed System Framework of Pedestrian Recognition

Our proposed the framework of pedestrian recognition is different from proposed system in [11]. The differences lie in the preprocessing stage. Firstly candidates region is selected using Klaman filter [14] in the matching stage. Secnodly we adopt fusing multi-feature, including HOG feature [17], LBP feature [18] and SIFT feature [19] to detect and extract human blobs, whereas the WRM method only uses static image information-HOG feature. Thirdly we segment the pedestrian using context-aware saliency detection algorithm [13] to relieve the effect of changing shape, resolution and appearance feature. Last we recognize the pedestrian by the WRM method [11]. These aforementioned steps are explained in the following subsections.

A. Marked Candidate Region in Klaman Filter

The Kalman filter is a tool that can estimate the variables of a wide range of processes. The Kalman filter not only works well in practice, but it is theoretically attractive because it can be shown that of all possible filters, it is the one that minimizes the variance of the estimation error. The Kalman filters are based on linear dynamic systems discretized in the time domain. They are modeled on a Markov chain built on linear operators perturbed by Gaussian noise. In order to effectively improved adaptability of pedestrian detection and reduced computation complexity, we adopt Kalman filter algorithm [14] to estimate the candidate region. Therefore we must model the process in accordance with the framework of the Kalman filter. This means specifying the following matrices: F_k is the state transition matrix which is applied to the previous state x_{k-1}, H_k is the observation matrix which maps the true state space into the observed space. At time k an observation (or measurement) z_k of the true state $\boldsymbol{x}_k$ is made according to

$$z_k = H_k x_k + v_k \tag{7}$$

where v_k is the observation noise which is assumed to be zero mean Gaussian white noise with covariance R_k. The true state at time k is evolved from the state at $(k-1)$ according to

$$x_k = F_k\, x_{k-1} + B_k\, u_k + w_k \tag{8}$$

where B_k is the control-input matrix which is applied to the control vector u_k; w_k is the process noise which is assumed to be drawn from a zero mean multivariate normal distribution with covariance Q_k.

Each moving object (here is pedestrian) is described by its centroid and rectangular window. For finding the candidate region, Kalman filter is consists of prediction and update stage. Parameter including state vector x_k, measurement vector z_k, state transition matrix F, measurement matrix H, in prediction stage is set as follows

$$x_k = [x_{0,k}, y_{0,k}, l_k, h_k, v_{x,k}, v_{y,k}, v_{l,k}, v_{h,k}]^T \tag{9}$$

$$z_k = [x_{0,k}, y_{0,k}, l_k, h_k]^T \tag{10}$$

$$F = \begin{bmatrix} 1 & 0 & 0 & 0 & \Delta t & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & \Delta t & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & \Delta t & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \Delta t \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{11}$$

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{12}$$

where the vector x_k contains all of the information about the present state, which is an 8 dimensional state vector. But we cannot measure x_k directly. Instead we measure z_k, which is a function of x_k that is corrupted by the noise. $(x_{0,k}, y_{0,k})$ is centroid coordinate of marked rectangular region. l_k and h_k are the width and height of marked rectangle at time k. $v_{x,k}, v_{y,k}, v_{l,k}, v_{h,k}$ represents the change speed of x and y direction of the centroid coordinate, and the height and the width of marked rectangle respectively.

The horizontal, vertical centroid coordinates and area of the i^{th} object in the k^{th} frame are respectively described as x^i, y^i and S_k^i, x_{k+1}^j, y_{k+1}^j and S_{k+1}^j represent the horizontal, vertical centroid coordinates and area of the j^{th} object in the $k+1^{th}$ frame respectively. Then we can get centroid matching degree $D(i,j)$ and area matching degree $A(i,j)$ as in (13)-(14).

$$D(i,j) = \frac{|\sqrt{(x_k^i - x_{k+1}^j)^2 + (y_k^i - y_{k+1}^j)^2}\,|}{Max_n\,|\sqrt{(x_k^i - x_{k+1}^n)^2 + (y_k^i - y_{k+1}^n)^2}\,|} \tag{13}$$

$$A(i,j) = \frac{|S_k^i - S_{k+1}^j|}{Max_n\,|S_k^i - S_{k+1}^n|} \tag{14}$$

So we denote cost function $C(i,j)$. The selected target region is candidate region looking for with marking out rectangular box, as in (15).

$$C(i,j) = \eta D(i,j) + \lambda A(i,j) \tag{15}$$

where $\eta + \lambda = 1$, we set $\eta = 0.3, \lambda = 0.7$ through a large number of experiments.

Fig. 2 illustrates the marked candidate region in the aerial video of different outdoor scene, e.g. square and grassland. The region of target pedestrian is accurately marked at different time step. This greatly decreases the subsequent

detection and segmentation of the candidate pedestrian for avoiding the search and segmentation of pedestrian in whole frame.

B. Detection Pedestrian Using Multi-Feature (HLS) Model and Multi-Scale Detection

We propose pedestrian detection by fusing multi-feature HLS model (Fig. 3). The HLS model not only keeps HOG feature, but also supplements LBP and SIFT feature for aerial video and camera movement problem. The HOG descriptor captures the most important cues of human body, such as head and shoulders in good detail. Moreover the HOG descriptor is less sensitive to light changes and small offset, which has robustness and real-time. The LBP descriptor inhibits the uneven light and eliminates shadow. The SIFT descriptor keeps invariant to rotation, scale and perspective changes. Therefore, fusing three features is a good choice for aerial scenes.

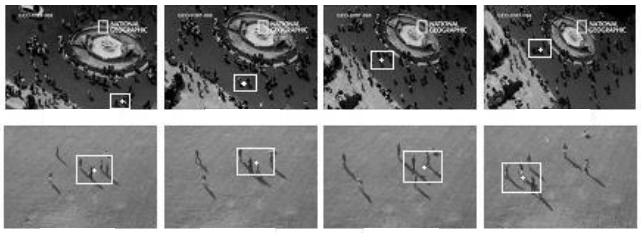


Fig. 2. Marked candidate region using Kalman filter (the white rectangle box is marked region, the white cross is the central of marked region, first and second row is square and grassland scene respectively).

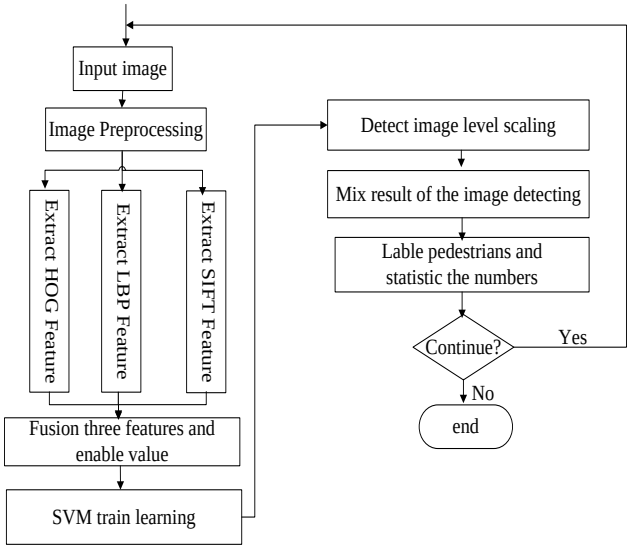


Fig. 3. Pedestrian detection using HLS model and multi-scale detection.

Different feature have different contribution to detection human body. We extract the local HOG, LBP and SIFT features of image based on certain size of block and cell. The key point of SIFT is different in terms of different size of detection window and different image. The key point is chose by Online Kernel Principal Component Analysis [20]. For example, given the size of detection window 64×128, the dimension of HOG, LBP and SIFT feature is 3780, 1856 and 800 respectively. We take a weighted mechanism based on the proportion of each feature in the actual detection process to combine the contribution of different feature. So the proposed HLS model is represented as follows:

$$f_{HLS}(w,h)=\alpha f_{HOG}(w,h)+\beta f_{LBP}(w,h)+\gamma f_{SIFT}(w,h) \quad (16)$$

where $a+\beta+\gamma=1$, w and h represent the width and height of detection image.

For detection, we train a support vector machine (SVM) classifier based on the HLS model. Furthermore we make use of the combining human detection result of different scale to mark the pedestrian in the original frame.

Given the frame, the multi-scale pyramid is built by automatic defining the number of layer S_n:

$$S_n = floor(\frac{\log(S_e / S_s)}{\log S_r})+1 \quad (17)$$

where $S_s=1$ denotes the starting scale, S_r is the rate of scale, S_e denotes the end scale of Pyramid.

$$S_e = \min\{W_{org} / W_{win}, H_{org} / H_{win}\} \quad (18)$$

where W_{win} and H_{win} respectively represent the width and height of detective window, W_{org} and H_{org} respectively represent the width and height of current frame. Then we take bilinear interpolation to adjust the size of each layer image of pyramid such that each layer image size keeps with the original input frame. The initial detection results are achieved by the learned SVM classifier on basis of the extraction corresponding features in the detection window and given sliding stride N_s of detection window for each layer. Let (x_i, y_i, s_i) be the detection position and scale for the i-th detection, w_i be the detection confidence, the detections are represented in 3-D space (position and scale space) as $p_i=(x_i, y_i, s_i)$, $i=1,2,\cdots,n$, where $s_i=\log(s_i)$ to ensure detections homogeneity in the 3-D space. Finally, borrowing the idea of literature [17], according to the list of initial results $p=(x, y, s)$ in different scale space, the uncertainty matrix H_i is computed for each detection point. Let $diag[H_i]$ represent the 3 diagonal elements of H_i.

$$diag[H_i]=[(\exp(S_i)\sigma_x)^2,(\exp(S_i)\sigma_y)^2,(\sigma_s)^2] \quad (19)$$

where $S_i=[S_s, S_sS_r, ..., S_n]$, the value of σ_s, σ_x and σ_y are the specified smoothing values by users. Scaling the smoothing values σ_x and σ_y by the term $\exp(S_i)$ increases the spatial uncertainty of the detected point. Then repeated calculation of the average offset vector until the aggregate to a specified pattern.

$$m(p)=H_h\frac{\nabla f(p)}{f(p)}\equiv H_h(p)[\sum_{i=1}^{n}w_i(p)H_i^{-1}p_i]-p \quad (20)$$

where the weighted estimate at a point p is given by $f(p)$, the gradient is $\nabla f(p)$, Let $H_h(p)$ be the weighted harmonic mean of the uncertainty matrices H_i computed at p. When $m(p)=0$, implying $\nabla f(p)=0$, the mode can be

iteratively estimated by

$$p_m = H_h(p_m)\left[\sum_{i=1}^{n} w_i(p_m) H_i^{-1} p_i\right] \quad (21)$$

According to the location and scale of the center of all pattern lists that provide the final detection result, target human is labeled in the rectangle box with size.

C. Pedestrian Segmentation Using Visual Saliency Detection

Because of the detected pedestrian is marked with a rectangle box, it contains some background regions with noises and target pedestrian with low contrast. Traditional algorithms of image segmentation only rely on color and appearance, which are less effective to extract the pedestrian region. However, the latter matching algorithm requires precise target, so the quality of segmentation directly affects the accuracy of recognition. The pedestrian recognition across aerial video faces the challenges problems of low quality image, high pose variations, minor availability of details, and the possibility of high density crowds. Human vision system is very robust for the object detection and segmentation in low quality image. In order to improve the accuracy of human recognition, the precision target pedestrian must be obtained like human visual system. Therefore we introduce visual saliency to segment the marked pedestrian region. When we look down from height, the target will be very small. Usually, we are only concerned about certain points, which can be treated as salient points, while most of images' foreground. Therefore, borrowing the human's perceptual and conceptual subjective cognitive ability, human body is segmented by context-aware saliency detection (CA) algorithm [13] in this paper. The algorithm is associated with surrounding environment to segment visual salient regions. Its segmentation results will not be affected by the change of shape, resolution and appearance. Therefore, the segmentation algorithm overcomes shortcomings of traditional segmentation algorithm, and is suitable for aerial image segmentation. The steps are as follows:

1) Local-global single-scale saliency

We divide image *I* of scale *r* into *n* patches with equal size. p_i and p_j present patches centered at pixel *i* and *j*, which are considered as salient points with respect to all other image patches, then the distance $d(p_i, p_j)$ between two patches is calculated.

$$d(p_i, p_j) = \frac{d_{color}(p_i, p_j)}{1 + c * d_{position}(p_i, p_j)} \quad (22)$$

where *c* is constant parameter, $d_{color}(p_i, p_j)$ is the Euclidean distance between the vectored patches p_i and p_j in CIE L*a*b color space, normalized to the range [0, 1]. $d_{position}(p_i, p_j)$ is the Euclidean distance between the positions of patches p_i and p_j that normalized by the larger image dimension. For each patch p_i, we search for the *K* most similar patches $\{q_k\}^K_{k=1}$ in the image, according to $d_{color}(p_i, p_j)$. Salient value of each patch can be calculated.

$$S_i^r = 1 - \exp\{-\frac{1}{K}\sum_{k=1}^{K} d(p_i^r, q_k^r)\} \quad (23)$$

2) Multi-scale saliency enhancement

Background patches are likely to have similar patches at multiple scale, e.g., in large homogeneous or blurred regions. Therefore we incorporate multiple scales to further decrease the saliency of background pixels. The details is considering the entire patches in the image I whose different scales, e.g. $R_q = \{r, ½r, ¼r\}$, as candidate neighbors. According to (23), we can calculate salient value in different scales, then the saliency map at each scale is normalized to the range [0, 1] and interpolated back to original image size. Furthermore, we get the average value of pixel *i* saliency at *M* scales via (24).

$$\bar{S}_i = \frac{1}{M}\sum_{r \in R_q} S_i^r \quad (24)$$

3) Including the immediate context

We search the cluster center that is the nearest from this pixel *i*, then final salient value is computed by its weight and salient value of pixel.

$$S_i = \frac{1}{M}\sum_{r \in R_q} S_i^r (1 - d_{foci}^r(i)) \quad (25)$$

where $d_{foci}^r(i)$ is Euclidean distance between pixel *i* and the nearest clustering center at scale *r*, normalized to the range [0, 1].

4) Getting final image

From the above three steps, we can get I_{gray} salient gray image, doing binary and removing noise I_{bw}, then get final image I_{seg}.

D. Pedestrian Alignment

For the camera angle differences and pedestrian various postures, the detective target is likely to tip over, even upside down. As a result, this will seriously reduce the accuracy of recognition. This paper uses the same strategy as WMR method using HST model (Head Shoulder and Torso) and AAM model (Active Appearance Model) algorithm [11].

$$M_{norm} = f(\beta) \times M_{real} \quad (27)$$

where M_{real} is a real matrix, M_{norm} is a normal matrix. $f(\beta)$ is a affine transformation function. β denotes rotation angle. Fig. 4 shows that the process of pedestrian alignment based on HST model and AAM algorithm. So we adaptively achieved the set of candidate $C=\{c_j, j=1...m\}$ from the marked region by the aforementioned marked, detection, segmentation and alignment steps.

E. Matching Processing

After the above preprocessing, we can get candidates. However, if the entire blob of pedestrian is used to match, which will not only need much more computing time, but also contain amount of noise, resulting in affecting recognition

accuracy.

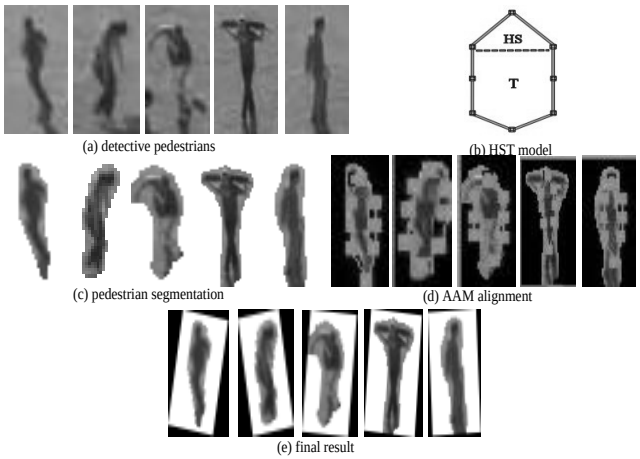


Fig. 4. The alignment process of HST and AAM algorithm.

Therefore we adopt the WRM method [11] to recognize the pedestrian in areial video. We find the candidate with the highest probability $P_{obj} = \max\{P(c_j)\}$, If $P_{obj} \geq T_{obj}$ we found the target; otherwise, the target would disappear, where T_{obj} is the target's threshold.

IV. Experimental Results

A large number of experiments are carried out to prove the effectiveness of the KMFS-WRM algorithm in aerial video, such as grassland, football field, square, and *et al*. Experiment environment is IntelCore2@2.0GHz, 2G memory, PC. In this paper, we use two standard libraries of aerial video (**UAV dataset** and **Public Aerial dataset**) which are provided by the University of Central Florida. Related parameters settings are as follows: window-size(20,40), HOG block-size(10,10), HOG cell-size(5,5), HOG bin(9), LBP cell-size(10,10), SIFT 15 key points, window-stride(5,5), scale rate $S_r = 1.02$, $a = 0.7, \beta = 0.1, \gamma = 0.2$, weight ratio $\zeta = 0.05$, $T_{obj} = 0.28$, $\tau = 1$. The width and height in video frame is 352 and 240 respectively, and frequency is 30 fps.

We demonstrate proposed pedestrian recognition framework by four experiments: pedestrian detection using proposed HLS model, segmentation using CA method, weight acquisition and comparing accuracy of pedestrian recognition between KMFS-WRM and original WRM algorithm.

A. Experiment of Pedestrian Detection Using Proposed HLS Model

Different feature has different contribution for the pedestrian detection. HLS model is the detection method by fusing the HOG, LBP and SIFT feature. How to determine the proportion of different feature is the significant problem. Fig. 7 shows the comparison of different weight of three features in terms of Human Misuse Detection Ratio (HMDR) and Human Omission Ratio (HOR). The formulas are as follows.

$$HMDR = \frac{FP}{TN + FP} \tag{27}$$

$$HOR = \frac{FN}{TP + FN} \tag{28}$$

where *TN* denotes the number of false object that is detected as false object, *FP* denotes the number of false object that is detected as positive object, *FN* denotes the number of positive object that is detected as false object, and *TP* denotes the number of positive object that is detected as positive object.

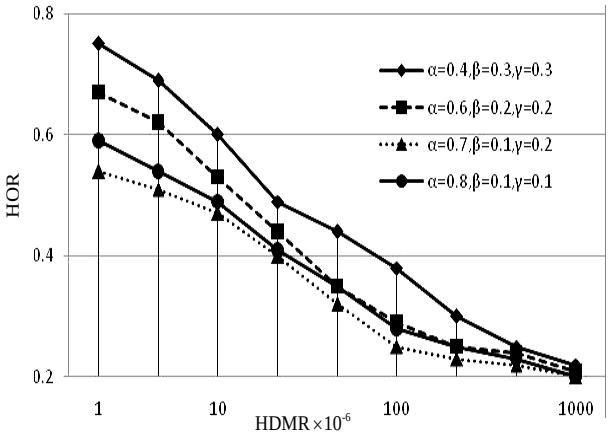


Fig. 5. Weight comparisons of different feature in terms of HMDR and HOR.

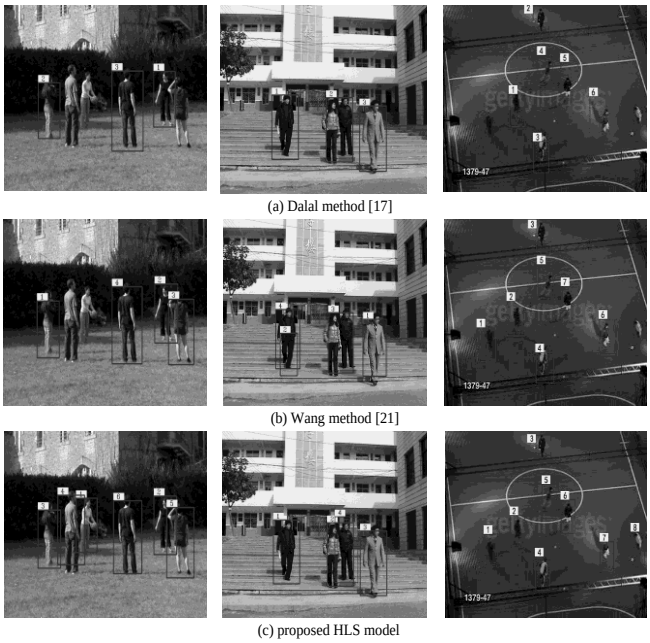


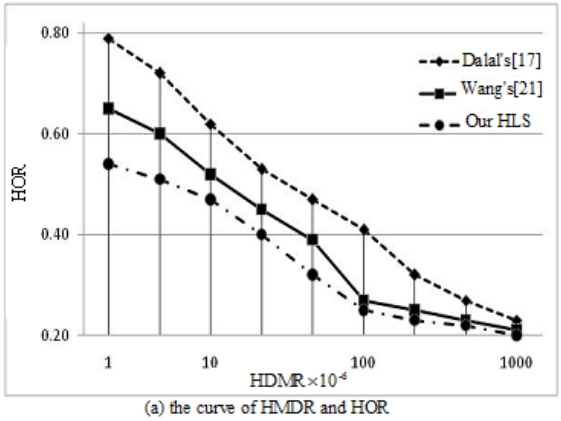
Fig. 6. The result of pedestrian detection using three different combinations of feature, white marked block represents the number of detected pedestrian.

Fig. 5 shows the weight comparison results of different feature. According to the characteristics of aerial video, we find the HOG descriptor is the most important in the detection of pedestrian, the SIFT descriptor is the second important and finally LBP feature. The HMDR and HOR is decreasing with the increasing ratio of HOG descriptor α, but the increasing tendency of HMDR and HOR occur when α=0.8. As a result we accepted that the ratio value of HOG (α), LBP (β) and SIFT (γ) descriptor is 0.7, 0.1, 0.2 respectively based on the minimum of HMDR and HOR in our experiment.

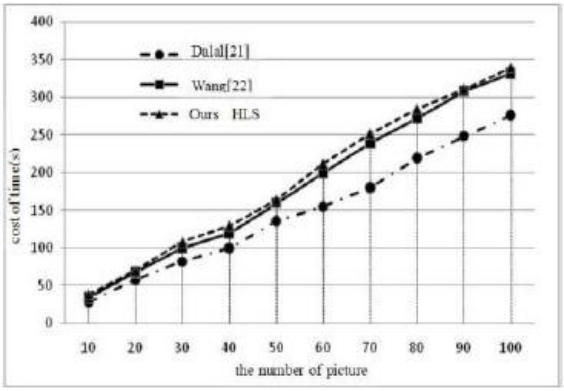
In order to further proving the accuracy of pedestrian detection by the combining different feature, we compared the detection method with only using HOG descriptor [17], the combining of HOG and LBP descriptor [21] and our proposed HSL model. Fig. 6 shows the detection result of three scenes: campus playground, teaching building and soccer field. Each

column represents a kind of scenes. In fact, there are six humans in the frame of first column, four humans in the frame of second column and eight humans in the frame of last column. Three humans, three humans and six humans are respectively marked with white marked block based on the proposed detection method by Dalal *et al*. [17] (Fig. 6(a)), four humans, four humans and six humans are respectively marked based on the proposed detection method by Wang *et al*. [21] (Fig. 6(b)), whereas six humans, four humans and eight humans are respectively marked in our method (Fig. 6 (c)). Obviously, there are pedestrian misuse and omission in Dalal's method [17] and Wang's method [21]. The experiment results demonstrated our proposed HLS model outperformance Dalal's and Wang's method since combining the superiority of different feature by different weight. This experiment shows that our multi-feature fusion is an effective way to improve the accuracy of pedestrian detection.

To confirm the robustness and efficiency of our proposed method that detects the pedestrian, we give two experiment of comparison with Dalal's method [17], Wang's method [21] and our method from the accuracy of detection pedestrian and computation complexity (Fig. 7). Firstly we illustrated the accuracy of detection pedestrian by the Receiver Operation Character (ROC) curve using HMDR and HOR (Fig. 7(a)). We can see from Fig. 7 (a), when the HDMR is same in three compared algorithms, Dalal's method has the highest of HOR, Wang's method is the second highest, and our approach is the lowest. The main reason lies in the different contribution of extracting different feature. In Dalal's method, only HOG descriptor is adopted. In Wang's method, using HOG and LBP descriptor, which lead to better presentment of pedestrian, to obviously improve the accuracy of pedestrian detection relative to Dalal's approach, but the combining of the two features is simple. The combination of SIFT, HOG and LBP feature based on the different weight greatly improve the accuracy of pedestrian detection. Secondly we analyses the computation complexity. With the increasing of feature dimension, the complexity also is increasing. Therefore Dalal's approach has the lowest computation time, the time of our and Wang's approach is slightly higher than Dalal's. Our method is basic consistent with Wang's method because using PCA method to reduce the dimension of adding SIFT feature. In summary, our proposed HSL model outperforms Dalal's and Wang's method from HDMR, HOR and computation complexity.



(a) the curve of HMDR and HOR



(b) computation time

Fig. 7. Comparison results of three approaches.

Fig. 8. The comparison of segmentation algorithm between KMFS-WRM and original WRM [11](row 1 source image, row 2 salient image, row 3 binary image, row 4 WRM results, row 5 KMFS-WRM results).

B. Experiment of Pedestrian Segmentation Using CA Model

Aerial image segmentation is shown in Fig. 8. The first group of images is source images gotten by pedestrian detection algorithm using HLS model. The second group of images is gray images obtained by context-aware saliency detection algorithm [13]. We can see many bright points, which are salient points, yet background region is darker. The third group of images is binary images by operating on the second group of gray images. The fourth group of images is segmented by traditional segmentation algorithm [11]. We find the result of segmentation is worse, and also many regions of pedestrian are occluded. However, using context-aware Saliency detection algorithm [13] overcomes the shortcomings, and gets a relatively complete target shown in the fifth group of images.

C. Experiment of Pedestrian Recognition

Fig. 9 shows accuracy of pedestrian recognition between our KMFS-WRM algorithm and Oreifej's original WRM algorithm [11]. There are three experimental results of different outdoor scene, including daytime soccer field, nighttime basketball court and square: the first row is the result of KMFS-WRM algorithm in each group, and the second row is the result of Oreifej's original WRM algorithm [11]. The first column in Fig. 9 represents the set of voters

(choosing 4 voters). The red rectangle in the second column is the marked candidate region using Kalman filter. The third column represents the set of candidates. We can get the final votes for each candidate by matching algorithm. Finally, candidate having the most votes is the target and is marked with the red rectangular box in the fourth column. The last column shows the zoom of target pedestrian. The attachment and arrow with red between the fourth column and the last column represent the correct identification of pedestrian, nevertheless the blue one represent the wrong identification.

As can be seen from Fig. 9, the recognition accuracy is very poor in WRM [11] because of using traditional segmentation algorithm, and then candidates' votes are quite similar and less discrimination. What's worse, the results of the first and second group are wrong. KMFS-WRM uses context-aware saliency detection algorithm and effect of segmentation is fine, improving the recognition accuracy. Moreover the number of candidates is fixed in supervised learning in WRM method [11]. But mostly, we don't need so many pedestrians, the more candidates, the greater calculated amount, causing real-time reduced. So we use Kalman filter to mark candidate region, and then matching target. Consequently, the number of candidates is decreased with reducing calculated amount and enhancing real-time, moreover the number of candidates is automatically determined in accordance with the scene content, as shown in Fig. 9.

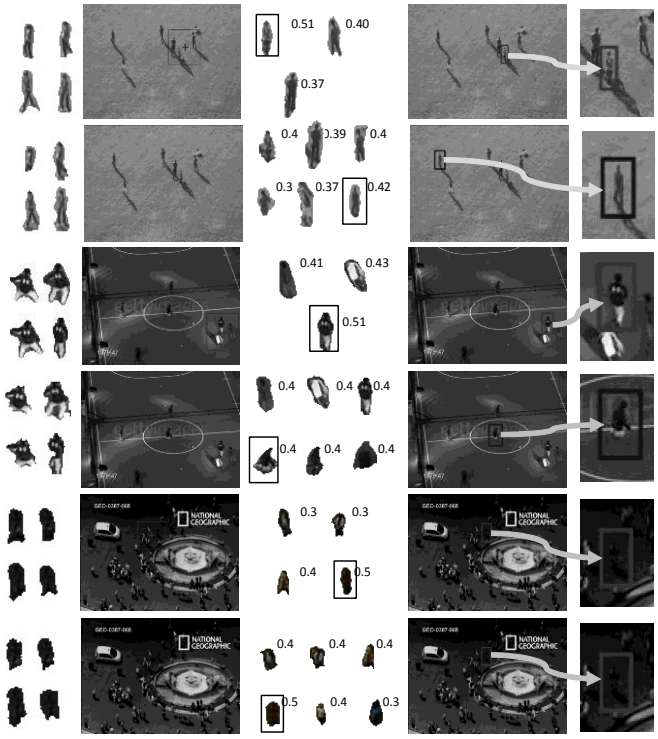


Fig. 9. The comparison of pedestrian recognition between the KMFS-WRM and Original WRM (total 3 groups, the first row is the result of KMFS-WRM and second row is Original WRM, red line with arrow is right and blue line with arrow is wrong).

In addition, we do 100 times experiments respectively using KMFS-WRM algorithm and original WRM [11] in every environmental video (grass, football field and square). Meanwhile, we record the consuming time in matching process and final result of recognition in each experiment. The statistical results are shown in Table I and Table II respectively. The WRM method implements the matching and recognition by dividing pedestrian' blob into *N* smaller regions using Mean Shift algorithm and extracting the feature, including the color histogram of region pixel in HSV color space and HOG descriptor of bounding rectangle of marked pedestrian. Then Oreifej *et al*. apply PCA on the feature space and extract the eigenvectors corresponding to the top 30 eigenvalues. Taking into consideration the matching entire blob of pedestrian, which will not only need much more computing time, but also contain amount of noise, resulting in affecting recognition accuracy, the proposed KMFS-WRM method use online kernel principal component analysis to dimensionality reduction. From the Table I, the KMFS-WRM uses Kalman filter to mark candidates' region and select candidates, which reduce the number of candidates and average time consuming in matching process over original WRM [11]. From the Table II, we can see, original WRM [11] applies traditional segmentation algorithm, which causes poor recognition accuracy, and however, the KMFS-WRM is opposite.

TABLE I: COMPARED THE TIME CONSUMING IN KMFS-WRM AND WRM [11] IN MATCHING PROCESS (UNITS SECOND)

	Grassland	Football field	Square
WRM	6.30	5,18	5.27
KMFS-WRM	3.59	4.01	3.97

TABLE II: COMPARED THE ACCURACY IN KMFS-WRM AND WRM[11] (TOTAL 100 TIMES)

	Grassland	Football field	Square
WRM	0.67	0.50	0.31
KMFS-WRM	0.82	0.71	0.49

V. CONCLUSION

A novel algorithm (KMFS-WRM) based on Kalman filter and saliency detection to recognize pedestrian in aerial video is proposed in this paper. The KMFS-WRM uses context-aware saliency detection to segment pedestrians and avoids the disadvantages of traditional segmentation algorithm in the preprocessing stage. Meanwhile, it applies Kalman filter to mark candidate's region and then match pedestrian, which not only reduces computational complexity, but also enhances self-adaptability. For KMFS-WRM and original WRM algorithm, we analyze and compare time consuming of matching process and recognition accuracy through a large number of experiments. The experimental result demonstrates that the KMFS-WRM outperforms WRM algorithms.

REFERENCES

[1] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance", in *Proc. IEEE Computer Vision and Pattern Recognition*, 2006, pp. 1528-1535.

[2] D. Anguelov, L. Kuang-chih, and S. B. Gokturk," Contextual identity recognition in personal photo albums", in *Proc. IEEE Computer Vision and Pattern Recognition*, 2007, pp. 1-7.

[3] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features", in *Proc. European Conference on Computer Vision*, 2008, pp. 262-275.

[4] H. M. Rara and A. A. Ali, "Face recognition at-a-distance using texture, Dense- and sparse-stereo reconstruction", in *Proc. 20th International Conference on Pattern Recognition*, 2010, pp. 1221-1224.

[5] M. A. Maloof, P. Langley, T. Binford, R. Nevatia, and S. Sage, "Improved rooftop detection in aerial images with machine learning," *Machine Learning*, vol. 53, no. 1-2, pp. 157-191, 2003.
[6] T. Zhao and R. Nevatia. "Car detection in low resolution aerial image," in *Proc. 8th IEEE International Conference on Computer Vision*, 2001, vol. 1, pp. 710-717.
[7] Y. Li, I. Atmosukarto, M. Kobashi, J. Yuen, and L. Shapiro. "Object and event recognition for aerial surveillance," in *Proc. SPIE-The International Society for Optical Engineering*, 2005, pp. 139-149.
[8] B. Nicolas, J. Viglino, and J. Cocquerez. "Knowledge based system for the automatic extraction of road intersections from aerial images," *International Archives of Photogrammetry and Remote Sensing*, vol. 88, no. 2, pp. 254-283, 2000.
[9] C. Vestri and F. Devernay. "Using robust methods for automatic extraction of buildings," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2001, vol. 1, pp. I-133-I-138.
[10] L. Wei and V. Prinet. "Building detection from high-resolution satellite image using probability model," in *Proc. Geoscience and Remote Sensing Symposium, IGARSS*, 2005, pp. 25-29.
[11] O. Oreifej, R. Mehran, and M. Shah, "Human identity recognition in aerial images," in *Proc. IEEE Computer Vision and Pattern Recognition*, 2010, pp. 709-716.
[12] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219-227, 1985.
[13] S. Goferman and L. Zelnik-Manor, "Context-aware saliency detection," in *Proc. IEEE Computer Vision and Pattern Recognition*, 2010, pp. 2376-2383.
[14] R. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, pp. 35-45, 1960.
[15] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, pp. 1-20, 2000.
[16] P. Lawrence and B. Sergey, "The anatomy of a large-scale hypertextual web search engine", in *Proc. the Seventh International Web Conference*, 1998, pp. 1-25.
[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, USA, 2005, vol. 1, pp. 886-893.
[18] M. Heikkila and M. Pietikainen, "A texture-based method for modeling the background and detecting moving objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 657-662, 2006.
[19] G. L. David, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.
[20] P. Honeine, "Online Kernel principal component analysis: A reduced-order model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1814-1826, 2012.
[21] X. Y. Wang, T. X. Han, and S. C. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. International Conference on Computer Vision*, 2009, vol. 1, pp. 32-39.

Chunping Liu was born in Chongqing, China, in 1971. She received her PhD degree at pattern recognition and artificial intelligence from Nanjing University of Science & Technology, China, in 2002. She was a visiting scholar in computer vision lab of University of Central Florida from 2010 to 2011.

She is now an associated professor of computer science, pattern recognition and image processing at the School of Computer Science & Technology in Soochow University, China. Her research interests include computer vision, image analysis and recognition, in particular in the domains of visual saliency detection, object detection and recognition and scene understanding. She has published more than 60 refereed journal articles and conference proceedings on image analysis, computer vision, and pattern recognition.

Fang Xu was born in Luan, China, in 1988. He received his bachelor degree from Anhui University of technology, Maanshan, China, in 2010. He is currently pursuing his M.S. degree at Soochow University, Suzhou, China. His research interests include background modeling, visual tracking, and image processing.

Xingbao Wang was born in Leuan, China, in 1988. He received his M.S. degree at Soochow University, Suzhou, China, in 2012. He is currently working in Anhui USTCiflytek Co., LTD. His research interests include video processing, background modeling, and pedestrian tracking.

Shengrong Gong received his M.S. degree from Harbin Institute of Technology in 1993 and Ph.D. degree from Beihang University in 2001.

He is now a professor and doctoral supervisors of the School of Computer Science and Technology, Soochow University. Currently he is a senior member of Chinese Computer Society, editors of communication journal, virtual reality professional of Chinese Society of image and graphics. He acted as chairman for 2010-2011 YOCSEF of the Academic Committee of Suzhou sub-forum. He got twice award of the Scientific and Technological Progress, and has published more than 100 academic articles. His research interests are image and video process, pattern recognition and computer vision.