The Research of Unmanned Aerial Vehicle Video Fusion Evaluation Method Based on Structure Similarity

Jun Xie[1, a], Xiuying Fan [1, b], Yingnan Liu[2, c], Zhonglin Xu [1, d] and Shuang Wen[1,e]

[1] Air Force Aviation University, Changchun, Jilin, 130021, China

[2] ChangChun Automobile Industry Institute, Changchun, Jilin, 130013, China

[a]xiejun1981@gmail.com, [b]fanxiuying@163.com, [c]liuyingnan2005@163.com, [d]xu_zhonglin@sina.com, [e]wenshuang123@163.com

Keywords: Unmanned aerial vehicle, Video fusion, Structure similarity, Evaluation factors

Abstract. Through carrying a variety of imaging sensors, the unmanned aerial vehicle began to show its flexibility and importance, image fusion technology becomes more and more indispensible. In this paper, the existing video fusion method and fusion performance evaluation method is analyzed, and an improved video fusion evaluation algorithm based on structure similarity was proposed. The simulation results were shown that the improved fusion performance evaluation method was better than the conventional method, and a method to solve the fusion performance evaluation problem was provided.

Introduction

Through carrying a variety of imaging sensors, the unmanned aerial vehicle's ability of autonomous navigation, battlefield reconnaissance, combat visual assessment and tracking target search task were improved. Unmanned aerial vehicle not only can adapt to changes in the complex environment conditions, but also can obtain abundant scene space information. At present the unmanned aerial vehicle imaging sensor can get visible light and infrared video image, and the video images are widely used in the process of practical application. Because of the weather, light and other environmental conditions, the single image sensor were restricted and it was not suitable for all-weather work, meanwhile, the imaging principle of each sensor were different, and the characteristic information of the images were also different.

Video Fusion

Image fusion technology can integrate two complementary information of different kinds of images, and the obtained fusion images is by integration of their respective advantages, the single sensor in application environment, using range and the limitation of the target acquisition were overcome, at the same time the image spatial resolution and clarity can be improved, the image understanding and recognition were facilitated, the using efficiency of image data were effectively improved. Video is a set of continuous image sequence according to the time sequence actually, it is using the principle of human visual temporarily leave to make the eye with movement feeling by playing a series of images. It is a kind of the most abundant information, intuitive, vivid and concrete bearing information of the media; Video fusion technology is through the integration of multiple visual sensors at the same time or different time to get on the same specific scenarios of video information, so as to enrich video details and enhance cognitive effect. Video fusion technology is widely used in military activity with the rapid development of science and technology. Video fusion and image fusion are the same purpose, in order to obtain the same scene or target with more accurate, comprehensive and reliable description of video images [1].

Improved Video Fusion Evaluation Algorithm

The algorithm is mainly based on the structure similarity information of the image. Due to the human eye vision system's sensitivity to the structure information of the image, and image structure information loss can well reflect the image distortion, so we can use the structure similarity information of the input images and fused image to evaluate the performance of image fusion algorithm. Structure similarity is defined by Wang [2,3], for image A and image B, the structure similarity is defined:

$$SSIM(A,B) = [l(A,B)]^{\alpha}[c(A,B)]^{\beta}[s(A,B)]^{\gamma} = (\frac{2\mu_A\mu_B + C_1}{\mu_A^{\ 2} + \mu_B^{\ 2} + C_1})^{\alpha}(\frac{2\sigma_A\sigma_B + C_2}{\sigma_A^{\ 2} + \sigma_B^{\ 2} + C_2})^{\beta}(\frac{\sigma_{AB} + C_3}{\sigma_A\sigma_B + C_3})^{\gamma} \quad (1)$$

In which the l (A, B), c (A, B) and s (A, B) are brightness, contrast and correlation coefficient respectively, for, μ_A and μ_B are the average of the image A and image B respectively, σ_A and σ_B are the variances of the image A and image B respectively, σ_{AB} is the covariance of image A and image B, α, β and γ can be adjusted according to the importance of each part, C_1, C_2 and C_3 are constant, they are used to avoid the condition of zero denominator. Make $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$, the Eq.1 come to:

$$SSIM(A,B) = (\frac{2\mu_A\mu_B + C_1}{\mu_A^{\ 2} + \mu_B^{\ 2} + C_1})(\frac{2\sigma_{AB} + C_2}{\sigma_A^{\ 2} + \sigma_B^{\ 2} + C_2}) \quad (2)$$

The video fusion performance evaluation methods based on structure similarity and human vision can evaluate video fusion performance from two aspects of information extraction of time and space and time consistency, and the evaluation results are more close to the subjective evaluation.Based on structure similarity of SSIM, the space fusion performance evaluation index were build, the time performance evaluation index were build according to the SSIM values of each frame difference image between fused video and input video; so the video fusion algorithm performance were evaluated comprehensive [2].

The specific implementation steps include four steps:

First, to build a single frame space fusion performance evaluation factors according to the SSIM values between each frame image of the fused video and the input video. The building formula is [4]:

$$Q_{S,t}(V_a, V_b, V_f) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(\lambda_a(w_{i,j,t})(SSIM(V_a, V_f \mid w_{i,j,t})) + \lambda_b(w_{i,j,t})(SSIM(V_b, V_f \mid w_{i,j,t}))}{\sum_{i=1}^{I}\sum_{j=1}^{J}(\lambda_a(w_{i,j,t}) + \lambda_b(w_{i,j,t}))} \quad (3)$$

In which the input videos are V_a and V_b, the fused video is V_f, the $I \times J$ is the size of each frame, $(w_{i,j,t})$ is the local window of the t frames in the image space of (i, j); $\lambda_a(w_{i,j,t})$ and $\lambda_b(w_{i,j,t})$ are the weights of input video V_a and V_b under the current window. The $SSIM(V_a, V_f \mid w_{i,j,t})$ and $SSIM(V_b, V_f \mid w_{i,j,t})$ are the structure similarity values between the fused video and the input video. It can be obtained from Eq.2:

$$\lambda_a(w_{i,j,t}) = \log\left[1 + \frac{\sigma_{Va}^2(w_{i,j,t})}{C_t(V_a)}\right] \qquad \lambda_b(w_{i,j,t}) = \log\left[1 + \frac{\sigma_{Vb}^2(w_{i,j,t})}{C_t(V_b)}\right] \quad (4)$$

In which the $\sigma_{Va}^2(w_{i,j,t})$ and $\sigma_{Vb}^2(w_{i,j,t})$ are the variance of V_a and V_b under the current local window $w_{i,j,t}$, it is used to represent video signal strength under the current window, $C_t(V_a)$ and $C_t(V_b)$ are the noise intensity of video image and the current frame image respectively.

Second, to build a single frame time fusion performance evaluation factors according to the SSIM values between each frame difference image of fused video images and the input video images.

$$Q_{T,t}(D_a, D_b, D_f) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(\xi_a(w_{i,j,t})(SSIM(D_a, D_f \mid w_{i,j,t})) + \xi_b(w_{i,j,t})(SSIM(D_b, D_f \mid w_{i,j,t}))}{\sum_{i=1}^{I}\sum_{j=1}^{J}(\xi_a(w_{i,j,t}) + \xi_b(w_{i,j,t}))} \tag{5}$$

In which D_a、D_b and D_f are the frame video images of V_a、V_b and V_f respectively, it can be calculated by the Eq.6:

$$D_I(i,j,t) = V_I(i,j,t) - V_I(i,j,t-1) \qquad i = a,b,f \tag{6}$$

$\xi_a(w_{m,n,t})$ and $\xi_b(w_{m,n,t})$ are the weights of input frame video image D_a and D_b under the current window; $SSIM(D_a, D_f \mid w_{i,j,t})$ and $SSIM(D_b, D_f \mid w_{i,j,t})$ are the structure similarity values between D_f and D_a、D_b under the current partial window respectively:

$$\xi_a(w_{m,n,t}) = 1 + \log(1 + \| v_a(m,n,t) \|) \qquad \xi_b(w_{m,n,t}) = 1 + \log(1 + \| v_b(m,n,t) \|) \tag{7}$$

In which $v_a(m,n,t)$ and $v_b(m,n,t)$ are the local motion vector of V_a and V_b in the current spatial-time location, $\|...\|$ is the motion vector norm.

Third, to combine the space and time performance evaluation factors, to build a single frame spatial- time fusion performance evaluation factor:

$$Q_t(V_a, V_b, V_f) = Q_{S,t}(V_a, V_b, V_f)^{\vartheta} \cdot Q_{T,t}(D_a, D_b, D_f)^{1-\vartheta} \qquad \vartheta \in [0, 1] \tag{8}$$

Fourth, the performance evaluation factors of all the single frame spatial-time were weighted average, to build a global spatial-time fusion performance evaluation factor:

$$Q(V_a, V_b, V_f) = \frac{\sum_{t=1}^{T} Q_t(V_a, V_b, V_f) \cdot \rho(t)}{\sum_{t=1}^{T} \rho(t)} \tag{9}$$

In which T is the number of frames in video image, $\rho(t)$ is the global frame weight:

$$\rho(t) = \max(\rho_a(t), \rho_b(t)) \tag{10}$$

In which $\rho_a(t)$ and $\rho_b(t)$ are the global frame weights of the current frame image in V_a and V_b:

$$\rho_a(t) = \log\left[1 + \frac{1 + c_{g,a}(t)}{1 + v_{g,a}(t)}\right] \qquad \rho_b(t) = \log\left[1 + \frac{1 + c_{g,b}(t)}{1 + v_{g,b}(t)}\right] \tag{11}$$

In which, $v_{g,a}(t)$、$c_{g,a}(t)$ and $v_{g,b}(t)$、$c_{g,b}(t)$ are the global movement rate and strength contrast of V_a and V_b in current frame image respectively [5].

The video fusion performance algorithm based on the structure similarity which described could evaluate fusion algorithm performance from two aspects of information extraction of time and space

and time consistency, and combined with the human visual characteristic to make the objective evaluation results more close to the human eye subjective evaluation results. The process of the fusion performance evaluation method was shown in Fig.1.

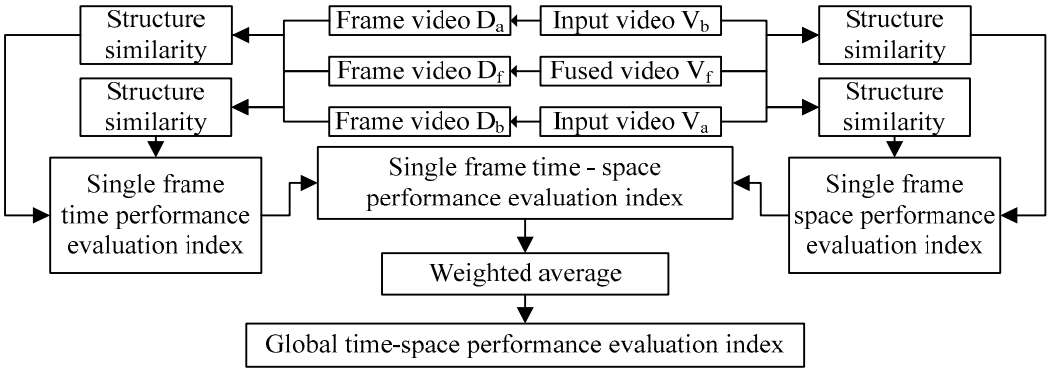

Fig.1 The process of the fusion performance evaluation method

The Algorithm Simulation Results and Analysis

In order to verify the validity and superiority of the algorithm, three different kinds of conventional multi-scale video fusion method were used, they are fusion methods of 3D-CWT, DWFT and DWT. In these methods, all of the low-pass subband coefficients were the average fusion rules, high-pass subband coefficients were taken absolute maximum fusion rules.

The videos used in the experiment were obtained from the Unmanned Aerial Vehicle. The three methods were used to fusion video sources respectively, and then the proposed fusion performance evaluation method was compared with the existing three methods. The results data were shown in Tab.1. We can see that the fusion performance which proposed was better than conventional methods.

Tab.1 The comparison results of the different fusion performance evaluation method

Video groups	Fusion method / Evaluation method	3D-CWT	DWFT	DWT
Video group1	Conventional method	0.3541	0.3690	0.3452
	Proposed IVFE method	0.7913	0.7830	0.7651
Video group2	Conventional method	0.2576	0.2780	0.2415
	Proposed IVFE method	0.8132	0.8040	0.7805
Video group3	Conventional method	0.2101	0.1821	0.1546
	Proposed IVFE method	0.7650	0.7541	0.7068

Conclusions

Aiming at the shortcomings of the existing video fusion performance evaluation method, we put forward an improved video fusion performance evaluation method based on structure similarity. The fused images were decomposed in the method, and the space-time consistency information of three dimensional was constructed from the input video images. The weighted average was made to all of the fusion performance evaluation factors and to build the global spatial and temporal consistency fusion performance evaluation factors. In combination with energy and three dimensional gradient structure tensor characteristics needs of local and global design parameters, makes the objective evaluation results more accurate and has better robustness for noise.

References

[1] W.R.Hendee, P. N.Well. *The Perception of Visual Information* (Springer, Berlin 1997).

[2] Z.Wang, A. C. Bov, H. R. She, Image Processing, vol.13(2004):p. 1-14.

[3] Z. Wang, Q. Li, The Optical Society of America, vol.24(2007): p. B61–B69.

[4] Q. Zhang,L. Wang,H. Li, Signal Processing, vol.92(2012):p. 912-925.

[5] A. A.Sto, E. P.Sim, Nature Neuroscience, vol.9(2006): p. 578–585.