

8. ASSESSING THE QUALITY OF SYNTHESIZED IMAGES

Previous Chapter presents several methods for the fusion of sets of multispectral images B_{kl} at a low spatial resolution l and sets of images A_h at a higher spatial resolution h but with a lower spectral content. These methods aim at synthesizing sets of multispectral images B^*_{kh} , which are as close as possible to the reality B_{kh} . The three properties that should be respected by the synthesized sets are listed in previous Chapter.

Producers, *i.e.* providers of fused products, and customers, *i.e.* users of such fused products, may hesitate to select one of these methods or fused products. Commercial softwares often propose several different methods and it is not obvious for non-specialists to select one method or another for a given case. It follows that usually producers often use methods, which are not the most suitable for their customers.

Several comparisons between methods have been published and are regularly published. However results poorly disseminate in the community and there is lack of knowledge among producers regarding these methods, their advantages and limits. The lack of standardization of protocols for comparison does not add to the clarity of the results. Some efforts have been made recently but a lot still remain.

Behind the choice of a method lie needs for quality. Not neglecting aspects related to software complexity, implementation and maintenance and computation time and other constraints, the quality of a fused product is the driving cause for the selection of a method in industrial systems and production lines.

Thus, the problem is twofold. Firstly, how to assess the *a priori* quality of a synthesized set of images B_h produced by a given fusion method? This may translate into: what is the typical quality one may expect by running a given method over given cases? The answer helps in selecting a method. Secondly, how to assess *a posteriori* the quality of a synthesized set of images effectively produced within a given industrial process?

Quality assessment needs a protocol. We will see later that the same protocol may answer both questions: the *a priori* and *a posteriori* assessment of quality. Such a protocol and the associated quantification of the quality may help in

- system requirements by providing a framework for users to better specify their needs for information;
- information communication by allowing producers, customers and other persons from all backgrounds to communicate the usefulness of an image to perform a task;
- and analysis by providing an instrument for developing other system performance tools or for assessing the effects of changes in the fusion methods or sensor design or image chain or production line on image quality.

A protocol for quality assessment should have very clear objectives. The objectives of the protocol discussed hereafter are the assessment of the performances of the fused products with respect to the three properties listed in previous Chapter. The typical approach for the assessment of the quality by the means of visual analysis performed by a panel of investigators is also reported. Such an approach is tailored to the needs and objectives of a specific community of users. The actual spatial resolution of the fused product was assessed in this way in the military community.

QUALITY ASSESSMENT NEEDS A REFERENCE

Quality is assessed with respect to a reference image. In the case of the assessment of a method (*a priori* assessment of fused products), sets of actual multispectral images B_h at high resolution h are usually available. The fused products B^*_{kh} are made from the images A_h and B_{kl} and are compared to these references B_{kh} through a visual analysis or computation of similarities and discrepancies, in an automated way or not.

WHAT TO DO IF NO REFERENCE IS AVAILABLE?

Such a reference is not always available and should be created. This is the usual case of the assessment of a fused product (*a posteriori* assessment). One of the most common approaches to this shortcoming consists in interpolating low resolution images B_{kl} up to the high resolution h , and assuming that these images constitute the reference. In any case, are the interpolated images representatives of what would be observed by a similar sensor with a higher resolution, and these interpolated images cannot constitute a valid reference. It follows that this approach is not valid and should not be used. It is in itself a paradox: if interpolated images are assumed to be the reference, why should one bother with fusion methods?

Other protocols try to avoid establishing images of reference, mostly by using some statistical quantities or features derived from the original data set and from the synthesized images. One example is the use of the histograms of the synthetic products, which are compared to the original ones. The histograms for images taken by the SPOT system over the city of

Barcelona (Spain) are presented in Figure 8.1. These images are displayed and discussed in following Chapter. On the upper half are the histograms of the original images P and $XS1$. For the latter, the resolution is 20 m only: it contains four times fewer pixels than P or the synthetic images XS^* . For the comparison, the histogram of $XS1$ has been normalized to the others by multiplying the number of pixels by four. Though the resolution is increased by a factor of two relative to that of $XS1$, the histograms of the images XS^*I synthesized by two different fusion methods are expected to be close to that of $XS1$ in shape. This is true for the histogram of the image XS^*I_{RWM} synthesized by the ARSIS-RWM method (lower right). Its highest frequency is close to four times that of the histogram of $XS1$. The modal values are the same and the shapes of both histograms are very similar. On the contrary, the histogram of the image XS^*I_{P+XS} synthesized by the P+XS method (lower left) is much closer to that of the image P , both in shape and in peak. It indicates the discrepancies between the actual image $XS1(10\text{ m})$ and XS^*I_{P+XS} and the spectral distortion induced by the P+XS method.

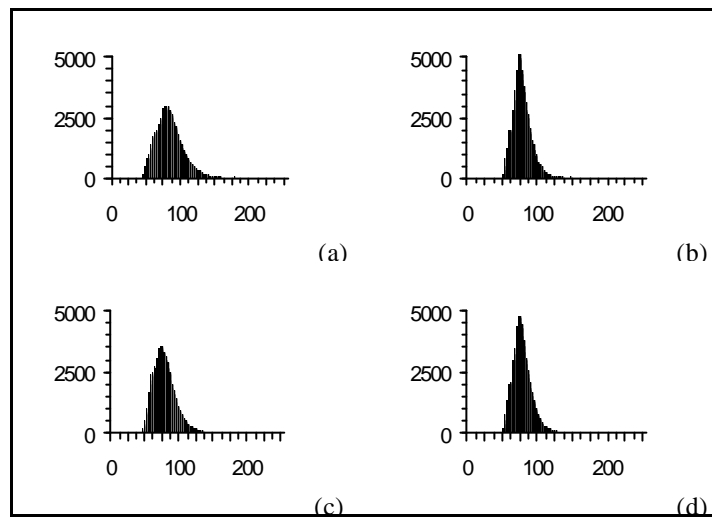


Figure 8.1. Comparison of histograms of original and synthetic SPOT images. Scene of Barcelona, Spain. (a) SPOT P , 10 m resolution; (b) SPOT $XS1$, 20 m resolution; (c) synthetic image ($P+XS$ method), 10 m resolution; (d) synthetic image (ARSIS-RWM method), 10 m resolution.

This comparison of histograms is a fairly good estimator of image quality, and is easy to handle. However, the effect of the spatial resolution upon the statistical properties of an image should not be neglected. Several published studies demonstrate the non-preservation of statistical distribution with the change in spatial resolution. This non-preservation depends upon the

observed type of landscape. The more energetic the high frequencies at scale h , the more dissimilar the statistical distributions at scales h and l . That means that we should not try to identify the statistical properties of a synthetic product to those of the original image. Therefore, any protocol based upon the comparison of statistical quantities (e.g., histogram, cumulative distribution, entropy etc.) is not valid.

Another approach found in the Earth observation domain is to compare land-use maps obtained after spectral (and possibly textural) classification of the fused products. This classification approach is valuable because land-use mapping is often the goal of satellite image processing. These maps are compared either to the map obtained from original low-resolution data (e.g., SPOT XS), or to ground truth. In the first case, the same assumption as above is made, that is that some statistical properties are preserved through the increase in resolution. More generally, classification greatly reduces the content in information; this reduction decreases the discrepancies between fusion methods. In the classification process, pixel spectral values are aggregated with their spectral neighbors. Hence, a small difference between the synthesized and the actual spectra at a given pixel may have an impact on classification ranging from null to significant. Furthermore, the results of the classification depend upon the type of landscape, its diversity, its heterogeneity, the time of observation, the optical properties of the atmosphere, the sensor system itself (including the viewing geometry), the type of classifier (supervised, unsupervised), and the classifier itself. Hence, this approach may not reflect the overall performance of a fusion method and should be avoided.

HOW TO CREATE A REFERENCE IMAGE?

Several authors have proposed an approach to create a reference image. It calls upon a change in scales and is as follows:

- two sets of images A_l and B_{kv} are created from the original sets of images A_h and B_{kl} . The image A_h is degraded to the low resolution l (A_l) and the images B_{kl} to the very low resolution v (B_{kv}) with $v=l(l/h)$. If $l=2h$, then $v=2l$;
- the fusion method is applied to the two sets of images, resulting into two sets of synthesized images B^*_{kh} at resolution h and B^*_{kl} at resolution l ;
- the original images B_{kl} serve as references. A comparison is performed between B_{kl} and B^*_{kl} by the means of visual analysis and analysis of the similarities and discrepancies;
- finally, the quality observed for the fused products B^*_{kl} is assumed to be close to the quality that would be observed if a reference at resolution h were present.

Such an approach alleviates the lack of "truth" images B_{kh} . This raises a question. How can the assessment of quality of the synthetic images be made at the highest resolution h based upon that made at the lowest resolution l ? In other words, how can one extrapolate the quality assessment made at the lowest resolution to the highest resolution?

Intuitively, one thinks that, except for objects having a size much larger than the resolution, the error should increase with the resolution, since the complexity of a scene usually increases as the resolution is getting better and better. That is, one may expect the error made at the highest resolution h to be greater than that at the lowest resolution l . However, several recent works have demonstrated the influence of the resolution on the quantification of parameters extracted from satellite imagery. Many works dealt with clouds (here the parameter is the cloud coverage), or address the problem of resolution in weather prediction and climate models. Others study how the values of a geographical parameter (e.g., the number and surface of lakes or agricultural lots in a region) vary as a function of the resolution. Some mathematical models have been constructed to explain such these changes in rather simple cases. All these studies demonstrate that the quality of the assessment of a parameter is an unpredictable function of the resolution. It is a very complex function of the relative power of the high frequencies and of the very high frequencies, *i.e.* objects that are unresolved at the resolution h , and of the distribution of these objects within the pixel. The multi-modality aspect adds to this complexity.

It follows that the quality of the synthetic images at the highest resolution h cannot be predicted from the assessments made with synthetic images at the lowest resolution l . However, we may rely on the results of several assessments performed at Ecole des Mines de Paris. They show that there is no clear relationship between the quality parameters obtained for the fused products B^*_{kh} and B^*_{kl} , or between B^*_{kl} and B^*_{kv} , as expected. Nevertheless, it has been often found that the quality was best at the resolution h (respectively l) relative to the resolution l (respectively v), and also that the ranking of a method relative to the others was the same at these resolutions. It does not prove that estimates should be better at the resolution h than at the resolution l . However, it seems reasonable to assume that the quality of the synthetic images at the highest resolution h is close to that at the lowest resolution l .

A GENERAL PROTOCOL FOR QUALITY ASSESSMENT

A protocol has been worked out, which is accepted by several professional organizations. It is simple to implement. It may become the standard approach agreed upon by all the producers of fused products whose scopes are in the frame of this discussion. It permits to alleviate the need for a reference image if not available and offers a complete checking of the three

properties¹. It can be used in any case, whether a reference image is available or not, and for evaluating products as well as methods. The general frame is as follows:

- the fusion method is applied to the original sets of images A_h and B_{kl} . It results into a new set of synthesized images B^*_{kh} at resolution h ;
- testing the first property: *any synthetic image B^*_{kh} , once degraded to its original resolution l , should be as identical as possible to the original image B^*_{kh} .* To achieve this, the synthetic image B^*_{kh} is spatially degraded to an approximate solution $(B^*_{kh})_l$ of B_{kl} . If the first property is true, then $(B^*_{kh})_l$ is very close to B_{kl} . The difference between both images is computed on a per-pixel basis. The fused products together with the difference image are visually compared to the original images B_{kl} in order to detect trends of error, if any. These trends may be related to the objects in the scene. Then some statistical quantities are computed to quantitatively express the similarities and discrepancies between both images;
- testing the second property: *any synthetic image B^*_{kh} should be as identical as possible to the image B_{kh} that the corresponding sensor would observe with the highest resolution h .* The second and third properties refer to B_{kh} , an image that would be sensed if the sensor had a better resolution h . This image is the reference image and is not always available; otherwise, all the above-cited methods would not have been developed. If a reference B_{kh} is available, the comparison is performed between B_{kh} and B^*_{kh} , using the same means, techniques and statistical parameters as for the first property. If a reference B_{kh} is not available, a change in scale is performed as described in the previous section for creating a reference. The images to compare are now the original images B_{kl} and the fused products B^*_{kl} . The comparison is made exactly in the same way than in the case of the availability of a reference. It is assumed that the quality attained for this reference at the resolution l is similar to that that would be attained at the resolution h ;
- testing the third property: *the multispectral set of synthetic images B^*_h should be as identical as possible to the multispectral set of images B_h that the corresponding sensor would observe with the highest resolution h .* As above, if the set of images B_h is not available, the comparison is performed between the sets B_l and B^*_l . As for all the properties, the comparison is made by the means of visual analysis and computation of similarities and discrepancies.

¹ L. Wald, T. Ranchin and M. Mangolini. *Fusion of satellite images of different spatial resolutions: assessing the quality of resulting images*. Photogrammetric Engineering & Remote Sensing, 63, 6, 691-699, 1997.

Depending upon the objectives of the assessment and of the available resources, the task of visual analysis will be more or less sophisticated and the computer analysis of the similarities and discrepancies will be more or less extensive. An example of experimentation for the assessment of several fusion methods is given in next Chapter.

THE IMPORTANCE OF THE SELECTION OF THE TEST IMAGES

The type of landscape or objects present within the image used to assess the quality of a synthesizing method has a strong influence upon the results. Obviously, if the objects of the scene are spatially homogeneous for scales ranging between h and l , any sound method will provide good results. In this case, the benefit of the fusion is questionable since interpolation methods and even duplication will lead to satisfactory results.

Whatever the method, the more predictable the changes in signal with the scale, the better the quality of the final product. Hence, scenes whose objects are self-similar for scales between h and l should be avoided for test cases, since they do not enhance the properties of a method.

Taking Earth observation as an example, over areas such as the ocean or large agricultural lots, which appear very homogeneous at, say, 20 m resolution, the error made in assuming that these areas are still homogeneous at, say, 10 m resolution, is small. On the contrary, urban areas or small agricultural lots are among the most difficult cases because they exhibit a large number of interwoven objects having different scales.

The particular case of the SPOT images of the city of Barcelona was examined. Barcelona is a large city located in northeast Spain, on the Mediterranean coast. Its harbor is the busiest in Spain. The scene is mostly comprised of urban districts, highways and railroads. It also exhibits small agricultural lots and mountainous areas covered by typical Mediterranean vegetation. The images are shown in next Chapter where several fusion methods are compared (Figs 9.1 and 9.2). Such an urban area has been selected for illustrating the comparison because it is certainly the most difficult type of landscape to deal with according to our knowledge. Urban areas often point out the qualities and drawbacks of algorithms because of the high variability of information in space and spectral band, induced by the diversity of features in both size and nature.

It was found that all information (100 percent), expressed as variance, of the homogeneous part covering the Mediterranean Sea, is borne by structures larger than 40 m for each of the three modalities. On the contrary, for the urban area, half the information (50 percent) is borne by structures having sizes less than 40 m. Urban areas do not possess self-similarity properties, though some parameters, such as the growth of city limits, can be approximated by fractal functions. In other words, structures observed at,

say, 10 m resolution, cannot be accurately predicted from their observations at lower resolution, say, 20 m. This is well-known by experienced image interpreters, and is also sustained by published mathematical evidence. The benefit of an image of a higher spatial resolution is the greatest in these cases. Hence, it is recommended that test images should mainly include such areas. Such cases also offer a large diversity of spectral signatures, which is helpful in judging the ability of a method to synthesize the spectral signatures during the change in spatial resolution.

The spectral heterogeneity of a scene may be characterized by the spectral diversity of the set B of images relative to the maximum possible number of spectra S_{max} . A heterogeneity parameter he can be defined:

$$he = S / S_{max} \quad [8.1]$$

where S is the number of spectra observed in the set B . If NP is the number of pixels of the images, then $S_{max} = NP$ and

$$he = S / NP \text{ and } 0 \leq he \leq 1 \quad [8.2]$$

The larger he , the more spectrally heterogeneous the scene.

A threshold cannot be given for he , separating suitable test cases for inappropriate ones. Actually, this parameter is not robust enough. Assume a scene that offers the same number of spectra when its sizes are slightly reduced. It results in increasing he but the difficulty in synthesizing images remains the same.

To avoid this problem, we define another quantity ho :

$$ho = 10^4 / S \quad [8.3]$$

The larger ho , the more spectrally homogeneous the scene. Hence ho characterizes the spectral homogeneity of the set B . The larger the number of modalities, the smaller ho . For the case of the SPOT images with three modalities, the author found ho values less than 1.4 and most often comprised between 0.2 and 0.4 for urban areas. These areas are considered as the most difficult test cases. For the case of Marseille discussed in Chapter 7 (Fig. 7.8), $ho=0.1$. One of the most difficult cases encountered by the author is the color image (R, G, B) of the baboon. This image is well-known in the community of researchers in image processing. It displays the very colored face of a mandrill and most of the information is contained in the very high frequencies. In this case, $ho=0.04$, *i.e.* the spectral homogeneity is dramatically low. ho decreases as the number of modality increases. A value of $ho=0.01$ was found for the case of a set of images taken in four spectral bands by the satellite SPOT-4. The landscape was made of several villages close each to the other surrounded by small agricultural lots and other vegetation patches exhibiting high frequencies.

This quantity h_o can be used to discriminate between test cases that permit to assess the properties of a fusion method and others. The smaller h_o , the more difficult the images to synthesize at a better resolution. A threshold of 0.4 can be set up from experience. Appropriate test cases should exhibit h_o values lower than this threshold.

ASSESSMENT BY A PANEL OF INVESTIGATORS

The visual analysis is a key to quality assessment. The objective comparison of the visual quality of multiple images is a difficult and lengthy task to handle. The human visual system is not equally sensitive to various types of distortion in an image. The perceived image quality is strongly dependent upon the observer and upon the thematic application. Standard protocols have been defined, in the field of television and image compression or Earth observation by airborne or space-borne instruments.

Several investigators are gathered together to perform such an assessment. Several sets of fused products are shown to these investigators, who judge some well-defined aspects of the images with respect to well-established criteria. Then their notations are weighted and further processed to obtain a mean opinion score defining the quality of the result. When it comes to the assessment of the quality of a set of multispectral images, the mass of data becomes very large. This dramatically increases the difficulty in computing a quantitative picture quality scale.

Such operations are relevant to the general problem of the assessment of the satisfaction of the customers regarding a given product. Similar experimentations are currently performed for industrial products. Conceptually, the assessment of fused products or of fusion methods is not different. Similar techniques for the selection of panels of users may be used, similar criteria may be employed, and similar mathematical procedures may be applied for the screening of the individual responses and the analysis of the results.

The panel should comprise as much as possible investigators. These investigators may be either trained persons or unaware persons depending on the purpose of the test. The larger the panel, the better since statistics will perform better on a large panel. However, a large panel is more costly in many aspects than a smaller one. The investigators should view images and perform the requested analysis in the same conditions: same type of color monitor, same monitor calibration, same distance of viewing, same surrounding illumination etc. Such assessments operations are very heavy to manage and accordingly, they cannot be performed on a routine basis.

The protocol of such experimentations is more or less the same and is as follows. A set of specifications is established regarding the quality of images. This set of specifications comprises a set of criteria to be respected

by the ideal product. Examples of criteria in the case of visual interpretation of images of scenes, natural or artificial, may be:

- colors should be as close as possible from colors perceived by the human eyes;
- objects of size T_0 or more should be detectable;
- objects of size T_1 ($T_1 > T_0$) or more should be identifiable;
- objects of size T_2 ($T_2 > T_1$) or more should be subjected to analysis.

A committee within the U.S. Government has established criteria for the interpretability of multispectral imagery in Earth observation, which may serve as references. Examples of such criteria are given in Table 8.1.

Then one product is selected among a series of standard products. This product is called the reference product. Its performances with respect to the above-mentioned criteria serve as references against which are compared the subjective valuation of the panel of users. If the score of a fused product is better than that of the reference product, the fused product is said to be better or to offer better performances than the reference product. Here, a standard product may be the images $B_{khInterp}$ resulting from an interpolation of the images B_{kl} from the resolution l to the resolution h . Another standard product may be a fused product produced by a well-known method (e.g., a projection and substitution method).

Then a panel of investigators is selected. The investigators assess each fused product versus the defined criteria with respect to the reference product. For each criterion, each investigator gives a note. The scale often comprises five notes: much worse performances, worse performances, similar performances, better performances, much better performances. It may comprise more notes, e.g. ranging from 0 to 10 or 0 to 100. When the references are loosely defined or even absent, the scale is often reduced to four notes:

- not satisfying (not relevant, not performant, not efficient), weak;
- not much satisfying (not much relevant, not much performant, not much efficient), rather weak;
- rather satisfying (rather relevant, rather performant, rather efficient), rather strong;
- very satisfying (very relevant, very performant, very efficient), very strong.

Once the human analysis performed, the individual notations are screened. Apparent inconsistencies of the answers (e.g., an increase in resolution should likely lead to an increase in the quality of the identification of objects) are looked for. Biased answers are rejected as well as those resulting from misunderstanding of the instructions, criteria, objectives of

the task or protocol. Cross-analyses are performed on the set of scores to discover irregularities. Finally the individual notations are weighted and mean scores are obtained. They qualify specific aspects of the fused product and its overall quality.

MS IIRS Level 1

- Distinguish between urban and rural areas.
- Identify a large wetland (greater than 100 acres).
- Delineate coastal shoreline.
- Detect major highway and rail bridges over water.
- Delineate extent of snow or ice cover.

MS IIRS Level 2

- Detect multilane highways.
- Determine water current direction as indicated by color differences.
- Detect timber clear-cutting.
- Delineate extent of cultivated land.
- Identify riverside flood plains.

MS IIRS Level 3

- Detect vegetation/soil moisture differences along a linear feature (suggesting the presence of a fence line).
- Identify major street patterns in urban areas.
- Identify shoreline indications of predominant water currents.
- Distinguish among residential, commercial, and industrial areas within an urban area.
- Detect reservoir depletion.

MS IIRS Level 4

- Detect recently constructed weapon positions based on the presence of revetments, berms, and ground scarring in vegetated areas.
 - Distinguish between two-lane improved and unimproved roads.
 - Detect indications of natural surface airstrip maintenance or improvements (e.g., runway extension, grading, resurfacing, etc.).
 - Detect landslide or rockslide large enough to obstruct a single-lane road.
 - Identify areas suitable for use as light fixed-wing aircraft (e.g., Cessna, Piper Cub, or Beechcraft), landing strips.
-

Table 8.1. Example of criteria related to interpretability of multispectral images taken from the US Government²

² *Multispectral imagery interpretability rating scale. Reference Guide. Image Resolution Assessment and Reporting Standards (IRARS) Committee, U.S. Government. February 1995.*

GROUND SAMPLE DISTANCE - RESOLUTION OF THE FUSED PRODUCT

Image resolution has a significant effect on interpretability of images. It can be defined as a ground sample distance (GSD), that is the smallest distance that can be measured accurately by the analysts. The fused product intends to simulate what should be observed with a sensor having the best spatial resolution and one may expect the ground sample distance measured in the fused product to be greater than the claimed resolution h .

In the course of the analysis, the investigators are asked to compare the GSD they measure on the fused product to that measured on the reference product. An effective ground sample distance (EGSD) is thus defined.

Experiments made for the US Department of Defense³ by the means of panels of image analysts show that perceived image quality is proportional to the logarithm of the GSD. The effective ground sample distance can be roughly predicted as a function of the spatial resolutions h and l of the high and low resolution images:

$$EGSD = l - 0.94(l-h) \quad [8.4]$$

where $EGSD$, h and l are expressed in meters. Another formulation was proposed

$$EGSD = (1.103 h) - (0.004 h^2) + (0.001 l^2) + 0.37 \quad [8.5]$$

Equation 8.4 better fits the observations made. The relative gain in resolution is constant (equal to 0.94 times the difference $l-h$) for the resolutions that have been studied (h and l less than 30 m). Table 8.2 gives some values of $EGSD$ computed from Equation 8.4 for several couples of resolutions (h, l).

Table 8.2 shows that the effective distance $EGSD$ is close to the spatial resolution h . The smaller the ratio l/h , the closer the $EGSD$ to h . This similarity between h and $EGSD$ demonstrates the benefits of the fusion of images.

The values in Table 8.2 are indicative. The effective distance $EGSD$ depends upon the fusion method employed to construct the products and of the properties of the sensors themselves, including the modulation transfer function, which may impact on the quality of the fused products, depending upon the methods.

³ J. Vrabel. *Multispectral imagery advanced band sharpening study*. Photogrammetric Engineering & Remote Sensing, 66, 1, 73-79, 2000.

$l(m)$	$h(m)$	$EGSD(m)$
30	10	11.2
20	10	10.6
20	5	5.9
4	2	2.1
4	1	1.2
2	1	1.1

Table 8.2 Predicted effective ground sample distance (EGSD).

COMPUTER-DERIVED MEASURES OF PERFORMANCES

The general protocol is based upon visual analyses of the fused products B_{kh}^* (respectively B_{kl}^*) with respect to the original images B_{kh} (respectively B_{kl}) and upon the computation of the difference between the fused product and the original images on a per-pixel basis. Statistical quantities help in summarizing the similarities and discrepancies between the sets of images. Such measures of performances estimated from these differences offer the benefits of quantitative values and the advantage of being automated in the production lines.

QUANTITATIVE ASSESSMENT FOR THE FIRST PROPERTY

An important point here is the way the synthetic image B_{kh}^* is degraded to $(B_{kh}^*)_l$. Some wavelet transforms have the ability to separate scales well, that is, to separate structures of small size from larger ones and, therefore, to simulate what would be observed by a lower resolution sensor. Many authors use an averaging operator on a window of 3 by 3 pixels or more. Such an operator does not have this ability in scale separation and is not as appropriate here. Other filtering operators should be used, some of them simulating a given modulation transfer function (MTF) of a sensor.

A comparison was made at École des Mines de Paris (T. Ranchin, personal communication) on a few scenes using some operators, such as a sine cardinal (sinc) kernel truncated by a Hanning apodisation function of size 13 by 13 pixels, a truncated Shannon function, a bi-cubic spline, a pyramid-shaped weighted average, and the wavelet transforms of Daubechies (1988, regularity of 2, 10 and 20). It showed relative discrepancies between the results on the order of a very few per cent. In conclusion, there is an influence of the filtering operator upon the results, but it can be kept very small provided the operator is appropriate enough.

The quantities that are computed from the differences between the two sets of images are similar to the first and second sets of criteria described under the second property below.

QUANTITATIVE ASSESSMENT FOR THE SECOND PROPERTY

The synthetic image B^*_{kh} (respectively B^*_{kl}) is compared to the reference image B_{kh} (respectively B_{kl}) by means of some criteria described below. The numerical comparison should be made preferably in physical units and in relative values. Thus, different tests made over different scenes may be compared. A difference is computed between B_{kh} and B^*_{kh} (respectively B_{kl} and B^*_{kl}). After visual inspection, the difference image is reduced to a few statistical parameters, which summarize it. There are a large number of candidate parameters. We have computed many for several tens of cases. We have retained some whose definitions are well-known to engineers and researchers and which clearly characterize the advantages and disadvantages of a method.

Two sets of criteria are proposed to quantitatively summarize the performance of a method in synthesizing an image in one spectral band. The first set of criteria provides a global view of the discrepancies between the original image B_{kh} and the synthetic one B^*_{kh} (respectively B_{kl} and B^*_{kl}). It contains:

- the bias, as well as its value relative to the mean value of the original image. Recall that the bias is the difference between the means of the original image and of the synthetic image. Ideally, the bias should be null;
- the difference in variances (variance of the original image minus variance of the synthetic image), as well as its value relative to the variance of the original image. This difference expresses the quantity of information added or lost during the enhancement of the spatial resolution. For a method providing too many innovations (in the sense of information theory), *i.e.*, "inventing" too much information, the difference will be negative because the variance of the synthetic image will be larger than the original variance. In the opposite case, the difference will be positive. In information theory, the entropy describes the quantity of information. However, we selected the variance difference because most researchers, engineers and practitioners are much more familiar with variance, and entropy and variance act quite similarly for our purpose. Ideally, the variance difference should be null;
- the correlation coefficient between the original and synthetic images. It shows the similarity in small size structures between the original and synthetic images. It should be as close as possible to 1;
- the standard deviation of the difference image, as well as its value

relative to the mean of the original image. It globally indicates the level of error at any pixel. Ideally, it should be null.

The error at pixel level may be more detailed. The absolute value of the difference and the absolute relative error are computed at each pixel. The absolute relative error is the absolute value of the difference between the original and synthetic values, divided by the original value. Then the histogram of the absolute values of the difference and the histogram of these relative errors are computed. Both can be seen as probability density functions. Therefore, we can compute the probability of having at a pixel an error or a relative error (in absolute value) less than a given threshold.

This probability denotes the error made at pixel level, and hence indicates the capability of a method to synthesize the small size structures. The closer to 100 percent the probability for a given error threshold, the better the synthesis. The ideal value is a probability of 100 percent for a null error, relative or not. Here, for reasons of computer precision, the lowest threshold "no relative error or null error" should be set to a very small value instead of zero. Values such as 0.001 or 0.001 percent can be used.

QUANTITATIVE ASSESSMENT OF THE MULTISPECTRAL QUALITY (THIRD PROPERTY)

Visual inspection may be made through color composites of, for example, the first three principal components of the set of images. Both color composites should agree visually. Most methods for color composites are using dynamical adjustment for color coding. If the sets of images are different, even slightly, then the color coding will be different for both composites and no comparison will be possible.

Practically, we recommend the following approach. For each modality or spectral band k , the reference images B_k and the fused images B^*_k are juxtaposed into a single computer file. Here the set of reference images is that used to test the second property (*i.e.* the set B_h or B_l). The principal components analysis as well as the color coding are performed on this set of juxtaposed files. If the number of modalities is less than or equal to three, there is no need to perform a principal components analysis. A color composite is computed by the means of the first three components, which usually contain most of the information. This color composite is split in order to retrieve the composite of the reference images on the one hand and the composite of the fused products on the other hand. If more than one fused product is to be assessed for the same scene, the concatenation (juxtaposition) should be performed with all sets of fused products. This approach guarantees that the color composites are comparable.

The color composites are displayed, simultaneously or alternatively, onto the screen and are compared to the reference composite and to the others.

The advantage of this visual assessment is that it does show trend in errors, if any, possibly related to features in the scene. The drawback of it is that it is a subjective assessment and also that this assessment may be limited either by physiological factors (e.g., color contrast perception by humans), or by technical factors (e.g., when a large number of modalities or spectral bands are present). In the latter case, and if the scene offers a large variety of objects, the color re-coding of the first three principal components reduces dramatically the differences between the sets of images B and B^* , particularly if these differences are random, *i.e.*, not related to specific features in the scene or to a spectral band or modality.

A quantitative assessment can be made using the following three additional sets of criteria in order to quantify the performance of a method to synthesize the spectral signatures during the change in spatial resolution.

The third set (numbered after the two sets described above for the second property) deals with the information correlation between the different spectral images taken two at a time. This dependence can be expressed by the correlation coefficients, with the ideal values being given by the set of reference images B . This is done for every pair among the N available modalities and the image A . As an example, for the case of the modality k , the correlation coefficient between each pair $(B_k, B_j, j=1..N)$ and (B_k, A) is computed and compared to the correlation coefficient for each pair $(B^*_k, B^*_j, j=1..N)$ and (B^*_k, A) . The correlation coefficients found for the fused products should be as close as possible to the correlation coefficients found for the reference images.

The fourth set of criteria partly quantifies the synthesis of the actual multi-modality or multispectral n -tuplets by a method, where n -tuple means the vector composed by each of the N modalities or spectral bands at a pixel. It comprises the number of different n -tuplets (*i.e.*, the number of spectra) observed in the reference set B and in the synthesized set B^* , as well as the difference between these numbers. A positive difference means that the synthesized images do not present enough n -tuplets; a negative difference means too many spectral innovations.

The previous criteria do not guarantee that the synthesized n -tuplets are the same as in the reference set B . The fifth and final set of criteria assesses the performance in synthesizing the actual n -tuplets. It deals with the most frequent n -tuplets, because they are predominant in multispectral classification. For a given threshold in frequency, only the n -tuplets having a frequency (relative number of pixels) greater than this threshold are used. The threshold is set to e.g., 0.01 percent, 0.05 percent, 0.1 percent, and 0.5 percent, successively. The greater the threshold, the lower the number of n -tuplets, but the greater the number of pixels exhibiting one of these n -tuplets. For each of the n -tuplets, the difference is computed between the

reference frequency and the one observed in the synthesized images. These differences are summarized by the following quantities:

- the number of actual n -tuplets, the number of coincident n -tuplets in the synthesized images, and the difference between these numbers, expressed in absolute and relative terms;
- the number of pixels in these n -tuplets, in absolute and relative terms;
- and the difference between the above number of pixels for the reference and synthesized sets of images, in absolute and relative terms.

This protocol has been applied to several cases. Its capabilities in characterizing the performances of methods and the quality of fused products have been demonstrated. The statistical parameters have proven their high value to summarize the similarities and discrepancies, still conveying enough details so that one may see at a glance the major merits and drawbacks of a method or of a fused product.

A GLOBAL ERROR PARAMETER FOR DESCRIBING THE QUALITY

There is a further need for a simple characterization of the quality of the product of the fusion process, which can be associated to each product and qualifies it. It would greatly help producers to select methods and improve their production lines, and customers to make their choice among products and to assess the impact of this quality on further processing.

The protocol discussed in this Chapter computes the differences between the synthesized images and the actual ones. These differences are summarized by various statistical quantities, which characterize the performance in synthesizing images in a given modality and the multi-modality signature, and especially the most frequent spectra. Published works often use statistical quantities, such as root mean square errors $RMSE(B_k)$. On the contrary, biases and mean values are given seldom.

These quantities as discussed before are very useful to fully understand the performances and properties of a method. However, experience shows that there are too many figures, which are of no help to the customers. There is a need for a quantity, which gives a quick insight of the quality. What we are looking for, is a number simple to understand which is a good indicator of the overall error of the fused product. The closer to 0 this number, the better the product. This quantity should fill three requirements:

First requirement. It should be independent of units, and accordingly of calibration coefficients and instrument gain. Customers seldom consider calibration coefficients. Some fusion methods can be applied to unitless quantities or to radiances. Consequently, the quality parameter should be independent of units.

Second requirement. This quantity should be independent of the number of spectral bands under consideration. This is a *sine qua non* condition to compare results obtained in various conditions.

Third requirement. This quantity should be independent of the scales h and l . This permits to compare results obtained in different cases, with different resolutions.

The following quantity was proposed to globally characterize the quality of the fused product. It was called total error and is given by:

$$Total\ error = \sum_{k=1}^N RMSE(B_k) \quad [8.6]$$

It is actually the sum over the N modalities of the root mean square errors (RMSE) for each modality k . The RMSE is that computed by the means of the reference set of images used for the testing of the second property. (*i.e.* B_{kh} or B_{kl}). It is defined as

$$RMSE(B_k) = \frac{1}{NP} \sqrt{\sum_{i=1}^{NP} (B_k(i) - B_k^*(i))^2} \quad [8.7]$$

where i is the current pixel and NP is the number of pixels. It is also equal to:

$$RMSE(B_k) = \sqrt{(bias)^2 + (standard\ deviation)^2} \quad [8.8]$$

This total error does not obey any of the three requirements. In particular it is sensitive to the changes from numerical counts to radiances. Another error was proposed⁴ in order to be able to compare errors obtained from different methods, different cases and different sensors. Let M_k be the mean value for the original spectral image B_k . Let M be the mean radiance of the N images B_k :

$$M = (1/N) \sum_{k=1}^N M_k \quad [8.9]$$

The relative average spectral error RASE is expressed in percent and characterizes the average performance of a method in the considered spectral bands:

⁴ T. Ranchin, and L. Wald. *Fusion of high spatial and spectral resolution images: the ARSIS concept and its implementation*. Photogrammetric Engineering & Remote Sensing, 66(1), 49-61, 2000.

$$RASE = \frac{100}{M} \sqrt{\frac{1}{N} \sum_{k=1}^N RMSE(B_k)^2} \quad [8.10]$$

The RASE mostly obeys the first and second requirements. Shortcomings arise in the case of uncalibrated images in different modalities with very different dynamics in gray levels. Further, the RASE does not obey the third requirement.

From this experience, another quantity is proposed. It is called ERGAS, after its name in French "*erreur relative globale adimensionnelle de synthèse*" that means relative adimensional global error in synthesis.

$$ERGAS = 100 \frac{h}{l} \sqrt{\frac{1}{N} \sum_{k=1}^N \left[\frac{RMSE(B_k)^2}{(M_k)^2} \right]} \quad [8.11]$$

It is more robust than the RASE with respect to calibration and changes of units. It also obeys the second requirement. The ratio h/l takes into account the various resolutions. For the same error ERGAS, the mean value of the relative $RMSE(B_k)$ increases as the ratio h/l decreases, since it is equal to:

$$RMSE(B_k) = \sqrt{\frac{1}{N} \sum_{k=1}^N \left[\frac{RMSE(B_k)^2}{(M_k)^2} \right]} \quad [8.12]$$

For example, if $h/l=1/2$ and $ERGAS=3$, the mean value of the relative $RMSE(B_k)$ is equal to 6 percent. If $h/l=1/4$, this mean value is equal to 12 percent for the same ERGAS. This recognizes the increase in difficulty when synthesizing images with large differences in resolutions h and l .

These various quantities were computed for several cases (see following Chapter). These cases comprise the application of various fusion methods on different sets of images acquired in various modalities with different scales h and l . The quality of each fused product was assessed as described in the previous pages. A global note was given to each product: bad or good.

The total error decreases as the RMSE for each modality k decreases. It is very sensitive to changes in units and in number of modalities. There is no evident relationship between the total error and the global note of quality. The total error cannot represent in a simple way the overall quality.

The relative average spectral error RASE behaves better. It offers a better tendency to decrease as the quality increases. It is independent of units provided they are the same for all bands. It is also independent of the number of bands provided the range of values for each band is constant.

However, like before, there is evident relationship between the error RASE and the global note of quality.

The error ERGAS exhibits a strong tendency to decrease as the quality increases. Thus, it is a good indicator of the quality. It behaves correctly whatever the number of bands is because it uses for each band the RMSE relative to the mean of the band. This definition makes also this quantity independent of the calibration or changes in units, allowing even changes from band to band.

Figure 8.2 displays the error ERGAS computed for several cases. These cases have been sorted out in two categories: bad or good. The labeling was made by the persons providing the cases to the author. Though based upon the protocol above-mentioned and numerical parameters, it has obviously a subjective aspect in the absence of an accepted global error parameter.

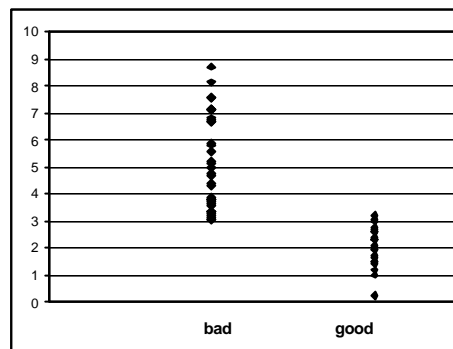


Figure 8.2. The error ERGAS for several cases of fusion

A striking feature in Figure 8.2 is the presence of a threshold. Cases of "good quality" exhibit values less than 3, or slightly greater, while the error ERGAS is larger than 3 for cases of "bad quality".

The existence of this threshold means that the error ERGAS is a good candidate for being the desired global error parameter. A fused product of good quality should exhibit an error ERGAS less than 3. This threshold corresponds to a mean value of the relative RMSE of 6 percent if $h/l=1/2$ and 12 percent if $h/l=1/4$ (Equation 8.12).

The error ERGAS provides a quick and accurate insight of the overall quality of a fused product. It behaves better than the other quality parameters. Since the error ERGAS reflects the conclusions of the different authors relative to the methods, it may serve to broadly assess the quality of a method. Very similar values of the error ERGAS are found for different cases, which have been declared satisfactory by their authors.

A threshold of satisfaction may be set to $ERGAS=3$ for a product. Below 3, the global error is small and the product is of good quality. Well above 3, the global error is large and the product is of lower quality. The quality decreases as the error $ERGAS$ increases.

Further investigations on the error $ERGAS$, or an equivalent error, would make possible in a near future for producers of fused products to deliver a standardized assessment of the quality of their products. This would allow them to better design and improve their production chains, and would allow customers to better select the products and improve their efficiency.