Construction of video mosaics using the Minimum Description Length

Maria-João Rendas
Lab. I3S, Sophia Antipolis, France
Email: rendas@i3s.unice.fr

***Abstract*— The paper presents work on the construction of video mosaics of the sea bottom, acquired by an autonomous underwater robot. The approach presented, based on the Minimum Description Length, while strongly grounded on probabilistic theory, is able to solve the image registration problem without need for user-defined probabilistic models (no need for neither world nor sensor models), presenting several advantages over traditional techniques. Mosaic construction is based on a dense locally estimated pure translational displacement field. The fact that the proposed metric for image co-registration does not require the presence of distinctive features (like corner-like structures), together with the fact that it is inherently robust to smooth illumination variations across the image plane, makes it particularly suitable for the underwater environment, where individual features present most often a strong resemblance, and where illumination variations are known to be frequent, due to the need to carry the light source onboard the platform. Even if a pure translational model is locally fit to each small image region, the method can globally cope with strong orientation variations of the platform, as the results present demonstrate.**
The paper details the formal presentation of the criterion used for local displacement estimation, and shows results of mosaic construction with real underwater images acquired under challenging conditions.

I. Introduction

Video mosaics of the sea floor acquired by autonomous vehicles, fusing distinct video frames taken by an onboard camera to produce a single image of the entire observed region, are useful representations in several distinct contexts. Their acquisition can be the ultimate goal of an underwater robotic mission, for end-users interested in *observing physical or biological oceanographic processes*, since they allow a global appreciation of the instantaneous characteristics of a given region; they can also be exploited to *study the dynamics of populations* of interest, by considering video mosaics of the same area taken at different times. Alternatively, video mosaics can be seen as *navigation aids*, being detailed maps of the explored workspace which the platform can exploit for navigation purposes.

In this paper we propose a new criterion for iterative mosaic update based on the Minimum Description Principle of Rissanen [1]. The criterion proposed here is generic (not restricted to a set of particular hypotheses concerning the operational region - like quasi-planarity - or to the motion of the robot - like naturally stabilized platforms), and can be applied to color or black & white images. With respect to previously proposed methods, it presents the following major differences:

1) Unlike most mosaicing techniques, which separate the (local) image registration from the (global) mosaic estimation steps (and often explicitely assess only the former), like [2], our approach iteratively and simultaneously solves the problem of global mosaic estimation and alignment of newly acquired data;
2) Unlike other methods (e.g. [3], [4]), it does not rely on detection (or even existence) of salient point-like features that can be unambiguously associated across video frames: the method processes globally the new image being fused into the mosaic;
3) It does not require the specification or identification of statistical distributions of the distinct regions of the images (like it was the case for the mosaicing technique that we proposed in [5]): use of the Minimum Description Length principle allows us to design an entirely data-driven algorithm;
4) We efficiently exploit the intrinsic discrete nature of video data, not having to resort to un-natural modelling approaches based on Gauss or exponential models, as most techniques do.

The Minimum Description Length (MDL) principle [1] is a criterion for solving model identification problems, where one amongst a set of candidate probabilistic families must be chosen as the preferred generative model associated to observed data. It bears relations to other approaches and tools used in general statistical estimation problems, such as Bayesian estimation techniques, Maximum Likelihood, or predictive coding, see e.g [6]. Originally motivated by model identification problems – where one of several competing models $\{\mathcal{H}_i\}_{i=1}^M$ of different complexities must be selected – as a way to define a penalty imposed to the models off larger complexity, it is now clear that its use is justified even if the compared models have the same number of degrees of freedom but possess different observability characteristics. In its most recent formulations [6], the MDL principle is stated as the search for the universal optimal code for the data in the set of candidate statistical models $\{\mathcal{H}_i\}_{i=1}^M$ which most efficiently compresses the data (and thus that most effectively captures the regularities in the data). Its most commonly used formulation considers a special sub-class of universal codes: two-part codes. These two part codes split the coding (actually the code length $L_i(X)$) of the data X in two parts: $L_i(X) = L_i^1(X) + L_i^2(X)$. The first part, $L_i^1(X)$, specifies

the element (distribution) $p^i \in \mathcal{H}_i\}$ of the probabilistic model, that is used to code the data, and the second part $L_i^2(X)$ uses the selected distribution p^i (the Maximum Likelihood estimate of the data distribution in the model $\mathcal{H}_i$) to provide a minimal length coding of the data in the model. The best model $\mathcal{H}_i$ is the one that leads to a smaller sum of these two terms.

We formulate the image registration problem as a *model estimation problem*, where the distinct models being compared correspond to distinct effectively observed sea floor areas (the integrated footprint of the video camera on the sea bottom). Note that at each new acquisition time the observed region is extended to the union - in a set theoretic sense - of the previously observed region and the new area observed by the frame that has been acquired).
As we show, use of the MDL criterion together with a non-parametric probabilistic modelling approach leads to an estimator based on the entropy of the empirical distribution of error between the observed images and the estimated mosaic. The derivation of the algorithm is free from requirement of user-defined parameters or thresholds, and requires only that the pixel noise distribution be centered at zero and symmetric. Note that the empirical error distribution is a global characteristic of the acquired images, and not only of carefully selected sub-regions in the data. Curiously, and even if this invariance has not been explicitly enforced, the estimator designed is invariant to global shifts in image intensity levels.

The paper presents in detail the derivation of the proposed criterion for displacement (or disparity) estimation (Section II) and the how local estimates using a simple translational model are used to obtain a global estimate of general deformation fields (Section III). Finally, it illustrates (Section IV) its performance, presenting several mosaics of real underwater image sequences acquired with an ROV[1] Phantom.

II. Estimation of displacement

In this section we assess the estimation of apparent image motion as the search for the best possible registration of two gray level images, i.e. of estimating the displacement of image regions as seen in consecutive frames of a video stream, $I(t_r)$ and $I(t_{r+1})$. We solve the problem by using the MDL principle. To simplify the derivation (and because it will be the model assumed locally to fully co-register the images), we consider that the two images can be registered by a simple translation, and that they have the same size $N_l \times N_c$. Consideration of more general registration models and of images of distinct sizes presents no conceptual difficulty. Since the temporal rate is not explicitly used by the algorithm, in this section we will use the simpler notation $I_1 = I(t_r)$ and $I_2 = I(t_{r+1})$. Let $\mathcal{IJ} = \{1, \ldots, N_c\} \times \{1, \ldots, N_c\}$ be the (discrete) image plane.

[1]Remotely Operated Vehicle

We assume that each image is a noisy version of an unknown "reference image" I_0, whose origin is made coincident with one of the observed images (I_1):

$$\begin{aligned} I_1(i,j) &= I_0(i,j) + w_1(i,j), \qquad (i,j) \in \mathcal{IJ} \\ I_2(i,j) &= I_0(i+\Delta_p, j+\Delta_q) + w_2(i,j), \qquad (i,j) \in \mathcal{IJ} \ . \end{aligned} \tag{1}$$

where the noises $w_1(i,j)$ and $w_2(i,j)$ are spatially white, and uncorrelated with each other.

The size of the "reference image" I_0 required to model the observed data depends on the *displacement vector* $\Delta = [\Delta_p \ \Delta_q]$, which is the parameter that we want to estimate.
Estimating Δ can thus be cast as a *model fitting problem*, where the dimension of the model (i.e., the number of independent parameters, in this case the size of the (unknown) "reference image" I_0) depends on the value of the parameter being estimated, and we can thus use the MDL (Minimum Description Length Principle) [1] to solve it.
According to MDL, the best estimate of Δ is the value corresponding to the model that leads to a *shorter code* for the data $I-1, I_2$:

$$\hat{\Delta}_{MDL} = \arg\min_{\Delta} L_{I_1,I_2}(\Delta), \tag{2}$$

where $L_{I_1,I_2}(\Delta)$ is the length of the most efficient code for I_1 and $I-2$ assuming model (1) and for a fixed value of Δ. We consider three-part codes, that encode the observed images in the following way:

1) first, the value of Δ is coded, using $L(\Delta)$ bits;
2) next, the estimated "reference image" $\hat{I}_0$ is coded, using the value of Δ in the observation model (1), requiring $L_{\hat{I}_0}$ bits. Note that the size of I_0 depends on the estimated amount of overlap between I_1 and I_2, and thus on the value of Δ: $L_{\hat{I}_0} = L_{\hat{I}_0}(\Delta)$;
3) finally, the regions of I_1 and I_2 that differ from the "reference image" I_0 (which is a function of Δ) are coded using an optimal code, with L_{I_1,I_2} bits. This optimal codelength is also estimated from the data, and depends on Δ: $L_{I_1,I_2} = L_{I_1,I_2}(\Delta)$

The total codelength for the two images is thus

$$L_{I_1,I_2}(\Delta) = L(\Delta) + L_{\hat{I}_0}(\Delta) + L_{I_1,I_2}(\Delta). \tag{3}$$

We now determine the different terms in the right-hand-side (*r.h.s.*) of eq. (3).

a) **Coding Δ:**
We code Δ with a fixed-length code (this is equivalent to having no priori preference for its value). We can thus drop the corresponding term from the estimation criterion (3).

b) **Coding the reference image $\hat{I}_0$:**
Assuming that Δ is known, and according to the observation model (1) we have either two independent observations of

pixel (i,j) of the reference image (for $(i,j) \in \mathcal{I}_p^\Delta \times \mathcal{I}_q^\Delta$):

$$\begin{aligned} I_1(i,j) &= I_0(i,j) + w_1(i,j), \quad , \\ I_2(i-\Delta_p, i-\Delta_q) &= I_0(i,j) + w_2(i-\Delta_p, j-\Delta_q), \end{aligned}$$

where we defined the *overlap intervals*

$$\mathcal{I}_p^\Delta = [\max(1, 1+\Delta_p), \min(N_l, N_l+\Delta_p)]\,,$$

(with an analogous definition for $\mathcal{I}_q^\Delta$), or just a single observation (for $(i,j) \in \mathcal{I}_1 = [1,n] \times [1,N_c] \setminus \mathcal{I}_{pq}^\Delta$),

$$I_1(i,j) = I_0(i,j) + w_1(i,j), \quad ,$$

(where we defined the *overlap region* $\mathcal{I}_{pq}^\Delta = \mathcal{I}_p^\Delta \times \mathcal{I}_q^\Delta$) and[2] for $(i,j) \in \mathcal{I}_2^\Delta = [1+\Delta_p, N_l+\Delta_p] \times [1+\Delta_q, N_c+\Delta_q] \setminus \mathcal{I}_{pq}^\Delta$

$$I_2(i-\Delta, j-\Delta) \quad = \quad I(i,j) + w_2(i-\Delta_p, j-\Delta_q) \ .$$

Assuming known the value of Δ, the Maximum Likelihood estimates of the reference image (that we denote here by $\hat{I}_\Delta$ for convenience of notation) for pixels $(i,j) \in \overline{\mathcal{I}_{ij}^\Delta}$ coincide with the observed images (I_1 or I_2). Assuming that the noise is uniformly distributed with a (discrete) symmetric distribution, $q(n) = q(-n)$, the estimates for pixels $(i,j) \in \mathcal{I}_{ij}^\Delta$ are the average of the two independent observations:

$$\hat{I}_\Delta(i,j) = \frac{1}{2}\left(I_1(i,j) + I_2(i-\Delta_i, j+\Delta_j)\right), (i,j) \in \mathcal{I}_{ij} \ .$$

Note that if the original image is coded in L bits (and thus with 2^L distinct values), the pixels $(i,j) \in \mathcal{I}_{ij}$ are defined in a set of larger cardinality (at most twice the alphabet of the original image), and can be coded with no more than $L+1$ bits. Knowledge of the values of Δ enables definition of a rule for scanning the reference image, and we need thus just to code the pixel intensities. The total number of bits required to code $\hat{I}_0$ is thus

$$\begin{aligned} L_{\hat{I}_0}(\Delta) &= 2\left(N_l \times N_c - n^\star(\Delta)\right) L + n^\star(\Delta)(L+1) \\ &= 2(N_l \times N_c)L - n^\star(\Delta)(L-1) \ , \end{aligned} \qquad (4)$$

where we defined $n^\star(\Delta) = \left|\mathcal{I}_{ij}^\Delta\right|$ to denote the size of the overlapping region. The first term in the *r.h.s.* of (4) is independent of Δ and will not contribute to our final estimation criterion.

c) **Coding the common part of images I_1 and I_2:**

To recover the original images, we must also code the difference of the original images with respect to the common part $\left\{I_0(i,j), (i,j) \in \mathcal{I}_{ij}^\Delta\right\}$. It is easily seen that given the reference image these two images are redundant, and that we need to code only one, for instance I_1. The best code is a Shannon code, with codelength

$$\begin{aligned} L_{I_1,I_2}(\Delta) &= L_{I_1;q}(\Delta) = -\log_2 p\left(I_1(i,j), (i,j) \in \mathcal{I}_{ij}^\Delta | \hat{I}_0\right) \\ &= -\sum_{(i,j)\in\mathcal{I}_{ij}^\Delta} \log_2 q\left(I_1(i,j) - \hat{I}_0(i,j)\right) \ , \end{aligned}$$

[2] $A \setminus B$ denotes set difference.

where $q(\cdot)$ is the distribution of the observation noise w in model (1). Using the expression for the estimate of the reference image, we obtain

$$\begin{aligned} L_{I_1;q}(\Delta) &= -\sum_{(i,j)\in\mathcal{I}_{ij}^\Delta} \log_2 q\left(\frac{I_1(i,j) - I_2(i,j)}{2}\right) \\ &= -\sum_{\epsilon_i \in \mathcal{E}} n(\epsilon_i) \log_2 q(\epsilon_i) \ , \end{aligned}$$

where $\mathcal{E}$ is the discrete set of possible values of the differences of pixels intensities, and $n(\epsilon_i)$ denotes the number of occurrences of value ϵ_i in the difference image.

It is easily checked that this codelength can be written as

$$L_{I_1;q}(\Delta) = n^\star(\Delta)\left(H(\hat{q}) - D(\hat{q}||q)\right) \ ,$$

where H is the Shannon entropy, D is the Kullback-Leibler distance (see [1]), and $\hat{q}$ is the empirical estimate of the distribution of the difference image pixels: $\hat{q}(\epsilon_i) = n(\epsilon_i)/\left(\sum_i n(\epsilon_i)\right)$. If we consider that the noise distribution q is unknown, including it in the model that is fitted to the observations, this codelength is minimized for $q = \hat{q}$, the empirical distribution. This implies that the codelength of the (overlapping) sub-image becomes simply

$$L_{I_1;\hat{q}}(\Delta) = n^\star(\Delta) H(\hat{q}) \ .$$

In this case, we must add to the total codelength $L_\Delta(I_1, I_2)$ in eq. (3) an additional term equal to the number of bits required to specify $\hat{q}$. The observed error sequence is a word of length $n^\star(\Delta)$, where each "letter" belongs to an alphabet of size $M \leq 2^{L+1}$. The number of possible types for these sequences is equal to

$$\begin{pmatrix} M + n^\star(\Delta) \\ n^\star(\Delta) \end{pmatrix} \ ,$$

implying that if M is known we can code the type with a fixed-length code of length

$$L_{\hat{q}}(\Delta) = \log_2 \begin{pmatrix} M + n^\star(\Delta) \\ n^\star(\Delta) \end{pmatrix} = \sum_{i=n^\star(\Delta)+1}^{M+n^\star(\Delta)} \log_2 i - \sum_{i=1}^{M} \log_2 i \ . \qquad (5)$$

Note that although not explicitly indicated, the value of M depends on the image registration, and thus the last term of the above expression must be kept in the determination of the optimization criterion. We should also include in the total codelength the cost of coding the value of M. However, this term will in general be much smaller than the other terms, and we neglect it in a first approximation.

Using eq. (5) in eq. (3) yields the total codelength required to code the two images using our three-part code:

$$\begin{aligned} L_\Delta(I_1, I_2) = \log_2(K) &+ 2(N_l \times N_c)L - n^\star(\Delta)(L-1) \\ &+ n^\star(\Delta) H(\hat{q}) \\ &+ \sum_{i=n^\star(\Delta)+1}^{M+n^\star(\Delta)} \log_2 i - \sum_{i=1}^{M} \log_2 i \ . \end{aligned}$$

We can finally present our estimation criterion as

$$\hat{\Delta} = \arg\min_{\Delta} n^{\star}(\Delta)\,[H(\hat{q}) - L + 1] + \sum_{i=n^{\star}(\Delta)+1}^{M+n^{\star}(\Delta)} \log_2 i - \sum_{i=1}^{M} \log_2 i \ .$$

We see that the criterion depends on Δ only through the *entropy* of the pixel-wise differences of the observed images and the *number of pixels that are put into correspondence*. It is thus robust with respect to *shifts* in the intensity value of the image, which do not affect neither the entropy $H(\hat{q})$ nor the value of M.

III. Deformation field

We discuss now how the displacement estimation criterion is used to built a global displacement field, and solve the image co-registration problem.

We allow for non-rigid deformations of the images when co-registering consecutive video frames, to be able to cope with situations where the robot is observing the sea bottom from a relatively short distance (for instance, as it may happen during a contour tracking), in which case the assumption of bottom planarity is not verified in many practical situations.

We start by estimating a dense displacement field over a discrete grid, by applying the algorithm presented in the previous section to a set of small windows inside the image plane. Let $\{p_n = (i_n, j_n)\}, n = 1, \ldots, N$ be the central pixels of the windows considered, and let (centered) set W denote the window used around each pixel, such that $W_n = W \oplus p_n$ is the set of pixels in the image plane covered by window W_n. (The exact configuration of this grid depends on the allowed computational complexity, and we will not discuss it here.) Denote by W_n^t the window n taken in image $I(t)$.

Using the algorithm of the previous section,we estimate a displacement field $\Delta_n, n = 1, \ldots, N$ by applying the estimation criterion (3) to windows $W_n^{t_r}$ and $W_n^{t_{r+1}}$. This displacement field that maps (co-registers) neighborhoods of image $I(t_{r+1})$ into their corresponding pixels in image $I(t_r)$. We interpolate/extrapolate vectors $\Delta_n, n = 1, \ldots$ to obtain a continuous vector field $\Delta(i, j)$ defined over the entire image plane. Examples of the results obtained are shown in the next Section.

Alternatively, we could consider a parametric model of the global disparity field (assuming for instance local planarity of the sea bed surface) and using it to compute the displacement field over the entire image. Our approach has the advantage of being robust to bad local estimates: those areas of the image that are not informative enough to result in good displacement estimates are generally those where there is no local texture, i.e., where the image is rather homogeneous. With our approach, wrong pixel correspondences will result in deformation of the image footprint, but not in degradation of the contents of the mosaic. On the contrary, point-like feature based methods are known to be very sensitive to wrong features associations, that induce large errors in the estimates of the parameters of the assumed deformation model, and thus on image co-registration.

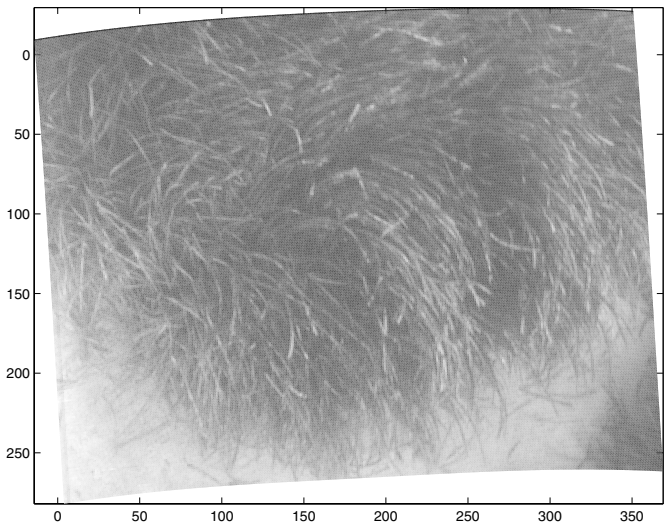

Fig. 3. Deformation of I_2 by the vector field estimated from values in the grid shown in Fig. 2.

Figure 1 displays the result of disparity estimation for two consecutive video frames acquired over a maerl bottom, with a bottom-looking camera installed in our Phantom platform during a sea-experiment in the Orkney Islands. The bottom is composed of mixtures of dead and alive maerl patches, creating a very textured visual pattern. For each element of the partition of the image frame W_n, we plot in each image a rectangle that encircles the *areas that are put in correspondence* by our local displacement estimation criterion. It can be seen that a consistent estimate of displacement is obtained across the entire image, confirming the hypothesis of nearly planar bottom and pure rotational motion of the robot. Note that one of the images is blurred by the platform's motion, and that the automatic gain adjustment of the camera imposed a variation of the gray levels between the two images. Our algorithm is robust with respect to both artifacts.

In Fig. 2 we show the result of applying the algorithm to pairs of images acquired with the Phantom (bottom-looking camera) during an experiment of contour tracking over a bank of Posidonia in Villefranche-sur-mer, South of France. In this experiment, the robot is highly maneuvering to track the crooked contour of the Posidonia bank. A small window size has been used, as it can be seen by the size of the rectangles overlaid in the plots. In spite of this, the algorithm is able to estimate the disparity field over the majority of the image, except when the windows have no significant local texture, or when the disparity field is equal (or larger) than the size of the windows used. The magenta lines show the vector field of displacement vectors. In Figure 3 we plot image I_2 deformed by the interpolated vector field. Comparison with image I_1 (see Fig. 4 shows a very good agreement of the two images.

In this section we considered the problem of estimating

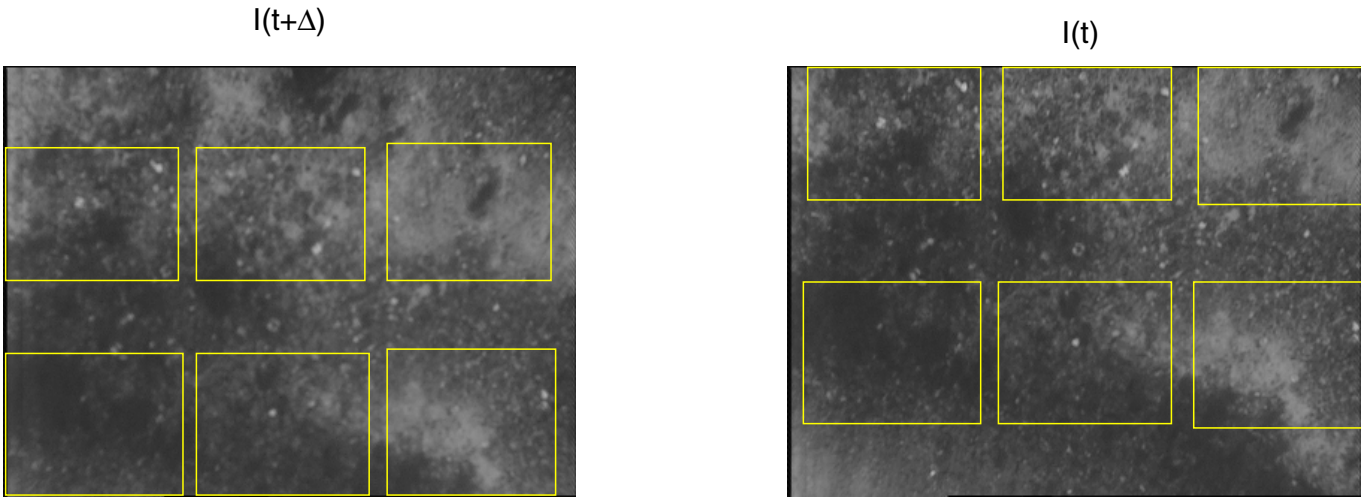

Fig. 1. Disparity estimation over a maerl field.

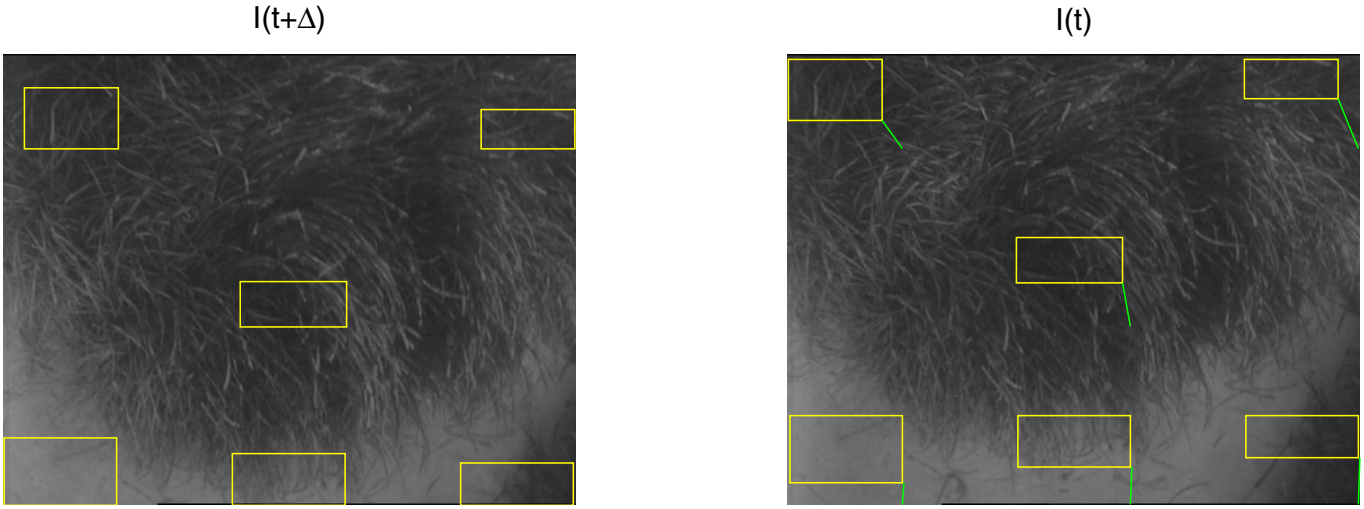

Fig. 2. Disparity estimation during a contour tracking (consecutive images, Posidonia bank).

the motion field between *consecutive frames*, over the entire images, based on its values at a *regular grid*. In the future, we will use local displacement estimation only from the regions that yield results to which a large confidence can be associated. This, together with a proper adjustment of the window size and image sampling scheme, adapted to the local characteristics of the image, are necessary requisites to be able to apply the approach proposed on-line, during actual observation of the underwater world.

IV. Results in real images

Figure 5 shows the result of co-registration of 5 consecutive images acquired during a real experiment with the ROV Phantom. In this experiment, the robot was (autonomously) tracking a contour between a Posidonia field (the strongly textured region in the image, where the leaves of Posidonia are visible) and a sandy bottom where a certain number of objects are present (dead Posidonia leaves, rocks, garbage,...). The crooked contour of the Posidonia bank induces in this experiment a strongly maneuvering behavior of the platform, with noticeable variation of the platform orientation (rotation) between consecutive frames. We can see that even if our approach locally fits a pure translational model, the overall rotational behavior is effectively recovered by working with small neighborhoods and allowing the displacement vector to vary across the image plane. The small altitude of the robot above sea bottom (compared to the height of the Posidonia leaves which was quite large at this time of year) imposes strong depth-induced variations on the displacement mapping, which is easy perceptible in the final mosaic obtained.

The previous example considered an highly textured region (the Posidonia leaves induce a high-frequency strongly contrasting pattern). Figure 6 illustrates the performance obtained by co-registring 10 frames of a sequence taken over the

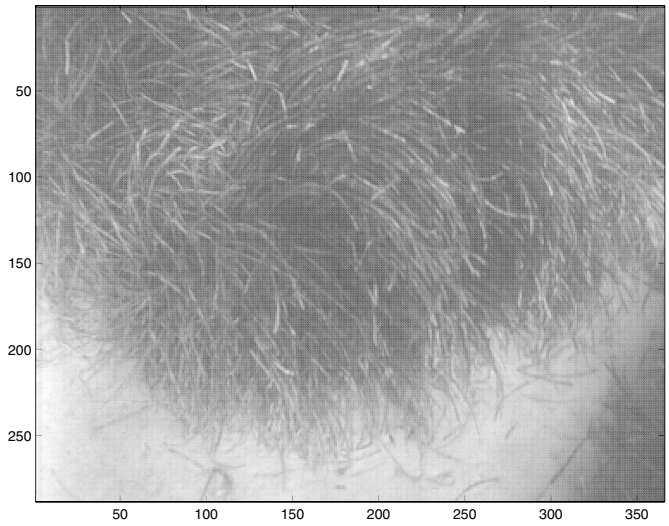

Fig. 4. Image I_1 in Fig. 3.

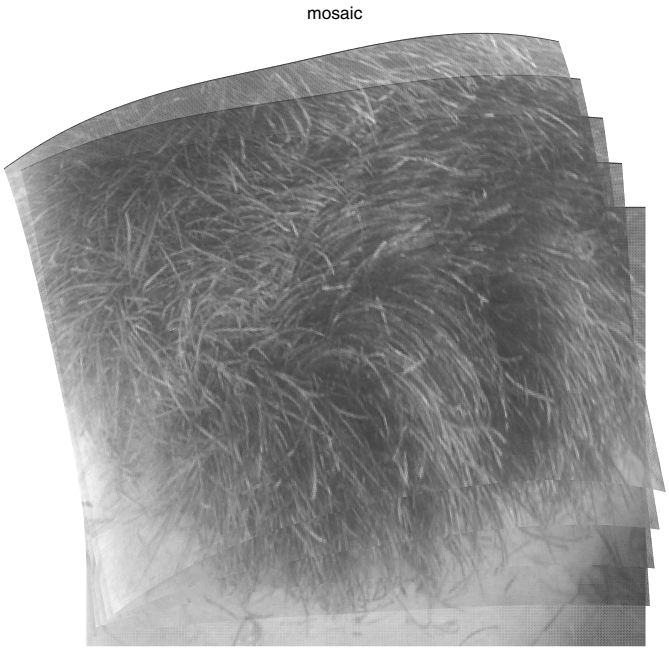

Fig. 5. Mosaic (Posidonia field).

boundary of a maerl field in Orkney. The perspective effects are very strong (the platform is moored, and balancing with the water motion), and the sandy pattern is very regular and of very low contrast. As this Figure shows, the estimated vector field, although working over reasonably large windows, was able to properly co-register the frames.

V. Conclusions

We presented a criterion for image registration that presents several advantages for the construction of underwater video mosaics. The criterion is obtained by formulating the image co-registration problem as a model selection problem, where

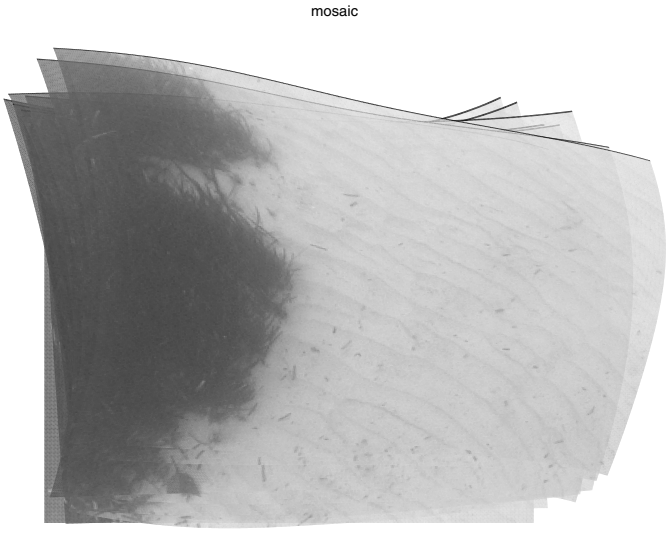

Fig. 6. Mosaic (border of a maerl patch).

the image overlap determines the complexity of the model. We used a three-part formulation of the message description length, which, by considering a completely non-parametric approach to the problem of statistical modelling, leads to a fitting metric dependent only (for pure translational models) on the number of image pixels that are put into correspondence and on the entropy of the pixel-wise differences of the co-registered sub-image. Results using real data in challenging conditions (varying depth and illumination conditions, strong rotational motions) demonstrate the appropriateness of our approach.

References

[1] Jorma Rissanen, Stochastic Complexity in Statistical Inquiry, World Scientific, Series in Computer Science-Vol. 15, 1989.

[2] Nuno Gracias, Sjoerd van der Zwaan, Alexandre Bernardino, Jos Santos-Victor, Mosaic Based Navigation for Autonomous Underwater Vehicles, J.Oceanic Eng., Vol 28, No. 4, October 2003.

[3] S. Negahdaripour and X. Xu, Mosaic-based positioning and improved motion-estimation methods for utomatic navigation of submersible vehicle. IEEE J Oceanic Eng. 27(1):79-99, 2002.

[4] C. T. Hsu, T. H. Cheng, R. A. Beuker, and J. K. Horng. Feature-based video mosaic. In Proc. Of ICIP 2000, pages 887-890, Vancouver, Canada, Sep. 2000.

[5] S. Rolfes and M.J. Rendas; Statistical Snakes: Application to tracking of Benthic Contours; 5th. IFAC Symposium on Intelligent Autonomous Vehicles; Lisbon, Portugal; July 2004.

[6] P. Grnwald, I.J. Myung, M. Pitt (editors). Advances in Minimum Description Length: Theory and Applications. 452 pages. MIT Press, April 2005.

[7] C. Barat, J. Rendas, "A robust visual attention system for detecting manufactured object in underwater video," Proc. IEEE Oceans 2006, Boston, September 2006.