17

Architectures for Image Fusion

Michael Heizmann[1] and Fernando Puente León[2]
[1]*Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Fraunhoferstraße 1, D-76131 Karlsruhe*
[2]*Institute of Industrial Information Technology, Karlsruhe Institute of Technology (KIT), Hertzstraße 16, D-76187 Karlsruhe*
Germany

1. Introduction

Image fusion can be defined as the combination of raw or processed images establishing the input information from different sources like cameras or other imaging sensors. Its aim is to obtain new or more precise knowledge which is the output information about the scene and which comprises, e.g., objects, events, or more complex situations. Depending on the task of the image acquisition, the output quantities can be images, features, or symbolic information such as decisions.

In automated visual inspection, the ultimate aims in most cases are to gain macroscopic geometrical information (e.g., length, width, or position of an object), to characterize the surface (e.g., reflectance properties, roughness, microstructure, occurrence of surface defects like dents, grooves or other marks), or to obtain volume properties of an object (e.g., material classification, degree of transparency, spatial distribution of components or defects). This information can then be used in various ways in industrial production engineering, e.g., for quality inspection or materials management.

In other application areas of image acquisition and processing, the tasks are similar: finally, some specific information about the observed scene is to be brought to light. Examples are autonomous vehicles, where the vision is one among the exteroceptive sensors serving to sense the surrounding of the vehicle and to recognize objects, and remote sensing, where the task is to reconstruct the properties of a remote object of interest (e.g., the Earths's surface) from acquired images.

However, many relevant scene properties cannot be determined automatically by evaluating just one image. Instead, the information of interest can often be captured in an image series by means of a properly designed imaging setup using homogeneous or inhomogeneous imaging sensors. The task of image fusion is then to collect and combine the desired information from the image series by means of an adequate extraction of the useful information.

This approach has a direct correspondence to the common visual examination performed by a human in everyday live: if a human is not able to determine the property of interest at first glance, he will alter the visual setup until this property is clearly visible with this setup or he is able to reconstruct the property in his mind.

Besides this essential justification for the use of image fusion in many situations, there are also several other task-dependent reasons:

– A higher accuracy and reliability of the inspection result can be obtained when redundant or complementary information is available. In this case, sensors which receive identical or comparable scene properties are required.
– A feature vector with a higher dimensionality than just visual intensities can be generated by evaluating distributed or orthogonal information. For this, sensors receiving different physical scene properties may be appropriate.
– The acquisition of information can be accelerated by simultaneous operation of multiple sensors of similar type.
– The costs for the acquisition of information can be reduced when several low-cost sensors are substituted for an expensive special sensor. In this case, image fusion is used for indirect measurement of the quantity of interest.

This contribution focusses mainly on theoretic considerations on the information content in image series and on systematic aspects of architectures for image fusion originating from the former considerations. The concepts presented in the contribution will be illustrated by means of several examples from automated visual inspection. They will demonstrate how these concepts can be transferred to efficient approaches for image fusion. That way, they offer a systematic approach for the conception of systems and algorithms of image acquisition and fusion, not only for automated visual inspection.

2. Acquisition of information

The basis of image fusion is established by imaging sensors delivering the data which contains the desired information of the scene. Even if there are many different types of imaging sensors with specific physical properties, they all feature some basic characteristics relevant to image fusion.

2.1 Reduction of information

The process of acquiring images can be divided into several stages: the radiance emitted from the scene is projected onto an imaging sensor by means of an optical setup, which is usually the lens, possibly equipped with optical filters. Then the imaging sensor converts this optical signal into a digitized image.
The processing chain from the scene radiance to the digitized image involves a reduction of information, which generally causes the mapping of the scene to be non-invertible. For example, the visual information emitted by the scene is reduced with respect to the following aspects when a matrix imaging sensor is used:

– Images have a finite support, i.e., they are limited spatially (by the field stop, which is usually defined by the sensor size) and temporally (even in the case that a temporal series—e.g., a video—be taken).
– The image acquisition is a projection in several respects: spatially, the three-dimensional scene is projected onto a two-dimensional imaging plane. The infinite-dimensional space of wavelengths is projected onto one spectral dimension (in the case of gray-value images), three (RGB images) or few more spectral dimensions. Finally, the exposure corresponds to a projection in the temporal dimension.
– The irradiance $E(\boldsymbol{\xi}, \tau)$ on the imaging sensor with continuous support at a certain time $\tau \in \mathbb{R}$ in the image plane $I \ni \boldsymbol{\xi} := (\xi, \eta)^{\mathrm{T}} \in \mathbb{R}^2$ is spatially and temporally integrated, sampled, and quantized, thus resulting in the digital image with reduced information content.

In terms of system theory, the irradiance $E(\boldsymbol{\xi},\tau)$ is convolved with the aperture function $A(\boldsymbol{\xi})$ of a single pixel and sampled with the pixel spacings $\Delta_1, \Delta_2 \in \mathbb{R}^2$ to obtain a spatially discrete image $g(\boldsymbol{x},\tau)$, $\boldsymbol{x} := (x,y) \in \mathbb{Z}^2$:

$$g(\boldsymbol{x},\tau) := g((n\Delta_1, m\Delta_2)^{\mathrm{T}},\tau) \quad \text{with} \quad g(\boldsymbol{\xi},\tau) \propto \left(E(\boldsymbol{\xi},\tau) \underset{\boldsymbol{\xi}}{**} A(\boldsymbol{\xi}) \right) \cdot D(\boldsymbol{\xi}), \tag{1}$$

where the operator $\underset{\boldsymbol{\xi}}{**}$ denotes the two-dimensional convolution with respect to $\boldsymbol{\xi}$, $D(\boldsymbol{\xi}) := \sum_i \sum_j \delta(\boldsymbol{\xi} - i\Delta_1 - j\Delta_2)$, $i,j \in \mathbb{Z}$, describes the grid pattern of the imaging sensor, $n,m \in \mathbb{N}$ are the pixel coordinates, and $\delta(x) = 1$ for $x = 0$, $\delta(x) = 0$ else. Since the pixels do not overlap, $\mathrm{supp}\{A(\boldsymbol{\xi})\} \cap \mathrm{supp}\{A(\boldsymbol{\xi} - i\Delta_1 - j\Delta_2)\} \overset{!}{=} \varnothing \; \forall i,j \in \mathbb{Z}\backslash\{0\}$ holds.

Analogously, the temporal exposure can be interpreted as a convolution of the image $g(\boldsymbol{x},\tau)$, which has a continuous temporal support, with a temporal exposure function and a sampling with the refresh rate in order to get the digital image $g(\boldsymbol{x},t)$ with discrete temporal support $t \in \mathbb{Z}$.

– Further disturbances are added to the useful information, e.g., caused by thermal noise of the imaging sensor or by atmospherical disturbances in the light path.

2.2 Characteristics of imaging sensors

Sensor systems and their resulting data can be classified with respect to several properties concerning the degree of conformity of the data's information content. In the case of image fusion, the sensor systems may be characterized by the following properties:

– *Commensurability*: due to the optical projection of the three-dimensional scene onto the two-dimensional image plane, the spatial dimensions of images taken by matrix sensors always have the same physical meaning. Therefore, matrix sensors provide spatially commensurable data. Only if the imaging sensors have different numbers of dimensions (e.g., when images from matrix sensors and from line sensors are to be compared), their respective data are spatially not commensurable. In the case of color sensors (e.g., with three color values representing a color dimension) resulting in three-dimensional data, commensurability with gray-value images can be effectuated by projecting the color dimension onto one gray-value, thus resulting in two-dimensional data.

– *Homogeneity*: if the sensors capture identical or comparable physical quantities of the scene, the sensors are called homogeneous. This is an important feature for practical reasons: after the images from homogeneous sensors have been registered such that their definition areas are properly aligned, the data of the images can be combined mostly without complex preprocessing, e.g., in a data fusion. If, in contrast, the sensors are not homogeneous, a preprocessing (such as feature extraction or classification) is usually necessary to properly link the information of the images.

  The homogeneity of different images depends on the kind of processing which has been applied to the images prior to the fusion step: for example, when the images of a stereo camera pair are evaluated, a depth map is obtained, which is an inhomogeneous information compared to the original intensity images.

  The notion of homogeneity for imaging sensors may also depend on semantic aspects of the image data: if, for example, the images of a spectral series are interpreted as simple intensities which are observed from the scene, the corresponding sensor systems may

be regarded as homogeneous. If, however, these images are used to characterize the scene material, they may be regarded as inhomogeneous, since they represent the spectral reflectance of the material in different bands, which can be interpreted as different physical quantities.

– *Virtuality*: in many applications, image series are recorded by a single imaging sensor which has been used several times with at least one varying illumination or acquisition parameter. Since in reality, the images are taken with the same physical sensor, the sensors referring to the images in the series are called virtual.

  The images of virtual sensors always differ in at least one imaging parameter: the acquisition time. Since this acquisition parameter is irrelevant for time-invariant scenes, virtual cameras can favorably be used to record redundant information, if no other illumination or acquisition parameter is varied during the acquisition of the image series.

– *Collocatedness*: if the positions of the imaging sensors, their orientations and pixel spacings together with the optical properties of the imaging lens are kept constant over the series, the sensors are called collocated. In consequence, the reproduction scale remains unchanged, and the images show identical areas of the scene. Collocated sensors are often realized as virtual sensors, when an illumination or acquisition parameter except the scene pose is varied for a time-invariant scene.

  A typical example of non-virtual collocated imaging sensors is a three-chip RGB-sensor, where the individual sensors for the three color values are located at the same optical position by means of a beam splitter.

  If the imaging sensors are not collocated, an image registration, e.g., by considering image features is usually required to align the images in the series by means of geometrical transforms (e.g., translation, rotation, scaling or projective transform; see, e.g., (Modersitzki, 2004)).

2.3 Image series

Once the images have been acquired, they establish the data foundation of image fusion in the form of image series $g(\boldsymbol{x}, \boldsymbol{p})$, which are functions of the discrete position vector $\boldsymbol{x}$ and a discrete parameter vector $\boldsymbol{p} := (p_1, p_2, \ldots, p_n)^{\mathrm{T}}$, $\boldsymbol{p} \in \mathbb{Z}^n$, $n \in \mathbb{N}$. The image series is obtained by taking images while the imaging parameters $p_1, p_2, \ldots$ are varied consecutively or simultaneously. Considering the acquisition of a time-invariant scene by means of a camera system, the variable parameters can refer either to the illumination or to the scene (Heizmann & Beyerer, 2005).

Useful illumination parameters comprise the position of the illumination sources relative to the scene (expressed, e.g., by the azimuth φ_{i} and the polar angle θ_{i} in a surface related coordinate system), the spatial distribution of irradiance (to describe homogeneous or structured illuminations), the illumination spectrum, its polarization state, and its coherence.

Observation parameters for common camera systems are the camera position and orientation relative to the scene (in terms of the extrinsic camera parameters (Faugeras & Luong, 2001) or the scene pose), the spectral sensitivity of the sensor system including spectral filters, the polarization sensitivity of the sensor system including polarization filters, parameters of the optical system such as the focus setting, the aperture or the focal length, and the exposure time. The classical intrinsic parameters (see, e.g., (Faugeras & Luong, 2001; Shapiro & Stockman, 2001)) are not considered as relevant parameters for image fusion in most cases, since they usually cannot be varied.

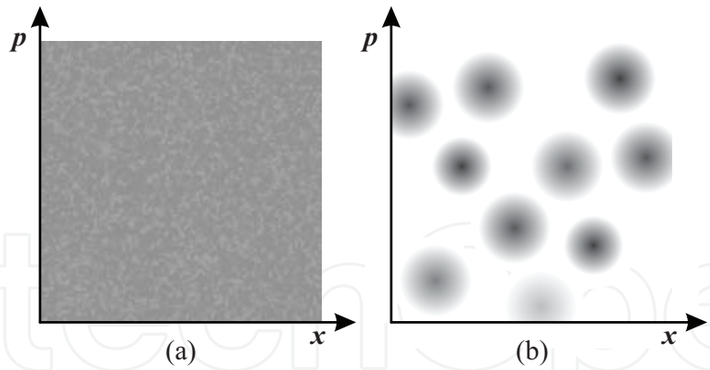


Fig. 1. Visualization of concurrent (a) and complementary (b) information: whereas in the case of concurrent information sources, the useful information (indicated as dark areas) is basically spread equally over the location-parameter domain (here simplified as one-dimensional domains each), it is concentrated in local regions for complementary information sources.

In consequence, the image series $g(\boldsymbol{x},\boldsymbol{p})$ establishes a multidimensional data object with a dimension for each varied parameter. An image $g(\boldsymbol{x},\boldsymbol{p}_i)$, $i \in \{1,\ldots,n\}$ which is sensed by the imaging sensor is a subspace of the image series for a fixed parameter vector $\boldsymbol{p}_i$.
In order to apply this definition of image series also in case that other sensor systems than cameras be used, the interpretation of the parameter vector $\boldsymbol{p}$ can be extended. If, for example, several inhomogeneous sensor systems should contribute to an image series, a nominally scaled parameter component can be used to distinguish the information sources.

3. Information content of image series

The information content in an image $g(\boldsymbol{x},\boldsymbol{p}_i)$, $i \in \{1,\ldots,n\}$, of the series $g(\boldsymbol{x},\boldsymbol{p})$ depends obviously on the imaging constellation which has been used during acquisition and, in consequence, the information content of the entire series depends on the variation of the parameter vector $\boldsymbol{p}$. Although it is often not possible to assign particular distributions of the information content in the image series to specific parameter variations, several elementary types of how information is distributed in image series can be identified. It is important to distinguish these types in order to identify suitable methods for fusing image series.

– *Redundant information*: in this case, the useful information is distributed similarly over all images of the series, i.e., it is spread equally over the location-parameter domain, see Fig. 1(a). In consequence, disturbances affect the useful information also in a similar manner. This type of relation can only be present when homogeneous sensors are used. A concurrent fusion, where all images contribute equivalently to the fusion result (e.g., by averaging over the image series), may be expedient to exploit the useful information.

  A typical example is noise reduction which can be achieved by image accumulation, i.e., averaging the intensity values for each location $\boldsymbol{x}$ over an image series acquired by homogeneous and collocated imaging sensors. If, for example, a stationary camera records n images disturbed by an additive white Gaussian noise, the observed images can be modeled as $g(\boldsymbol{x},i) = d(\boldsymbol{x}) + r(\boldsymbol{x},i)$, $i \in \{1,\ldots,n\}$, where the useful information $d(\boldsymbol{x})$ is deterministic and represents the desired scene property, and the additive noise $r(\boldsymbol{x},i)$ is the realization of a random process $R(\boldsymbol{x},i)$ with mean value $\mathrm{E}\{R(\boldsymbol{x},i)\} = 0\ \forall \boldsymbol{x},i$ and

variance $\mathrm{E}\{R(\boldsymbol{x},i)R(\boldsymbol{x},j)\} = \delta_i^j \sigma_R^2 \; \forall \boldsymbol{x},i,j$ with $\delta_i^j = 1$ for $i = j$, $\delta_i^j = 0$ else. While each original image $g(\boldsymbol{x},i)$ contains the full additive noise, i.e., $\sigma_G^2 = \sigma_R^2$, the variance of the noise in the accumulated image $f(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{x},i)$ is reduced to $\sigma_F^2 = \frac{1}{n}\sigma_R^2$.

– *Complementary information*: this relation applies if the useful information from homogeneous sensors is concentrated in the location-parameter domain such that, for a given location of the scene, only few images of the series are meaningful. Hence, the useful information is concentrated in local areas of the location-parameter domain, see Fig. 1(b). In order to merge the useful information, a complementary fusion is appropriate, where the contribution of an individual image to the fusion result depends on its local content of useful information. The local concentration of useful information may be caused by an inhomogeneous influence of disturbances, but it may also originate from the illumination-scene-sensor interdependence, when the susceptibility of the sensor depends on local differences in the constellation of illumination, scene, and camera (e.g., when the scene distance is varied in a focus series).

  An example is given below in Subsect. 5.1: in order to generate an image with synthetically enhanced depth of field, a focus series is taken. The useful information is identified by determining the parameter values (i.e., the scene distances) for each image location leading to a locally optimal focus indicator (Heizmann & Puente León, 2003; Puente León, 2002). Only these areas in the location-parameter domain are then used to build the fusion result.

  The same approach can be used for enhancing the image contrast by fusing illumination series (Heizmann & Puente León, 2003).

– *Distributed information*: this is the case when the useful information from homogeneous or inhomogeneous sensors is distributed over the series such that basically only the evaluation of all images allows a statement on the properties of interest. In contrast to redundant information, a single image alone does not contain enough information to conclude on the desired information. In comparison to complementary information, the useful information is not locally concentrated in the location-parameter domain. The inference from the image series to the useful information implies an image evaluation or interpretation, i.e., the image data must be transferred to a higher abstraction level (see Subsect. 4.2). Consequently, at least a feature extraction must be accomplished prior to the image fusion.

  As examples, distance maps can be generated by fusing stereo image series (Gheţa, Frese, Krüger, Saur, Heinze, Heizmann & Beyerer, 2007) or by fusing at least three images with varied directional illumination (photometric stereo) (Horn & Brooks, 1989). In the former case, the image values are interpreted as different views on the same scene which can be matched by means of epipolar geometry (Hartley & Zisserman, 2004). In the latter case, the image values are interpreted as response of the local surface shape and reflectance to the direction and intensity of the incident light. In both cases, a single image would not be enough to conclude on the desired distance map.

  A third example is given by methods of texture classification and surface inspection, when significant features are only obtainable by evaluating the entire series (Xin et al., 2004; Heizmann, 2006; Pérez Grassi & Puente León, 2007). Meaningful features for classifying topological textures can be developed especially when illumination series with varied azimuth of a directional illumination source are used, since the observed radiance of a topological texture strongly depends on the constellation of illumination. Examples in which distributed information is used are given in Subsections 5.2 and 5.3.

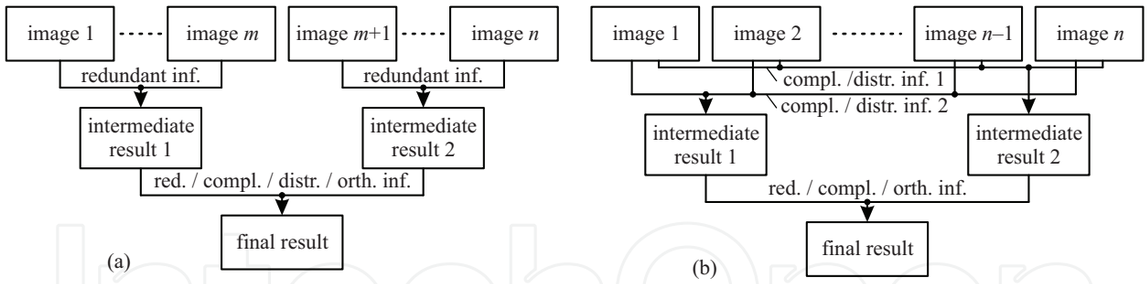


Fig. 2. Exemplary evaluation schemes for different combinations of information distributions in an image series: (a) When disjoint subsets of the image series contain redundant information, these subsets can be fused in a first stage to intermediate fusion results. The redundant, complementary, distributed or orthogonal information in these intermediate results is then fused in a second stage to the final result. (b) In the case of the image series being evaluated in more than one way, the respective information, which is usually complementary or distributed, is fused in a first stage to intermediate results, which are then fused to the final result.

– *Orthogonal information*: in this case, the images to be fused contain information on disjoint properties of the scene. Orthogonal information can be gained when inhomogeneous sensor systems, which deliver different physical properties of the scene, are deployed. It can also originate from different processing methods applied to data from basically homogeneous sensors, e.g., when reflectance information is used to generate a depth map by means of photometric stereo, the information in the resulting depth map is orthogonal to any of the original reflectance images. Since the information in the depth map and in a reflectance image is not directly linkable, a sensible fusion can only take place on the abstraction level of features or classification results.

  A typical example is the combination of 3D data of the scene and its visual appearance. Whereas the 3D data contains information on the spatial arrangement of the scene, the visual appearance is mainly governed by the reflectance of the scene. Their fusion implies the abstraction from 3D and visual data to object points featuring a position (specified in the 3D data) and a reflectance (observed in the visual image).

These basic types of distributions of the information content in the image series to be fused can also be combined in many ways. A typical combination appears when disjoint subsets of the set of all images originate from homogeneous sensors and contain redundant information, which is fused in a first stage, see Fig. 2(a). The information of the intermediate fusion results, which can be redundant, complementary, distributed or orthogonal, is then fused in a second stage to obtain the final result.
Another typical case of combination occurs when the information content of the image series can be exploited in several ways, see Fig. 2(b). Here, the usually complementary or distributed information which is extracted by the different processing methods represents intermediate fusion results, which are then fused to the final result.
The latter type of information processing usually appears when multivariate image series, in which the images differ in more than one imaging parameter, are evaluated. As an example, when images from differently positioned cameras with different focus settings are used to capture a scene, combined stereo and focus series are recorded (Gheţa et al., 2006; Gheţa, Frese, Heizmann & Beyerer, 2007). Such image series can be evaluated by exploiting the stereo and the focus information separately, e.g., by means of the approaches depth from stereo and

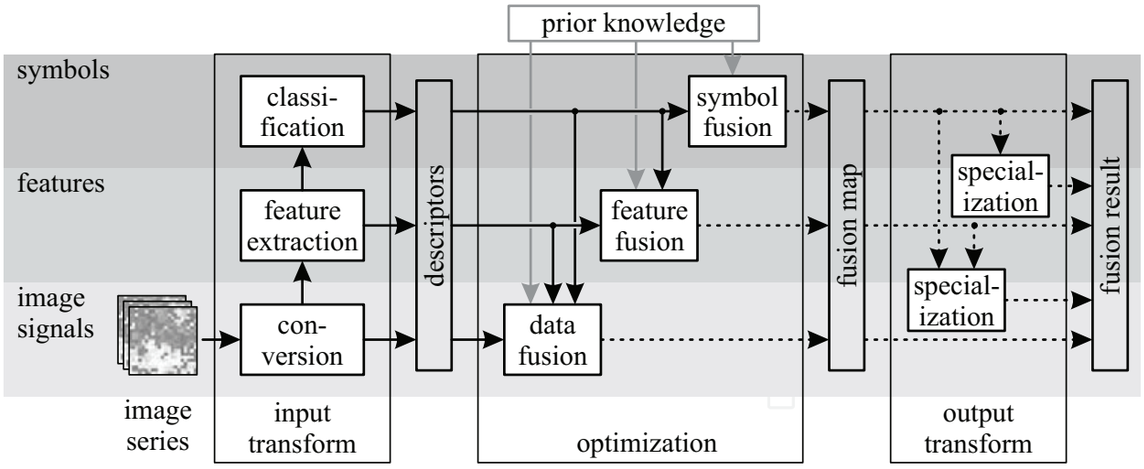


Fig. 3. General concept for image fusion (image series are indicated with continuous lines, single images are indicated with dotted lines): first, the images of the series are processed separately in the input transform in order to obtain meaningful descriptors of the useful information. During the optimization step, the useful information is selected from the descriptors. In this step, beneficial information from higher abstraction levels and prior knowledge can be included. The resulting fusion map is then converted into the desired fusion result by an output transform. During the input and the output transform, a change of the abstraction level can take place: whereas in the input transform, the image data may be lifted to a higher abstraction level in order to extract the useful information, the abstraction level of the fusion map as result of the optimization step may be lowered to obtain the desired fusion result.

depth from focus. Stereo and focus evaluations both use distributed information in the series, but each one with reference to the respective information content.

4. General concept for image fusion

Many approaches of image fusion can be traced back to a common concept with respect to an underlying processing scheme and abstraction levels involved, see Fig. 3. Although in many specific realizations, some of the processing steps or abstraction levels may be missing or cannot be strictly assigned to a step or level mentioned here, this general concept is justified since it can help to analyze existing fusion approaches and to synthesize new fusion approaches by suitably combining existing methods.

4.1 Processing scheme

In many image fusion approaches, a general processing scheme can be identified leading from the acquisition of the image series to the fusion result containing the concentrated useful information (Heizmann & Puente León, 2007), see Fig. 3.

Starting from the recorded images series, the first step is to transform the images into signals on an abstraction level where the actual combination of the useful information will take place. The aim of this input transform is to map the information in the image series onto significant descriptors such that the relevant information content becomes manifest for the following optimization step. The transform may include a conversion of the image data within the abstraction level of images or processing steps of feature extraction or

classification in order to obtain information on higher abstraction levels. Since the descriptors may belong to different domains such as the spatial domain, frequency domain, parameter domain, parameter frequency domain etc., operators suitable for these domains must be used. Common operators to extract significant image features in the input transform comprise geometrical, intensity, Fourier, wavelet or morphological transforms, principal component analysis, cross-correlation, or local operators, and may not only refer to spatial dimensions, but also to any other parameter dimension.

In the second step, an optimization is performed to select the useful information from the transformed image series. By means of a suitable quality criterion, the descriptors obtained by the input transform are assessed and combined to form a fusion map which contains the desired fused information and which is afterwards used in the output transform. The optimization takes place in the descriptor domain reached by the input transform. During the optimization process, prior knowledge (e.g., known constraints, physical laws, and required or desired properties of the fusion result) which is related to the fusion task must be included in order to ensure a consistent result. Common methods for the optimization step comprise linear and nonlinear operators, energy minimization methods, Bayesian statistics, Kalman filtering, and many other methods used in pattern recognition. An example of a classification-based optimization is given in Subsect. 5.3.

To establish a comprehensive formulation of the optimization problem, energy functionals have shown to be a universal approach (Clark & Yuille, 1990; Beyerer et al., 2008). To express relevant information contained in sensor data and prior knowledge, energy terms $E_k(r(\boldsymbol{x}),.)$ are introduced. They are modeled such that the relevant information is reflected in monotonic functions, which take lower values for more desirable properties of the fusion result $r(\boldsymbol{x})$ or intermediate descriptors. The optimization task is expressed for the energy functional

$$E(r(\boldsymbol{x})) := \sum_k \lambda_k E_k(r(\boldsymbol{x}),.)\,, \qquad k = \{1,\ldots,n\}, \qquad \lambda_k > 0\,. \tag{2}$$

The desired optimal fusion result is then obtained by minimizing the energy functional $E(r(\boldsymbol{x}))$ with respect to the fusion result $r(\boldsymbol{x})$. The energy formalism has several advantages: the fusion task is represented implicitly and compactly, all kinds of information and constraints can be included by introducing suitable energy terms, and the relevance of different contributions can be considered explicitly by adjusting the weights λ_k. However, the main drawback of the energy formalism is that there exists no universally applicable method for minimizing the energy functional $E(r(\boldsymbol{x}))$. Suitable minimization approaches strongly depend on the information used for the fusion, and hence, an approach found suitable for a specific task is hardly transferable to a different task.

In the last step, the fusion map representing the fused information is used as a construction plan for building the fusion result. To obtain the fusion result in the desired form, the fusion map is converted to the desired abstraction level. Depending on the domains of the descriptors and the optimization result in comparison to the abstraction level of the desired fusion result, a specialization may be used to lower the abstraction level. Examples of output transforms are the use of the fusion map as a look-up table or the trivial identity, e.g., in the case of depth maps which are obtained from fusing focus series, see Subsect. 5.1, or in the example of defect detection presented in Subsect. 5.2.

4.2 Abstraction levels

The fusion of the information in an image series can take place on different abstraction levels, see Fig. 3. In the following, three main abstraction levels are introduced: the level of the

image data itself, the level of features which are extracted from the image data and which are usually used to describe scene properties, and the level of symbols which can be obtained as results of a classification step. Apparently, the assignment of a specific fusion approach to a distinct abstraction level cannot always be strict, since, e.g., the image values themselves may be interpreted as features in certain cases. The differentiation of three abstraction levels is rather intended to demonstrate how common methods of image processing and pattern recognition fit into the introduced concept for image fusion.

– *Data-level fusion* (pixel-level fusion): in this case, the combination of information takes place on the level of the image data itself, i.e., on the intensity values, without any abstraction step. The precondition of a data fusion is that the image series contains redundant or complementary information and that the images have been recorded with homogeneous sensors, such that the image intensities refer to the same physical properties of the scene.

  A typical objective is to obtain a fusion result with an image quality which is better with respect to some quality criterion, for example by means of a concurrent (e.g., to reduce the sensor noise) or a complementary fusion (e.g., to synthetically enhance the depth of field).

– *Feature-level fusion*: here, the combination of information takes place on the level of features which must have been extracted from the image series before. These features may be generated from single images (e.g., local texture features) or by a simultaneous evaluation of the entire series (e.g., the variance of intensities for an image location within a series). While the latter case applies to distributed information recorded with homogeneous sensors, possibly inhomogeneous sensors which allow the extraction of at least comparable features are sufficient in the former case.

  A first typical task of feature-level fusion is to improve the extraction of image features, e.g., with respect to their accuracy or reliability. A second typical task is to gain access to information (e.g., features) which is distributed over the series.

  An example of the latter task is the fusion of images obtained from a camera array which is equipped with different spectral filters. A possible approach for the spatial and spectral assessment of the scene is given by a region based fusion: the properties of regions in the image series are used as features and fused with respect to their disparity (to obtain the spatial reconstruction) and the spectral intensities (to obtain a spectral characterization of the scene) (Gheţa et al., 2010).

– *Symbol-level fusion* (decision-level fusion): symbolic information, which is gained by a preceding feature extraction and a classification from single images, is combined. A fusion on this abstraction level imposes the least restrictions to the relation of information in the image series and to the choice of sensors: any type of relation and inhomogeneous sensors are admissible, if the symbolic information resulting from the respective classification approaches is connectable.

  The objective of symbol-level fusion is mainly to improve the accuracy and reliability of a classification, e.g., for defect detection or object recognition. In this case, the symbolic information is given by object hypotheses established by single sensors, which are fused to a consolidated hypothesis.

During the fusion on a certain abstraction level, it may be necessary or reasonable to use information from a higher abstraction level which has been processed from single images prior to the actual fusion step.
As an example, in order to fuse complementary information, the areas in the location-parameter domain containing the desired useful information must be identified, if

they are not a priori known. This identification requires a criterion which is usually at least on the abstraction level of features. In the application example of Subsect. 5.1, the criterion for identifying focussed image regions is a contrast measure on the level of features.
An important practical issue concerns the question of which abstraction level should be chosen to solve a given fusion task. There exists a tradeoff between a moderate implementation effort and an optimal exploitation of the useful information in the image series: concerning the necessary development and implementation effort, a fusion approach on a higher abstraction level is often easier to realize in comparison to lower abstraction levels. For the abstraction of single images to a higher level, standard algorithms of image processing and pattern recognition can be used in most cases. The optimization is then often performed by relatively simple operations of logical combination.
However, the quality of a fusion result obtained on an abstraction level which is as low as possible is mostly superior to the results obtained on higher abstraction levels. This can be traced back to the modification and potential reduction of the useful information during the abstraction step applied to the single images. When the information fusion is performed on a lower abstraction level, the processing which must be individually adapted to the information contained in the images and to the fusion task can ensure that the useful information is conserved for the fusion result at the best.

5. Examples

5.1 Fusion of focus series

The following example is used to demonstrate the concepts shown above, see Fig. 4: in order to evaluate a firing pin print for an automated database search, a visual image which shows the impression in detail and a depth map reflecting the spatial structure of the impression are needed (Heizmann & Puente León, 2003; Puente León, 2002). As information sources, images taken with a camera which is equipped with a macroscopic lens are to be used. Whereas a focussed image cannot be taken from the whole impression due to the limited depth of field Δ_z of the macroscope used, the depth map is a feature image which is not directly accessible from the single images.
Both tasks can be solved by applying methods of image fusion to a focus series: a series of n images $g(\boldsymbol{x},z)$, $z \in \{z_1,\ldots,z_n\}$ with varied distance z of the camera to the scene is taken such that each surface point is depicted in focus in at least one image, i.e., $z_{i+1} - z_i \leq \Delta_z \; \forall i \in \{1,\ldots,n-1\}$, and each surface point is mapped onto the same element of the imaging sensor in all images of the series. Thus, the sensors of the focus series are commensurable, homogeneous, virtual, and collocated with respect to the image plane.
In the input transform, significant descriptors for focussed imaging must be extracted from the focus series. The useful information to describe a visually sharp image is the local contrast $v(\boldsymbol{x},z)$, which can also be used to determine the in-focus plane for each image point and thus the desired depth map. A suitable descriptor for the useful information is therefore located on the abstraction level of features and determined for each image, e.g., by using the local variance.
In the optimization step, the useful information is selected from the descriptors of the images. For both tasks, the scene distance leading to the maximal local contrast represents the desired information and must be determined for each image point, i.e., $\bar{d}(\boldsymbol{x}) := \arg\max_z v(\boldsymbol{x},z)$. Although this preliminary depth map indicates the optimal scene distances based on the sensor information, it is not yet satisfactory, since high depth steps and estimation errors may occur, e.g., in surface regions with faint texture. At this point, prior knowledge stating that

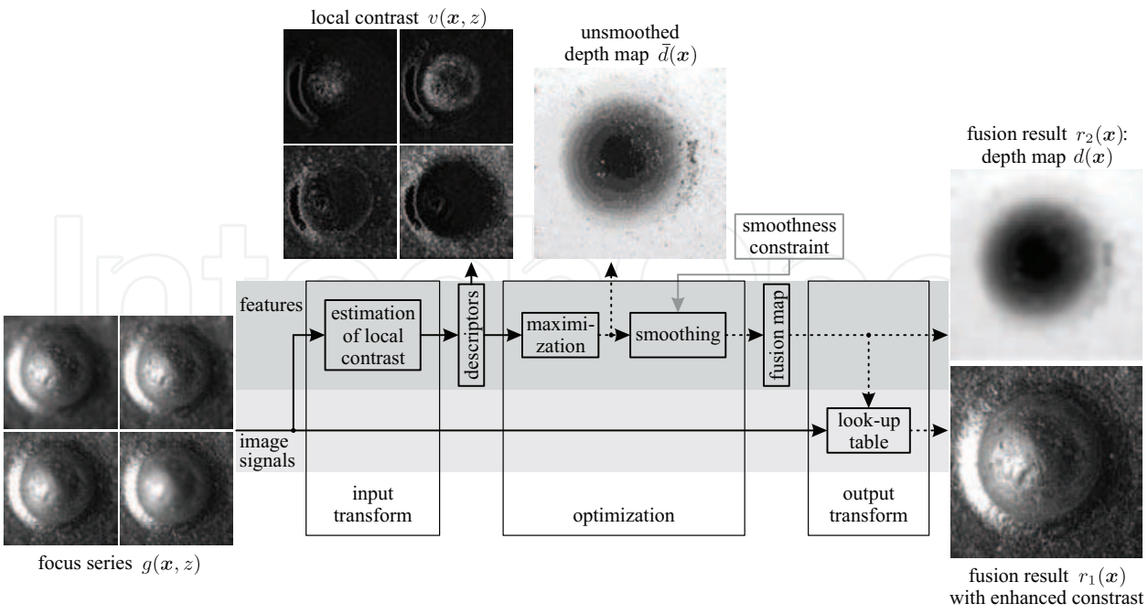


Fig. 4. Fusion of a focus series in order to obtain an image with synthetically enhanced contrast and a depth map (image series are indicated with continuous lines, single images are indicated with dotted lines): from the original image series $g(\boldsymbol{x},z)$, the local contrast $v(\boldsymbol{x},z)$ establishing a suitable descriptor of the useful information is estimated in the input transform (lighter areas have higher contrast). In the optimization step, a preliminary depth map $\bar{d}(\boldsymbol{x})$ is distilled first from the series of contrast images, which is then combined with prior information to a smoothed depth map $d(\boldsymbol{x})$ constituting the fusion map (darker areas are farther away than lighter areas). To obtain the desired image with enhanced depth of field, the fusion map is used as look-up table to compose the fusion result $r_1(\boldsymbol{x})$ from the focus series. The depth map itself represents the second desired fusion result $r_2(\boldsymbol{x})$.

the maximal inclination of the object surface is limited is incorporated in the fusion process by smoothing the preliminary depth map. The smoothed depth map $d(\boldsymbol{x})$ is then the result of the optimization step, i.e., the fusion map. It shows where the desired useful information is concentrated in the location-parameter domain.

The aim of the last step is to transform the fusion map into the desired result on the respective abstraction level. In order to construct the image with synthetically enhanced depth of field, the fusion map on the feature level must be specialized. To that aim, it is used as a lookup table: for each image point $\boldsymbol{x}$, the intensity value from the image with the respective scene distance $d(\boldsymbol{x})$ is selected from the focus series in order to form the fusion result, i.e., $r_1(\boldsymbol{x}) := g(\boldsymbol{x},d(\boldsymbol{x}))$. The second desired result, the depth map, is the fusion map, since it reflects just the vertical position of the in-focus plane, i.e., $r_2(\boldsymbol{x}) := d(\boldsymbol{x})$.

In this example, the construction of the image with enhanced depth of field can be classified as a fusion of complementary information. Although the actual combination of the input information takes place on the level of image signals, the evaluation and optimization of the useful information is performed on the abstraction level of features. The determination of the depth map, however, uses distributed information which is fused on the level of features.

5.2 Detection of defects based on illumination series

The second example is concerned with the detection of defects on membranes of pressure sensors. It is based on fusion of illumination series, and yields a feature image as the fusion result. The field of inspection is about 10 mm^2; the defects themselves are in the order of a few hundredths of a square millimeter. Figure 5(a) shows an example of a defective membrane illuminated with diffuse light. It features several local defects, whose actual position can be determined by comparing this image with the fusion result in Figure 5(c). It is obvious that images taken with a diffuse illumination hardly allow to discern intact regions from defective areas. Consequently, an inspection strategy based on an analysis of such images—i.e. without employing any fusion methods—is not likely to succeed.

Figure 5(b) shows eight of the $n = 16$ images of the original illumination series, which was obtained by rotating azimuthally a light source in steps of $\Delta\varphi = 22.5$ degrees. Due to the geometry of the machined membrane surface, images obtained with an illumination angle differing by 180 degrees one from another have a similar appearance. If now a faultless surface point $\boldsymbol{x}_0$ is observed at a certain illumination angle φ_0, its intensity $g(\boldsymbol{x}_0, \varphi_0)$ is particularly high, if the illumination direction is perpendicular to the machining grooves. Each facet of a groove acts then as a mirror that reflects the light incident from a direction perpendicular to the local direction of the groove. Consequently, the intensity signal $g(\boldsymbol{x}_0, \varphi)$ obtained for a varying illumination angle φ features a characteristic shape that allows to discern whether the point $\boldsymbol{x}_0$ belongs to a defective region or not.

By harmonic analysis of the signals $g(\boldsymbol{x}, \varphi_i)$ of the series, a suitable feature image can be defined as a measure $m(\boldsymbol{x})$ of the local defects:

$$m(\boldsymbol{x}) = \frac{|G(\boldsymbol{x}, f_\varphi = 1)|}{|G(\boldsymbol{x}, f_\varphi = 1)| + |G(\boldsymbol{x}, f_\varphi = 0)|}, \tag{3}$$

where

$$\begin{aligned} G(\boldsymbol{x}, f_\varphi) &:= \mathrm{DFT}_\varphi\{g(\boldsymbol{x}, \varphi)\} \\ &= \sum_{i=0}^{n-1} g(\boldsymbol{x}, \varphi_i) \cdot \exp\left(-\mathrm{j}2\pi \frac{i f_\varphi}{n}\right) \end{aligned} \tag{4}$$

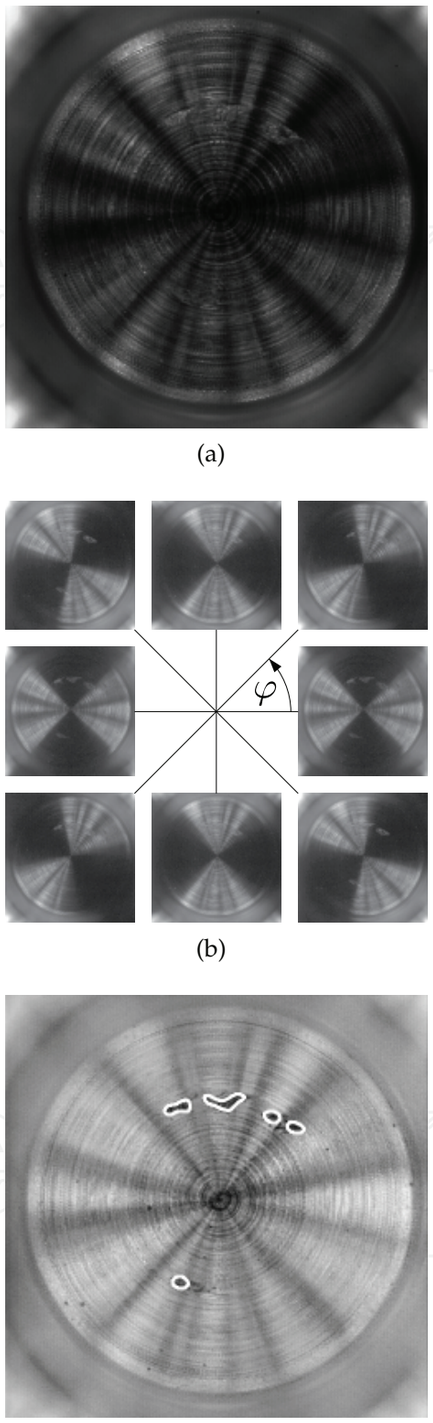
denotes the one-dimensional discrete Fourier transform (DFT) of the series with respect to the illumination angle φ.

Equation (3) computes a feature based on the comparison of two frequency components with respect to the image intensities at the location $\boldsymbol{x}$: the fundamental oscillation, and the DC component. It is easy to recognize that the values of $m(\boldsymbol{x})$ are all within the range $[0,\ 1]$, and that the ratio 0.5 is obtained when the energy of both components is the same. A value of $m(\boldsymbol{x})$ higher than 0.5 means that the fundamental oscillation—i.e. a non-defective groove texture—dominates potential defects at the location $\boldsymbol{x}$. Otherwise, $\boldsymbol{x}$ is likely to be a local defect.

Equation (3) performs the initial transform, after which an optimization and—if needed—a final transform could be performed. However, in the present case both the optimization and the output transforms are trivial, since

$$r(\boldsymbol{x}) = m(\boldsymbol{x}) \tag{5}$$

holds.



(a)

(b)

(c)

Fig. 5. Detection of defects on membranes of pressure sensors: (a) Image of a defective membrane obtained with diffuse illumination; (b) Image series of the defective membrane; (c) Fusion result with highlighted defects.

Figure 5(c) shows the feature image $r(\boldsymbol{x}) = m(\boldsymbol{x})$ obtained through fusion of the image series of Figure 5(b). The fusion result clearly highlights several defective areas, which appear darker than the faultless regions. For a better interpretation of the results, the results of a further defect detection step have been overlaid. To this end, an edge detection method based on a Laplacian-of-Gaussian (LoG) filter according to (Beyerer & Puente León, 1997) has been used.

5.3 Classification of defects based on illumination series

The method presented in the last subsection is based on a reflection model describing the intensities of a certain object or defect. Thus, both the design and computational effort will necessarily increase, if different types of defects need to be distinguished. Instead of a single feature image $m(\boldsymbol{x})$, a suitable set of features $\{m_i(\boldsymbol{x})\}$ will be necessary to discern the classes of defects in the feature space.

This example presents an alternative based on the systematic extraction of local features and a subsequent classification. A major advantage of this approach is that, after a suitable set of features has been defined, an arbitrary number of defects can be distinguished.

Additionally, if the extracted features are invariant against transforms considered to be irrelevant (e.g., translation, rotation, scaling or intensity), the computational costs remain acceptable. Moreover, thanks to the generalization capabilities of classifiers, a higher tolerance in the case of a class mismatch can be expected.

A common approach to construct a feature m out of an image $g(\boldsymbol{x})$ that is invariant against a certain transformation group T is integrating over this group:

$$m := \int_T f(t\{g(\boldsymbol{x})\})\,\mathrm{d}t = \int_P f(t(p)\{g(\boldsymbol{x})\})\,\mathrm{d}\boldsymbol{p}\,. \tag{6}$$

This equation is known as the Haar integral. The function $t \in T$ is a transformation parameterized by the vector $p \in P$, where P is the parameter space and f is an arbitrary, local kernel function.

In the following, we will focus on a single application scenario, in which different classes of varnish defects on wood surfaces are to be detected and classified. To achieve a maximal contrast, the pictures of the surface are taken under directional illumination, which is realized by means of a distant point light source, whose direction is described by a fixed elevation angle ϑ and a variable azimuth φ.

In this example, we aim at extracting invariant features with respect to the two-dimensional Euclidean motion, which involves rotation and translation in $\mathbb{R}^2$. The parameter vector of the transformation function is $\boldsymbol{p} = (\tau_x, \tau_y, \phi)^{\mathrm{T}}$, where τ_x and τ_y denote the translation parameters in x and y direction, and ϕ is the rotation parameter. The compactness and finiteness of this group guarantee the convergence of the integral (Schulz-Mirbach, 1995).

To process all images of the series by the Haar integral, the illumination azimuth φ needs to be added to the parameter vector $\boldsymbol{p}$, and the kernel function has to be extended accordingly to the third dimension of the input data. However, these modifications are beyond the scope of this paper. For a more comprehensive discussion of the feature extraction, we refer to (Pérez Grassi & Puente León, 2008). The resulting features do not only exhibit all invariant properties discussed above, but they also are invariant with respect to both illumination and contrast.

After extracting a set of ten features, a classification is performed by a Support Vector Machine. Since the inspected surfaces may feature areas with and without defects, the presented method is applied locally. To this end, the series of images is subdivided into local windows of 32×32

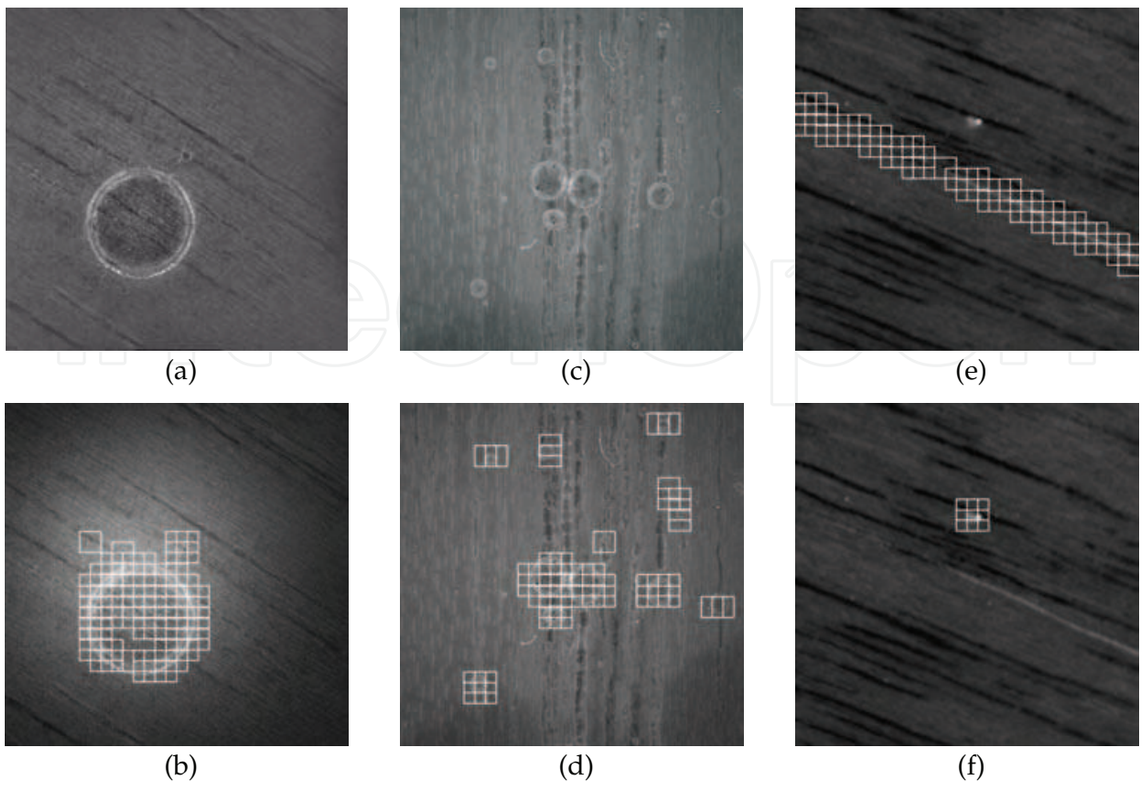


Fig. 6. Classification of different types of varnish defects on wood surfaces: (a) Surface with two craters of different radii (the smaller crater is located in the image center); (b) Classified crater regions based on image (a); (c) Varnished surface with bubbles and other varnish defects; (d) Classified bubble regions based on image (c); (e) Surface with bubble and fissure with classified fissure regions; (f) Surface with bubble and fissure with classified bubble regions.

pixels with a spatial overlap of 50 %. A 10-dimensional feature vector $\boldsymbol{m}$ is extracted from each window according to a list of kernel functions showing the same structure but with different parameters (Pérez Grassi & Puente León, 2008).
Five different classes were defined to train the system: no defect, bubble, ampulla, fissure, and crater. A group of 20 series of images of different wooden surfaces featuring different defects constituted the training list. To test the performance of the system, a disjoint list of series of images was used. Figure 6 shows a representative selection of the obtained classification results based on illumination series consisting of $n = 8$ images.
Figure 6(b) shows the inspection results for a surface showing two craters of different radii (Fig. 6(a)). Both defects were correctly classified. However, the results do not yield any information about the size of the detected defects, since craters of different sizes belong to the same class. The next example illustrates the detection of bubbles. Although the original image Figure 6(c) does not contain only bubbles, but also elongated defects, the classifier is able to distinguish between the different types of defects (Fig. 6(d)).
Finally, the Figures 6(e) and (f) show the results for a surface featuring a bubble and a fissure. The invariant approach introduced in this subsection is able to recognize both defects at one time using the same group of kernel functions.

6. Conclusion

Image fusion offers powerful tools to obtain desired information from a scene by using image series. The main precondition is to find an imaging setup with at least one varied acquisition parameter—illumination or observation parameter—such that the resulting image series contains the useful information in the form of redundant, complementary, distributed or orthogonal information. Once the image series has been acquired, a suitable procedure of image fusion can often be reduced to a standard concept for image fusion. It comprises a processing scheme consisting of an input transform, which converts the sensor information into significant descriptors, an optimization, which selects the useful information from the descriptors to generate a fusion map, and an output transform, which converts the fusion map into the desired form. The processing may take place on different abstraction levels—the levels of image signals, features, and symbols—, incorporating many common methods of image processing and pattern recognition.

7. References

Beyerer, J., Heizmann, M., Sander, J. & Gheţa, I. (2008). *Image Fusion: Theory and Applications*, Academic Press, Amsterdam, chapter Bayesian Methods for Image Fusion, pp. 157–192.

Beyerer, J. & Puente León, F. (1997). Detection of defects in groove textures of honed surfaces, *International Journal of Machine Tools and Manufacture* 37(3): 371–389.

Clark, J. J. & Yuille, A. L. (1990). *Data fusion for sensory information processing systems*, Kluwer Academic Publishers, Boston/Dordrecht/London.

Faugeras, O. D. & Luong, Q.-T. (2001). *The geometry of multiple images*, MIT Press, Cambridge (MA).

Gheţa, I., Frese, C. & Heizmann, M. (2006). Fusion of combined stereo and focus series for depth estimation, *in* C. Hochberger & R. Liskowsky (eds), *INFORMATIK 2006, Informatik für Menschen, Beiträge der 36. Jahrestagung der Gesellschaft für Informatik e.V. (GI)*, Vol. 1, Gesellschaft für Informatik (GI), Bonn, pp. 359–363.

Gheţa, I., Frese, C., Heizmann, M. & Beyerer, J. (2007). A new approach for estimating depth by fusing stereo and defocus information, *in* R. Koschke, O. Herzog, K.-H. Rödiger & M. Ronthaler (eds), *INFORMATIK 2007, Informatik trifft Logistik, Beiträge der 37. Jahrestagung der Gesellschaft für Informatik e.V. (GI)*, Vol. 1, Gesellschaft für Informatik (GI), Bonn, pp. 26–31.

Gheţa, I., Frese, C., Krüger, W., Saur, G., Heinze, N., Heizmann, M. & Beyerer, J. (2007). Depth map estimation from flight image series using multi-view along-track stereo, *in* A. Grün & H. Kahmen (eds), *Proceedings of 8th International Conference on Optical 3-D Measurement Techniques*, Vol. 2, Zürich, pp. 119–125.

Gheţa, I., Höfer, S., Heizmann, M. & Beyerer, J. (2010). A novel approach for the fusion of combined stereo and spectral series, *in* D. Fofi & K. S. Niel (eds), *Image Processing: Machine Vision Applications III*, Proceedings of SPIE Volume 7538. Paper No. 7538 0G.

Hartley, R. & Zisserman, A. (2004). *Multiple view geometry in computer vision*, 2nd edn, Cambridge Univ. Press, Cambridge.

Heizmann, M. (2006). Techniques for the segmentation of striation patterns, *IEEE Transactions on Image Processing* 15(3): 624–631.

Heizmann, M. & Beyerer, J. (2005). Sampling the parameter domains of image series, *in* E. R. Dougherty, J. T. Astola & K. O. Egiazarian (eds), *Image Processing: Algorithms and*

Systems IV, Vol. 5672 of *Proceedings of SPIE/IS&T Electronic Imaging*, pp. 23–33.

Heizmann, M. & Puente León, F. (2003). Imaging and analysis of forensic striation marks, *Optical Engineering* 42(12): 3423–3432.

Heizmann, M. & Puente León, F. (2007). Fusion von Bildsignalen, *tm – Technisches Messen* 74(3): 130–138. (in German).

Horn, B. K. P. & Brooks, M. J. (1989). *Shape from shading*, MIT Press.

Modersitzki, J. (2004). *Numerical methods for image registration*, Oxford University Press.

Puente León, F. (2002). Komplementäre Bildfusion zur Inspektion technischer Oberflächen, *tm — Technisches Messen* 69(4): 161–168. (in German).

Pérez Grassi, A. & Puente León, F. (2007). Translation and rotation invariant histogram features for series of images, *in* R. Koschke, O. Herzog, K.-H. Rödiger & M. Ronthaler (eds), *INFORMATIK 2007, Informatik trifft Logistik, Beiträge der 37. Jahrestagung der Gesellschaft für Informatik e.V. (GI)*, Vol. 1, Gesellschaft für Informatik (GI), Bonn, pp. 38–43.

Pérez Grassi, A. & Puente León, F. (2008). Invariante Merkmale zur Klassifikation von Defekten aus Beleuchtungsserien, *Technisches Messen* 75(7-8): 455–463.

Schulz-Mirbach, H. (1995). *Anwendung von Invarianzprinzipien zur Merkmalgewinnung in der Mustererkennung*, VDI Verlag, Düsseldorf.

Shapiro, L. G. & Stockman, G. C. (2001). *Computer vision*, Prentice Hall, Upper Saddle River (NJ).

Xin, B., Heizmann, M., Kammel, S. & Stiller, C. (2004). Bildfolgenauswertung zur Inspektion geschliffener Oberflächen, *tm — Technisches Messen* 71(4): 218–226. (in German).

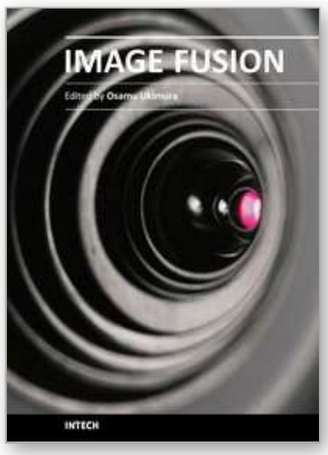


Image Fusion
Edited by Osamu Ukimura

ISBN 978-953-307-679-9
Hard cover, 428 pages
Publisher InTech
Published online 12, January, 2011
Published in print edition January, 2011

Image fusion technology has successfully contributed to various fields such as medical diagnosis and navigation, surveillance systems, remote sensing, digital cameras, military applications, computer vision, etc. Image fusion aims to generate a fused single image which contains more precise reliable visualization of the objects than any source image of them. This book presents various recent advances in research and development in the field of image fusion. It has been created through the diligence and creativity of some of the most accomplished experts in various fields.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen