GRAPH MATCHING USING SIFT DESCRIPTORS
An Application to Pose Recovery of a Mobile Robot

Gerard Sanromà[a], René Alquézar[b] and Francesc Serratosa[a]
[a]*Departament d'Enginyeria Informàtica i Matemàtiques, URV*
Av. Paisos Catalans 26, Campus Sescelades, 43007 Tarragona, Spain
[b]*Institut de Robòtica i Informàtica Industrial, CSIC-UPC*
Llorens Artigas 4-6, 08028 Barcelona, Spain

Keywords: Graph matching, SIFT, Pose recovery.

Abstract: Image-feature matching based on Local Invariant Feature Extraction (LIFE) methods has proven to be successful, and SIFT is one of the most effective. SIFT matching uses only local texture information to compute the correspondences. A number of approaches have been presented aimed at enhancing the image-features matches computed using only local information such as SIFT. What most of these approaches have in common is that they use a higher level information such as spatial arrangement of the feature points to reject a subset of outliers. The main limitation of the outlier rejectors is that they are not able to enhance the configuration of matches by adding new useful ones. In the present work we propose a graph matching algorithm aimed not only at rejecting erroneous matches but also at selecting additional useful ones. We use both the graph structure to encode the geometrical information and the SIFT descriptors in the node's attributes to provide local texture information. This algorithm is an ensemble of successful ideas previously reported by other researchers. We demonstrate the effectiveness of our algorithm in a pose recovery application.

1 INTRODUCTION

Visual odometry is used in mobile robotics to measure the spatial displacement experienced by a robot given the images taken at each location. In many SLAM systems it is a key point to estimate the robot trajectory in open-loop (V. Ila and Andrade-Cetto, 2009), (V. Ila and Sanfeliu, 2007), (Ila et al., 2010).

A data association between the images is needed in order to estimate the spatial displacement of the robot. Image feature matching based on local invariant features extraction (LIFE) has proven to be successful, and SIFT (Lowe, 2004) is one of the most effective. In SIFT, each feature is represented by its location and orientation on the image and a descriptor vector retaining information relative to the local texture. Such descriptors are invariant at a certain extent to changes in scale, rotation and illumination, making them suitable for matching images from the same scene under varying pose and environmental conditions. Features are then associated according to the closeness between their descriptors.

There exist a number of approaches aimed at enhancing the data association computed using local descriptors. Some examples are ICP (Besl and McKay, 1992), RANSAC (Brown and Lowe, 2003) and Graph Transformation Matching (Aguilar et al., 2009). What all these approaches have in common is that they use the geometrical information to reject a subset of erroneous matches (outliers).

Graphs are general-purpose structures aimed at representation where features are represented by nodes and the relations between them by edges. More in the topic of the present paper, Aguilar et al. (Aguilar et al., 2009) have recently presented an approach to use graph-based representations to the same end. To give some details, they build two K-nearest-neighbour graphs with the keypoints of the two images that have been matched (i.e., edges are placed joining a keypoint with the K nearest neighbours in space). The non-matched keypoints are discarded so, they begin with two isomorphic graphs. At each iteration, the algorithm removes the pair of matched nodes most structurally dissimilar and re-computes the K-nn structure (in both graphs). The process ends when two topologically identical graphs are obtained. Graph Transformation Matching has been recently used for recovering the pose of a mobile robot from a set

Sanromà G., Alquézar R. and Serratosa F. (2010).
GRAPH MATCHING USING SIFT DESCRIPTORS - An Application to Pose Recovery of a Mobile Robot.
In *Proceedings of the International Conference on Computer Vision Theory and Applications*, pages 249-254
DOI: 10.5220/0002829702490254

of known 2D views using epipolar geometry (Frank-Bolton et al., 2008).

The main limitation of the outlier rejectors is that they are unable to produce additional useful matches different from the initial set.

In this paper we present a novel iterative graph matching algorithm aimed at evolving an initial set of correspondences computed with the locally based SIFT method, to a kind of compromise between the constraints imposed by both the SIFT descriptors and the structural relations. Unlike the approaches described above, our method is able of including additional useful matches (those that satisfy the new combined constraints). This is the first graph matching algorithm that we have knowledge using structural relations along with SIFT descriptors.

We demonstrate the effectiveness of our method in a pose recovery application. Our method gets less error with more matches.

The organization of this paper is as follows: In section 2 we introduce some preliminary concepts. In section 3 our graph matching algorithm is presented. In section 4 we describe the experiments discuss the results, and in section 5 the conclusions are given.

2 PRELIMINARIES

Definition 1. *A* **Graph** *G (attributed relational graph) is a 3-tuple $G = (V,E,Z)$ where V is a set of vertices (also called nodes), $E \subseteq V \times V$ is a set of edges, where $e \in E$, $e = (v_i, v_j)$ is an edge joining nodes $v_i, v_j \in V$, and Z is a set of vectors, where $z_i \in Z$ is a vector of attributes associated to node $v_i \in V$.*

Although edges may contain attributes, we focus on the binary case where edges exist (1) or not (0).

Definition 2. *A* **matching matrix** *S is a binary matrix defining an injective mapping from a data-graph $G_D = (V_D, E_D, Z_D)$ to a model-graph $G_M = (V_M, E_M, Z_M)$. Hence, an element $s_{ij} \in S$ is set to 1 if node $v_i \in V_D$ is matched to node $v_j \in V_M$, and 0 otherwise. On the other hand, matching node $v_a \in G_D$ to node* NULL *(no node) means to put the a-th row of S to zeros.*

A number of *SIFT keys* are extracted from an image through the local invariant feature extraction method described in (Lowe, 2004).

Definition 3. *Each* **SIFT key** *$P_i = \left(X_i^T, R_i^T, U_i^T\right)^T$ is composed by its 2D location in the image $X_i = (x_i, y_i)$, its gradient magnitude and orientation $R_i = (r_i, \alpha_i)$ and a descriptor vector of length 128, $U_i = (u_{i,1}, \ldots, u_{i,128})$ with information the local texture on the image.*

Let $P_i^s, i = 1,\ldots,n$ and $P_j^d, j = 1,\ldots,m$ be the *SIFT keys* of a source and destination images, respectively.

Definition 4. *A* SIFT key P_k^s *from the source image is* **positively SIFT matched** *to a* SIFT key P_l^d *from the destination image if* $\text{dist}\left(U_k^s, U_l^d\right) = \min\left(\text{dist}\left(U_k^s, U_j^d\right)\right),\ j = 1,\ldots,m$ *and* $\frac{\text{dist}\left(U_k^s, U_l^d\right)}{\text{dist}\left(U_k^s, U_{l2}^d\right)} < \rho$, *where* $\text{dist}(\bullet)$ *is the Euclidean distance, U_{l2}^d is the descriptor of the destination image with the second smallest distance from U_k^s, and $0 < \rho \leq 1$ is a ratio value controlling the tolerance to false positives.*

Let $P_i^{left}, i = 1,\ldots,n$ and $P_j^{right}, j = 1,\ldots,m$ be the *SIFT keys* of a left and right images of a scene, obtained with stereo-vision. Let f be a function such that, $f(i) = j$ means that P_i^{left} is positively SIFT matched to P_j^{right}, and $f(i') = 0$ means that $P_{i'}^{left}$ is not matched to any. Let $X_l^{3D} = (x_l, y_l, z_l)^T$, $l = 1,\ldots,t$, be the 3D coordinate-vectors obtained through stereo triangulation of the local 2D coordinate-vectors X_k^{left} and $X_{f(k)}^{right}, \forall k$ s.t. $f(k) \neq 0$.

Definition 5. *A* **K-nearest-neighbour** *(K-nn)* **SIFT graph** *G is a 3-tuple $G = (V,E,Z)$, where V is a set of nodes associated to each X_l^{3D}, $Z = \left\{U_k^{left} \mid f(k) \neq 0\right\}$, is a set of SIFT descriptor-vectors associated to the nodes, and E is a set of edges where for each v_i there is an edge (v_i, v_{ip}) that join v_i with its K closest neighbours v_{ip}, $p = 1\ldots K$ in the space of the coordinate-vectors X_l^{3D}.*

3 A NOVEL GRAPH MATCHING ALGORITHM

Let $G_D = (V_D, E_D, Z_D)$ and $G_M = (V_M, E_M, Z_M)$ be a data and a model graph, respectively. In this section, we present an iterative graph matching algorithm, aimed at driving an initial estimate of the best matching matrix $S^{(1)}$ through the space of matching configurations, in the direction fixed by a new set of constraints aimed at representing a compromise between SIFT attributes and structural relations. This algorithm is built from an ensemble of previously reported ideas by other researchers (Gold and Rangarajan, 1996) (Luo and Hancock, 2001) (Cross and Hancock, 1998).

3.1 A Measure of Structural Consistency

It is a well-known strategy to state that a match from a node $v_a \in V_D$ to a node $v_\alpha \in V_M$ is more likely to occur as more nodes adjacent to v_a are assigned to nodes adjacent to v_α (Luo and Hancock, 2001) (Gold and Rangarajan, 1996).

We define a *hit* as a node $v_b \in V_D$ adjacent to v_a that is matched to a node $v_\beta \in V_M$ adjacent to v_α.

(Luo and Hancock, 2001) used the EM algorithm to iteratively find the maximum likelihood estimate of the matching matrix S. They used a probability model based on the Bernoulli distribution in order to accomodate *hits* and *no hits* with fixed probabilities $(1-P_e)$ and P_e (being P_e the probability of error).

On the other hand, (Gold and Rangarajan, 1996) proposed an iterative algorithm to solve the assignment problem using graduated nonconvexity. They used the compatibility measure between links to gauge the *hits*.

Interestingly, both approaches (Luo and Hancock, 2001) and (Gold and Rangarajan, 1996) according with their respective frameworks (i.e., Expectation-Maximization and graduated nonconvexity) ended up maximizing a similar expression.

We adopt the mentioned expression as our measure of structural consistency for the match $v_a \rightarrow v_\alpha$. This expression is

$$Q_{a\alpha} = \exp\left[\mu \sum_{b\in V_D}\sum_{\beta\in V_M} D_{ab}M_{\alpha\beta}s_{b\beta}\right] \qquad (1)$$

where D and M are the adjacency matrices of G_D and G_M, respectively (i.e., D_{ab} means that there is an edge joining v_a and v_b; and $M_{\alpha\beta}$ means that there is an edge joining v_α and v_β), $s_{b\beta} \in S$ is an element of the $n \times m$ current matching matrix S and $\mu > 0$ is a control parameter.

The presented expression is the exponential of the number of *hits* for a match $a \rightarrow \alpha$, weighted by a parameter μ.

In (Gold and Rangarajan, 1996), μ controls the convexity to avoid poor local minima. A high value of μ tends to exaggerate the difference of the highest values with respect to the others. On the other hand, in (Luo and Hancock, 2001), $\mu = \ln[(1-P_e)/P_e]$. A high value of P_e means not to penalize too much the structural errors. This has sense, since increasing the value of P_e (decreasing the value of μ) has the effect of smoothing the differences among the values.

3.2 A Measure of Similarity between SIFT Attributes

Up to this point, we have shown how to measure the contribution of matching one node to another with regards to the structural relations.

We propose the inverse of the distance as the measure of similarity of the SIFT attributes from two nodes. More formally, we define the similarity of matching node $v_a \in V_D$ to node $v_\alpha \in V_M$ with regards to the SIFT attributes as

$$R_{a\alpha} = \frac{1}{\text{dist}\,(z_a^D, z_\alpha^M)} \qquad (2)$$

where $dist\left(z_a^D, z_\alpha^M\right)$ is the Euclidean distance between SIFT descriptors $z_a \in Z_D$ and $z_\alpha \in Z_M$.

The advantages of this measure are that we can easily reformulate the ratio criterion of definition 4 so as to obtain the same results as the original SIFT matching, while having turned a distance to a similarity function.

3.3 A Combined Measure of Consistency and Similarity

We propose a combined measure of the consistence of a match inspired in the work of simultaneous graph matching and alignment by Cross and Hancock (Cross and Hancock, 1998). In their work, they combined the structural relations with attributes information of the 2D coordinate positions to recover both the configuration of matches and the spatial transformation. Since the SIFT descriptors have a constant value throughout the process, we are only interested in recovering the correspondences.

Our combined expression for gauging the consistence of matching the node $v_a \in V_D$ to node $v_\alpha \in V_M$ is:

$$W_{a\alpha} = Q_{a\alpha}R_{a\alpha} \qquad (3)$$

where $Q_{a\alpha}$ is the structural consistency coefficient described in equation (1) and $R_{a\alpha}$ is the SIFT similarity coefficient described in equation (2).

The use of the multiplication to combine the measures due to both the local information and the surrounding matches is closely related to the idea of Probabilistic Relaxation (R.A. and W., 1983).

With this measure to hand, we define the matrix Ω of combined coefficients as:

$$\Omega = \begin{pmatrix} W_{11} & \dots & W_{1m} \\ \vdots & W_{a\alpha} & \vdots \\ W_{n1} & \dots & W_{nm} \end{pmatrix} \qquad (4)$$

3.4 A Cleaning Heuristic

A cleaning heuristic is needed to obtain a binary $n \times m$ *matching matrix* (definition 2) S that selects the matches corresponding to the highest coefficients from the continuous matrix Ω. Since it may not exist an exact isomorphism between two graphs G_D and G_M, we also need a criterion to match nodes $v_a \in G_D$ to *NULL*.

Borrowing the idea of the ratio criterion of the *positive SIFT matches*, we propose the following cleaning procedure:

1. Initialize S to an $n \times m$ matrix of zeros. Let $\Omega' = \Omega$. Set all $W'_{a,\alpha} \in \Omega'$ to zero except those $W'_{a,k}$ from each row a of Ω' s.t. $W'_{a,k} = \max\left(W'_{a,\alpha}\right)$, $\alpha = 1,\ldots,m$ and $W'_{a,k}/W'_{a,k2} > \frac{1}{\rho}$, where $W'_{a,k2}$ is the second highest element in the a-th row of Ω'. Note that we use the same ratio value $0 \leq \rho \leq 1$ as in definition 4 to control the acceptance rate.
2. Find the maximun element $W'_{a,\alpha} \in \Omega'$ and activate the corresponding match $s_{a\alpha} \in S$
3. Set to zeros the row and column of Ω' where $W'_{a,\alpha}$ belongs to
4. Repeat steps 2-3 until Ω' becomes a matrix of zeros

3.5 The Algorithm

Let G_D and G_M be two *K-nn SIFT graphs* obtained from two pairs of stereo-images, with n and m nodes respectively.

Our algorithm for matching G_D to G_M is the following:

1. Initialize the matching matrix at the first iteration $S^{(1)}$ to be the result of applying the cleaning heuristic of subsection 3.4 to the matrix of SIFT similarity coefficients computed using equation (2). Note that the structural information has no influence in the computation of the initial *matching matrix*, and that it becomes an injective set of *positive SIFT matches*.
2. At iteration i, compute the $n \times m$ matrix of combined coefficients $\Omega^{(i)}$ as described in equations (3) and (4)
3. Compute the $n \times m$ matching matrix $S^{(i+1)}$ applying to $\Omega^{(i)}$ the cleaning heuristic described in subsection 3.4
4. Increment the iteration number i and repeat steps 2-4 until convergence of the matching matrix

In the next section we give further details about the empirical behaviour shown by the present algorithm with regards to the convergence.

4 EXPERIMENTS AND DISCUSSION

We have designed a pose recovery experiment to evaluate the effectivenes of three image-features matching methods. The errors obtained estimating the displacement of the robot using the matches selected by each method are taken as a measure of effectiveness of the method.

We have 130 stereo-image pairs taken at different places during a mobile robot outdoor route. We have the ground truth positions and orientations of the robot at these places, computed with high precision using the SLAM system presented in (Ila et al., 2010), (V. Ila and Andrade-Cetto, 2009). These 130 images are divided into two sets, the origin set and the destination set, with 65 images each, so that one image of the origin set is matched against one image of the destination set (thus, we carry out 65 matching experiments).

To summarize, for each experiment we have two pairs of stereo-images (origin and destination) each one with its set of SIFT keys (features). Each feature is associated to a 3D coordinate position (through stereo-triangulation), relative to the camera of the robot at that place. We also have the ground truth positions and orientations of each place.

Given an association between the features at the origin and destination places, we make an estimation of the destination pose (position and orientation), as described in (V. Ila and Sanfeliu, 2007).

We assume that the higher the error between the estimated and the ground truth destination poses is, the worse the features association is.

We have compared the approach presented in this paper (denoted as Graph Matching with SIFT) to SIFT matching (definition 4) (Lowe, 2004) and Graph Transformation Matching (Aguilar et al., 2009).

We have used a maximum of 80 features in each experiment when available. This has represented around a 50% of the total available data. The features have been selected among the most salient (regarding the gradient magnitude of the SIFT keys). We have built a *K-nn SIFT graph* (definition 5) for each pair of stereo-images using the values of $K = 7$ in the Graph Transformation Matching method and $K = 21$ in our method. The value of $\mu = 0.15$ from equation (1) has been used in our method (the values used in all the methods have been carefully chosen to perform well).

Figures 1 and 2 show the mean errors among the 65 experiments of each method at an interval of matching acceptance ratios ranging from 0.4 to 1. Although lower ratio values often lead to less error (better quality matches), the analysis at values lower than 0.4 has not too much interest, since a significant number of matching experiments return not enough (or not at all) results to recover the spatial transformations.

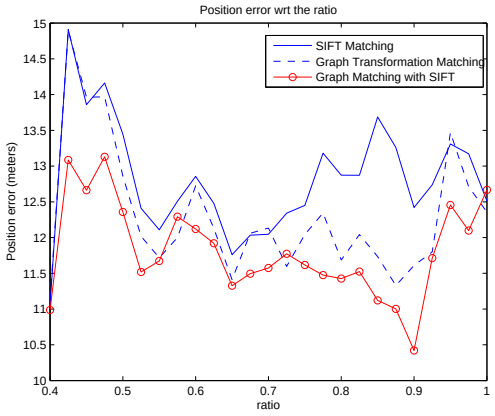

Figure 1: Position errors w.r.t. the acceptance ratio.

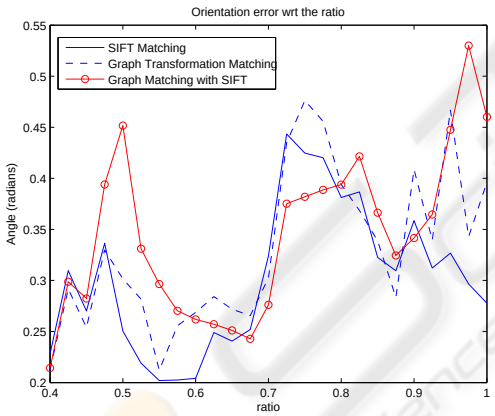

Figure 2: Orientation errors w.r.t. the acceptance ratio.

Figure 3 shows the mean number of matches returned by each method at each acceptance ratio.

We have empirically observed two different behaviours with regards to the convergence of the present algorithm. The stable case, where the matching matrix reaches a certain configuration that remains stable along iterations. In this case, we stop our algorithm at the first iteration where the matching matrix does not change. The unstable case, where the matching matrix evolves until a point where it starts to loop indefinitely between two different configurations. In this case, we stop our algorithm and we arbitrarily choose one of both configurations as we consider that they are equally likely solutions. Both

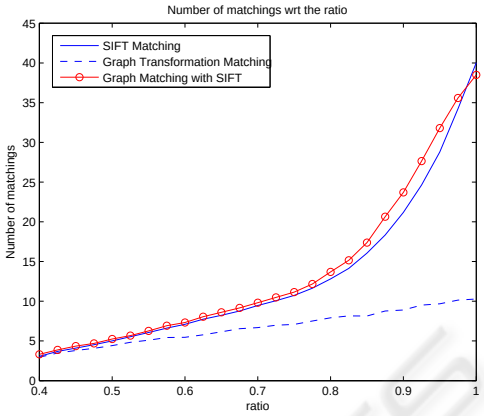

Figure 3: Number of matches returned w.r.t. the acceptance ratio.

cases are observed nearly in the same number of experiments. Figure 4 represents the mean number of iterations needed by our method to stop (regarding the mentioned criteria). The maximum number of iterations permitted is 20.

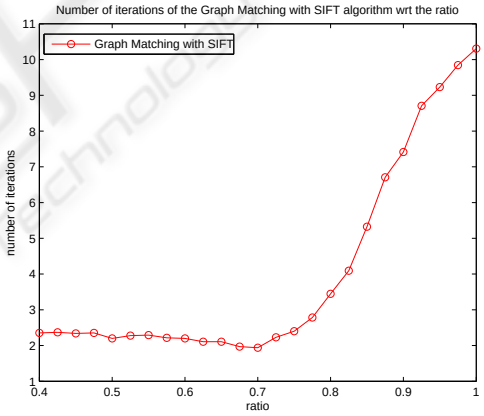

Figure 4: Number of iterations until stop.

As is shown in figure 1, the proposed approach demonstrates to perform significantly better than the others with regards to the position estimation, in the database used in this experiment. On the other hand, as we see in figure 2, orientation errors are not as good as it would be expected. It is worth noting that the Graph Transformation Matching method also experiments a performance decreasing with respect to SIFT matching in orientation recovery. We need to further study this fact, since both approaches are aimed at the improvement of the SIFT matching. Figure 3 show evidences that our method is not only supposed to remove outliers, but also to introduce additional useful matches. It can be observed that, while the positive SIFT matches added above an acceptance ratio of 0.65 ($\approx$ 7 matches) do nothing but to deteriorate

its efficiency, our method improves its efficiency until a threshold of 0.9 ($\approx$ 24 matches), where it actually performs optimally. On the other hand, it is clear that the enhancement introduced by the Graph Transformation Matching is, as expected, based on the rejection of the SIFT outliers.

5 CONCLUSIONS

We have presented a new attributed graph matching algorithm that combines the local texture information of the SIFT descriptors with the higher level information of the graph structure to derive a set of matches. Unlike the SIFT enhancements based on outlier rejection, our approach aims to both eliminate erroneous matches and add new useful ones. We have evaluated three different approaches to image-feature matching in a pose recovery application: our method, SIFT matching (Lowe, 2004), and a graph-based outlier rejector run on the positive SIFT matches (Aguilar et al., 2009). In the methods that use graphs, we have used the 3-Dimensional positional information attached to each feature to build the K-nn SIFT graphs.

In the position estimation experiments, our approach has been superior than the others. With a higher number of correspondences than SIFT matching, our method gets even a lower positional error than the outlier rejector. In conclusion, our method gets more and better matches.

On the other hand, both our method and the outlier rejector perform worse than SIFT matching in orientation recovery. This seems contradictory since both methods are designed as an enhancement of SIFT matching. We therefore need to further study this fact.

ACKNOWLEDGEMENTS

We want to acknowledge Juan Andrade-Cetto and Viorela Ila for providing us with the stereo images from the robot route, the code for triangulating the 3D feature positions and the ground truth poses used to compute the errors.

This research was partially supported by Consolider Ingenio 2010, project CSD2007-00018, by the CICYT project DPI 2007-61452 and by the Universitat Rovira i Virgili.

REFERENCES

Aguilar, W., Frauel, Y., Escolano, F., and Martinez-Perez, M. E. (2009). A robust graph transformation matching for non-rigid registration. *Image and Vision Computing*, 27:897–910.

Besl, P. J. and McKay, N. D. (1992). A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2).

Brown, M. and Lowe, D. G. (2003). Recognising panoramas. *Proceedings of the International Conference on Computer Vision*.

Cross, A. D. J. and Hancock, E. R. (1998). Graph matching with a dual-step em algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11).

Frank-Bolton, P., Alvarado-Gonzalez, A. M., Aguilar, W., and Frauel, Y. (2008). Vision based localization for mobile robots using a set of known views. In *Proceedings of Advances in Visual Computing (LNCS)*, volume 5358 of *4th International Symposium on Visual Computing*, pages 195–204.

Gold, S. and Rangarajan, A. (1996). A graduated assignment algorithm for graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(4).

Ila, V., Porta, J. M., and Andrade-Cetto, J. (2010). Information-based compact pose slam. *IEEE Transactions on Robotics*, 26(1). In press.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2).

Luo, B. and Hancock, E. R. (2001). Structural graph matching using the em algorithm and singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10).

R.A., H. and W., Z. S. (1983). On the foundations of relaxation labeling proecesses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(3).

V. Ila, J. P. and Andrade-Cetto, J. (2009). Reduced state representation in delayed state slam. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Saint Louis*, pages 4919–4924.

V. Ila, J. Andrade-Cetto, R. V. and Sanfeliu, A. (2007). Vision-based loop closing for delayed state robot mapping. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego*.