Information Fusion 24 (2015) 54–71

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Multisensor video fusion based on higher order singular value decomposition

Qiang Zhang [a,b,*], Yabin Wang [b], Martin D. Levine [c], Xiaoqing Yuan [b], Long Wang [d]

[a] Key Laboratory of Electronic Equipment Structure Design (Xidian University), Ministry of Education, Xi'an, Shaanxi 710071, China
[b] Center for Complex Systems, School of Mechano-electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China
[c] Center for Intelligent Machines, Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 2A7, Canada
[d] Center for Systems and Control, College of Engineering, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Available online 22 October 2014

Keywords:
Video fusion
Video tensor
Higher order singular value decomposition
Surfacelet Transform

ABSTRACT

With the ongoing development of sensor technologies, more and more kinds of video sensors are being employed in video surveillance systems to improve robustness and monitoring performance. In addition, there is often a strong motivation to simultaneously observe the same scene by more than one kind of sensor. How to sufficiently and effectively utilize the information captured by these different sensors is thus of considerable interest. This can be realized using video fusion, by which multiple aligned videos from different sensors are merged into a composite.

In this paper, a video fusion algorithm is presented based on the 3D Surfacelet Transform (3D-ST) and the higher order singular value decomposition (HOSVD). In the proposed method, input videos are first decomposed into many subbands using the 3D-ST. Then the relevant subbands from all of the input videos are merged to obtain the corresponding subbands of the intended fused video. Finally, the fused video is constructed by performing the inverse 3D-ST on the merged subband coefficients. Typically, the spatial information in the scene backgrounds and the temporal information related to moving objects are mixed together in each subband. In the proposed fusion method, the spatial and temporal information are actually first separated from each other and then merged using the HOSVD. This is different from the currently published fusion rules (e.g., spatio-temporal energy "maximum" or "matching"), which are usually just simple extensions of static image fusion rules. In these, the spatial and temporal information contained in the input videos are generally treated equally and merged by the same fusion strategy. In addition, we note that the so-called "scene noise" in an input video has been largely ignored by the current literature. We show that this noise can be distinguished from the spatio-temporal objects of interest in the scene and then suppressed using the HOSVD. Clearly, this would be very advantageous for a surveillance system, particularly one dealing with scenes of crowds.

Experimental results demonstrate that the proposed fusion method exhibits a lower computational complexity than some existing published video fusion methods, such as the ones based on the structure tensor and the pulse-coupled neural network (PCNN). When the videos are noisy, a modified version of the proposed method is shown to perform better than specialized methods based on the Bivariate-Laplacian model and the PCNN.

1. Introduction

Video surveillance plays an important role in modern society and has been widely applied in many fields [1]. Traditionally, the color video sensor has been the only modality employed. It can work well under ideal conditions but is inadequate in certain cases, such as scenes with poor lighting and smoke or dust [2]. These issues can often be addressed by simultaneously employing multiple modality aligned video sensors to capture the contents of the same scene [2–4]. Thus, how to sufficiently and efficiently utilize the information captured from these different sensors is of considerable interest. In this paper, we discuss how this can be realized using video fusion, by which multiple aligned videos from different sensors are merged into a composite. The fused video contains more useful information than any of the individual input videos and can be used to better interpret the scene [2,3,5].

* Corresponding author at: P.O. Box 183, Department of Automatic Control, Xidian University, No. 2 South TaiBai Road, Xi'an, Shaanxi Province 710071, China. Tel.: +86 029 88231936.
E-mail address: qzhang@xidian.edu.cn (Q. Zhang).

http://dx.doi.org/10.1016/j.inffus.2014.09.008
1566-2535/

Fig. 1. Illustration of video fusion for scene surveillance. The first and second rows illustrate several frames from an infrared and a visible light video, respectively, both capturing the same scene. The last row illustrates the corresponding frames from the fused video, which is obtained by our proposed method. As shown in the last row, the background images (extracted from the visible light video) and the moving targets (extracted from the infrared video) are well-integrated in the fused video. Moreover, the noise is also suppressed during the fusion process.

Fig. 1 illustrates an example of video fusion for scene surveillance. As shown in the first row, the moving persons are quite visually evident in the images taken with an infrared video camera. However, the scene environment (e.g., the building and the road) is virtually invisible. In the second row of images, taken by a conventional video camera, we can hardly observe that there are also two men (or moving targets) in the scene. By fusing the two input videos using the method proposed in this paper, we are able to obtain a new processed video, in which the moving targets from the infrared camera and the background images (or the environment of the scene) from the visible light camera are well integrated. This is achieved without object detection. As indicated in the last row of Fig. 1, the fused video shows that there are two men walking across the scene, one walking in front of the building and the other towards it.

Numerous fusion methods, especially at the signal-level, have been proposed in the literature [6–8] to achieve a result similar to that of Fig. 1. However, most of them are only applicable to static images even though current surveillance systems are based on video. Thus dynamic images and video fusion would seem to be more desirable [9].

Since the existing video fusion algorithms are founded on individual frames, they are in fact independently fused frame by frame [2,9,10]. Thus the spatial information in videos has dominated the literature on video fusion, perhaps more aptly referred to as image fusion. Recently, some algorithms [11–13] have been proposed based on the non-separate tri-dimensional Multi-Scale Transform (3D-MST). Examples are the 3D Surfacelet Transform (3D-ST) [14], the 3D Uniform Curvelet Transform [15] and the 3D Shearlet Transform [16]. On the other hand, the temporal information in the input videos has usually been ignored during the fusion process [11,13]. As opposed to the approaches using individual frames, such fusion algorithms would simultaneously integrate multiple aligned video frames. Generally, we would expect that these methods would exhibit superior performance for extracting spatio-temporal information.

An important issue in this regard is how to actually merge the different subband coefficients of the input videos, i.e., the fusion rule. As with image fusion algorithms based on the 2D-MST, this is also central to video fusion methods based on the 3D-MST. We note that most of image fusion rules currently in the literature could also be extended to fuse videos from the standpoint of the spatio-temporal domain. An example would be the spatio-temporal energy matching fusion rule in [11]. However, a video obviously contains moving objects as well as stationary ones. And the temporal features generally arouse more attention than the spatial ones [17]. Nevertheless, these fusion rules treat spatial and temporal information similarly by using the same fusion strategies for both.

We have proposed an alternative approach in [12], where the fusion rule was based on a spatio-temporal structure tensor [18]. In this work, eigenvalue decomposition was first performed on the spatio-temporal structure tensor matrices. The resulting subbands of the input videos were then filtered into three types of regions (i.e., regions containing mainly (1) background spatial information, (2) moving objects or (3) non-salient spatio-temporal information). This was followed by a different fusion strategy specifically designed for each type of region. Such an approach produces better performance than some static image fusion rules but the improvement is at the cost of a greatly increased computational complexity.

In this paper, we suggest a novel video fusion algorithm based on the 3D-ST[1] and higher order singular value decomposition (HOSVD) [19–21]. As shown in Fig. 2, the proposed method contains three parts. First it employs the 3D-ST as the MST to decompose input videos into different subbands. Then corresponding subbands from each input video are fused and, finally, reconstructed by performing the inverse 3D-ST. This approach is different from [12] in that the identification of the spatial or temporal information is achieved globally rather than pixel by pixel, which greatly reduces the computational complexity.

In fact, as one of the more efficient tensor decomposition techniques, the HOSVD has been widely employed in many fields, such as image denoising [22], face recognition [23], and texture analysis [24]. Also, two image fusion methods based on the HOSVD were proposed in [25,26]. In the former, the authors constructed an image tensor using input image frames and employed the HOSVD to obtain a set of basis images. Then the fused image was determined by optimizing the projective coefficients of these basis images. In the latter, the authors exploited the HOSVD to define

[1] Other 3D-MSTs, such as the 3D Uniform Curvelet and Shearlet Transforms, could also be used.