Spatio-Temporal Quasi-Flat Zones for Morphological Video Segmentation

Jonathan Weber[1], Sébastien Lefèvre[2], and Pierre Gançarski[1]

[1] University of Strasbourg - LSIIT
{j.weber,gancarski}@unistra.fr
[2] University of South Brittany - Valoria
sebastien.lefevre@univ-ubs.fr

Abstract. In order to face the various needs of users, user-driven segmentation methods are expected to provide more relevant results than fully automatic approaches. Within Mathematical Morphology, several user-driven approaches have been proposed, mostly relying on the watershed transform. Nevertheless, Soille (IEEE TPAMI, 2008) has recently suggested another solution by gathering puzzle pieces computed as Quasi-Flat Zones (QFZ) of an image. In this paper, we study more deeply this user-driven segmentation scheme in the context of video data. Thus we also introduce the concept of Spatio-Temporal QFZ and propose several methods for extracting such zones from a video sequence.

Key words: Quasi-flat zones, video segmentation, segmentation personalization

1 Introduction

Following the increase of textual and then image data in personal databases and Web repositories, we are currently facing the same evolution with video data. Many video processing schemes or related use cases require a prior segmentation to get the objects-of-interest to be further processed. However, the segmentation of a given video is often not unique and depends on user needs. Thus it is necessary to rely on a segmentation method able to provide a personalized result.

Video segmentation methods designed within the framework of Mathematical Morphology may be clustered in two categories: automatic methods [1, 2] which do not require any user interaction (apart from parameter settings) and interactive methods [3] (but also video extension of [4]) where the user has to draw some markers over objects-of-interest in order to drive the segmentation process. Results returned by automatic methods are then not adapted to user needs and often face the problem of over-segmentation. Interactive methods are more time-consuming but provide a personalized result. Another solution to solve the problem of segmentation personalization is to provide an over-segmentation which will then be reduced by the user through region merging in order him to obtain the expected segmentation. Image over-segmentation may be achieved using flat zones [5] but it then results in an extreme over-segmentation. Quasi-Flat Zones have been introduced in order to reduce this over-segmentation while keeping interesting properties of flat zones. QFZ are based on a less restrictive criteria to build the regions, thus leading to larger regions, while keeping a low computational cost and region borders able to represent most of the frontiers between the objects-of-interest. Besides, Soille [6] notices that QFZ are not really segmentation methods but rather methods which split an image into puzzle pieces. Identifying the QFZ is then a preprocessing step in an image segmentation process based on merging of puzzle pieces. This merging may be driven by the user, thus solving the problem of segmentation personalization. Let us observe however that there is no definition of QFZ for video sequences yet.

In this article, we recall the QFZ definition in the framework of logical predicate connectivity introduced by Soille [7, 8]. We then extend this definition to video sequences, and study how it can be applied to video segmentation personalization. Finally we give some conclusions and indicate future research directions.

2 QFZ-based image segmentation

2.1 Logical Predicate Connectivity

QFZ rely on the concept of α-connected path. A path is said α-connected if all paths between any pair of its pixels are Lipschitz-continuous, thus leading to the following definition.

A path $\mathcal{P}$, considering a neighbourhood N, and composed of n pixels $(p_0, p_1, ..., p_{n-1})$ is an α-connected path (α-$\mathcal{P}$) if and only if:

$$\forall i \in [0, n-2], p_i \in N(p_{i+1}) \text{ and } |f(p_i) - f(p_{i+1})| \leq \alpha \tag{1}$$

This notion let us define the most simple QFZ, i.e., α-connected zones [9] which will be noted here α-CC. An α-CC is defined as:

$$\alpha\text{-}CC(p) = \{p\} \cup \{Q \mid \forall q \in Q, \alpha\text{-}\mathcal{P}(p,q) \neq \emptyset\} \tag{2}$$

The α-CC of a pixel p is then the set of pixels to which it is linked through an α-connected path. Let us observe that flat zones are a particular case of α-CC with $\alpha = 0$. The α-CC have the following hierarchical property which will be useful later in this paper:

$$\forall \alpha' \leq \alpha, \alpha'\text{-}CC(p) \subseteq \alpha\text{-}CC(p) \tag{3}$$

Segmenting an image into α-CC results in an under-segmentation. If α is set too high, it will lead to a chaining effect, which may result on a single QFZ for the whole image (this depends of the image under consideration and the α value of course). In order to counter this problem, new QFZ definitions based on α-CC have been elaborated. In a goal of unification of existing works, Soille and Grazzini [7, 8] have proposed a theoretical framework called logical predicate connectivity. We recall that a logical predicate P return true when the parameter satisfies the predicate, false otherwise. They define a new kind of QFZ (noted $(P_1, ..., P_n)$-CC here) which lead to QFZ satisfying all the n logical predicates.

Various predicates may be involved, for instance: global range predicate which checks if the difference between minimal and maximal values of pixels within a QFZ is below a threshold (ω); connectivity index which is the ratio between the number of 2-pixels α-connected paths and the number of 2-pixels paths within a QFZ. This predicate is verified if the index is higher than a threshold (β). The $(P_1, ..., P_n)$-CC thus consists in seeking, for each pixel p, the largest α-CC satisfying all the predicates. Thanks to the property 3, we know that if $\alpha' < \alpha$ then α'-$CC(p)$ is less or equal to α-$CC(p)$. When predicates are not verified for a given value of α, we can use this property to decide to decrement α in order to check if the predicates are verified for a lower value and to loop until finding the maximal value of α for which all the predicates are verified:

$$
\begin{aligned}
&(P_1, ..., P_n)\text{-}CC(p) = \\
&\qquad \bigvee \left\{ \alpha'\text{-}CC(p) \;\middle|\; \begin{array}{r} \forall k \in \{1, ..., n\} \;\; P_k\left(\alpha'\text{-}CC(p)\right) = true \\ \forall \alpha" \leq \alpha', \forall q \in \alpha'\text{-}CC(p), \;\; P_k\left(\alpha"\text{-}CC(q)\right) = true \end{array} \right\} \qquad (4)
\end{aligned}
$$

This theoretical framework is adapted to methods ensuring the unicity property. Indeed we are seeking the largest α'-CC verifying all logical predicates. It is thus not possible to consider methods which do not provide a unique QFZ segmentation. More than only a framework to unify existing definitions, the $(P_1, ..., P_n)$-CC also allows to elaborate new QFZ definitions. Three predicates are currently used within the QFZ: local range (α), global range (ω) and connectivity index (β). In the framework introduced by Soille and Grazzini, it is possible to include predicates related to other features (perimeter, area, etc.) but also to more complex descriptors (texture, gradient, etc.) as long as these predicates fulfill the condition defined in Eq. (4).

Some clues to using QFZ in multivariate images have been given by Soille [6]: α is assumed to be a vector with the same value in all components. Then, α may be easily ordered through a total ordering (decrementing $\alpha = (3, 3, 3)$ gives $\alpha = (2, 2, 2)$). Global range predicate is processed similarly, and is true only if it is verified marginally for all bands.

In the sequel of this article, we will denote by QFZ the colour QFZ built using $(P_1, ..., P_n)$-CC with only the global range predicate and a given α.

2.2 Filtering

QFZ suffer from the transition region problem. Transition regions are regions between two objects where a staircase phenomenon occurs on edge pixel values. This is due to the image discretization process and the subsequent value interpolation. This artefact leads to an over-segmentation near to the edges which will then be made of tiny QFZ. Some solutions have been proposed to solve this problem. Soille and Grazzini [8] define transition regions as QFZ containing only transition pixels. Every pixel which is not a local extremum is considered as a transition pixel. All QFZ corresponding to transition regions are removed, and remaining QFZ are enlarged using a region growing algorithm [10]. After the removal of these regions, the amount of flat zones is reduced significantly.

The solution proposed by Soille and Grazzini does not depend of any parameter and relies on a precise definition of a transition region. But from our experiments, we have observed that many regions of a few pixels remains after applying their strategy. These regions do not fit with the definition of transition regions, but are still sources of a high over-segmentation. Thus, a more efficient and robust filtering method is still lacking.

Other authors have proposed QFZ filtering methods using a QFZ minimal area thresholding step. Angulo and Serra [11] suggest to merge QFZ characterized by an area lower than a given threshold with the most similar neighbouring QFZ. With this method, no more transition region is present in the final segmentation. Zanoguera [12] removes QFZ with an area below a given threshould (thus including transition regions) before applying a Watershed transform to enlarge remaining QFZ in areas where small QFZ have been removed. Soille [13] proposes a filtering method based on an iterative increase of the minimal area, followed at each iteration by both a region growing algorithm relying on QFZ with area greater or equal to minimal area and an image simplification algorithm. The simplified image will then be segmented into QFZ at the next iteration. This process is repeated until filtered QFZ become stable. Following some ideas introduced in these methods, we have also design a filtering method. It relies on the Seeded Region Growing (SRG) algorithm [10] but we apply it on the QFZ rather than on the pixels. To do so, we consider a minimal area threshold similarly to existing approaches. We set all QFZ with an area greater or equal to this threshold as seeds for the SRG algorithm which is applied on the region adjacency graph. We thus obtain a much more reduced over-segmentation compared to the result obtained without filtering. The highest the area threshold is, the more reduced the over-segmentation is. But in the same time, it is much more probable to obtain an under-segmentation of some objects-of-interest. Our region growing being applied on QFZ rather than on pixels, the proposed method requires a low computational cost.

3 Extension to video data

3.1 Limits of a $3D$ straight extension

The most direct extension of QFZ to video sequences is to consider a video sequence as a 3-D spatio-temporal cube. We can reuse the existing definitions, thus changing only the neighbourhood considered (spatio-temporal rather than purely spatial).

Computing the $(P_1, ..., P_n)$-CC in $3D$, we obtain a higher spatial oversegmentation than in $2D$. On the *carphone* sample (Fig. 2.a) for $\alpha = \omega = 20$ (we will use these values in the sequel), we obtain on average $4441CC$ per frame in $2D$ vs. $6779CC$ per frame in $3D$ ($55040CC$ on the full sample). Indeed, by analysing the video sequences in $3D$, the considered neighborhood contains more pixels and therefore an α-CC contains more pixels (see *chaining effect* discussed above). This naturally increases the risk of violating one of the considered predicates. Thus, the largest α-CC satisfying all predicates is often produced with a small α value. This leads to tiny QFZ of only a few pixels, while such QFZ are unusable for video segmentation.

3.2 Sequential processing of spatial and temporal dimensions

As $3D$ approach is not suitable for video processing, we consider rather the $2D+t$ approach. In this approach, we successively (and no more jointly) consider the spatial and temporal dimensions, as illustrated in Fig. 1. We discuss here first the spatial to temporal $(2D + t)$ approach and then the temporal to spatial $(t + 2D)$ approach.

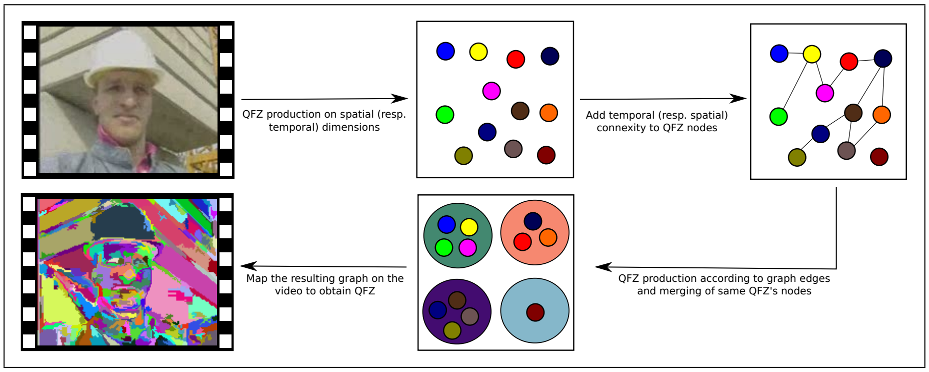
QFZ production on spatial (resp. temporal) dimensions
Add temporal (resp. spatial) connexity to QFZ nodes
Map the resulting graph on the video to obtain QFZ
QFZ production according to graph edges and merging of same QFZ's nodes

Fig. 1. Video Quasi-Flat Zones production by separated processing of spatial and temporal dimensions.

With the spatial to temporal approach, QFZ are first built on each frame independently. Then they are considered as nodes of a graph which are valued (here we consider the QFZ mean value). Edges are then introduced to temporally connect QFZ from successive frames and overlapping spatial coordinates. Each edge is valued by the difference between related node values. The new QFZ are the largest connected components of nodes whose connecting edges have a value less or equal to α and which do not violate any predicate. We observe that $(P_1, ..., P_n)$-CC produced significantly fewer regions in $2D + t$ $(23926CC)$, thus reducing the extreme segmentation we noted in $3D$ $(55040CC)$. This can be explained by the distinct computing of the two dimensions (spatial and temporal). In $2D+t$ the first processing (spatial only) produces spatially wider QFZ reducing the spatial over-segmentation. But, then the second processing (temporal) introduces temporal over-segmentation. This is due to the predicate constraints: the regions being spatially more extensive, they are less homogeneous and therefore may have significantly different means which will violate a predicate during the temporal computing of $(P_1, ..., P_n)$-CC.

With the temporal to spatial approach, QFZ are first built for each spatial coordinate independently, according to the temporal dimension. After this temporal processing, we therefore obtained an extreme spatial over-segmentation since, for each frame, each pixel belongs to a different QFZ. Similarly to the $2D+t$ case, we consider QFZ as nodes of a graph and apply the same process as previously, but considering here the spatial dimension instead of the temporal one. We note that $(P_1, ..., P_n)$-CC produces fewer regions $(16, 830CC)$ than the $2D + t$ approach due to a smaller temporal over-segmentation.

Let us observe that the $(P_1, ..., P_n)$-CC highlights an interesting phenomenon. Due to their different order when processing spatial and temporal dimensions, the approaches $2D+t$ and $t+2D$ induce different over-segmentations: a reduced spatial but high temporal one for the former, and a higher spatial but reduced temporal one for the latter. Nevertheless, both approaches provide better results than the $3D$ approach. Selecting between $2D + t$ and $t + 2D$ depends on the video under consideration. It may seem better to use the first approach with short videos of high-resolution, and to use the second for long videos of lower-resolution. Moreover, let us note that the spatial and the temporal processing are both relying on $(P_1, ..., P_n)$-CC, which guarantees the uniqueness of the result. Thus they also ensure this fundamental property.

3.3 Filtering

The filtering methods presented in section 2.2 can be extended to video data. As far as our method based on a minimum area threshold is concerned, adaptation depends on the chosen approach. For the $3D$ approach, we could trivially extend the method and no longer consider a minimum area but a minimum volume. However, if considering a minimum volume would be effective in the context of truly three-dimensional images, it is not suitable for video that are spatio-temporal and not purely spatial. Indeed, assuming a minimum volume, a QFZ having few pixels in a spatial area, but on many frames, would be kept despite the fact it is probably not an object but a part of an object. Thus, we use a threshold of minimum mean area, the mean area being computed as follows:

$$A_{mean} = \frac{\#\text{ QFZ pixels}}{\#\text{ frames where the QFZ is present}} \tag{5}$$

For the $2D + t$ approach, we may also use this definition, but we rather use the filtering after the $2D$ processing. So we apply the filtering area not at the end of the process, but rather after the first stage (i.e., prior to temporal processing). Doing so allows us to have fewer QFZ to be processed during the temporal step.

We could do the same for the $t + 2D$ approach by filtering QFZ produced by the temporal processing. But setting a threshold on the minimum number of frames in which a pixel must belong to the same QFZ would have been tricky and very probably without any sense. Therefore we filter the QFZ in the same manner as for the $3D$ approach, once the QFZ building is achieved.

Similarly to the image filtering, the oversegmentation is here strongly reduced. Indeed, setting the threshold of minimum mean area to 10 pixels, we get $980CC$ for $2D+t$ and $319CC$ for $t+2D$. In addition, we get few or no under-segmentation, which makes obtained QFZ relevant for segmentation.

Filtering by minimum area threshold is very effective in reducing QFZ over-segmentation in video sequences. We obtain a very substantial over-segmentation reduction while maintaining the QFZ quality. By combining the definitions of video QFZ and the filtering area, we obtain an effective method of video pre-segmentation. This pre-segmentation can be used by QFZ merging methods to obtain user-personnalized segmentation.

4 User-driven video segmentation

Encountered in image segmentation, over-segmentation is even more present when dealing with video segmentation. For instance, when processing the sample *carphone* (Fig. 2.a) with the Predictive Watershed [2], we obtained about 2000 regions. This problem happens obviously also with segmentation by quasi-flat zones: segmenting the same sequence by $(P_1, ..., P_n)$-CC $t+2D$, with $\alpha = \omega = 20$ and a minimal area of 10 also provides an over-segmentation ($319CC$). Moreover, the resulting segmentation is not personnalized. This drawback may be solved by relying on user interaction. Such an interaction aims the user to both customize the segmentation and reduce over-segmentation. User-driven segmentation is a well-known principle in Mathematical Morphology, and has been recently used by the watershed from propagated markers method [3]. In the context of QFZ, as indicated by Soille [6], it may correspond to the assembling of puzzle pieces.

We suggest a new principle for user-interactivity in video segmentation by defining a QFZ segmentation guided by markers. First, a base QFZ segmentation is produced. Then, the user draws markers on the video data. Thus, he customizes the segmentation by indicating his objects-of-interest. QFZ beneath the markers are considered as the seeds of a Seeded Region Growing algorithm [10]. The region growing will then merge the different QFZ according to their distance in terms of color, which can be related to an α parameter in the QFZ context. Since the user only see original video (and not QFZ), it is possible that several markers are found over the same QFZ. In this case, we consider that there are two possibilities: either the marker has been ill-drawed or the QFZ is ill-segmented. Here, we assume that the user has well-drawn the marker and that the QFZ has to be corrected. To do so, this QFZ is segmented using Seeded Region Growing with the user's markers as seeds. Thus, ill-segmented QFZ is corrected: it both improves the accuracy of the initial over-segmentation and solve the problem of having multiple markers over the same QFZ.

In order to evaluate the relevance of our proposal, we conducted some experiments on the *carphone* sequence. We compared the $(P_1, ..., P_n)$-CC $2D+t$ and $t+2D$ methods to the Marker-Based Watershed known as the standard interactive segmentation method of Mathematical Morphology. We also compared interactive QFZ to the Seeded Region Growing involved in our method, in order

Method	**Parameters**	**Precision**		
		(a)	**(b)**	**(c)**
$(P_1,...,P_n)$-$CC\ 2D+t$	$\alpha=\omega=20$	**0.94**	1.10	0.81
$(P_1,...,P_n)$-$CC\ 2D+t$	$\alpha=\omega=30$	1.00	0.77	**0.72**
$(P_1,...,P_n)$-$CC\ t+2D$	$\alpha=\omega=20$	2.92	0.92	1.45
$(P_1,...,P_n)$-$CC\ t+2D$	$\alpha=\omega=30$	6.25	**0.34**	0.84
Marker-Based Watershed		7.77	2.72	1.76
Seeded Region Growing		8.73	3.23	2.62

Table 1. Comparison of frontier precision with different markers (a) a few points on the median frame, b) heavy markers one the median frame, c) heavy markers on three frames)

to show how our method benefits from such algorithm and what it offers compared to a direct processing of Seeded Region Growing. In this perspective, we used three different settings of markers (Fig. 2). The results of these experiments are presented in Tab. 1. We denote here by precision the average spatial distance (in pixels) between the frontiers of the resulting segmentation and those of the reference segmentation. We used two sets of parameters (α, ω) for $(P_1,...,P_n)$-CC to show that our method is robust to parameter settings (which have besides not been optimized). Let us observe that, whatever the markers used, our method always provides better results than the other two approaches. However, like the other two interactive methods, the interactive QFZ segmentation is very sensitive to the markers given by the user.

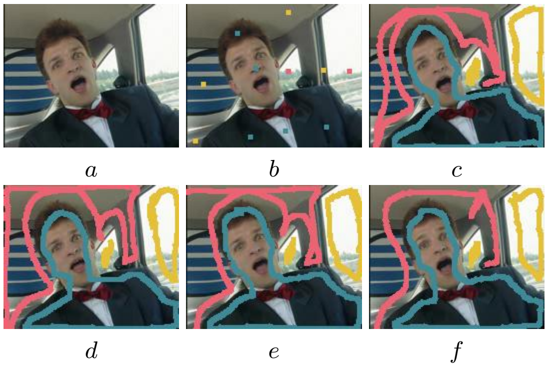

Fig. 2. Markers on sequence *carphone* (a) frame 7 of extract from *carphone* sequence, b) few points on frame 7, c) heavy markers on frame 7, def) heavy markers on frame 3,7 and 10)

We also compared our method to a more recent method, the Watershed from Propagated Markers [3], for which we used the binding of markers and the region-based motion propagation. As the objective here was to compare our method to a recent interactive approach in similar conditions (here the time required for the user), we did not allow the marker correction by the user and marked only the first frame. The results are given in Tab. 2 and show that marker-based $(P_1, ..., P_n)$-CC in these conditions is more accurate than the Watershed from Propagated Markers.

Method	**Parameters**	**Precision**
$(P_1, ..., P_n)$-CC $2D + t$	$\alpha = \omega = 20$	0.81
$(P_1, ..., P_n)$-CC $2D + t$	$\alpha = \omega = 30$	0.73
$(P_1, ..., P_n)$-CC $t + 2D$	$\alpha = \omega = 20$	0.93
$(P_1, ..., P_n)$-CC $t + 2D$	$\alpha = \omega = 30$	**0.71**
Watershed From Propagated Markers		2.02

Table 2. Comparison of frontier precision between $(P_1, ..., P_n)$-CC and watershed from propagated markers

5 Conclusion

In this paper, we proposed both an extension of QFZ to video sequences and an interactive method for assembling these QFZ in order to build a user-personnalized segmentation. The separate processing of spatial and temporal dimensions improves the segmentation compared to a straight three-dimensional processing of video data. The proposed method for assembling QFZ according to the user's needs is intuitive and provides good results compared to other existing methods.

Our future work will focus on improving the markers. Indeed, the video is currently only marked before the processing. However, it seems relevant to be able to correct markers (like what is done in [3]) in order to iteratively improve the segmentation. Indeed the eventual correction of some QFZ will improve the over-segmentation at each iteration. Moreover, most of the computational cost of our approach is linked to the initial QFZ segmentation: the marker-based QFZ merging is very efficient because it is performed on the QFZ adjacency graph, unlike other interactive segmentation methods that restart all the segmentation process when modifying markers (cf. Marker-based Watershed and Seeded Region Growing). We also consider to apply video QFZ on other data spaces, such as optical flow values instead of pixel values. Moreover, we plan to improve the QFZ merging process by using other features than only the mean color.

Finally, as our method is based on a graph reduction process, we would like to design a machine learning scheme to understand how to perform segmentation from this reduction process. The idea is here to perform first a learning of some videos marked by the user to then enable the system to segment unmarked, but simply over-segmented with QFZ, video sequences.

References

1. Agnus, V.: Segmentation spatio-temporelle de squences d'images par des oprateurs de morphologie mathmatique. PhD thesis, Universit Louis Pasteur, Strasbourg (2001)
2. Chien, S.Y., Huang, Y.W., Chen, L.G.: Predictive watershed: a fast watershed algorithm for video segmentation. IEEE Transactions on Circuits and Systems for Video Technology **13**(5) (May 2003) 453–461
3. Flores, F., Lotufo, R.: Watershed from propagated markers: An interactive method to morphological object segmentation in image sequences. Image and Vision Computing **28**(11) (2010) 1491–1514
4. Rivest, J.F., Beucher, S., Delhomme, J.: Marker-controlled segmentation: an application to electrical borehole imaging. Journal of Electronic Imaging **1**(2) (1992) 136–142
5. Serra, J., Salembier, P.: Connected operators and pyramids. In: Proceedings of SPIE, Non-Linear Algebra and Morphological Image Processing,. Volume 2030. (1993) 65–76
6. Soille, P.: Constrained connectivity for hierarchical image partitioning and simplification. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(7) (July 2008) 1132–1145
7. Soille, P.: On genuine connectivity relations based on logical predicates. In: Proceedings of the 14th International Conference on Image Analysis and Processing, Washington, DC, USA, IEEE Computer Society (2007) 487–492
8. Soille, P., Grazzini, J.: Constrained connectivity and transition regions. In: Proceedings of the 9th International Symposium on Mathematical Morphology and Its Application to Signal and Image Processing, Berlin, Heidelberg, Springer-Verlag (2009) 59–69
9. Nagao, M., Matsuyama, T., Ikeda, Y.: Region extraction and shape analysis in aerial photographs. Computer Graphics and Image Processing **10**(3) (1979) 195–223
10. Adams, R., Bischof, L.: Seeded region growing. IEEE Transanction on Pattern Analysis and Machine Intelligence **16**(6) (1994) 641–647
11. Angulo, J., Serra, J.: Color segmentation by ordered mergings. In: Proceedings of the IEEE International Conference on Image Processing. (2003) 125–128
12. Tous, F.Z.: Segmentation interactive d'images fixes et de squences vido base sur des hierarchies de partitions. PhD thesis, Ecole des Mines de Paris (2001)
13. Soille, P.: Constrained connectivity for the processing of very-high-resolution satellite images. International Journal of Remote Sensing **31**(22) (2010) 5879–5893