Discrimination of soil texture and contour recognitions during archaeological excavation using Machine Learning

IOP Conf. Series: Materials Science and Engineering **364** (2018) 012042 doi:10.1088/1757-899X/364/1/012042

Discrimination of soil texture and contour recognitions during archaeological excavation using Machine Learning

I Cacciari[1], G F Pocobelli[2], S Cicola[2], S Siano[1]

[1] Istituto di Fisica Applicata "Nello Carrara", Consiglio Nazionale delle Ricerche, v. M. del Piano 10, 50019 Sesto Fiorentino (FI), Italy
[2] Cooperativa Archeologia, v. L. La Vista, 5, 50133 Firenze, Italy
i.cacciari@ifac.cnr.it

Abstract. The most time-honoured tool for understanding the processes of the human past is represented by archaeological excavation. By examining an area at discrete temporal periods, archaeologists are literally able to look backwards in time: they can analyse incomplete material records in order to understand and reconstruct the cultural history of an area at particular moments in time. Since the digging process destroys the site forever, great care must be paid during both the excavation and the documentation. In general, after a stratum has been completely excavated, both the floors and walls are cleaned and made ready for documentation. Photos of both the sides and bedrock of a given excavation are collected, and several sketches of what the archaeologists have seen in the trenches are made. In these drawings are delineated the features and shapes of artefacts on the horizontal plane. In addition, depending on the colours and similarities of the textures, drawing are also made of the archaeological layers. This approach is time-consuming, is affected by human ability, and does not make possible a prompt digitization of the results. Within this context, the automatized identification of archaeological stratigraphy during excavation work is welcomed by archaeologists. Here, a k-means unsupervised machine learning algorithm has been used for colour clustering digital images of excavation sites. The algorithm that we have developed attempts to enhance the colour similarity while keeping the colours separate one from another as much as possible. The main idea is that pixels belonging to the same colour cluster are a part of the same layer. Once the layer has been identified, a statistical approach based on Haralick features is used to characterize each strata in terms of texture. Unsupervised machine learning combined with texture analysis could become a good practice in speeding up the documentation work of archaeologists and paving the way towards the creation of an "automated archaeologist".

1. Introduction

One of the key concept in modern archaeological theory and practice is stratigraphy. Although it derives from geology, it refers to the long-term build-up of consecutive layers of soil material that are due to both human and geological activities. Since archaeological findings are generally located below the surface of the ground, the most important tool for understanding the human past is represented by a controlled exploration of the layers below the surface: in other words, by means of excavation. After the said excavation has been completed, the archaeologist studies these layers in order to comprehend the historical processes of the site formation. In modern archaeology, lithological criteria are used to characterize the basic archaeological levels: these generally include colour, texture, hardness of the soil, etc. Since excavations come across different layers at various elevations under the surface, one of the main tasks of the archaeologist during an excavation is to distinguish the various layers. Once a layer

has been completely excavated, both the wall and the floor of the trench are cleaned and then prepared for documentation. In general, archaeologists take photos of both the sides and the bedrock of the excavation, and then make a sketch of what they see. The extent, shape of the features, artefacts and layers in the horizontal plane are delineated by means of these drawings. However, this part of the archaeologist's work is not only considerably time-consuming and may be affected by human skill, but it also complicates the digitalization of the results. Therefore, any attempts at automatizing the identification of layers during excavation work are welcomed by the community. The aim of this work is to explore the possibility of creating an "automated archaeologist" [1] who can offer a prompt guide for "human archaeologists" in recognizing the archaeological layer during the excavation and in providing digital images of it. This, in turn, would simplify the drawing step and hence speed up the documentation of excavations. The "automated archaeologist" that we are imagining is based on image clustering created by Machine Learning (ML). This approach has been considered suitable for classifying regions (of excavation images) that may be considered similar (cluster identification). The potential of ML has been demonstrated in a variety of different fields [2-7]. However, very few examples of ML used in archaeology are known in the literature. Nonetheless, it has been applied to assessing the complementary value of geochemistry and machine-learning on predictive modelling in archaeology [8], in order to offer a complete methodology for the classification of archaeological data [9], as well as to a content-based image retrieval system for historical glass and to an automatic system for medieval coin classification [10].

In this work, ML has been used for speeding up layer identification that provides digital images of the archaeological excavation site. Among the different types of ML tasks we have considered unsupervised learning and, in particular, image clustering. The algorithm that we have developed is based on k-means. This makes it possible to enhance colour similarity and to keep colours separate from one another as far as possible: in other words, it divides the original image into k regions in which similar data points are grouped together, and these groups differ significantly one from the other. The work is divided as follows: section 2 introduces the images considered for colour clustering, the k-means algorithm, and the texture analysis performed in order to characterize the colour clusters obtained. Section 3 describes the experiments performed on the images and the results obtained. Lastly, Section 4 presents our conclusions as well as future perspectives.

2. Materials and methods

2.1. Images for colour clustering

In this work, two types of images have been colour clustered: namely, a mockup and several archaeological sites. The mockup was prepared in such a way as to simulate an excavation site characterized by different colours and textures. A green plastic dustsheet was spread under a wooden box; the box was then filled with soil (A) and peat (B1-B6) (figure 1).

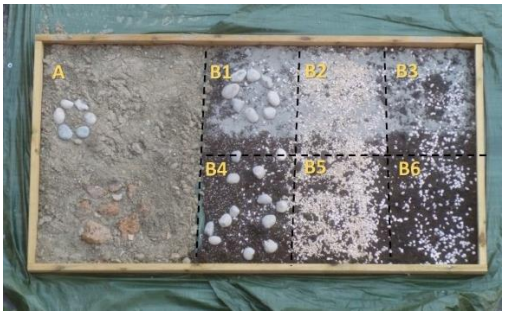

Figure 1. Mockup prepared in order to simulate an excavation site characterized by different texture and colours. A) six pebbles were placed on an almost uniform soil layer. A peat layer was partially covered by: B1) sand and a circle of 8 pebbles, B2) sand and a large quantity of gravel, B3) sand and a small quantity of gravel, B4) randomly-placed pebbles, B5) a large quantity of gravel, B6) a small quantity of gravel.

A pebble circle was introduced on the soil layer (A) to simulate the case of a uniform soil background in which the anthropic environment was marked out by areas with different textures and

colours (pebbles). Moreover, six different areas were prepared on the peat background to simulate different combinations of texture and colour. A digital camera was used to image the scene with a 1280x960 resolution. Three actual archaeological sites were also considered for colour clustering. The images were photographed during excavation campaigns in central Italy. On the first site, a portion of ancient stone pavement (stratigraphic unit) emerges from the background of the excavation site. The second site represents a portion of a building, two orthogonal walls of which are recognizable. The third site represents another portion of that particular building: in the background are bricks that probably resulted from the collapse of the roof. The excavation limit is also visible.

2.2. Archaeological texture

In general, the archaeological surface may contain minute variations in colour, texture, composition and hardness. Some of them have a visual nature and others, a tactile nature. In archaeology, texture is considered a surface attribute that has a visual or tactile variety that may characterize its appearance. On the other hand, in image processing, texture describes the amplitude patterns and quantifies the spatial arrangement of colour or intensities in an image or in a selected portion of one. Within this framework, one of the most effective tools for quantifying the perceived texture of an image is based on the grey-level co-occurrence matrix (GLCM, [11]).

The elements of the co-occurrence matrix measure the number of times that different combinations of pixel pairs of a specific grey level occur in an image for various directions (ϕ) and different distances (δ) [12]. Given an $M \times N$ neighbourhood of an input image with G grey levels, let $f(m,n)$ be the intensity at pixel (m,n) of the neighbourhood. The element (i,j) of the GLCM is consequently defined as follows:

$$P(i,j|\Delta x,\Delta y) = \frac{1}{(M-\Delta x)(N-\Delta y)} \sum_{n=1}^{N-\Delta y} \sum_{m=1}^{M-\Delta x} A \tag{1}$$

where

$$A = \begin{cases} 1 & \text{if } f(m,n)=i \text{ and } f(m+\Delta x, n+\Delta y)=j \\ 0 & \text{elsewhere} \end{cases} \tag{2}$$

and $\delta=(\Delta x^2+\Delta y^2)^{1/2}$, $\phi=arctg(\Delta y/\Delta x)$. In this way, the matrix element P contains the second-order statistical probability values for changes between grey levels i and j at a particular displacement distance d and at a particular angle θ. For this study, we considered images with G=256 and δ=1. A normalized symmetrical matrix was computed by summing up the four matrices ϕ=0°, 45°, 90°, 135° and normalized by dividing each entry by the total number of pixel pairs. This made it possible to avoid a dependency of direction. The corresponding normalized co-occurrence values are in the [0, 1] interval, and this enabled them to be considered as probabilities. A number of textural features can be extracted from the co-occurrence matrices. Here, we considered energy to be defined as follows:

$$E = \sum_i \sum_j P_{i,j}^2. \tag{3}$$

It returns the greyscale distribution homogeneity of images, and measures the textural uniformity of an image. Information on the random nature of the spatial distribution is then supplied. Energy assumes its highest values when grey-level distribution has either a constant or a periodic pattern. In a homogeneous image, very few dominant grey-tone transitions are expected, and the corresponding co-occurrence matrix has fewer entries of a larger magnitude, thus resulting in a sizable value for the energy feature. Because this feature is generally useful for highlighting continuity and geometry, it was considered suitable for this work [13, 14].

Moreover, the size of the neighbourhood partly determines the success of a texture-based image analysis. If the window size is too small, not enough spatial information can be extracted to distinguish between various different features; on the contrary, if it is too large, it could overlap different features and introduce spatial errors [15]. In this work, the same aspect ratio of the original image was used for

neighbourhood areas. For each neighbourhood size considered here, ten calculations chosen randomly from different areas of the same cluster were performed and their mean value was considered. The associated errors were estimated as the maximum deviations.

2.3. K-means colour clustering algorithm

ML is a powerful tool for solving a variety of problems, ranging from the visualization of high dimensional and cluster identification to pattern recognition. The basic aim of ML is to learn to identify automatically valid and potentially useful patterns and to make intelligent decisions based on data. In this work, we have considered unsupervised ML. This approach makes it possible to model a hidden structure or distribution in unlabelled data in order to learn more about given data. Clustering represents one of the main approaches of unsupervised learning: it assigns a set of inputs into subsets called clusters, so that each subset ideally shares some common characteristic and is able to place any new input within the appropriate cluster. Therefore, clustering can recognize different patterns in a dataset. In image processing, ML is used to divide a digital image into different regions for border detection or object recognition. In particular, it has been applied for different purposes in medicine [16-20], biology [21, 22], agriculture [23, 24], geophysics [25, 26], remote sensing [27, 28], security and crime detection [29], marketing and consumer analysis [30, 31], document clustering [32, 33], and automatic image annotation [34].

One of the simplest unsupervised algorithms that can be used to solve a clustering problem in digital images is represented by k-means. The algorithm developed in this work is based on k-means. It follows a simple and rapid way to cluster a digital image by means of a certain number of clusters established a priori. For each cluster, the main idea is to define a centroid (barycentre). The next step is to consider each point in the dataset and to associate it with the nearest centroid. This step is concluded, and an early groupage is completed, when no point is pending. New k centroids are then recalculated as the barycentre of the k clusters obtained in the previous step. The centroids are calculated according to the Euclidean distance between the colour dimensions and the centroids. Once new k centroids have been calculated, a new linking has to be established between the same dataset and the nearest new centroid (minimum distance). The iteration stops when the location of the k centroids no longer changes.

As described above, clustering enables dataset points to be grouped with some similarity along a dimension, while the points that differ from each other are kept further apart. In the case of digital images, the dimension used is generally colour, because the human vision system chooses colour, rather than shapes and texture, as its main discriminant feature. In this work, we have considered RGB, HSL and CieLab colour space. Images of excavation sites have been clustered using k-means in such a way that the different regions of the image are marked by k colours and the boundaries are revealed by separating the different regions. The outputs of the algorithm are k images in which all the non-zero pixels represent the object in the cluster. By assigning an 8-bit number to each pixel in a cluster, a composite image (in a false colour) is then produced: this helps the drawing of the layer contour using standard edge-detection techniques.

3. Results and discussions

The number of clusters (established a priori) is one of the main issues of the k-means algorithm: a poor colour clustering and, hence, a poor archaeological level identification may be the result of an underestimation of the k number. The best number of colours should be recognized by performing some preliminary tests. To this end, we performed colour (RGB) clustering on a portion of area A of the mockup, using different k values. In the case of k=2, it was observed that both the plastic dustsheet and the pebbles circles were considered as belonging to the cluster of the same colour, and the wooden box and soil, to the other. The wooden box in the corresponding composite was scarcely visible, and this underestimation of the colours number determined a mistake in the identification of the contour. The same held true for k=3 and k=4. The best colour clustering (for the portion of area A considered) was achieved with k=5. In this case, the anthropic environment (pebbles) was clearly separated from the

background (dustsheet and soil). The contour of pebbles circle and of the wooden box could thus be drawn accurately.

In these preliminary tests we considered RGB colour space. In general, situations are likely to be found in which other colour spaces might be more suitable for colour clustering. To this end, we also considered HSL (Hue Saturation Lightness) and CieLab [35] for colour clustering of the portion of area A in the mockup. The results obtained with k=5 are presented in figure 2. It is interesting to note that, in this case, the best clustering was obtained with RGB colour space, because all the details were clearly separated. In both clusterings with HSL and CieLab colour spaces, the green dustsheet was considered to belong to the same cluster of the upper part of the wooden box, and the white pebbles, to the left part of the wooden box. The results of these preliminary tests led us to make two considerations: first, not only the k number, but also the colour space, may have impact on the cluster contour; secondly, since these two features should be tested before drawing the contour, this preliminary tests could be considered to be part of a good practice.

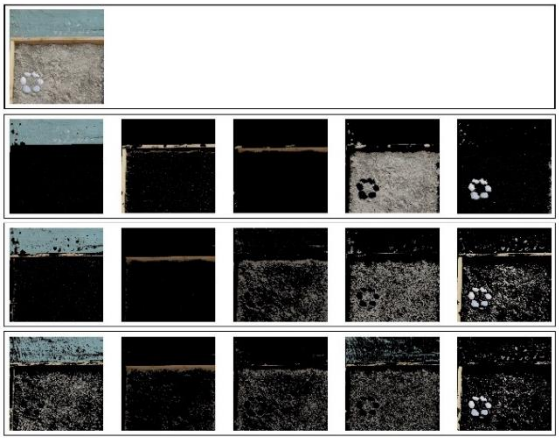

Figure 2. First row: portion of the mockup (original image), from second to fourth rows, RGB, HSL and CieLab colour clustering with k=5 respectively.

Analogous tests (number of colours and colour space) were performed on the other areas of the mockup: for the B areas, the best colour clustering was achieved with RGB and k=4. In particular, in the B1 case, the colour clustering has clearly separated the anthropic environment (pebbles) from the background. This represents a demanding case for colour clustering, because there is no strong colour difference between sand, gravel background, and the pebbles circle as there is in the A and B4 areas. This result may suggest that the algorithm could be useful also in cases in which the main differences are in the texture, rather than in the colour. In the B2 case, the background is fairly uniform (a sand layer partially covers the gravel); the texture is practically uniform, but is made with materials of two colours. The clustering brought out the two materials quite well, even if it was not possible to draw a contour. The same held true also for cases B3, B5 and B6, in which different combinations of sand, gravel and peat were considered for the background.

The texture clustering is considered by archaeologists to be equally interesting as colour clustering. Once the image of an archaeological excavation site has been colour clustered, it is remarkable for characterizing each cluster in terms of mean texture.

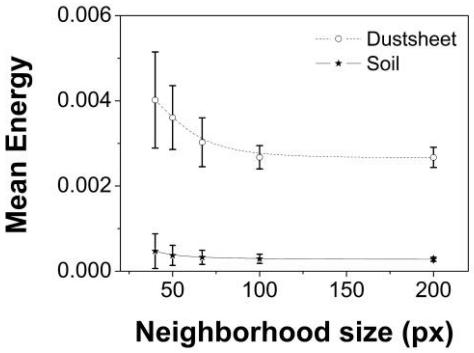

Figure 3. Mean energy calculated with different neighbourhood sizes in the dustsheet and soil areas in a portion of area A.

To this end, the energy (as one of the Haralick features that can be obtained from GLCM) was calculated with different neighbourhood sizes in the dustsheet and soil regions (in a portion of area A of the mockup). The results obtained are summarized in figure 3. With smaller neighbourhood sizes, the mean energy was greater than with ones of a larger size, and a downward trend was then established; then, with larger neighbourhood sizes, the mean energy values tended to be constant. In this work, we have suggested to use these values to characterize each colour cluster and, hence, the archaeological level. It is interesting to note that the dustsheet had greater energy than the soil. This corresponded to a greater degree of uniformity in the image and, therefore, to a smoother texture.

Moreover, colour clustering was also performed on images of two excavation sites. Figure 4 shows the composite image (RGB, k=3) and the original image (site 1) with, overlaid, the regional contours obtained by means of edge-detection. Even if the processing of the image is of poor quality, the stones of the pavement have been highlighted quite well. Analogously for site 2, the two portions of wall have been contoured quite well (figure 5).

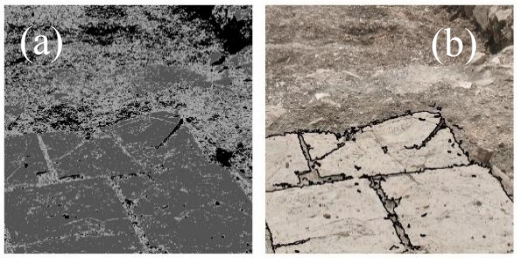

Figure 4. Excavation site 1: (a) composite image obtained with k=3 and RGB colour space; (b) original image with overlaid contours of the layer.

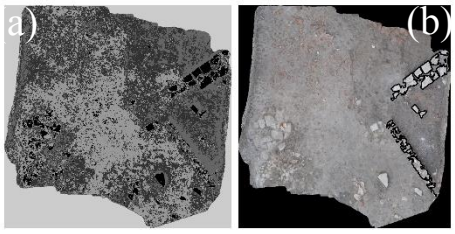

Figure 5. Excavation site 2: (a) composite image obtained with k=4 and RGB colour space; (b) original image with overlaid contours of the archaeological layer.

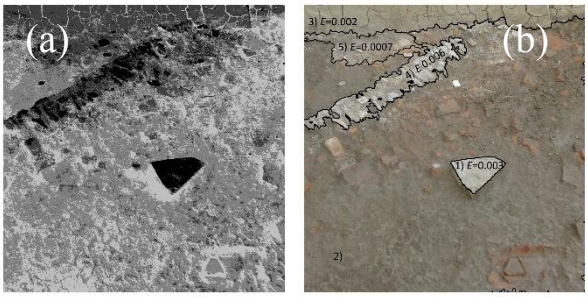

Figure 6. Excavation site 3: (a) composite image obtained with k=5 and RGB colour space; (b) original image with overlaid contours of layers and corresponding mean energy values.

The case of site 3 is particularly interesting. In figure 6 two layers are contoured: the excavation limit (3), a portion of a wall (4) and a region with a different texture (5) placed between (4) and (3). The stone (1) is also identified. This is interesting, since the algorithm may spread its potential to findings that are not actually levels, but are of natural interest to archaeologists.

Moreover, the colour clustering of site 3 highlights a layer (1) that seems to have a smoother texture than that of the background (2) (figure 6). The colour clusters have been characterized in terms of texture by calculating the corresponding mean energy. As each neighbourhood area should be included in the colour clusters, it is not possible to perform a texture analysis with larger-sized neighbourhoods. For this reason, for the image of site 3 (1100x1120 pixel) we have considered an area of 44x45 pixels. In addition, the results of the mean energy calculations are reported in figure 6. This site is extremely interesting, because different combinations of texture and colours can be observed. The results show that the two layers (labelled as cluster 4 and cluster 5) have approximately the same texture, even if they are chromatically different. Clusters 3 and 5 seem to be similar in terms of colour, but show different textures.

4. Conclusions

In this work, we aimed to explore the use of machine learning to automatize some tasks of archaeologists. This approach could help in identifying archaeological levels during an excavation campaign and in reducing the time spent for digitalizing the results. To this end, an unsupervised ML algorithm has been developed to colour-cluster digital images of archaeological excavation sites. The results obtained with a k-means algorithm and edge detection represent the first demonstration that layers can be readily and properly identified. The layers contoured have been characterized in terms of textural uniformity by calculating the corresponding energy. This proves that the combination of ML and texture analysis can serve to become a valid practice for speeding up the documentation work of archaeologists.

5. Acknowledgements

The present work has been carried out within the framework of the Archeo 3.0 project funded by the Tuscan Region (POR FESR 2014-2020).

6. References

[1] Barceló J A 2007 Proc. of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology, 413-17
[2] Dua S, Du X *Mining and Machine Learning in Cybersecurity*, ed. Auerbach Publications (Boston, USA)
[3] Gyorfi L, Ottucsak G, Walk H 2012 *Machine Learning For Financial Engineering*, ed. H. Walk (London: Imperial College Press)
[4] Clifton D A 2016 *Machine Learning for Healthcare Technologies*, ed D A Clifton (London: Institution of Engineering and Technology)
[5] Cracknell M J, Reading A M 2014 *Comput. Geosci.* **63** 22-33

[6] Harvey A S, Fotopoulos G 2016 *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **XLI-B8** 423-30
[7] Lary D J, Alavi A H, Gandomi A H, Walker A L 2016 *Geosc. Front* **7** 3-10
[8] Oonk S and Spijker J 2015 *J. Archaeol. Sci.* **59** 80-88
[9] Hörr C, Lindinger E and Brunnett G 2014 J. *Comput. and Cult. Herit.* **7** 1-23
[10] van der Maaten L, Boon P, Lange G, Paijmans H and Postma E 2006 *Computer Applications and Quantitative Methods in Archaeology* ed J. T. Clark and E.Hagemeister Proc. of the 34th Conf, United States
[11] Haralick R. M., Shanmugam K and Dinstein I 1973 *IEEE Trans. Syst. Man Cybern. Sys.* **3** 610-21
[12] Nixon M, Aguado A 2008 *Feature extraction & Image Processing* ed M. Nixon (Oxford: Oxford Press)
[13] Preethi G and Sornagopal V 2014 *Int. Conf. on Green Comput. Com. and Elect. Eng.* 1-6
[14] West B P, May S. R, Eastwood J. E and Rossen C 2002 *The Leading Edge* **21** 1042-49
[15] Toennies K D 2012 *Guide to Medical Image Analysis: Methods and Algorithms* (London: Springer-Verlag)
[16] Halberstadt W and Douglas T S 2008 *Comput. Biol. Med.* **38** 165-70
[17] Liao L, Lin T and Li B 2008 *Patt. Recogn. Lett.* **29** 1580-88
[18] Hill J, Corona E, Ao J, Mitra S and Nutter B 2014 *Advanced Computational Approaches to Biomedical Engineering* ed. P. Saha, U. Maulik, S. Basu (Berlin: Springer)
[19] Fouad S, Randell D, Hisham A G, Landini M G 2017 *Ann. Conf. Med. Image Understanding and Analysis* 767-79
[20] Govindaraj V, Vishnuvarthanan A, Thiagarajan A, Kannan M, and Murugan P R 2016 *J. Clin. Exp. Neuroimmunol.* **1** 1-10
[21] Hruschka E R, Campello G B and de Castro L N 2006 *Inf. Sci.* **176** 1898-1927.
[22] Xu Y, Wu J, Yin C.-C and Mao Y 2016 *PLOS*
[23] Papajorgji P, Chinchuluun R, Lee W S, Bhorania J and Pardalos P M 2009 *Advances in Modeling Agricultural Systems* ed. P. J Papajorgji, P. M. Pardalos (New York: Springer)
[24] Chinchuluun R, Lee W S, Bhorania J, Pardalos P. M 2008 *Advances in Modeling Agricultural Systems* ed. P. J Papajorgji, P. M. Pardalos (New York: Springer)
[25] Song Y-C, Meng H-D, O' Grady M J and O'Hare G M P 2010 *Computat. Geosci.* **14** 263-71
[26] Sidorova V S 2008 *Lect. Notes Comput. Sc*. **18** 693-99
[27] Bi H, Sun J and Xu Z 2017 *IEEE Trans. Geosci. Remote Sens.* **55** 3531-44
[28] Hassanzadeh A, Kaarna A and Kauranne T 2017 *Scandinavian Conf. on Image Anal.* 169-80
[29] Grubesic T H 2006 *J Quant Criminol* **22** 77-105
[30] Wang Y-J and Lee H-S 2008 *Inf Sci* **178** 1087-97
[31] Li J, Wang K and Xu L 2009 *Ann. of Operations Res*. **168** 225-45
[32] Cai X and Li W 2011 *Inf Sci* **181** 3816-27
[33] Carullo M, Binaghi E and Gallo I 2009 *Pattern Recognit. Lett.* **30** 870-76.
[34] Favorskaya Lakhmi M, Jain C and Proskurin A 2016 *New Approaches in Intelligent Image Analysis*, ed. R. Kountchev, K. Nakamatsu (Springer)
[35] Oleari C 1998 *Misurare il coloure* (Milan: Hoepli)