Journal of Electronic Imaging

JElectronicImaging.org

Real-time unmanned aerial vehicle tracking of fast moving small target on ground

Junhua Yan
Jun Du
Yong Young
Christopher R. Chatwin
Rupert C. D. Young
Philip Birch

SPIE. IS&T imaging.org

Junhua Yan, Jun Du, Yong Young, Christopher R. Chatwin, Rupert C. D. Young, Philip Birch, "Real-time unmanned aerial vehicle tracking of fast moving small target on ground," *J. Electron. Imaging* **27**(5), 053010 (2018), doi: 10.1117/1.JEI.27.5.053010.

Real-time unmanned aerial vehicle tracking of fast moving small target on ground

Junhua Yan,[a,*] Jun Du,[a] Yong Young,[a] Christopher R. Chatwin,[b] Rupert C. D. Young,[b] and Philip Birch[b]
[a]Nanjing University of Aeronautics and Astronautics, College of Astronautics, Nanjing, China
[b]University of Sussex, School of Engineering and Informatics, Brighton, United Kingdom

Abstract. To solve the problems of occlusion and fast motion of small targets in unmanned aerial vehicle target tracking, an adaptive algorithm that fuses the improved color histogram tracking response and the correlation filter tracking response based on multichannel histogram of oriented gradient features is proposed to realize small target tracking with high accuracy. The state judgment index is used to determine whether the target is in a fast motion or an occlusion state. In the fast motion state, the search area is enlarged, and the color optimal model that suppresses the suspected area is used for rough detection. Then, redetection in the location of multiple peaks in the rough detection response is carried out using the correlation filter to accurately locate the target. In an occlusion state, the model stops updating, the search area is expanded, and the current color model is used for rough detection. Then, redetection in the place of multiple peaks in the rough detection response is carried out using the correlation filter to accurately locate the target. Experimental results show that the proposed method can track small targets accurately. The frame rate of the proposed method is 40.23 frames/s, indicating usable real-time performance. [DOI: 10.1117/1.JEI.27.5.053010]

Keywords: unmanned aerial vehicle; small target; fast motion; occlusion; correlation filter; global color histogram.

Paper 180307 received Apr. 6, 2018; accepted for publication Aug. 16, 2018; published online Sep. 14, 2018.

1 Introduction

During unmanned aerial vehicle (UAV) target tracking, the target is far away from the camera; hence, the target pixel size in the image (the number of pixels occupied by the target) is small. In addition, when the UAV moves swiftly, the camera is actively adjusted, and the target position shift between adjacent frames may exceed 20 pixels. In a target tracking review article with more than 1000 citations,[1] the above two cases are classified as low resolution and fast motion, respectively. Low resolution can also cause the target to be blocked easily. These factors make it challenging to track the ground moving small target accurately and in real time.

When the target occupies a small number of pixels, limited feature information is obtained from target pixels. High-level features that are capable of more powerful feature expression are favored in such circumstances to ensure the robustness of the tracking method. Danelljan et al.[2] effectively improved the tracking effect of the method in their paper[3] using multichannel color features instead of grayscale features. However, a single color feature is not sufficient for capturing all illumination changes. Henriques et al.[4] used the multichannel histogram of oriented gradient (HOG)[5] to represent the target, which can well represent the local shape feature of the target. Hence, the tracking effect of the correlation filtering was significantly improved. However, trackers based on HOG often perform poorly when the target has movements or serious deformations. Experiments showed that the above two single feature models cannot favorably cope with small targets, resulting in target drifting. Danelljan et al.[6] won the 2016 VOT-challenge with a comprehensive combination of the multichannel color feature, a well-trained convolution neural network (CNN) feature, and the HOG feature. However, due to the limited number of online training samples in the target tracking, the over-dimensioned feature vector easily leads to over fitting; this method requires the updating of more than 800,000 model parameters with each use, making it difficult to fulfill the real-time requirement of target tracking.

Expanding the search area to obtain a larger sampling area is one of the ways to deal with fast moving targets. However, the amount of computation is increased and the false alarm rate rises due to the introduction of objects similar to the target. To cope with fast movements of the target, Ma et al.[7] introduced an online random fern classifier, which is similar to training learning detection,[8] to redetect targets. However, the redetection module is based on the grayscale features, so it is difficult to achieve good redetection performance in a large area. Zhang et al.[9] used multiple trackers as an expert group to conduct semisupervised loss judgment on the expert group's tracking results to select the optimal tracking result and improve the reliability of the tracker. However, it is still difficult for the method to deal with disturbing objects in the search area based on a single grayscale feature. Additionally, each frame requires multiple tracking and detection, making it difficult to achieve real-time performance. Zhu et al.[10] used edge boxes[11] to obtain areas with more closed edge information as a global candidate area instead of using a local search area. However, when the target is small, its edge information is relatively limited, making it difficult for edge boxes to accurately locate the target. In addition, the edge box method requires sampling of a large number of areas to improve

*Address all correspondence to: Junhua Yan, E-mail: yjh9758@126.com

the probability that the target is detected, compromising real-time performance.

Small targets are easily obscured, which increases the difficulty of tracking. Jia[12] used a local sparse representation of the target to cope with partial occlusions of the target. Zhao et al.[13] used an innovative keypoint matching-based tracker to handle the partial occlusion problem, yet these two methods cannot cope with relatively large occlusion sizes. Also, the average frame rate of this method on the OTB2013[1] dataset is 8.5 frames/s, not satisfying the real-time requirement of target tracking. In addition, small targets that lack information are not suitable for local sparse representation. Kalal et al.[8] introduced the online random fern classifier to redetect targets, but the redetection module is based on simple grayscale features, so it is difficult to obtain good redetection results. In the case that the target is completely obscured, Yan et al.[14] used the Kalman filter method to estimate the target position to achieve the target tracking, though the position estimation method cannot accurately estimate how the target would separate from the occlusion.

In this paper, to solve the problems of fast motions and occlusions of the target in UAV target tracking, an adaptive algorithm that fuses the improved color histogram tracking response and the correlation filter tracking response based on multichannel HOG features is proposed to realize stronger feature expression for small targets. The state judgment index is used to determine whether the target is in a fast motion or an occlusion state. In the fast motion state, the search area is enlarged, and the color optimal model that suppresses the suspected area is used for rough detection. Then, redetection in the location of multiple peaks in the rough detection response is carried out using the correlation filter to accurately locate the target. In the occlusion state, the model stops updating, the search area is expanded, and the current color model is used for rough detection. Then, redetection in the location of multiple peaks in the rough detection response is carried out using the correlation filter to accurately locate the target. The block diagram of real-time UAV tracking of fast moving small target on ground is shown in Fig. 1.

2 Target Tracking Method by Fusing Two Tracking Models

The HOG is a statistical feature based on the local gradient direction, which cannot cope with target deformations well. The global color distribution of the target does not change greatly with target deformations. Therefore, the global color feature can better deal with target deformations. By contrast, the color feature cannot deal with illumination changes very well, whereas the HOG uses gamma correction to normalize the contrast of the original image and can better deal with illumination changes. The color feature and the HOG complement each other. Hence, a tracking model based on the fusion of these two features is expected to represent the small target more powerfully and track it more accurately. A flowchart of the proposed target tracking method by the fusion of two tracking models is shown in Fig. 2.

2.1 *Tracking Response of the Correlation Filter Model Based on Local Multichannel HOG Features*

The correlation filter tracking method based on multichannel local HOG features is divided into the training stage and the detection stage. In the training stage, the optimal correlation filter is obtained by training the sample set, and the optimal filter is updated according to the fast updating strategy. Multichannel local HOG features are extracted for each pixel in the local search area of the previous frame, which are then used to form a matrix, and the rows and columns of the matrix are cyclically shifted to obtain a training sample set. According to the characteristics of the circulant matrix, the discrete Fourier domain was used to solve the correlation filter instead of the Ridge regression to avoid matrix inversion, reducing the complexity of the algorithm by several orders of magnitude and achieving real-time performance.[4] In the detection stage, multichannel local HOG features are extracted for each pixel in the local search area of the current frame, which are then used to form a matrix, and the rows and columns of the matrix are cyclically shifted to obtain a to-be-detected sample set. The correlation filter response score for each sample set is obtained according to the updated optimal filter, and the coordinates of the sample with the highest score are set as the center location.

2.1.1 *Characteristics of the circulant matrix*

In this section, the one-dimensional (1-D) single channel signal, which is methodologically similar to the two-dimensional (2-D) multichannel signal, is used to describe the acceleration characteristic of the circulant matrix.[4] Suppose that the 1-D single channel signal is represented by a vector of $n \times 1$, denoted as $x = [x_0, x_1, x_2, Kx_{n-1}]$, then the circulant $\mathbf{X}$ is obtained by cyclic shift $C(\mathbf{X})$ of $\mathbf{x}$, shown as

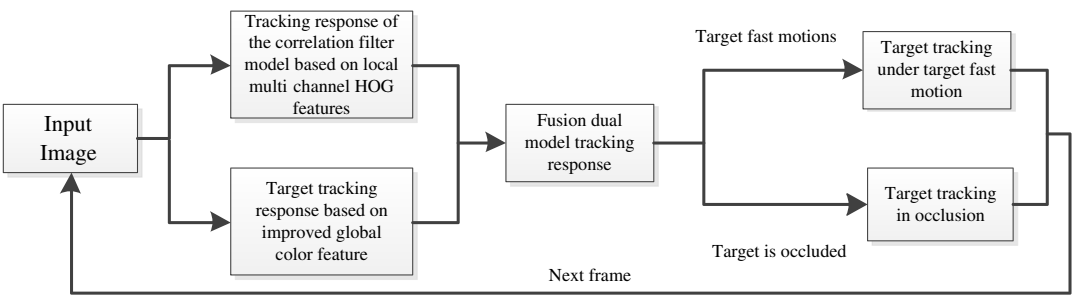

Fig. 1 Block diagram of real time UAV tracking of fast moving small target on ground.

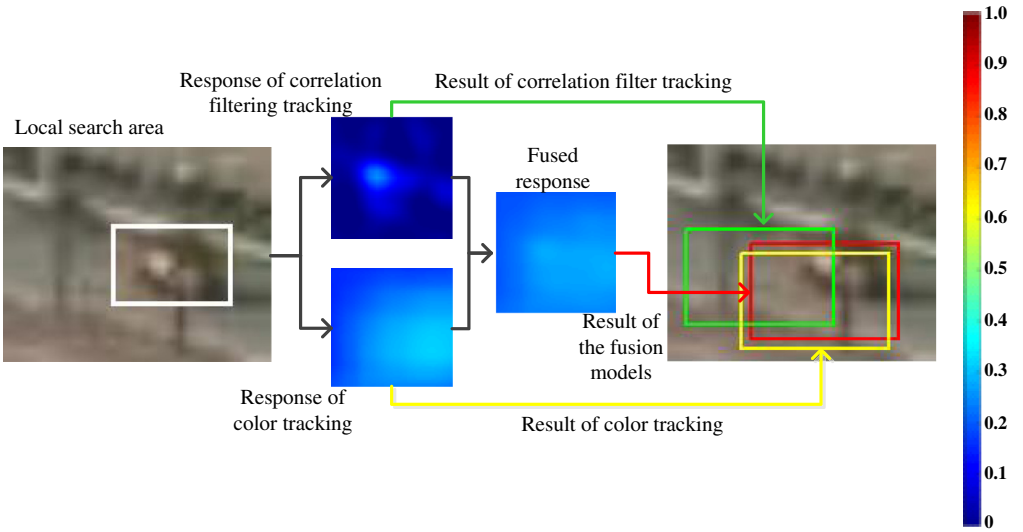

Fig. 2 Flowchart of target tracking method by fusing two tracking models.

$$\mathbf{X} = C(\mathbf{x}) = [\,\mathbf{x}_0 \quad \mathbf{x}_1 \quad \cdots \quad \mathbf{x}_q \quad \cdots \quad \mathbf{x}_{n-1}\,] = \begin{bmatrix} x_0 & x_1 & \cdots & x_{n-1} \\ x_{n-1} & x_0 & \cdots & x_{n-2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n-q} & x_{n-q+1} & \cdots & x_{n-q-1} \\ \vdots & \vdots & \vdots & \vdots \\ x_1 & x_2 & \cdots & x_0 \end{bmatrix}. \tag{1}$$

Circulant $\mathbf{X}$ is the training sample set. Each row vector $\mathbf{x}_q$ is a sample, and its corresponding label vector is $\mathbf{y}$:

$$\mathbf{y} = [y_0, y_1, y_2, \cdots y_q, \cdots y_{n-1}]^T \times \left(y_q = \begin{cases} 0 & x_q \text{ is negative sample} \\ 1 & x_q \text{ is positive sample} \end{cases} \right). \tag{2}$$

The goal of training is to find a function $f(\mathbf{X}) = \mathbf{w}^T\mathbf{X}$ that minimizes the squared error between samples x_q and their label value y_q, as shown in Eq. (1). Here, λ is a regularization parameter that controls overfitting. Note that

$$\min_{\mathbf{w}} \sum_q [f(\mathbf{x}_q) - y_q]^2 + \lambda \|\mathbf{w}\|^2, \tag{3}$$

where $\mathbf{w}$ is the coefficient to be solved for. The linear regression least squares of $\mathbf{w}$ can be computed as follows:

$$\mathbf{w} = (\mathbf{X}^H\mathbf{X} + \lambda I)^{-1}\mathbf{X}^H\mathbf{y}, \tag{4}$$

where the superscript H is the conjugate transpose. Directly solving Eq. (4) involves matrix inversion, which demands a huge amount of computation, compromising the real-time performance of the tracking method.

To reduce computational complexity, Eq. (4) is transformed into the frequency domain. According to Ref. 15, the circulant matrix is diagonalized by the discrete Fourier transform matrix $\mathbf{F}$:

$$\mathbf{X} = C(\mathbf{x}) = \mathbf{F}\,\mathrm{diag}(\hat{\mathbf{x}})\mathbf{F}^H, \tag{5}$$

where $\hat{\mathbf{x}}$ represents the discrete Fourier transformation of $\mathbf{x}$, $\hat{\mathbf{x}} = F(\mathbf{x})$. Here,

$$\begin{aligned} \mathbf{w} &= [\mathbf{F}\,\mathrm{diag}(\bar{\hat{\mathbf{x}}})\mathbf{F}^H\mathbf{F}\,\mathrm{diag}(\hat{\mathbf{x}})\mathbf{F}^H \\ &\quad + \lambda\mathbf{F}\,\mathrm{diag}(\delta)\mathbf{F}^H]^{-1}\mathbf{F}\,\mathrm{diag}(\bar{\hat{\mathbf{x}}})\mathbf{F}^H\mathbf{y} \\ &= [\mathbf{F}\,\mathrm{diag}(\bar{\hat{\mathbf{x}}}\odot\hat{\mathbf{x}} + \lambda\delta)\mathbf{F}^H]^{-1}\mathbf{F}\,\mathrm{diag}(\bar{\hat{\mathbf{x}}})\mathbf{F}^H\mathbf{y} \\ &= \mathbf{F}\,\mathrm{diag}\left(\frac{\bar{\hat{\mathbf{x}}}}{\bar{\hat{\mathbf{x}}}\odot\hat{\mathbf{x}} + \lambda\delta}\right)\mathbf{F}^H\mathbf{y} \\ &= C\left[F^{-1}\left(\frac{\bar{\hat{\mathbf{x}}}}{\bar{\hat{\mathbf{x}}}\odot\hat{\mathbf{x}} + \lambda\delta}\right)\right]\mathbf{y}. \end{aligned} \tag{6}$$

In Eq. (6), $\mathrm{diag}(\bar{\hat{\mathbf{x}}})\mathrm{diag}(\bar{\hat{\mathbf{x}}}) = \mathrm{diag}(\bar{\hat{\mathbf{x}}}\odot\hat{\mathbf{x}})$; δ is an all-1 vector and is omitted in the following equations $\bar{\hat{\mathbf{x}}}$ represents the conjugation of $\hat{\mathbf{x}}$.

Then, use the following according to the convolution property of the circulant matrix discussed in Ref. 16:

$$\mathcal{F}[C(\mathbf{x})\mathbf{y}] = \mathcal{F}(\tilde{\mathbf{x}} * \mathbf{y}) = \overline{\mathcal{F}(\mathbf{x})}\odot\mathcal{F}(\mathbf{y}),$$

where $\tilde{\mathbf{x}}$ represents the reverse order of $\mathbf{x}$. A Fourier transform is carried out on both sides of Eq. (6) to solve for $\mathbf{w}$:

$$\mathcal{F}(\mathbf{w}) = \overline{\mathcal{F}\left[\mathcal{F}^{-1}\left(\frac{\bar{\hat{\mathbf{x}}}}{\hat{\mathbf{x}}\odot\bar{\hat{\mathbf{x}}} + \lambda}\right)\right]}\odot\mathcal{F}(\mathbf{y}), \tag{7}$$

$$\mathbf{w} = \mathcal{F}^{-1}\left(\frac{\hat{\mathbf{x}}\odot\hat{\mathbf{y}}}{\bar{\hat{\mathbf{x}}}\odot\hat{\mathbf{x}} + \lambda}\right). \tag{8}$$

2.1.2 *Correlation filter tracking method based on multichannel HOG feature*

Training stage: The local searching area D_t for training is selected by setting the center pixel P_{t-1} of the target tracking box of the previous frame I_{t-1} as the center of D_t. The width and height of D_t are W and H, respectively, which are twice

the width and height of the target of the previous frame.[17] The multichannel local HOG feature $dh^N_{w,h}$ is extracted for each pixel (w,h) in D_{t-1}, where N is the number of channels of the feature. A $W \times H$ matrix $\mathbf{DH}^N$ is constructed using $dh^N_{w,h}$. Each element $dh^N_{w,h}$ in the matrix is an N-dimensional vector. Training samples $\{\mathbf{DH}^N_{w,h} | w \in \{0,1,\ldots,W-1\}, h \in \{0,1,\ldots,H-1\}\}$ are generated by a cyclic shift operation on $\mathbf{DH}^N$. Training samples are used to train the optimal correlation filter h^N_{cf} so that it has the highest filtering response to the sample centered on (w,h) in D_{t-1}. The training process is a ridge regression process. Its purpose is to minimize the loss, as shown in Eq. (9):

$$\arg\min_{h^N_{cf}} \sum_{W,H} \left(\left\| \sum_{n=1}^{N} h^n_{cf} * \mathbf{DH}^n_{w,h} - g_{w,h} \right\|^2 + \lambda \sum_{n=1}^{N} \|h^n_{cf}\|^2 \right), \tag{9}$$

where $*$ represents the convolution operation, $\mathbf{DH}^n_{w,h}(n = 1,2,\cdots N)$ is the component of the sample in each channel, and $h^n_{cf}(n = 1,2,\cdots N)$ is the component of the correlation filter h^N_{cf} on each channel. Here, $g_{w,h}$ is the ideal 2-D Gaussian response corresponding to $\mathbf{DH}^N_{w,h}$. $\| \sum_{n=1}^{N} h^n_{cf} * \mathbf{DH}^n_{w,h} - g_{w,h} \|^2$ represents the loss function, and $\lambda \sum_{n=1}^{N} \|h^n_{cf}\|^2$ is the regular item to prevent overfitting, which must be >0. Note that λ is a regularization parameter and is assigned the optimal value 0.001 derived in Ref. 4. The idea discussed in Sec. 2.1.1 is applied to solve Eq. (9), and the optimal correlation filter of the previous frame in the frequency domain is obtained:

$$\hat{h}^n_{cf} = \frac{\hat{g} \odot \hat{\mathbf{DH}}^n}{\sum_{n=1}^{N} (\overline{\hat{\mathbf{DH}}^n} \odot \hat{\mathbf{DH}}^n) + \lambda} = \frac{A^n}{B + \lambda}. \tag{10}$$

The optimal filter is updated according to the fast updating strategy proposed in Ref. 16:

$$A^n_t = (1-\eta) A^n_{t-1} + \eta \bar{\hat{g}}_{t-1} \odot \hat{\mathbf{DH}}^n_{t-1}, \tag{11}$$

$$B_t = (1-\eta) B_{t-1} + \eta \sum_{i=1}^{N} \overline{\hat{\mathbf{DH}}^n_{t-1}} \odot \hat{\mathbf{DH}}^n_{t-1}, \tag{12}$$

where η is an update parameter, which determines the update rate. A larger η means a greater impact of the current frame on the module, indicating faster model update. In this paper, η is assigned the optimal value of 0.01 derived in Ref. 16.

Detection stage: The local searching area D_t for training is selected by setting the center pixel P_{t-1} of the target tracking box of the previous frame I_{t-1} as the center of D_t. The multichannel local HOG feature $dh^N_{w,h}$ is extracted for each pixel (w,h) in D_{t-1}, where N is the number of channels of the feature. A $W \times H$ matrix $\mathbf{DH}^N$ is constructed using $dh^N_{w,h}$. Each element $dh^N_{w,h}$ in the matrix is an N-dimensional vector. Detecting samples $\{\mathbf{DH}^N_{w,h} | w \in \{0,1,\ldots,W-1\}, h \in \{0,1,\ldots,H-1\}\}$ are generated by a cyclic shift operation on $\mathbf{DH}^N$. According to the updated optimal filter, the correlation filtering response score of each sample is obtained as follows:

$$S_{cf}(w,h) = \mathcal{F}^{-1}\left\{ \frac{\sum_{n=1}^{N} A^n_t \hat{\mathbf{DH}}^n_{w,h}}{B_t + \lambda} \right\}. \tag{13}$$

The position of the target center pixel x in D_t is set to be the coordinates of the point $(w_{\max}, h_{\max})$ with the highest response score, and the correlation filtering tracking response score is $S_{cf}(x)$.

2.2 *Target Tracking Response Based on Improved Global Color Feature*

In the color histogram tracking, the probability of the pixel x belonging to the target in the current local search area D_t is obtained by constructing the target normalized RGB color histogram and looking up the table. According to the normalized color histogram Hist_{fg} of the foreground and the normalized color histogram Hist_{bg} of the background of the current frame, the probability $p_{fg}(x)$ that the pixel x belongs to the foreground and the probability $p_{bg}(x)$ that the pixel x belongs to the background are, respectively, calculated.

$$p_{fg}(x) = \text{Hist}_{fg}(i_x), \tag{14}$$

$$p_{bg}(x) = \text{Hist}_{bg}(i_x), \tag{15}$$

where i_x indicates that the pixel x belongs to the i'th bin in the color histogram.

According to Ref. 18, the probability that pixel x belongs to the target in the search area is denoted as

$$p(x \in O | D_t) = \frac{p_{fg}(x)}{p_{fg}(x) + p_{bg}(x)}. \tag{16}$$

To adapt the representation to changing object appearance and illumination conditions, we update the object model on a regular basis using linear interpolation $P_t(x \in O|D_t) = \eta_c P(x \in O|D_t) + (1-\eta_c) P_{t-1}(x \in O|D_t)$, with a learning rate η_c.

The probability integral graph I in the search area D_t is calculated, and the response score $S_{\text{hist}}(x)$ of the target box in D_t with pixel x as the center and the target size area as the size of the box is obtained:

$$S_{\text{hist}}(x) = I(i+W/2, j+H/2) + I(i,j) - I(i+W/2, j) - I(i, j+H/2), \tag{17}$$

where W and H are the width and height of the current target, respectively, and (i,j) represents the horizontal and vertical coordinates of the pixel x.

The position of the target center pixel x in D_t is set to be the coordinates of the point $(i_{\max}, j_{\max})$ with the highest response score, and the color tracking response score is $S_{\text{hist}}(x)$.

If the target is relatively small, drifting to areas with a similar color is likely to happen. To cope with drifting, the current method suppresses areas with suspected color similarities to reduce interference from these areas.

When the response score $S_{\text{hist}}(x)$ of the box area satisfies Eq. (18), it is considered to be a suspected area:

$$S_{\text{hist}}(x_{\text{dis}}) \geq \theta_0 \max[S_{\text{hist}}(x)], \theta_0 \in [0,1], \tag{18}$$

where x_{dis} represents the central position of the suspected rectangular area and θ_0 is the threshold parameter, which is arbitrarily set to be 0.8 here.

The suspected area is sorted according to its response score. The normalized color histogram set $\{\text{Hist}_{\text{dis}}^n | n = 1 \cdots N\}$ for the first N suspect areas is calculated. Then, the probability that pixel x belongs to each suspected area is calculated, followed by recalculation of the probability that pixel x in D_t belongs to the target as shown in Eq. (19):

$$P_t(x \in O|D_t) = \frac{p_{fg}(x)}{p_{fg}(x) + p_{bg}(x) + \frac{1}{N}\sum_{n=1}^{N} p_{\text{dis}}^n(x)}. \quad (19)$$

Then, the color tracking response score $S_{\text{hist}}(x)$ of the target tracking box in the search area D_t is recalculated using Eq. (17).

To test whether the suppression of color suspicious areas can effectively reduce interference from suspected areas, a comparative experiment on images with small targets and color-like areas in the UAV123 dataset is carried out. Some experimental results are shown in Fig. 3.

The second column in Fig. 3 is the probability map of pixels belonging to the target without suppression. Probability values at the target area are high, yet those of color suspicious areas are also high, causing interference to target tracking. The third column in Fig. 3 is the probability map of pixels belonging to the target with suppression. Responses in suspected areas are suppressed. The decrease in probability values at the target area in the map is less than that at suspected areas, which makes the probability value of the target more prominent. Thus, experiments show that suppression of color-like areas can effectively reduce interference from suspected areas.

2.3 *Fusion Dual Model Tracking Response*

Adaptive fusion of the improved color histogram tracking response and the correlation filter tracking response based on multichannel HOG feature was carried out to determine the center position P_t of the target tracking box in the current frame I_t:

$$P_t = \arg\max_{x \in D_t} S_f(x), \quad (20)$$

$$S_f(x) = S_{cf}(x) + f[S_{\text{hist}}(x)], \quad (21)$$

where $S_f(x)$ is the tracking response score at x of the fusion dual model. When there are many suspected areas, the target tracking box, which is determined by the target tracking response based on improved global color feature, is likely to drift to areas with a similar color. To guarantee the exact location of the target tracking box, which is determined by the fusion dual model tracking response, we need to reduce the color tracking response score. Therefore, the value of the score is reduced to lower the impact of the color tracking response on the overall fusion probability. Hence, the color tracking response score $S_{\text{hist}}(x)$ is adaptively adjusted as follows:

$$f[S_{\text{hist}}(x)] = \begin{cases} [S_{\text{hist}}(x)]^{1/2}, & N = 1 \\ S_{\text{hist}}(x), & 1 < N \le 2 \\ [S_{\text{hist}}(x)]^{(1+N/2)}, & 2 < N \le 5 \end{cases}, \quad (22)$$

where N is the number of suspected areas. The color tracking response score is affected by the number of suspected areas. When N is large, it is considered that the color tracking response score is not credible enough. Hence, the value of the score is reduced to lower the impact of the color tracking response on the overall fusion probability. On the contrary, if N is small, indicating that the color tracking response score is credible, it is appropriate to increase its value for better tracking results. The aim of these modifications is to achieve better tracking results.

Target tracking results of the adaptive fusion dual model are compared with target tracking results of single models as shown in Fig. 4. In each image in Fig. 4, the green box demonstrates the tracking result of the correlation filter model

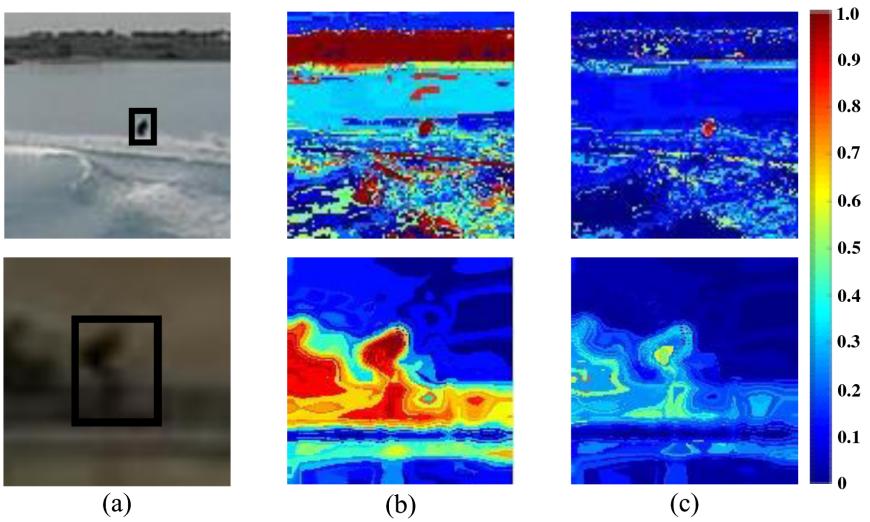

Fig. 3 Color tracking responses with and without color suspicious area suppression: (a) original image (the target is within the black box), (b) probability map of pixels belonging to the target without suppression, and (c) probability map of pixels belonging to the target with suppression.

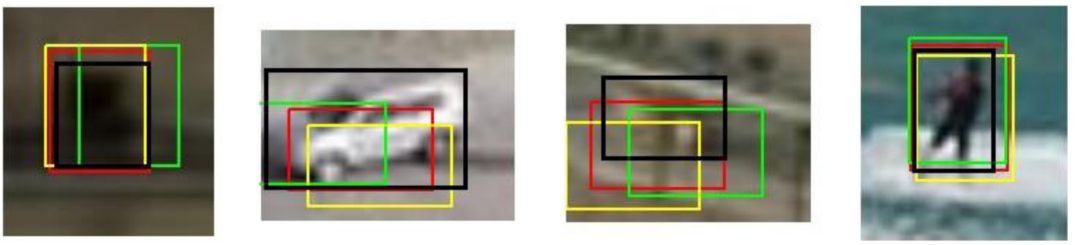

Fig. 4 Comparison of target tracking results of the adaptive fusion dual model and two single models.

Table 1 Comparison of OS scores of target tracking results of the adaptive fusion dual model and two single models.

Tracking method	Correlation filter tracking	Color tracking	Fusion two model tracking
OS of the first image	0.469	0.746	0.750
OS of the second image	0.431	0.353	0.508
OS of the third image	0.238	0.157	0.395
OS of the fourth image	0.710	0.666	0.721

based on multichannel HOG features. The yellow box demonstrates the tracking result of the improved global color histogram feature model. The red box demonstrates the tracking result of the proposed adaptive fusion dual model. The black box marks the true location of the target.

The overlap score (OS) is used to measure the accuracy of target tracking results. OS is calculated as follows:

$$\mathrm{OS} = \frac{B_{gt} \cap B_t}{B_{gt} \cup B_t}, \tag{23}$$

where B_{gt} represents the true target position and B_t represents the target location identified by the tracking method. Higher OS scores indicate higher accuracy.

Figure 4 and Table 1 show that the proposed target tracking method that fuses multiple tracking models can achieve better OS, indicating higher tracking accuracy.

3 Fast Moving or Occluded Target Tracking

In the process of target tracking, rapid movement of the target leads to rapid location changes of the target in the video. Consequently, the target easily moves out of the local search area, resulting in tracking failure. In addition, the small size of the target makes it an easy victim of occlusion. In this paper, a tracking method that copes with target fast motions and target occlusions is proposed.

3.1 *Target Tracking under Target Fast Motion*

In a target tracking review article with >1000 citations,[1] the two cases described above are classified as fast motion and low resolution. A target is considered to be in the fast motion state when its position offsets >20 pixels between adjacent frames.

$$(i_{t-1} - i_{t-2}) > 20 or (j_{t-1} - j_{t-2}) > 20, \tag{24}$$

(i_{t-1}, j_{t-1}) and (i_{t-2}, j_{t-2}) are the coordinates of the centers P_{t-1} and P_{t-2}, respectively, of the tracking box in frame $t-1$ and the frame $t-2$, respectively.

3.1.1 *Update the model parameters of the correlation filter model and color feature model*

When the target is in the fast motion state, model parameters of the correlation filter model and the color histogram model need to be adjusted to cope with changes in target posture and illumination.

The correlation filter model parameters A^n and B, the foreground histogram feature Hist_{fg}, and the fusion response peak value $\max(S_f)$ of the frame, which is 2FR (FR indicates video frame rate) frames before the current frame, form a set which is then divided into L segments in time order A_l^n, B_l, and $\mathrm{Hist}_{fg(l)}$ corresponding to the frame with the largest response peak value from each segment; these are selected to form an expert group $[A_l^n, B_l, \mathrm{Hist}_{fg(l)}]$. Weighted summation was performed on the members of the expert group to obtain the optimal correlation filter model and the foreground histogram feature needed in color tracking. Taking into account the temporal correlation between frames in a video sequence, greater weights are assigned to parameters in frames that are closer to the current frame:

$$A^n = \frac{2}{L(L+1)} \sum_{l=1}^{L} l \times A_l^n, \tag{25}$$

$$B = \frac{2}{L(L+1)} \sum_{l=1}^{L} l \times B_l, \tag{26}$$

$$\mathrm{Hist}_{fg} = \frac{2}{L(L+1)} \sum_{l=1}^{L} L \times \mathrm{Hist}_{fg(l)}. \tag{27}$$

3.1.2 *Redetect to find the true target*

To track the fast moving target in real time, the local search area in color tracking is expanded to $2D_t$, and color primary detection is performed to obtain the color tracking response score $S_{\mathrm{hist}}(x)$.

In color tracking, the true target area can be mistaken as a suspected area and, hence, be suppressed. Also, when the true target area and the suspected object are close, the suspected object can be omitted because it yields a subpeak response closed to the peak response area. In either case, multiple peak areas appear in the final color tracking response, and the highest peak response does not necessarily

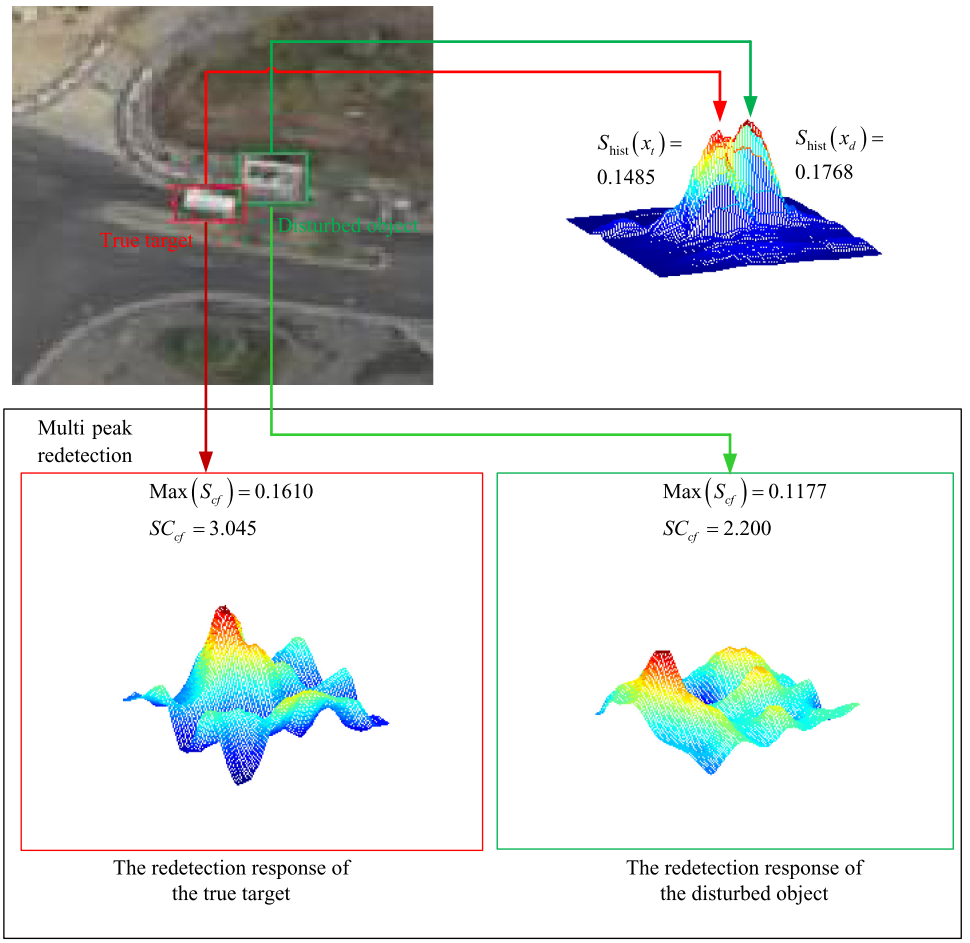

Fig. 5 Schematic of multipeak redetection.

represent the true target. For instance, as shown in Fig. 5, in the upper left image, the red box marks the true target, whereas the green box is the suspected object, and the response value of the suspected object in the color tracking response is higher than that of the true target.

To accurately track the true target, this paper uses the optimal correlation filter obtained in Sec. 3.1.1 to redetect the multipeak position in the color tracking response to determine the true target. The multipeak position $\{x_p^n | n \in 1, \ldots, N_1\}$ of the color tracking response is determined first, where N_1 is the number of peak positions. The multipeak position is determined using Eq. (28):

$$S_{\text{hist}}(x_p) > \theta_1 \max[S_{\text{hist}}(x)], \tag{28}$$

where θ_1 is assigned 0.8. Then, the local search area $\{D_t^n | n \in 1, \ldots, \mathrm{N}_1\}$ is selected in x_p^n. The optimal correlation filter is used to redetect the local area to obtain the multipeak redetection response set $\{S_{cf}^n(x) | n \in 1, \ldots, N_1\}$, where the peak position of $\max[S_{cf}^n(x)]$ is the target center position P_t.

3.2 *Target Tracking in Occlusion*

3.2.1 *Judge the degree of occlusion of the target*

Fusion dual model target tracking response $[S_f(x)]$ is shown in Fig. 6. Original images are shown in the first row in Fig. 6. The target is a pedestrian and is obscured by the car during tracking. The fusion tracking response maps $[S_f(x)]$ corresponding to the local search area are shown in the second row. In the first column of Fig. 6, the target is not occluded and there is only a single peak corresponding to the target in the tracking response map. Except for the sharp peak at the center of the target, the rest of the map is relatively smooth. In the second column, the target is partially occluded, and there are many peaks in the tracking response map, with no single maximum peak value and with large fluctuations. In the third column, the target is more occluded, and an additional peak appears in the tracking response map, with an even larger overall fluctuation.

To cope with this problem, an indicator OCC for judging the degree of occlusion of the target is proposed:

$$\mathrm{OCC} = \frac{\max[S_f(x)] - \min[S_f(x)]}{\operatorname{mean}\left\{\sum_{W,H} |S_f(x) - \min[S_f(x)]|\right\}}, \tag{29}$$

where W and H represent the width and the height, respectively, of the response map corresponding to the local search area. This indicator reflects the degree of smoothness of the response map and the confidence level that the peak is in the center of the target.

In the process of target tracking, the value of OCC, which is used to judge the degree of occlusion of the target, is

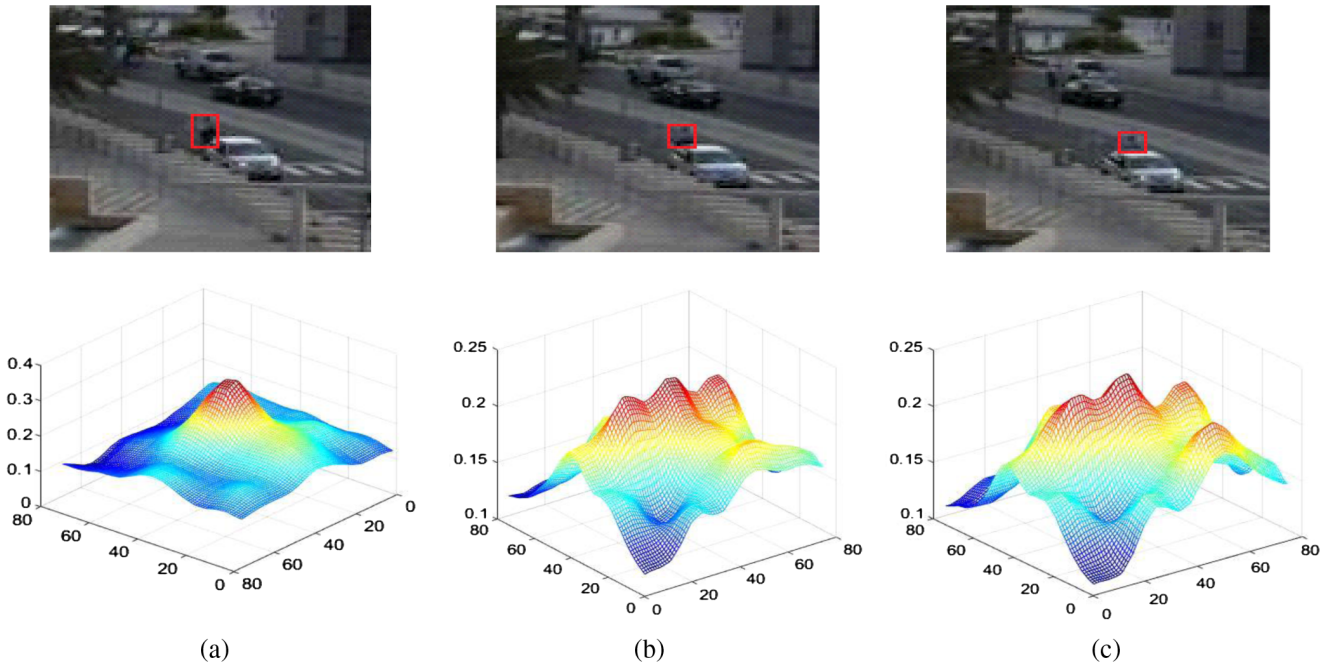

Fig. 6 Fusion of dual model tracking result and tracking response map: (a) OCC = 3.0472, (b) OCC = 2.2965, and (c) OCC = 2.2811.

shown in the third row of Fig. 6. The first column has the largest OCC value. The second column has a smaller OCC value, and the third column has an even smaller OCC value. Figure 5 shows that the OCC value can be used to judge the degree of occlusion of the target.

In this paper, when the OCC value of the I_t frame is less than β times the OCC value of the I_{t-1} frame, the target is considered occluded. Here, β is assigned 0.8.

3.2.2 *Redetect the occluded target*

A small target is easily obscured. When it is judged that the target is occluded by the occlusion indicator OCC, the local search area is expanded to $2D_t$. First, the color primary detection is performed to determine the multipeak position $\{x_p^n | n \in 1, \ldots, N_1\}$, where N_2 is the number of peak positions. The multipeak position is determined by Eq. (30):

$$S_{\text{hist}}(x_p) > \theta_2 \max[S_{\text{hist}}(x)], \tag{30}$$

where θ_2 is assigned a value of 0.7.

Then, the local search area $\{D_t^n | n \in 1, \ldots, N_1\}$ is selected in x_p^n. The preocclusion correlation filter is used to redetect the local area to obtain the multipeak redetection response set $\{S_{cf}^n(x) | n \in 1, \ldots, N_1\}$, where the peak position of $\max[S_{cf}^n(x)]$ is the target center position P_t.

4 Flowchart of the Proposed Method

The flowchart of the proposed real-time UAV tracking of a fast moving small target on the ground is shown in Fig. 7.

In the target tracking process, first, the center of the local search area D_t in the current frame I_t is set to be the center position P_{t-1} of the target tracking result of the I_{t-1} frame. The correlation filter tracking response $S_{cf}(x)$ and the color tracking response $S_{\text{hist}}(x)$ of the pixel x in D_t are calculated, and the score $S_f(x)$ is obtained by adaptively combining the two responses. The peak position of $S_f(x)$ is the same as the target center position P_t of the I_t frame. In the fast motion state, the proposed method uses the optimal correlation filter to redetect the multipeak position of the color tracking response to determine the true target. The optimal correlation filter is used to redetect the local area to obtain the multipeak redetection response set $\{S_{cf}^n(x) | n \in 1, \ldots, N_1\}$. The peak position of $\max[S_{cf}^n(x)]$ is the target center position P_t. In the occlusion state, the color primary detection is performed. The pre-occlusion correlation filter is used to redetect the local area to obtain the multipeak redetection response set $\{S_{cf}^n(x) | n \in 1, \ldots, N_1\}$ in the multipeak position. The peak position of $\max[S_{cf}^n(x)]$ is the target center position P_t.

5 Experimental Results and Analysis

A set of video sequences containing small targets, fast motion, and occlusion characteristics from the database UAV123 are selected for the experiment, including a total of 15 groups and 6611 images.[19] The image size is 1280×720 pixels. The tracking targets include people, cars, and other objects. All targets have fine manual annotation. The proposed method of this paper is compared with eight other state-of-the-art methods, including the CN tracker that uses color attributes as effective features,[2] the KCF tracker that uses the multichannel HOG feature,[4] the DSST tracker that relieves the scaling issue using the feature pyramid and the three-dimensional (3-D) correlation filter,[16] the LCT tracker that uses the online random fern classifier as the redetection component for long-term tracking,[7] the DAT tracker that uses the color histogram feature and suppresses the background area,[20] and the Staple tracker that fuses the color tracker and correlation filter tracker linearly.[18] The above-mentioned six methods have outstanding tracking results, and the speed of tracking meets the real-time requirement. Also, the MEEM[8] tracker

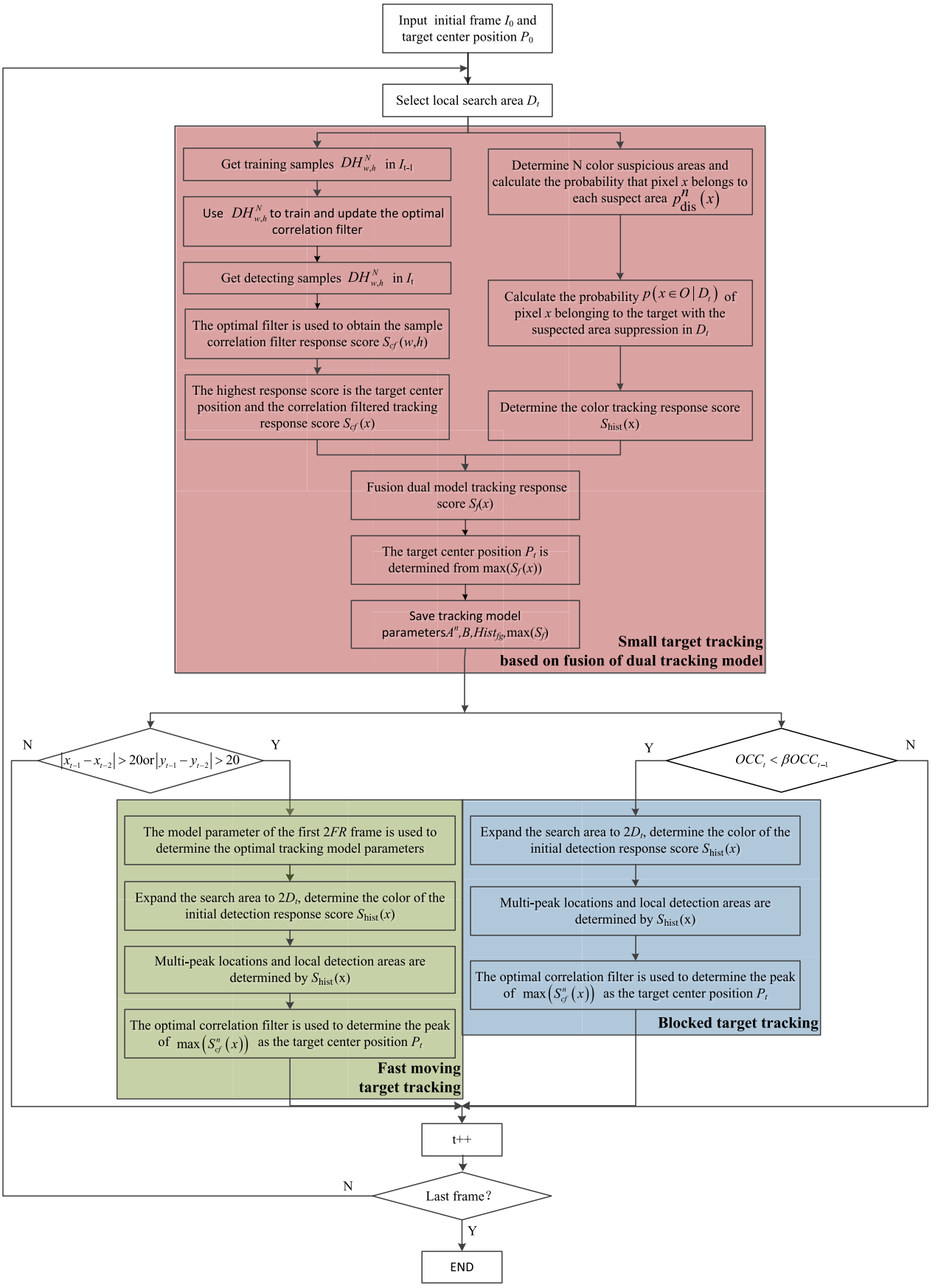

Fig. 7 Flowchart of real time UAV tracking of a fast moving small target on ground.

that uses the multiple tracker expert group to realize fast tracking, and the 2016 VOT Challenge champion CCOT[6] that uses the feature of deep convolution neural network are included in the comparison experiments.

All experimental results and related performance evaluations were obtained using the same data and initialization conditions. Experimental environment: Matlab2016; experimental platform: 3.60 GHz, Intel i7 CPU, 64-bit win7 operating system, with 8GB of memory.

Datasets come from Ref. 21.

5.1 *Comparison of Tracking Results for the Different Methods*

Tracking performance of the proposed method and CN, KCF, DSST, DAT, LCT, Staple, MEEM, and C-COT are compared in the video sequence set in which the target is small, fast moving, and occluded as shown in Fig. 8. The white box is the location of the true value of the target, which is used to compare with the tracking position obtained by the algorithm.

The targets in the first and second rows of Bike2 and Truck4 were <200 pixels in size, and the target in the third row in Car11 were <100 pixels in size. Experimental results show that the tracking boxes of the other eight methods easily lose the target or drift to suspected objects. The proposed method is able to better characterize small targets with little feature information because of the adaptive fusion of multifeature models, improving the success rate of target tracking. In the third and the fourth rows of Bike3 and Car11, the targets are blocked and the other eight methods were unable to deal with occlusion, resulting in failure in tracking. The proposed method efficiently judges whether the target is occluded and initiates the corresponding tracking method when occlusion is detected, ensuring successful target tracking. In the fifth and the sixth rows of Wakeboard5 and Car14, the between-frame target position distance is >20 pixels. The other eight methods lost the target under this situation. The proposed method efficiently judges whether the target is in fast motion and initiates the corresponding tracking method when fast motion is detected, ensuring successful target tracking. In the seventh and the eighth rows of Car13 and Truck3, there are strong interfering objects near the true target, and most of the eight methods failed to track. The proposed method suppressed the suspected areas effectively and greatly reduced the interference from the suspected areas. It can resist the impact of strong interfering objects on small targets and track the target successfully.

5.2 *Performance Comparison Experiment*

5.2.1 *Experiment of overlap success rate*

If the overlap score of the tracking result of the I_t frame is beyond a given threshold, it is considered that the proposed method has successfully tracked the target in the I^t frame. The overlap success rate[1] is the ratio of the number of successful tracking frames to the total number of frames. The overlap score is defined in Eq. (23).

The comparison of the overlap success rate of the proposed method with that of the other eight methods is shown in Fig. 9. In this paper, the area under curve (AUC) of the overlap success rate curve was used to evaluate the performance of the tracking methods because it is considered

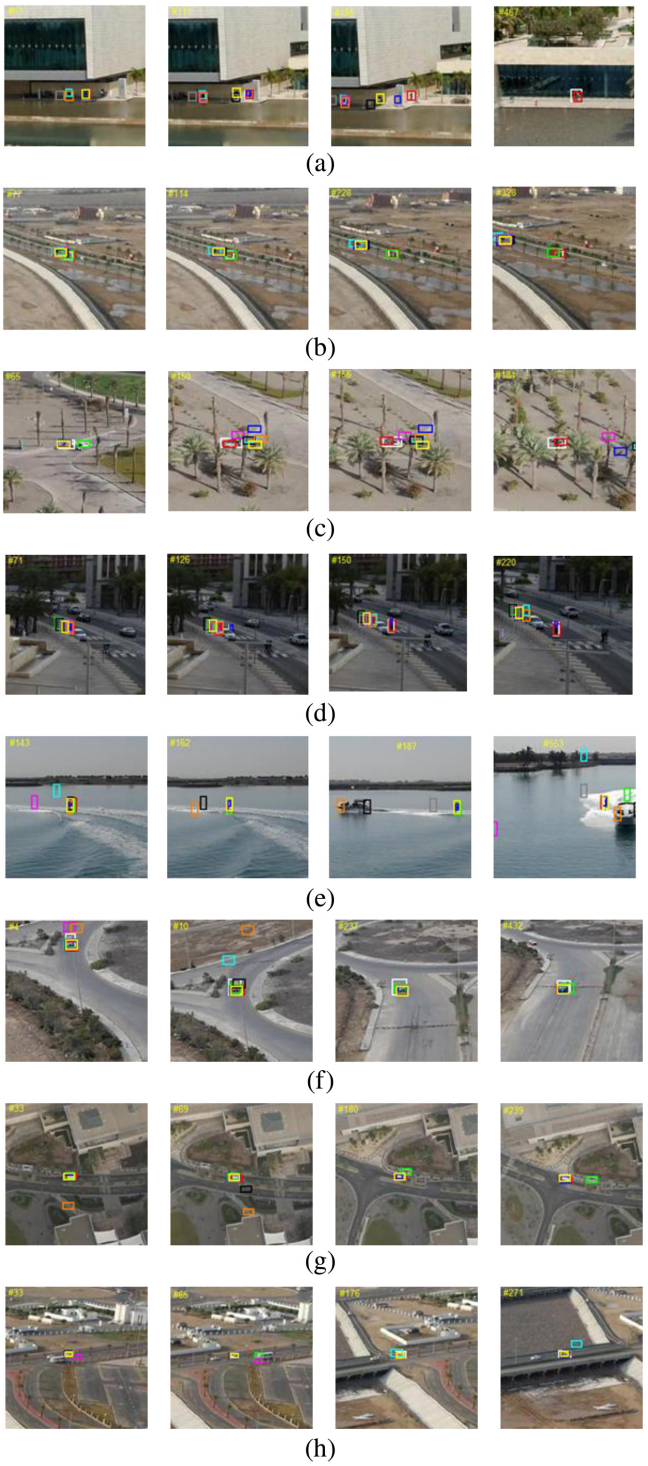

Fig. 8 Comparison of tracking results of the proposed method and state state-of-the-art tracking methods: (a) Bike2, (b) Truck4, (c) Car11, (d) Bike3, (e) Wakeboard5, (f) Car14, (g) Car13, and (h) Truck3.

a more accurate evaluation of the overall tracking performance. The AUC values of all methods tested are listed after each method name in the figure legend of Fig. 9.

As shown in Fig. 9, the proposed method in this paper has the highest AUC, indicating that the performance of the proposed method has a high overlap success rate. When the overlap threshold is <0.5, the overlap success rate of the proposed method is substantially higher than that of the other

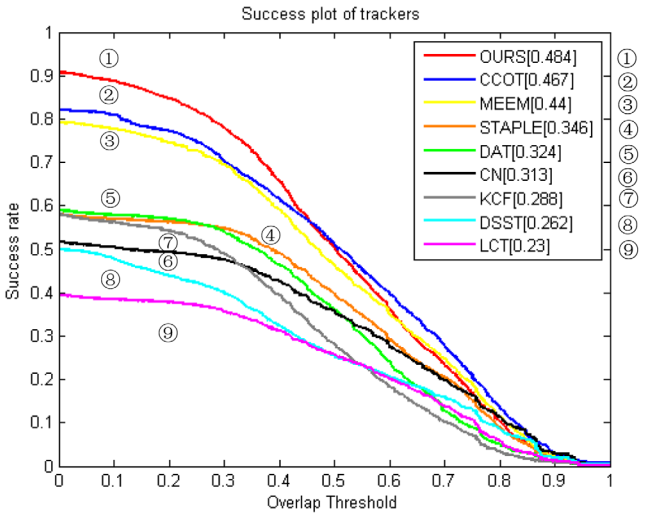

Fig. 9 Comparison of the overlap success rate of the proposed method with the other eight methods.

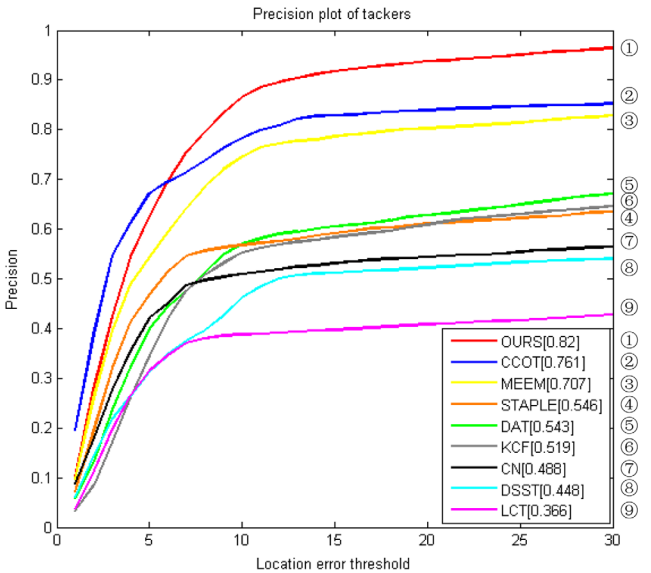

Fig. 10 Comparison of the distance precision rate of the proposed method with the other eight methods.

methods. However, the success rate of the proposed method is slightly lower than that of the CCOT when the overlap threshold is high. This is because the proposed method mainly aims at small targets; hence, it does not adopt a complex scale adaptive strategy.

5.2.2 *Experiment of distance precision rate*

If the Euclidean distance between the center of the I_t frame tracking result and the given target center is within a given location error threshold, it is considered that the proposed method has tracked the target precisely in the I_t frame. The distance precision rate is the ratio of the number of precise tracking frames to the total number of frames. The comparison of the distance precision rate of the proposed method with that of the other eight methods is shown in Fig. 10. The horizontal axis denotes the location error threshold and the vertical axis denotes the distance precision rate. The AUC value is again used as the evaluation index because it more accurately evaluates the overall performance of the methods. The AUC values of all methods tested are listed after each method name in the figure legend of Fig. 10.

As shown in Fig. 10, the proposed method in this paper has the highest AUC, indicating that the performance of the proposed method has a high distance precision rate. When the location error threshold is <5, the distance precision rate of the proposed method is substantially higher than that of the other methods. However, the success rate of the proposed method is slightly lower than that of the CCOT when the location error threshold is small. This is because the proposed method mainly aims at small targets; hence, it does not adopt a complex scale adaptive strategy.

5.2.3 *Experiment of average center location error*

The center location error is the average Euclidean distance between the center of the tracking result and the given target center. Table 2 shows the center location error of the proposed method and the other eight methods.

Table 2 shows that the average center location error of the proposed method is much smaller than that of the other eight methods. It shows that the tracking performance of this method is better than that of the other methods.

5.2.4 *Comparison of the real-time performance between methods*

The proposed method is compared with the other eight methods for real-time performance. The frames per second (fps) is

Table 2 Comparison of the center location error of the proposed method with that of the other eight methods.

Tracker	Proposed	CCOT	MEEM	STAPLE	DAT	KCF	DSST	CN	LCT
ACLE	7.27	25.82	29.39	63.18	70.58	79.43	88.32	89.75	138.80

Table 3 Comparison of the frames per second of the proposed method with that of the other eight methods.

Tracker	Proposed	CCOT	MEEM	STAPLE	DAT	KCF	DSST	CN	LCT
Fps	40.23	2.58	6.17	66.47	20.33	142.62	32.62	87.79	23.11

used to evaluate real-time performance. The fps of each method is shown in Table 3.

According to Table 3, when compared with CCOT, MEEM, DAT, DSST, and LCT, the proposed method has higher fps, indicating that the proposed method has better real-time performance. The fps of methods STAPLE, KCF, and CN are higher than the proposed method; however, the proposed method performance is superior to them due to its multifeature model and strategies for coping with small target fast motion and occlusion.

6 Conclusion

An adaptive algorithm that fuses the improved color histogram tracking response and the correlation filter tracking response based on multichannel HOG features is proposed to realize small target tracking with high accuracy. The state judgment index is used to determine whether the target is in fast motion or an occlusion state. In the fast motion state, the search area is enlarged, and the color optimal model that suppresses the suspected area is used for rough detection. Then, redetection in the place of multiple peaks in the rough detection response is carried out using the correlation filter to accurately locate the target. In the occlusion state, the model stops updating, the search area is expanded, and the current color model is used for rough detection. Then, redetection in the place of multiple peaks in the rough detection response is carried out using the correlation filter to accurately locate the target. The proposed method of this paper is compared with the other state-of-the-art methods using the UAV123 dataset. Experimental results show that the proposed method can accurately track a fast moving small target in real time. The fps of the proposed method is 40.23 indicating good real-time performance. In this paper, single target tracking is studied. In future research, multitarget tracking will be studied. Based on multitarget time-domain information and airspace information, an accurate real-time tracking method for UAV multitarget tracking will be developed.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61471194 and 61705104), Science and Technology on Avionics Integration Laboratory and Aeronautical Science Foundation of China (20155552050), and the Natural Science Foundation of Jiangsu Province (BK20170804).

References

1. Y. Wu, J. Lim, and M. H. Yang, "Online object tracking: a benchmark," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2411–2418 (2014).
2. M. Danelljan et al., "Adaptive color attributes for real-time visual tracking," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1090–1097 (2014).
3. J. F. Henriques et al., "Exploiting the circulant structure of tracking-by-detection with kernels," in *European Conf. on Computer Vision*, pp. 702–715 (2012).
4. J. F. Henriques et al., "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(3), 583–596 (2015).
5. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 886–893 (2015).
6. M. Danelljan et al., "Beyond correlation filters: learning continuous convolution operators for visual tracking," in *European Conf. on Computer Vision*, pp. 472–488 (2016).
7. C. Ma et al., "Long-term correlation tracking," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 5388–5396 (2015).
8. Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(7), 1409–1422 (2012).
9. J. Zhang, S. Ma, and S. Sclaroff, "MEEM: robust tracking via multiple experts using entropy minimization," *Lect. Notes Comput. Sci.* **8694**, 188–203 (2014).
10. G. Zhu, F. Porikli, and H. Li, "Beyond local search: tracking objects everywhere with instance-specific proposals," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 943–951 (2016).
11. C. L. Zitnick and P. Dollár, "Edge boxes: locating object proposals from edges," *Lect. Notes Comput. Sci.* **8693**, 391–405 (2014).
12. X. Jia, "Visual tracking via adaptive structural local sparse appearance model," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1822–1829 (2012).
13. Q. Zhao et al., "Object tracking via kernel-based forward-backward keypoint matching," *Proc. SPIE* **10225**, 1022504 (2017).
14. J. H. Yan et al., "Target tracking with improved CAMShift based on Kalman predictor," *J. Chin. Inertial Technol.* **22**(4), 537–542 (2014).
15. R. M. Gray, "Toeplitz and circulant matrices: a review," *Found. Trends Commun. Inf. Theory* **2**(3), 155–239 (2005).
16. M. Danelljan et al., "Discriminative scale space tracking," *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1561–1575 (2017).
17. H. K. Galoogahi, T. Sim, and S. Lucey, "Multi-channel correlation filters," in *IEEE Int. Conf. on Computer Vision*, pp. 3072–3079 (2014).
18. L. Bertinetto et al., "Staple: complementary learners for real-time tracking," in *IEEE Conf. on Computer Vision and Pattern Recognition* (2016).
19. M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," *Lect. Notes Comput. Sci.* **9905**, 445–461 (2016).
20. H. Possegger, T. Mauthner, and H. Bischof, "In defense of color-based model-free tracking," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2113–2120 (2015).
21. Image and Video Understanding Lab of King Abdullah University of Science and Technology, "UAV123 Dataset," 2017, https://ivul.kaust.edu.sa/Pages/Dataset-UAV123.aspx

Junhua Yan is a professor at Nanjing University of Aeronautics and Astronautics. She has been an academic visitor at University of Sussex (October 31, 2016–October 30, 2017). She received her BSc and MSc degrees and PhD from Nanjing University of Aeronautics and Astronautics in 1993, 2001, and 2004, respectively. She is the author of more than 40 journal papers and has 5 patents. Her current research interests include image quality assessment, multisource information fusion, target detection, tracking, and recognition.

Jun Du is a graduate student at Nanjing University of Aeronautics and Astronautics. Her research interest is target tracking. She received her BSc degree from Nanjing University of Aeronautics and Astronautics in 2016.

Yong Young received his BSc degree and MSc degree from Nanjing University of Aeronautics and Astronautics in 2015 and 2018, respectively. His research interest is target tracking. Currently, he is working at HIKVISION in China.

Christopher R. Chatwin is a director of the IISP research group, he has published a total of 220 journal papers, 230 conference papers, 10 book chapters, 2 books, and 112 technical reports. His patents (PCT-PN-WO03/073366, US10/504771, JPN2003-571984, EU03708323.5, and UK2389901) form the basis for new patent application in labelling technology for brand protection. He is a course convenor for the new MSc in security technologies and systems. He is a PhD external-examiner at the universities of Cambridge, Hull, Glasgow, Liverpool, Cairo, Singapore, Ghent, and Lulea.

Rupert C. D. Young is a reader and the head of the department. He has published a total of 120 journal papers and 133 conference papers. He is an editor of the *Asian Journal of Physics*. He is in the organizing committee and has chaired sessions for the Optical Pattern Recognition conference over the last 10 years, as a part of the annual SPIE Defense and Security Symposium, Orlando, Florida. He is a member of the OSA and SPIE.

Philip Birch is a senior lecturer. He has published a total of 70 journal papers and 80 conference papers, and has made major contributions to the research group in optoelectronics and image processing. He left for 3 years to work as a project manager in a start-up company called Spiral Scratch Ltd. He also acted as the Liverpool University KTP facilitator and developed links with Sheffield Hallam University.