6DOF Entropy Minimization SLAM

Juan Manuel Sáez and Francisco Escolano
Robot Vision Group, Departamento de Ciencia de la Computación e Inteligencia Artificial
Universidad de Alicante, Ap.99, E-03080, Alicante, Spain
Email: {jmsaez,sco}@dccia.ua.es

***Abstract*— In this paper, we propose and validate an entropy minimization algorithm for solving the SLAM problem in the 6DOF case with semi-sparse (stereo) data. The proposed SLAM solution relies on both an efficient and robust strategy for ego-motion estimation and an effective global rectification strategy. Our global rectification method is scalable because it relies on dynamically compressing actions, in order to reduce the number of variables to optimize, and thus on integrating/fusing observations. We have implemented a wearable stereo device that runs the SLAM algorithm in real time and we have tested such implementation both in indoor and outdoor scenarios. Our experiments show that action compression is a critical element for yielding acceptable and efficient solutions to the global optimization problem in the 6DOF case.**

I. Introduction

The Simultaneous Localization and Mapping (SLAM) problem consists of the estimation of both the relative positions of the environmental features with respect to the observer (map) and the position of the observer itself (trajectory in the map). Such problem has been recognized as a fundamental one in robotics because its solution is key for endowing robots with real autonomous capabilities. Thus, as robots must be able of estimating/learning the map of a given environment (either indoor or outdoor) as the robot traverses it, SLAM algorithms must grasp a collection of (usually noisy) observations taken by a sensor (laser, 3D sweeping laser, or stereo camera) and elicit a globally consistent (free of undesirable drift) interpretation of the environment as efficiently as possible in order to facilitate online learning.

When the used sensors yield dense enough data for building 3D maps of the environment (typically 3D sweeping laser range finders [1][2][3][4][5]) the problem of simultaneously computing the map and observer poses may be formulated in terms of maximizing a log-likelihood function and then using an EM-like algorithm to obtain, at least, a local maximum. In some of these algorithms, the core of SLAM is the registration of successive point clouds by means of an Iterative Closest Point (ICP) [6][7] algorithm, embodied in a computational scheme which enforces of a global-consistency criterion [8].

On the other hand, when the sensor provides a sparse set of features (2D laser range finders) Extended Kalman Filters (EKFs) relying on simple data-association mechanisms are used for solving the registration problem between features in consecutive views. The two fundamental weaknesses of this approach (a quadratic complexity with the number of features, and the collapse of the filter when data-association fails) are partially solved in the recent FastSLAM approach [9][10][11]. Although the sensor used in these cases are typically range (laser) sensors, the sparse approach has been recently brought to the computer vision arena [12][13]. However, in these latter algorithms, if one wants real-time solutions, the registration (structure from motion) problem must be circumvented because solving it accurately implies intensive batch processing (off-line). Consequently, only short sequences can be tracked without errors. However, there are recent experiments embodying structure-from-motion in the filter[14].

Computer vision solutions are desirable because cameras (specially monocular ones) are cheaper than 2D (and obviously 3D) lasers. However, in the middle of the SLAM spectrum we may find stereo-based solutions which compute semi-sparse 3D maps. In [15], where 3D information yields 2D maps, these maps are represented with a occupancy grid models. In other cases, like in [16], stereo data even allow the computation of 3D planes although some manual guidance is necessary when there is no data. In [17], stereo vision is fused with inertial information in order to recover 3D segments. On the other hand, in [18][19] 3D landmarks based on scale-invariant image features are used to compute the map. Such a computation relies on estimating the ego-motion of the robot, tracking the landmarks using the odometry for prediction, and finally superimposing the landmarks to obtain the map. However, such an approach is not globally consistent because it only relies on local estimations. This is why it is extended in [20], although the global consistency proposed exploits the closing-the-loop constraint for performing backwards corrections.

Besides considering the dense, sparse and semi-sparse approaches it is important to point out that most of the latter solutions revised above are constrained to 3 Degrees of Freedom (DOF), that is, robots are assumed to be confined in a plane. However, 6DOF solutions are demanded in many contexts like humanoid robots (QRIO) [21], walking robots with more than two legs [22], and underwater/aquatic robots [23]. To date there have been few attempts to migrate SLAM algorithms to the full 6DOF case. For instance, in [24] the FastSLAM/Rao-Blackwellized particle filter has been applied to 6DOF SLAM with stereo vision; the EKF approach has been chosen for monocular SLAM in the 3D space by tracking a reduced set of features [13]; and the relaxation of an ICP process has been followed in [25] in the context of a mine mapping application. The limited results obtained to date are due to the significant increase of complexity in 6DOF with respect to the 3DOF case. In this new context, data-association and matching/registration methods are more prone to failures

and this implies a later computational effort in recovering (if possible) a globally-consistent solution. This is why dense or semi-sparse solutions, using sensors from which one may implement stable data-association and matching methods, seem to offer a promising approach.

Following this semi-sparse approach we have proposed recently a stereo-based approach which solves the registration problem (egomotion) in real time and embodies this mechanism in a global-consistent (and typically off-line) estimation of the map [26] without the need of imposing the closing-the-loop constraint (although other assumptions like the assumption of being in a plane-parallel or Manhattan environment are applied). In our latest version, global consistency is enforced by minimizing an entropy-based criterion with a randomized algorithm. The observer, typically a small mobile robot, was confined to the XZ (horizontal) plane, that is, only 3 DOF were considered. However, in this paper we address the problem of migrating these mechanisms to a wearable stereo device. In order to do so in near real time, there are three key questions to solve: (i) Considering 6DOFs both in the egomotion and in the global-rectification algorithms; (ii) Enabling the scalability of the algorithm, in terms of both space and computation requirements, using a variable time resolution strategy for integrating (assimilating) observations; and (iii) Implementing the latter mechanisms into a real situated device. The effective extension of the egomotion algorithm, has been addressed in an early version of the wearable device in the context of helping visually impaired people [27] but SLAM results were only acceptable in indoor environments and also in controlled conditions. Here, in this paper, we face the general 6DOF case in depth.

The technical details of the approach are described in Section II (egomotion) and in Section III (map building, rectification, and variable reduction using an information-based measure). In Section IV we present some indoor and outdoor 6DOF SLAM experiments performed by a human wearing our wearable stereo prototype. Finally, in Section V we present our conclusions and future work.

II. Egomotion/Action Estimation

In 3D space, the pose of the camera at time t is given by 6 parameters (degrees of freedom, DOF) $\mathbf{p}_t = [x_t, y_t, z_t, \theta_t^X, \theta_t^Y, \theta_t^Z]^T$. The goal of egomotion is to estimate the incremental action $\mathbf{a}_t = \delta\mathbf{p}_t$ describing how the current pose has been derived from the previous one $\mathbf{p}_{t-1}$, but only using visual cues. The four steps for solving action estimation are (i) feature extraction, (ii) feature matching, (iii) matching refinement, and (iv) transformation estimation (see Fig. 1).

A. Feature Extraction

Let $\mathbf{C}_t(x, y, z)$ the 3D point cloud observed from the $t-$th pose, and let $\mathbf{I}_t(u, v)$ the right intensity image of the $t-$th stereo pair (reference image). For the sake of both efficiency and robustness, instead of considering all points $\mathbf{M}_t = [x_t, y_t, z_t]^T \in \mathbf{C}_t$ we retain only those points whose projections $\mathbf{m}_t = [u_t, v_t]^T \in \mathbf{I}_t$ are associated to strict local maxima of $|\nabla\mathbf{I}_t|$. These points define the constrained cloud $\tilde{\mathbf{C}}_t(x, y, z)$. The same holds for $\mathbf{C}_{t-1}(x, y, z)$ and $\mathbf{I}_{t-1}(u, v)$.

B. Feature Matching

In order to find matchings between points $\mathbf{M}_t = [x_t, y_t, z_t]$ and $\mathbf{M}_{t-1} = [x_{t-1}, y_{t-1}, z_{t-1}]$ we will measure the similarity between the local appearances in the neighborhoods of their respective projections $\mathbf{m}_t$ and $\mathbf{m}_{t-1}$. As we need certain degree of invariance to change of texture appearance, matching similarity $S(\mathbf{M}_t, \mathbf{M}_{t-1})$ relies on the Pearson correlation ρ (illumination invariance) between the log-polar transforms $\mathbf{LP}$ (local orientation invariance) of the windows $\mathbf{W}_{\mathbf{m}_t}$ and $\mathbf{W}_{\mathbf{m}_{t-1}}$ centered on both points, that is, we must maximize the score

$$S(\mathbf{M}_t, \mathbf{M}_{t-1}) = |\rho(\mathbf{Z}_t, \mathbf{Z}_{t-1})|, \tag{1}$$

being $\rho(\mathbf{Z}_t, \mathbf{Z}_{t-1}) \in [-1, 1]$ the correlation coefficient of the random variables associated to the grey intensities of the log-polar mappings:

$$\mathbf{Z}_t = \mathbf{LP}(\mathbf{W}_{\mathbf{m}_t}), \mathbf{Z}_{t-1} = \mathbf{LP}(\mathbf{W}_{\mathbf{m}_{t-1}}). \tag{2}$$

In order to ensure the quality of the match, we reject candidates with: (i) low strength $S(\mathbf{M}_t, \mathbf{M}_{t-1}) \leq S_{min}$; (ii) low distinctiveness, that is, exists also $\tilde{\mathbf{M}}_t$ satisfying $S(\mathbf{M}_t, \mathbf{M}_{t-1})/S(\mathbf{M}_t, \tilde{\mathbf{M}}_{t-1}) = R_{min} \approx 1$; and (iii) unidirectionality, that is, for $\mathbf{M}_t$, the match maximizing $S(\mathbf{M}_t, \mathbf{M}_{t-1})$ is $\tilde{\mathbf{M}}_{t-1}$, but for this latter one, the match maximizing $S(\mathbf{M}_{t-1}, \mathbf{M}_t)$ is $\tilde{\mathbf{M}}_t \neq \mathbf{M}_t$.

C. Matching Refinement

Despite considering the three later conditions, the matching process is prone to outliers. Thus, after computing the best matches for all points, we proceed to identify and remove potential outliers. Suppose that the $i-$th point $\mathbf{M}_t^i$ of $\tilde{\mathbf{C}}_t$ matches the $j-$th point $\mathbf{M}_{t-1}^j$ of $\tilde{\mathbf{C}}_{t-1}$, and similarly $\mathbf{M}_t^k$ matches $\mathbf{M}_{t-1}^l$. Let $d_{ik} = ||\mathbf{M}_t^i - \mathbf{M}_t^k||$ and $d_{jl} = ||\mathbf{M}_{t-1}^j - \mathbf{M}_{t-1}^l||$. Let also D_{ikjl} be the maximum between the ratios d_{ik}/d_{jl} and d_{jl}/d_{ik}. Then, in order to preserve structural coherence it is better to retain matches where D_{ikjl} is close to the unit and remove the others. More globally, in order to consider whether $\mathbf{M}_t^i$ matches $\mathbf{M}_{t-1}^j$ or not, we evaluate the quantity

$$D_{ij} = \frac{\sum_k \sum_l D_{ikjl}}{|\mathcal{M}|}, \tag{3}$$

where $\mathcal{M}$ is the current set of matches, that is, for testing whether a given match should be removed or not, we consider the averaged sum of its maxima.

Leaving-the-worst-out is an iterative process in which we remove the match in $\mathcal{M}$, and their associated points, with higher D_{ij} and then proceed to re-compute, in the next iteration, the maxima for the rest of matches. We stop the process when either such a deviation reaches σ_{min}, being σ_{min} sufficiently small, or a minimum number of matches $|\mathcal{M}|_{\min}$ is reached.

Comparing this criterion with the statistical filter proposed in [28], we consider structural differences within the views (independently of the relative position between them) instead of considering structural differences between the views.

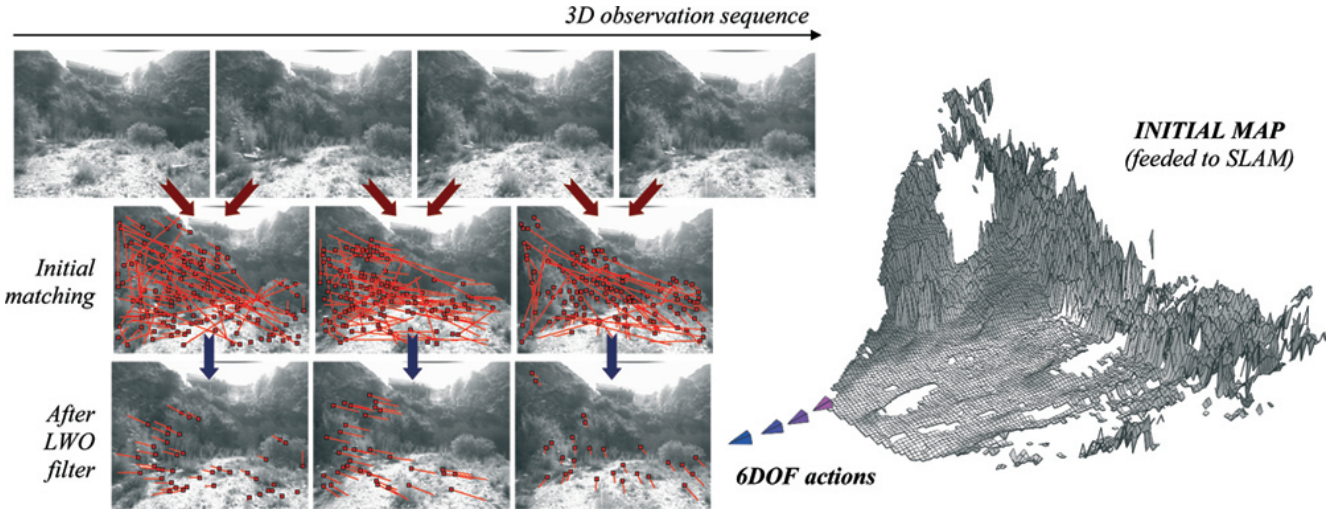

Fig. 1. Egomotion process. Left: Sequence of observations(top), initial matching each two consecutive observations (center), and refinement with LWO (bottom). Right: Initial map prior to rectification.

D. Transformation Estimation

The purpose of the leaving-the-worst-out process is to provide a set of good-quality matches in order to face action estimation directly. The idea is to perform both the refinement and action estimation once, that is, to avoid an interleaved EM-like estimation process. Then, let $\mathbf{R}_t$ and $\mathbf{t}_t$ the 3×3 rotation matrix and 3×1 translation vector, respectively, associated to action $\mathbf{a}_t$. Paying attention to the constrained 3D point clouds $\tilde{\mathbf{C}}_t(x, y, z)$ and $\tilde{\mathbf{C}}_{t-1}(x, y, z)$, and given that each point $\mathbf{M}_t^i$ in the first cloud matches point $\mathbf{M}_{t-1}^j$ in the second one, the optimal action is the one yielding the transformation (rotation and translation) that maximizes the degree of alignment between both clouds, that is, the one that minimizes the usual quadratic energy function

$$E(\mathbf{B}) = \sum_i \sum_j \mathbf{B}_{ij} ||\mathbf{M}_{t-1}^j - (\mathbf{R}_t \mathbf{M}_t^i + \mathbf{t}_t)||^2, \quad (4)$$

being $\mathbf{B}_{ij}$ binary matching variables (1 when $\mathbf{M}_t^i$ matches $\mathbf{M}_{t-1}^j$ and 0 otherwise). In order to minimize the latter function we perform a conjugate gradient descent with an adaptive step through the space of incremental actions.

III. Map Building and Rectification

Given a trajectory $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_{N-1}$ of size N, and a sequence of estimated actions $\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_{N-1}$ with length N, being $\mathbf{a}_0 = \mathbf{p}_0$, an initial approximation of the 3D map comes from superimposing all the point clouds with respect to the referential pose $\mathbf{p}_0$. We call the aggregation of all observations in a common reference system (initial map) as $\mathbf{A}$ when it relies on the complete point clouds. However, as this map accumulates the errors produced at local action estimations (errors due to the latter algorithm, or even to the absence of 3D cues when non-textured parts of the environment are observed), it is desirable to provide a consistency criterion and an updating strategy that exploits it in order to obtain a globally-consistent map.

A. Consistency Criterion

Our criterion for measuring the global consistency of the complete (not reduced) point cloud of the current map $\mathbf{A}$ is the minimization of the energy:

$$E(\mathbf{A}) = E_{global}(\mathbf{A}) + E_{align}(\mathbf{A}) \quad (5)$$

being

$$E_{global}(\mathbf{A}) = H(\mathbf{q}_{XYZ}) \quad (6)$$

$$E_{align}(\mathbf{A}) = H(\mathbf{q}_{XY}) + H(\mathbf{q}_{XZ}) + H(\mathbf{q}_{YZ}) \quad (7)$$

where $H(.)$ denotes the entropy of the argument. In the first term E_{global}, $\mathbf{q}_{XYZ}$ is the probability density of the 3D point cloud, that is, the distribution

$$\mathbf{q}_{XYZ}(x, y, z) = \frac{q(x, y, z)}{\sum_{x,y,z} q(x, y, z)} \quad (8)$$

associated to normalizing, after a proper discretization of the XYZ volume, and for each $x \in X$, $y \in Y$ and $z \in Z$, the sum $q(x, y, z)$ of all points $\mathbf{M}^i = [x^i, y^i, z^i]$ in the current map $\mathbf{A}$ satisfying $x^i \approx x$, $y^i \approx y$ and $z^i \approx z$.

The underlying rationale is that maximizing the correlation between observations is related to the minimization of the entropy of $\mathbf{q}_{XYZ}$ (uncertainty reduction). For instance, let $\mathbf{C}_{t-1}$ and $\mathbf{C}_t$ two consecutive complete point clouds before estimating their optimal relative pose (action) $\mathbf{a}_t$ which is given by the transformation $(\mathbf{R}_t, \mathbf{t}_t)$. As such transformation minimizes the error defined in Eq. 4, it also maximizes the correlation $\mathbf{C}_{t-1} * \mathbf{C}_t^{'}$, being $\mathbf{C}_t^{'} = \{\mathbf{R}_t \mathbf{M}_t^i + \mathbf{t}_t | \mathbf{M}_t^i \in \mathbf{C}_t\}$. This means that zones with similar structure in $\mathbf{C}_{t-1}$ and $\mathbf{C}_t^{'}$ will coincide in the partial map $\mathbf{A}_{t-1,t} = \mathbf{C}_{t-1} \cup \mathbf{C}_t^{'}$ and thus the $\mathbf{q}_{XYZ}$ distribution associated with that map will have many zones with high frequencies coming from the sum of coincident points. Thus, $H(\mathbf{q}_{XYZ}) \equiv H(\mathbf{A}_{t-1,t})$ will be the lower with respect to other configurations with few coincidences due to other transformations because few coincidences

yield a near uniform (maximal entropy) distribution. However as it is difficult to have zero error during egomotion (and this is why rectification is needed) sub-optimal alignments must be corrected (for instance by minimizing entropy).

On the other hand, the second term of Eq. 5, E_{align}, is only applicable to plane-parallel environments, that is, in environments where the main planes (walls, doors, floor, ceiling, and so on) are either parallel or orthogonal (Manhattan worlds). This allows, for instance, to correct a typical straight corridor that appears slightly curved if we only minimize E_{global}. In order to do so, we compute the 2D marginal distributions $\mathbf{q}_{XY}$, $\mathbf{q}_{XZ}$, $\mathbf{q}_{YZ}$ of $\mathbf{q}_{XYZ}$ with respect to each of the 3 main planes (see Fig. 2). Intuitively, if a corridor is perfectly orthogonal to the XZ (horizontal) plane, we will observe well-defined (crisp) projected walls and $H(\mathbf{q}_{XZ})$ will be minimal with respect to any other rotation. And an equivalent reasoning follows for the two other projections. Therefore, this second term, that we call alignment term, prefers maps yielding crisp projections and this is why curved corridors may be rectified. In order to accelerate the process, the first observation (reference system) should be approximately aligned with the main building directions.

The alignment term could be removed in a non plane-parallel environment ($E(\mathbf{A}) = E_{global}(\mathbf{A})$). In this case, the energy function has less information and is more difficult to find the best trajectory.

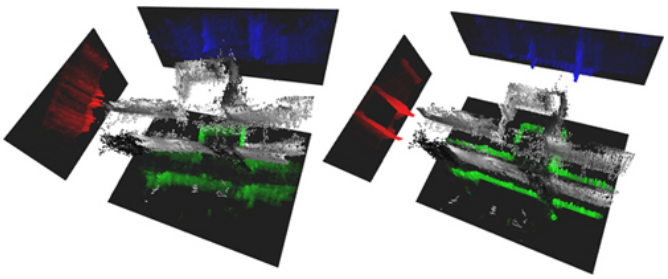

Fig. 2. Marginal distributions. Incorrect (left) and correct (right) alignment with the main planes.

Alternatively to use the grid and computing $\mathbf{q}_W$ as in Eq. 8, which is a biased but time-efficient solution (linear cost with the number of points), we could use a Parzen windows approach, like the one proposed in [29], to estimate the densities. However, the Parzen approach for density estimation has a quadratic cost although it produces unbiased solutions. Other entropy estimation approaches [31] were discarded for the same practical reason.

B. Information-driven Quasi-random Update

The underlying idea of our map-updating strategy is to modify all actions $\mathbf{a}_t$ simultaneously in order to obtain a new map $\mathbf{A}^{new}$. However, in practice, and in order to ensure the convergence of the optimization problem, we only modify simultaneously $K < N$ actions in each iteration, randomly selected (in our experiments, we use a 10% of N). The new trajectory is performed adding a random change to the selected actions. Each new action $\mathbf{a}_t^{new}$ is perturbed by randomly adding or subtracting ϵ_t. That is, $\mathbf{a}_t^{new}$ is a random variable following the uniform distribution $U(\mathbf{a}_t + \epsilon, \mathbf{a}_t - \epsilon)$ with $\epsilon = [\epsilon_x, \epsilon_y, \epsilon_z, \epsilon_{\theta^x}, \epsilon_{\theta^y}, \epsilon_{\theta^z}]^T$, whose scale must be carefully specified, attending to the scales of robot motion.

Considering all new actions $\mathbf{a}_0^{new}, \mathbf{a}_1^{new}, \ldots, \mathbf{a}_{N-1}^{new}$ simultaneously in the algorithm implies considering new poses $\mathbf{p}_0^{new}, \mathbf{p}_1^{new}, \ldots, \mathbf{p}_{N-1}^{new}$, and consequently a new map $\mathbf{A}^{new}$. Such a map is accepted if it reduces the energy, that is, if $E(\mathbf{A}^{new}) < E(\mathbf{A})$. Otherwise, none of the new actions are applied, we retain the current map $\mathbf{A}$, and a new iteration of the algorithm begins. Conceptually, this is closed to the Particle Filtering approach, but here we maintain a single path (instead of a population of paths), and also such path may change completely along time. Furthermore, for each new observation/action, we perform I_{max} (constant) iterations of this algorithm, following an *AnyTime* scheme. This strategy produces a suboptimal solution in a limited time.

Facing the full 6DOF case means that, as each new observation introduces 6 new variables intro the SLAM problem, the complexity of the optimization problem to solve grows so much that it is not possible to solve the global rectification in real time. Furthermore, the egomotion problem is harder and its associated error also grows. In the 3DOF case, practical considerations have motivated the latter randomized strategy instead of a more optimal one like the use Metropolis-Hastings samplers. We have also initially discarded a numerical gradient descent algorithm, because of the high number of energy evaluations for each gradient estimation. As we will show in the experimental section, the solution found is good enough for avoiding a significant global drift, which is specially critical in the 6DOF case.

Moreover, when working in 6DOF, we are also forced to introduce a mechanism for reducing the number of variables in order to solve SLAM in near real-time. In order to do that we test whether a given action $\mathbf{a}_t$ (6 variables) can be deleted or not from the trajectory. Deleting an action implies: (i) superimposing clouds $\mathbf{C}_{t-1}$ and $\mathbf{C}_t^{'}$, being $\mathbf{C}_t^{'}$ the transformed $\mathbf{C}_t$ according to $(\mathbf{R}_t, \mathbf{t}_t)$; and (ii) combining $\mathbf{a}_{t+1}$ with $\mathbf{a}_t$ into a new action $\mathbf{a}_{t+1}^{'}$. Then, for testing the delete condition we measure the mutual information [32] between the clouds:

$$I(\mathbf{C}_{t-1}; \mathbf{C}_t^{'}) = H(\mathbf{C}_{t-1}) + H(\mathbf{C}_t^{'}) - H(\mathbf{C}_{t-1}, \mathbf{C}_t^{'}), \quad (9)$$

which is the reduction of uncertainty of $\mathbf{C}_{t-1}$ due to the knowledge of $\mathbf{C}_t^{'}$. Approximating the joint entropy $H(\mathbf{C}_{t-1}, \mathbf{C}_t^{'})$ by $H(\mathbf{C}_{t-1} \cup \mathbf{C}_t^{'}) \equiv H(\mathbf{A}_{t-1,t})$ we obtain $\hat{I}(\mathbf{C}_{t-1}; \mathbf{C}_t^{'})$, and therefore we will remove action $\mathbf{a}_t$ when

$$\hat{I}(\mathbf{C}_{t-1}; \mathbf{C}_t^{'}) > \Theta, \quad (10)$$

being Θ a threshold. The latter approximation is dictated by efficiency because it relies on computing the entropy of one-dimensional distributions instead of computing joint entropies (two-dimensional). We have experimentally found a quasi-linear relation between both magnitudes for high-values of mutual information (for instance when considering temporally

adjacent views which usually share many information). However, at low values of mutual information what is conserved is the shape of both functions (see Fig. 3).

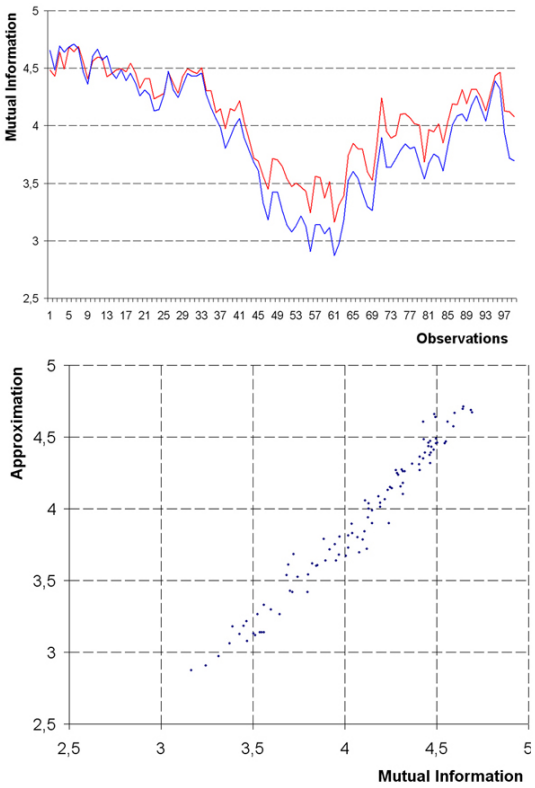

Fig. 3. Relating mutual information (MI) and the proposed approximation (PA). Top: MI (in red) and PA (in blue) between consecutive observations of a sequence of 100. Bottom: PA vs MI showing a quasi-linear dependence.

An alternative variable-reduction mechanism could rely on removing more past actions than recent ones (for instance by using a binary tree). However, the temporal approach yields maps which are more rigid at the beginning than at the end of the trajectory, and this may prevent closing a global loop. On the other hand, the mutual-information criterion provides a selective way of removing actions according to the amount of information (or statistical dependence) between the observations. The more information (dependence) the greater the rate of removing (compression) will be.

Finally, when the number of new observations since the last test is over N_{int} (currently $N_{int} = 50$) we perform the deletion test and proceed to remove the actions fulfilling it (satisfying Eq. 10) and thus N is reduced. Later tests may imply the removal of other actions not fulfilling the current test but which will satisfy it after a proper rectification between successive tests.

IV. Experiments and Discussion

A. The Wearable Device

In order to perform SLAM in 6DOF we have built a wearable stereo device consisting of a Digiclops stereo trinocular camera connected through a IEEE1394 firewire port to a small-sized laptop (Acer TM 382 Tci, with an Intel Centrino 1.6 Ghz processor) that we carry in a little knapsack (see Fig. 4). We obtain 320×240 stereo images producing clouds of 10.000 points on average which are typically reduced to 400 points in the constrained clouds. Given our previous experimental evaluations of the 3D estimation error for the Digiclops system, our maximum range is 12 meters, being the averaged error associated to such a distance below 0.91 meters.

Fig. 4. The wearable stereo device (top). Computer and Digiclops stereo camera (bottom).

B. Common Parameters

In this paper we present three SLAM experiments (one indoor and two outdoor). For all of them, the parameters for egomotion estimation are: $|\mathbf{W}| = 7 \times 7$ (the size of the appearance windows) in outdoor experiments and $|\mathbf{W}| = 5 \times 5$ in indoor experiments (sparser data requires smaller windows), $S_{min} = 0.8$ (minimal strength of a match), $R_{min} = 0.95$ (minimal distinctiveness ratio), $\sigma_{min} = 0.005$ and $|\mathcal{M}|_{\min} = 10$ (variance limit and minimal number of matches for stopping the leaving-the-worst-out process).

On the other hand, the parameters for map building and rectification are: $\epsilon_x = \epsilon_y = \epsilon_z = 0.05m$, $\epsilon_{\theta^x} = \epsilon_{\theta^y} = \epsilon_{\theta^z} = 2.86^o$. The fraction K/N of simultaneously selected actions in the quasi-random update process is 0.1 (10% of N). As we have indicated previously, we proceed to integrate views each $N_{int} = 50$ observations. A new observation is accepted when the estimated action is good enough, that is, when some rotation is above 15 degrees or some translation is higher than 0.5 meters. Egomotion is performed whenever the camera produces a new observation. However, the global rectification step is applied each new observation, with $I_{max} = 50$ iterations.

Fig. 5. Examples of images corresponding to observations in the first (left), second (center) and third (right) experiments.

TABLE I
EXPERIMENTS SUMMARY

Measures	Exp. #1	Exp. #2	Exp. #3
`egomotion_time`	372ms	326ms	297ms
`rectification_time`	1256ms	1518ms	1414ms
`information_thr`	3.5	4.0	4.0
`compression_time`	5060ms	4486ms	4655ms
`compression_rate`	83%	70%	73%

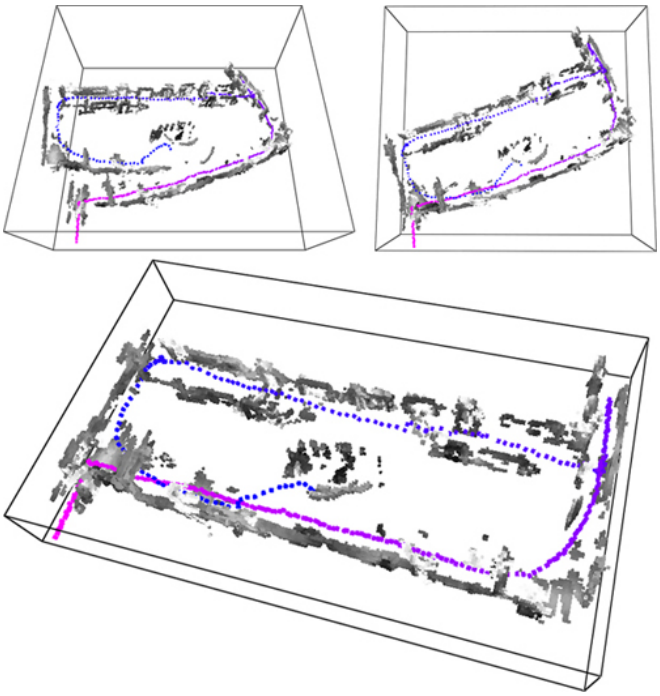

Fig. 6. Experiment #1: Indoor. Top-left: Egomotion (no rectification) results. Top-right: SLAM results without Manhattan assumption. Bottom: SLAM results exploiting the alignment terms.

C. Indoor Results

In order to test our SLAM algorithm in indoor sequences we have chosen an scenario where the egomotion algorithm works in critical conditions: a low-textured environment with moving obstacles (see a typical observation in the left part of Fig. 5). The wearable device closes a loop after a long trajectory (333 observations along 151.38m). The results are showed in Fig. 6 (in all experiments the trajectory is coded as a color gradient from blue –initial poses– to purple –final poses–). When using only egomotion there is a significant 3D drift which avoids further loop closing. However, when global rectification is applied but without exploiting the Manhattan assumption, the loop is closed but the main corridors appear curved instead of straight as they really are. Finally, when we use the complete energy function, the loop is closed and straight corridors are recovered.

In Table I, we present a quantitative summary of this and the other two experiments: `egomotion_time` is the averaged time of egomotion estimation (over number of observations) ; `rectification_time` is the averaged time for performing $I_{max} = 50$ rectification iterations (over the number of observations); `information_thr` is the Θ minimal value for compressing two consecutive views ; `compression_time` is the averaged time for performing the mutual information test each $N_{int} = 50$ new observations (over the number of compression steps) ; `compression_rate` is the percentage of compression $(N - N^{'})/N \times 100$ being N the number of original actions and $N^{'}$ the number of final actions at the end of the experiment. In the case of Experiment #1 we obtain a very high compression ratio due to a quasi-straight trajectory following the straight corridors.

D. Outdoor Results

Experiments #2 and #3 are devoted to analyze how our SLAM algorithm works in outdoor environments. Typical images of these experiments are showed in Fig. 5 (center and right respectively). In these cases the stereo sensor typically provides richer textures. Consequently, denser 3D data are provided and thus the performance of both the egomotion and rectification algorithms is improved with respect to the indoor case. However we cannot exploit the Manhattan assumption in these unstructured environments.

In Experiment #2 the wearable device follows a smooth ascending path and then it descends, following the same way, towards the starting point. We take 179 observations along 110.69m. The results are showed in Fig. 7. Although the first and the last poses should coincide if the trajectory is correctly estimated, using only egomotion the final overlap is missed and the Euclidean distance between the starting pose and the final one (pose error) is 4.489m. Using the SLAM algorithm without any compression the final overlap is recovered and the pose error is reduced to 2.891m. However, running the SLAM algorithm with compression the pose error is reduced, even more, to 2.094m. With respect to Experiment #1 we obtain similar computing times (see Table I). However, in order to obtain a compression rate (70%) comparable to the one obtained in the indoor case (83%) we have increased the mutual-information threshold from 3.5 to 4.0. Increasing such threshold is necessary in order to avoid excessive compression in outdoor scenarios, where consecutive views typically share more information (they are denser than in the indoor case). Excessive compression yields too rigid trajectory and thus suboptimal solutions.

Finally, Experiment #3 (see results in Fig. 8) is a more complex one in which we obtain 344 observations along

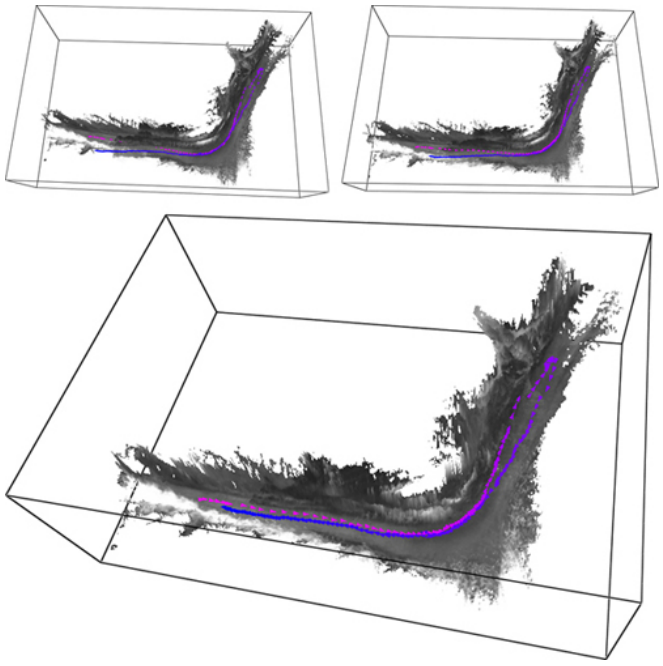

Fig. 7. Experiment #2: Outdoor, following a smooth path. Top-left: Egomotion (no rectification) results. Top-right: SLAM results without deleting any action (no-compression). Bottom: SLAM results exploiting compression.

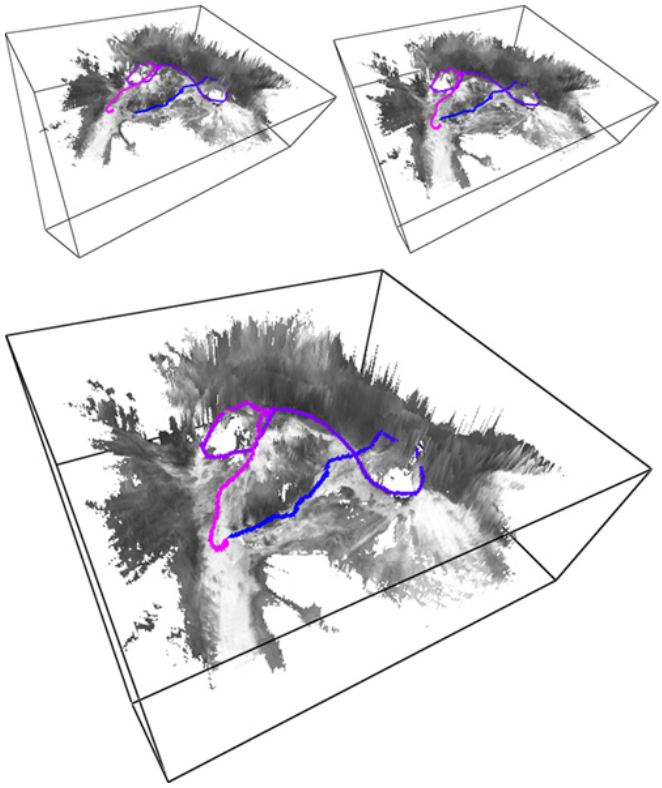

Fig. 8. Experiment #3: Outdoor, following an up-and-down path in a dry river bed. Top-left: Egomotion (no rectification) results. Top-right: SLAM results without deleting any action (no-compression). Bottom: SLAM results exploiting compression.

104.32m. In this case, we walk through the bed of a dry river following a winding trajectory, imposed by a uneven terrain. There are two small sub-loops and a global loop (because the first and last poses should coincide). As in Experiment #2, we have considered three cases: (i) only egomotion, (ii) SLAM without compression, (iii) full SLAM with compression. In the first case, we obtain a pose error (5.609m) higher than the one in Experiment #2, because of the higher number of observations. When using SLAM without compression the latter error is significantly reduced (2.664m) given the high amount of overlapped data due to the loops. Finally, when using compression we obtain a negligible error of 0.873m. As in Experiment #2 the mutual-information threshold is set to 4.0 and the compression rate is similar (73%). The paradox of having less error with more compression is only apparent: the role of compression is to reduce the complexity of the problem, making the minima of the energy function more achievable.

As we have seen in the results for experiments #2 and #3, compression plays a fundamental role in providing acceptable solutions for the SLAM problem. As the complexity of the problem increases each new observation, compression is a good way of reducing variables. Furthermore the $I_{max} = 50$ iterations can be performed faster than in the non-compression case, as a consequence of fusing observations. For instance, in Fig. 9 we show that the temporal gain of compression grows with the number of observations.

Apparently, from Table I one may think that the algorithm takes more than $6.5s$ to process an observation. However, compression time should be taken into account only every N_{int} observations, and from Fig. 9 we have that the processing

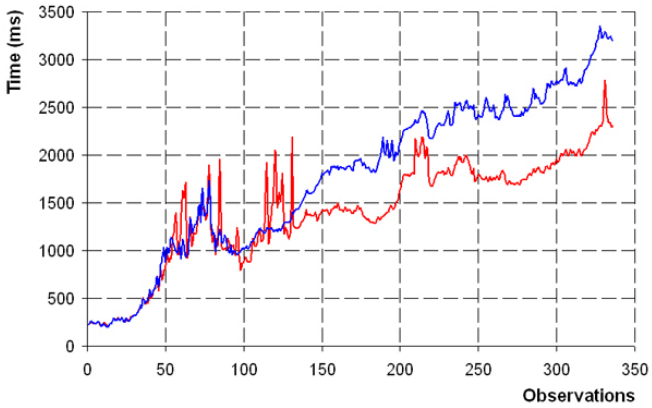

Fig. 9. Experiment #3: Global rectification times with (red) and without (blue) compression.

time per observation is below $2.5s$ when using compression. Moreover, the travelling speed of the subject is about $0.5m/s$.

V. Conclusions and Future Work

In this paper we have embodied a SLAM solution into a wearable stereo device for serving on-line metric (map) and positional (localization). This research is key for endowing many robots like humanoids, aquatic robots and multi-legged ones the capability of performing SLAM in their natural 6DOF contexts. In this regard, our proposal introduces three basic

elements: (i) a real-time egomotion estimation integrating 3D and 2D (appearance) information; (ii) a randomized algorithm for global rectification by entropy minimization (information maximization); and (iii) an information-based action compression strategy for reducing the number of variables and achieving real-time performance in global rectification.

The key element in our approach is the mechanism for reducing variables. However, in the current version action compression is lossy and actions cannot be de-compressed when needed. Consequently, suboptimal solutions may arise. Furthermore, the threshold for compressing actions is fixed beforehand. Our future work will address the two later questions: (i) develop a flexible optimization method for enabling action compression and de-compression (which should include a test for avoiding blind de-compressions) and (ii) design an strategy for learning or even updating the mutual-information threshold online.

Moreover, we should perform a detailed study about the convenience of incorporating a gradient descent for global rectification and also, we will explore the connections between our randomized method and Particle Filters.

ACKNOWLEDGMENT

This research is partially funded by the project TIC2002-02792 of the Spanish Government.

REFERENCES

[1] S. Thrun, W. Bungard, and D. Fox, "A real-time algorithm for mobile robot mapping with applications to multi-robot and 3D mapping," in *Proceedings of IEEE International Conference on Robotics and Automation*, San Francisco, April 2000.

[2] Y. Liu, R. Emery, D. Chakrabarti, W. Bungard, and S. Thrun, "Using EM to learn 3D models with mobile robots," in *In Proceedings of the International Conference on Machine Learning*, Williams College, Massachusetts, June 2001.

[3] C. Martin and S. Thrun, "Real-time acquisition of compact volumetric maps with mobile robots," in *Proceedings IEEE International Conference on Robotics and Automation*, Washington D.C., May 2002.

[4] D. Hähnel, W. Burgard, and S. Thrun, "Learning compact 3D models of indoor and outdoor environments with a mobile robot," *Robotics and Autonomous Systems*, no. 44, pp. 15–27, 2003.

[5] H. Surmann, N. A., and J. Hertzberg, "An autonomous mobile robot with a 3D laser range finder for 3D exploration and digitalization of indoor environments," *Robotics and Autonomous Systems*, no. 45, pp. 181–198, 2003.

[6] P. Besl and N. McKay, "A method for registration of 3D shapes," *IEEE Transactions on Pattern Analysis and Machine Intellitence*, vol. 2, no. 14, pp. 239–256, 1992.

[7] S. Rusinkiewicz and M. Levoy, "Efficient variants of the ICP algorithm," in *International Conference on 3D Digital Imaging and Modeling*, Quebec City, Canada, June 2001.

[8] D. Huber and M. Hebert, "Fully automatic registration of multiple 3D data sets," *Image and Vision Computing*, vol. 21, no. 7, pp. 637–650, 2003.

[9] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "FastSLAM: A factored solution to the simultaneous localization and mapping problem," in *Proceedings of the AAAI National Conference on Artificial Intelligence*, Edmonton, Canada, July 2002.

[10] D. Hähnel, D. Fox, W. Burgard, and S. Thrun, "A highly efficient FastSLAM algorithm for generating cyclic maps of large-scale environments from raw laser range measurements," in *Proceedings the IEEE International Conference on Intelligent Robots and Systems*, Las Vegas Hotel, Nevada, October 2003.

[11] S. Thrun, M. Montemerlo, D. Koller, B. Wegbreit, J. Nieto, and E. Nebot, "FastSLAM: An efficient solution to the simultaneous localization and mapping problem with unknown data association," in *To appear*, 2004.

[12] A. Davison and D. Murray, "Simultaneous localization and mapping using active vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 7, no. 24, 2002.

[13] A. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *Proceedings of the IEEE International Conference on Computer Vision*, Nice, October 2003.

[14] J. Meltzer, R. Gupta, Y. M.H., and S. Soatto, "Simultaneous localization and mapping using multiple view feature descriptors," in *Proceedings of International Conference on Intelligent Robots and Systems*, Sendai, Japan, September 2004.

[15] D. Murray and J. Little, "Using real-time stereo vision for mobile robot navigation," *Autonomous Robots*, vol. 8, no. 2, pp. 161–171, 2000.

[16] M. Iocchi and K. Konolige, "Visually realistic mapping of a planar environment with stereo," in *Proceedings of International Conference on Experimental Robotics*, Honolulu, Hawaii, December 2000.

[17] J. Lobo, C. Queiroz, and J. Dias, "World feature detection using stereo vision and inertial sensors," *Robotics and Autonomous Systems*, no. 44, pp. 69–81, 2003.

[18] S. Se, D. Lowe, and J. Little, "Vision-based mobile robot localization and mapping using scale-invariant features," in *In Proceedings of the IEEE International Conference on Robotics and Automation*, Seoul, Korea, May 2001.

[19] ——, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *International Journal of Robotics Research*, vol. 21, no. 8, pp. 735–758, 2002.

[20] ——, "Vision-based mapping with backward correction," in *Proceedings of the International Conference on Intelligent Robots and Systems*, Switzerland, October 2002.

[21] K. Okada, T. Ogura, A. Haneda, and M. Inaba, "Autonomous 3d walking systenm for an humanoid robot based on visual step recognition and 3D foot step planner," in *Proceedings of IEEE International Conference on Robotics and Automation*, Barcelona, April 2005.

[22] B. Grassmann, F. Zacharias, J. Zöllner, and R. Dillmann, "Localization of walking robots," in *Proceedings of IEEE International Conference on Robotics and Automation*, Barcelona, April 2005.

[23] G. Dudek, M. Jenkin, C. Prahacs, A. Hogue, J. Sattar, P. Giguere, A. German, H. Liu, S. Saunderson, A. Ripsman, L.-A. Torres, E. Milios, Z. P., and R. I., "A visually guided swimming robot," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Edmonton, August 2005.

[24] R. Sim, P. Elinas, M. Griffin, and J. Little, "Vision-based slam using the rao-blackwellised particle filter," in *IJCAI Workshop on Reasoning with Uncertainty in Robotics (RUR)*, Edimburgh, July 2005.

[25] A. Nüchter, H. Surmann, K. Lingermann, J. Hertzbert, and T. S., "6D SLAM with an application in autonomous mine mapping," in *Proceedings of IEEE International Conference on Robotics and Automation*, New Orleans, April 2004.

[26] J. M. Sáez and F. Escolano, "Entropy minimization SLAM using stereo vision," in *Proceedings of the International Conference on Robotics And Automation*, Barcelona, Spain, April 2005.

[27] J. M. Sáez, A. Peñalver, and F. Escolano, "First steps towards stereo-based 6DOF SLAM for the visually impaired," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, San Diego, June 2005.

[28] Z. Zhang, "Iterative point matching for registration of free-form curves and surfaces," *International Journal of Computer Vision*, vol. 13, no. 2, pp. 119–152, 1994.

[29] D. Erdogmus, K. Hild II, M. Lazaro, I. Santamaria, and J. Principe, "Adaptive blind deconvolution of linear channels using Renyi's entropy with Parzen estimation," *IEEE Transactions on Signal Processing*, vol. 52, no. 6, pp. 1489–1298, 2004.

[30] J. Beirlant, J. Dudewicz, L. Gyorfy, and E. Van Der Meulen, "Non-parametric entropy estimation: an overview," *International Journal of Mathematical and Statistical Sciences*, vol. 6, no. 1, pp. 17–39, 1997.

[31] P. Viola and W. Wells, "Alignment by maximization of mutual information," in *Proceedings of the IEEE International Conference on Computer Vision*, Cambridge, Massachusetts, June 1995.