Available online at www.sciencedirect.com
ScienceDirect
Information Fusion 8 (2007) 208–216
INFORMATION FUSION
www.elsevier.com/locate/inffus

Subjective tests for image fusion evaluation and objective metric validation

Vladimir Petrović *

Imaging Science and Biomedical Engineering, University of Manchester, Oxford Road, M13 9PT Manchester, United Kingdom

Received 13 September 2004; received in revised form 5 May 2005; accepted 7 May 2005
Available online 11 July 2005

Abstract

This paper focuses on the methodology for perceptual image fusion assessment through comparative tests and validation of objective fusion evaluation metrics. Initially, the theory of subjective fusion evaluation, adopted practice and methods to gauge relevance and significance of individual trials are examined. Further in this context, the methodology, experiences and results of a series of specific, subjective preference tests aimed at relative evaluation of fusion algorithms are presented. Test conditions and experimental procedure are described in detail and a number of explicit fusion metrics derived from the subjective test data are proposed. Relative fusion quality, fusion performance robustness (to content) and personal preference are all assessed by the metrics as different aspects of general image fusion performance. Finally, the methodology for subjective validation of objective fusion metrics using the reported test procedures is presented. In particular, explicit subjective–objective validation algorithms are defined and applied to a range of established objective measures of fusion performance in order to evaluate their subjective relevance.

Keywords: Image fusion evaluation; Fusion metric validation; Perceptual trials; Relative performance

1. Introduction

Multiple sensor modalities allow for increased robustness and enhanced performance in a growing range of modern imaging applications. In order to fully exploit the additional information they provide in the resulting multiple scene representations however, considerable processing effort is required. Furthermore, when the intended user is a human operator, displaying multiple image modalities simultaneously leads to confusion and information overload, while integrating information across a group of users is almost impossible [1–3]. As an answer to this problem information fusion has attracted a considerable amount of research attention [4]. In particular, the focus was on signal-level image fusion [1,5–9], an approach of the lowest abstraction level of the Global Fusion System Architecture [10] that aims to deal with the information overload by reducing the amount of image data used for viewing or further processing without forgoing the benefits of multisensor information. Signal-level fusion algorithms fuse (combine) multiple image modalities into a single fused image with an explicit aim to preserve in this image the content value from all available sensor modalities.

Past experiences have shown that, although it is relatively straightforward to fuse images, e.g. by simply averaging them, assessing the performance of fusion algorithms is much harder in practice. Consequently, whereas a plethora of image fusion algorithms have appeared in the literature only a handful of publications deals with its evaluation [11–19]. One of the main issues that has limited research in this area from its onset has been a distinct lack of image data both to test advanced algorithms and validate their performance relative to

* Tel.: +44 161 275 5140; fax: +44 161 275 5145.
E-mail address: v.petrovic@man.ac.uk

1566-2535/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.inffus.2005.05.001

one another and to formulate, test and validate efficient objective performance metrics. Moreover, the latter category also requires additional data in the form of fused images evaluated and annotated with known fusion quality or any other measure relating to fusion performance.

So far, subjective or perceptual image fusion evaluation trials, in which an audience of potential users is employed to evaluate a fusion system, have been the most reliable and trusted method of fusion assessment. A number of evaluation trials focusing on various aspects of fusion performance and different scenarios (applications) were reported in the literature [2,3,11–15]. Despite the difficulties in obtaining good validation data however, a small number of objective fusion performance metrics have also emerged [17–20]. These metrics are by enlarge theoretically based, however no effort has so far been made to evaluate them independently.

This paper deals with the methodology of relative, subjective evaluation of image fusion algorithms and subjective validation of objective fusion assessment metrics. In Section 2, an overview of existing methods and reported trials focuses on the theory, practical aspects and categorisation of subjective fusion evaluation as well as ways to gauge the significance and relevance of individual trials. Experiences and results of a number of specific subjective fusion evaluation preference trials are reported in Section 3. A number of explicit fusion metrics derived from the subjective results that assess a number of distinctive aspects of fusion for display are also proposed in this section. Subjective validation of a number of established objective fusion performance metrics is proposed through a number of subjective–objective validation methods in Section 4. Finally, Section 5 concludes with a discussion on the proposed subjective trial methodology and objective metric validation and provides some general directions for future development of image fusion assessment.

2. Perceptual image fusion evaluation

Subjective or perceptual evaluation trials have long been established as a reliable method for general image and video quality assessment with well established experimental procedures and practice [22,23]. With adequate modifications this approach has also been adopted for the assessment of image fusion results in a handful of perceptual fusion assessment studies reported thus far [2,3,11–15]. Within this context, the trials in which an audience of intended users evaluates fused images under tightly controlled conditions either by comparing them to each other or by performing specific visually oriented tasks provide a robust evaluation of fusion algorithms. Moreover, perceptual trials also provide the much needed "ground truth" for evaluation and validation of objective fusion metrics. The significance and consequently the impact of individual trials on fusion applications and fusion research in general however is determined by number of distinct factors such as their credibility and relevance. These factors are in turn determined by practical and organisational issues that define the conduct of tests such as test conditions and procedures, the role and size of the audience, as well as data on which the evaluation is based.

The relevance of a subjective test to a particular fusion application is defined by the role of the audience and tests can be either: (i) active or task related (quantitative) or (ii) passive or descriptive (qualitative). Most trials reported thus far belong to the former category [2,3,11–15], such as the trials by the US army conducted to evaluate the potential advantage of fused imagery in helicopter night flight [12,13] in which experienced pilots performed a number of pre-defined flying and unrelated visual tasks (e.g. target detection and recognition) in real and simulated flight conditions using fused and single sensor imagery. In another important contribution Toet et al. [2,11,14,15] performed series of perceptual tests, on both still and fused sequences, in which the subjects performed a variety of vision related tasks involving situational awareness in a number of different surveillance scenarios. Such active trials evaluate fusion performance quantitatively by measuring the subjects' performance in completing the given tasks (time taken and results accuracy). In [2] the average time taken to locate and identify a specific target in a fused scene and the probability of correct classification and false alarms formed a measure of how useful the fused image is as a visual cue. Such specificity makes active trials highly relevant in many active applications such as piloting, night time driving and security. Additionally, all tests can be directly formulated in accordance with the demands of a particular application to obtain highly relevant evaluation results.

Conversely, in passive tests, observers are simply asked to directly rate or rank fused images based on their impression, resulting in a qualitative evaluation. The narrower scope of factors considered in such tests restricts their relevance somewhat to general surveillance and applications that involve a general visual display. Such tests are useful however, in gauging the comfort in using a particular fusion driven display configuration. In [12,13] for example the pilots were also asked to rate the performance of each of a number of fused and single modality display configurations. The results consequently demonstrated the advantage of fused displays [12]. Relatively simple test procedures also make passive tests more practical for general algorithm validation [8,24] and as an accessory to a more complex fusion evaluation [16].

The level of control on the test conditions and equipment used such as display configuration, lighting, user interface etc. further impacts the credibility and