Dimensionality Reduction Using Automatic Supervision for Vision-Based Terrain Learning

Anelia Angelova
Computer Science Dept.
California Institute of Technology
Email: anelia@vision.caltech.edu

Larry Matthies, Daniel Helmick
Jet Propulsion Laboratory
California Institute of Technology
lhm, dhelmick@jpl.nasa.gov

Pietro Perona
Electrical Engineering Dept.
California Institute of Technology
perona@vision.caltech.edu

Abstract— **This paper considers the problem of learning to recognize different terrains from color imagery in a fully automatic fashion, using the robot's mechanical sensors as supervision. We present a probabilistic framework in which the visual information and the mechanical supervision interact to learn the available terrain types. Within this framework, a novel supervised dimensionality reduction method is proposed, in which the automatic supervision provided by the robot helps select better lower dimensional representations, more suitable for the discrimination task at hand. Incorporating supervision into the dimensionality reduction process is important, as some terrains might be visually similar but induce very different robot mobility. Therefore, choosing a lower dimensional visual representation adequately is expected to improve the vision-based terrain learning and the final classification performance. This is the first work that proposes *automatically supervised* dimensionality reduction in a probabilistic framework using the supervision coming from the robot's sensors. The proposed method stands in between methods for reasoning under uncertainty using probabilistic models and methods for learning the underlying structure of the data.**

The proposed approach has been tested on field test data collected by an autonomous robot while driving on soil, gravel and asphalt. Although the supervision might be ambiguous or noisy, our experiments show that it helps build a more appropriate lower dimensional visual representation and achieves improved terrain recognition performance compared to unsupervised learning methods.

I. INTRODUCTION

We consider the problem of learning to recognize terrain types from color imagery for the purposes of autonomous navigation. This is necessary because different terrains induce different mobility limitations on the vehicle. For example, the robot might get stuck in sand or mud, so it has to learn to avoid such terrains. Visual information is used as a forward-looking sensor to determine the terrain type *prior* to the robot entering the terrain, so that a better planning can be done. In this paper the robot learns *automatically* using its own mechanical measurements while traversing the terrains. In particular, the amount of robot slip is used as supervision for learning different terrain types and the robot's mobility on them.

Learning fully automatically is important, because in the context of autonomous navigation huge amounts of data are available and providing manual supervision is impractical. To avoid manual labeling, the so-called *self-supervised* learning methods have proposed to use the vehicle's sensors as supervision for learning [4], [11], [13], [16], [19]. The key idea of self-supervised learning is that one of the sensors can provide the ground truth for learning with another sensor and the underlying assumption is that the former sensor can be reliably clustered or thresholded [4], [11], [13], [16].

However, some signals obtained from the robot do not necessarily provide a unique clustering into well separable classes, but can be still useful for providing supervision. For example, different terrain types might induce similar robot mobility, i.e. the supervision might be *ambiguous*. In the particular case of slip, which is slope dependent, the robot can have the same slip on flat ground but different slip when traversing slopes. Our previous work [3] proposed a unified learning framework for this case, but its limitation is that the visual representation is low dimensional and the method can become numerically brittle or require prohibitive amounts of training data for higher dimensional inputs. Robotics applications often need to process data obtained from multiple sensors which is high dimensional. In particular, feature representations of visual data are typically of high dimensionality, especially if fine distinctions between terrains need to be done or a lot of intra-class variability has to be accommodated.

To cope with high dimensional input spaces, we propose to use the supervision, automatically obtained by the robot, to affect the dimensionality reduction process. The intuition is that two visually similar terrains which are not normally discriminated in the visual space, and are mapped to the same cluster in the lower dimensional space, might be discriminated properly after introducing the supervision. In our case the mechanical supervision is in the form of robot slip and might be ambiguous or noisy. To solve the problem in this setup, we present a probabilistic framework in which the mechanical supervision provided by the robot is used to learn the representation and classification of terrain types in the visual space automatically. This essentially means having the supervision help choose more appropriate and meaningful, with respect to the learning task, low dimensional projections of the initial visual data. Most previous dimensionality reduction techniques are completely unsupervised [17], [21], whereas here we propose to learn a more useful lower dimensional visual representation which at the same time allows for better discrimination of terrains determined to be different by the automatic mechanical supervision from the robot. The significance of the approach is

that a fully automatic learning and recognition of terrain types can be performed *without* using human supervision for data labeling. Moreover, the method allows the supervision signal obtained by the robot to be noisy or ambiguous, i.e. it might not have a one-to-one correspondence to the visual data.

II. PREVIOUS WORK

Learning to recognize terrains from vision and to determine their characteristics regarding traversability or robot mobility has been widely applied for autonomous vehicles [11], [16], [24]. However, current methods are not automated enough and human supervision or some other heuristics are still needed to determine traversability [9], [16]. Recently, the concept of learning from the vehicle's sensors, referred to as *learning from proprioception* [16], or *self-supervised learning* [4], [13], [19], has emerged. This idea has proved to be particularly useful for extending the perception range [4], [9], [16], [19] which is crucial to increasing the speed and efficiency of the robot [4]. Self-supervised learning approaches require good separability in the space of sensor responses, so that a unique terrain class assignment for each example is obtained. The latter is not always possible, e.g. driving at slower speed cannot produce definitive enough vibration patterns to discriminate terrains [6].

Dimensionality reduction techniques have also become very popular in robotics applications, because the input visual data is of high dimensionality and more efficient representations are needed [8], [12], [22]. Most previous dimensionality reduction methods are unsupervised [7], [17], [21], as they have been intended for data representation. However, in our robotics application, where additional mechanical sensor measurements are available, it is more rational to use them as supervision in selecting better lower dimensional data representation. Some recent work has proposed to include prior information into the dimensionality reduction framework, for example, by using known class labels [20] or by assuming the projections of some examples are given [25]. In our case, the supervision, i.e. the knowledge about class-membership, is fairly weak and neither of these approaches can be applied.

This work extends the probabilistic formulation for dimensionality reduction using Mixture of Factor Analyzers (MoFA) [7], [12], [17] with the major distinction that additional measurements, obtained independently by the robot, are used as supervision in the dimensionality reduction process. Moreover, in [17], [12] the lower dimensionality representation is observed (obtained by applying the unsupervised dimensionality reduction algorithm Isomap [21] prior to learning), whereas here it is unknown and needs to be learned. The particular application addresses recognizing terrain types and inherent mobility related to robot slip using visual input, similar to [2], with the difference that learning is done with automatic supervision, provided by the robot, and does not need manual labeling of terrain types, as in [2]. Being able to predict certain mechanical terrain properties remotely from only visual information and other sensors onboard the vehicle has

significant importance in autonomous navigation applications, because more intelligent planning could be done [16], [24].

III. PROBLEM FORMULATION

Consider the problem of predicting the mobility characteristics Z of the robot in each map cell of the forthcoming terrain using as input the visual information $\mathbf{x} \in \Omega$ in the cell and some information about the terrain geometry $\mathbf{y} \in \Phi$, e.g. local terrain slope (Ω is the visual space, Φ is the space of terrain slopes). The input variables $\mathbf{x}$ and $\mathbf{y}$ can be obtained by the robot from a distance, which will allow the prediction of the output variable from a distance too. Let us denote the function that needs to be evaluated as $Z = F(\mathbf{x}, \mathbf{y})$.

This problem can be reduced to recognizing the terrain type from visual information. That is, we can assume that there are a limited number (K) of terrain types that can be encountered and that on each terrain type the robot experiences different behavior (e.g. mobility):

$$F(\mathbf{x}, \mathbf{y}) = f_j(\mathbf{y}), \qquad \text{if } \mathbf{x} \in \Omega_j \qquad (1)$$

where $\Omega_j \in \Omega$ are different subsets in the visual space, $\Omega_i \cap \Omega_j = \emptyset, i \neq j$ and $f_j(\mathbf{y})$ are (nonlinear) functions which work in the domain Φ and which change their behavior depending on the terrain. In other words, different mobility behaviors occur on different terrain types which are determined by visual information. Now the question is how to learn the mapping $Z = F(\mathbf{x}, \mathbf{y})$ from training data $D = \{(\mathbf{x}_i, \mathbf{y}_i), z_i\}_{i=1}^N$, where $\mathbf{x}_i$ are the visual representations of patches from the observed terrain, $\mathbf{y}_i$ are the terrain slopes, and z_i are the slip measurements when the robot traverses that terrain.

The input space X, representing the visual data, can be of a very high dimension, which impedes working with it. Instead, we work with a lower dimensional embedding U of the input space X. For that purpose we need to learn the embedding $R : X \rightarrow U$ itself. As the learning of this mapping requires prohibitive amount of data whenever the input is high dimensional, we assume, similar to [7], [12], that it takes a particular form. Namely:

$$\mathbf{x} = \Lambda_j \mathbf{u}_j + \mathbf{v}_j \qquad \text{for } \mathbf{x} \in \Omega_j \qquad (2)$$

where Λ_j is the projection matrix and $\mathbf{u}_j, \mathbf{v}_j$ are normally distributed: $\mathbf{u}_j \sim \mathcal{N}(\mu_j, \Sigma_j)$, $\mathbf{v}_j \sim \mathcal{N}(\eta_j, \Psi_j)$. That is, we assume that a locally linear mapping is a good enough approximation for patches that belong to the same terrain class.

Figure 1 visualizes the problem when measurements of slip as a function of terrain slope are used as supervision. Robot slip is a measure of the lack of progress and is essentially the complement of robot mobility [2]. The measurements in Figure 1 are obtained from actual robot traversals and are computed as the difference between Visual Odometry (VO) based pose estimates [15] and the pose estimates from the kinematic model of the robot. The mechanical slip measurements are received completely automatically, as only the vehicle's sensors are needed to compute slip. A nonlinear model can approximate the slip behavior as a function of
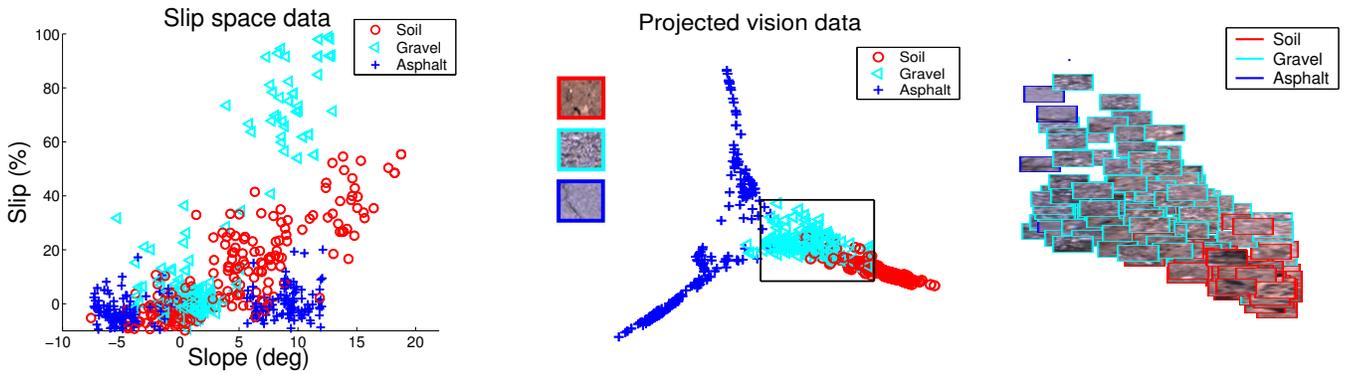
Fig. 1. Left: Slip measurements to be used as automatic supervision in our learning setup. Each training example consists of an image patch represented as a high dimensional point and a corresponding slip measurement represented as a function of the estimated slope angle. Middle: Lower dimensional projections of the visual data, obtained by the unsupervised dimensionality reduction algorithm Isomap [21]. The rectangle is expanded to the right and visualizes the original image patches. The ground truth terrain types in this figure are provided by human labeling, but our system works without human supervision and relies on the goodness-of-fit of nonlinear slip models to the slip measurements as automatic supervision to learn the terrain representation (dimensionality reduction), terrain classification, and the nonlinear slip models from the available training data.

slope for each terrain type. These models essentially act as supervision, but they are unknown and have to be learned from the data. The slopes can be easily estimated by the robot remotely using range data from stereo, ladar, etc., and a tilt sensor on the robot, which is readily available from the IMU, for example. We consider only the slip in the forward motion direction as dependent on the longitudinal slope, similar to slip measurements done for the Mars Exploration Rover [14], which is a simpler and more straightforward representation of slip than in [2]. This representation is also more convenient for using the slip measurements as supervision during learning. After the robot has learned how to visually discriminate the terrains, it is conceivable to learn more complex slip models using additional input variables (e.g. both longitudinal and lateral slopes, roughness, etc.), as in [2].

Figure 1 also shows the vision part of the input data, represented as described in Section V-B, projected into 2D by using the unsupervised dimensionality reduction algorithm Isomap [21]. As seen, there is a significant overlap between terrain classes which have visually similar patches. Because of the overlap, performing unsupervised, purely vision-based classification is not sufficient. So, to be able to learn to correctly discriminate these terrains and predict a potentially different mobility behavior on them, some form of supervision is needed. The key idea is that the dimensionality reduction process can also take advantage of the supervision information obtained from additional mechanical sensors.

The main problem in our formulation is that the slip signal to be used as supervision can be of very weak form and using slip measurements as supervision cannot be reduced to supervised learning, as in [4], [11]. In particular, because of the nonlinearity of the slip models $f_i(\mathbf{y})$, it is possible that some of the models overlap in parts of their domain (i.e. for some $i, j, i \neq j$, $f_i(\mathbf{y}) \equiv f_j(\mathbf{y})$, for $\mathbf{y} \in \Phi_0$, for some $\Phi_0 \subseteq \Phi$). For example, several terrains might exhibit the same slip for $\sim 0°$ slope, as seen in Figure 1, or simply two visually different terrain types might have the same slip behavior. Since

some of the supervision (for some of the training examples) is inherently ambiguous, the slip supervision signals cannot be directly clustered into well separable classes. However, if two terrains exhibit different slip behavior for any slope range, the supervision should still be able to force a better discrimination in the visual space, even though not all examples can definitively exercise supervision. The intuition is that examples for which the supervision signal is strong will propagate it to the examples of ambiguous supervision in the same class through their visual similarity. Finally, as the supervision is collected automatically by the robot's mechanical sensors, it is rather noisy. To cope with noisy and ambiguous supervision signals necessitates a framework which allows reasoning under uncertainty.

To summarize, our goal is to learn the function $Z = F(\mathbf{x}, \mathbf{y})$ from the available training data $D = \{\mathbf{x}_i, \mathbf{y}_i, z_i\}_{i=1}^N$. Thus, after learning, the mechanical behavior z for some query input example $(\mathbf{x}_q, \mathbf{y}_q)$ will be predicted as $z = F(\mathbf{x}_q, \mathbf{y}_q)$. We do not want to use manual labeling of the terrain types during training, so the slip measurements z_i, which are assumed to have come from one of several unknown nonlinear models, act as the only supervision to the whole system. The main problem is that using the mechanical measurements as the only ground truth, or supervision, we have to learn both the terrain classification and the unknown nonlinear functions for each terrain. In particular, a combinatorial enumeration problem needs to be solved as a subproblem, which is known to be computationally intractable [10]. Furthermore, the supervision is noisy and ambiguous.

IV. PROBABILISTIC FRAMEWORK FOR DIMENSIONALITY REDUCTION USING SUPERVISION

To solve the problem defined in Section III, we propose a probabilistic framework (Section IV-C) which performs dimensionality reduction and terrain classification by using automatic supervision and which can cope with both noisy and ambiguous supervision. A maximum likelihood estimation
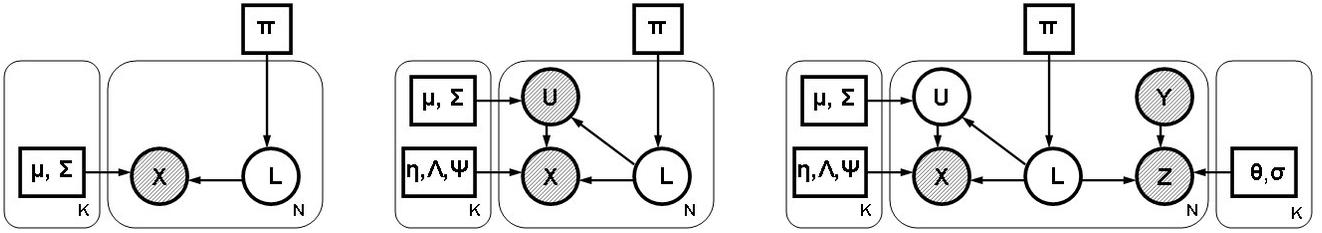
Fig. 2. Left: Graphical model of unsupervised clustering in the initial visual space. Middle: Graphical model of unsupervised dimensionality reduction based on MoFA [12], [17] (see also [7]). Right: Graphical model of automatically supervised dimensionality reduction in which mechanical measurements obtained automatically from the robot are used as supervision (proposed in this paper). The automatic supervision influences the selection of appropriate low dimensional representations and helps learn the distinction between different terrain types. The observed random variables are displayed in shaded circles.

will be done in this framework. To ease the exposition, we first describe two related probabilistic models.

A. Unsupervised clustering

The most straightforward approach to learn to classify examples corresponding to different terrains is to apply unsupervised learning (clustering). The corresponding graphical model is shown in Figure 2, left. The parameters μ_j, Σ_j are the means and covariances of each of the K clusters of visual data X and π_j are the prior probabilities of each class. The indicator variables L are *latent*, i.e. hidden, and are added to simplify the inference process; they define the class-membership of each training example, i.e. $L_{ij} = 1$ if the ith training example $\mathbf{x}_i$ belongs to the jth class. The model is used to learn the parameters of each class and the classification boundaries between them. However, inference in high dimensional spaces is numerically brittle and is limited by the amount and the diversity of the available training data.

B. Unsupervised dimensionality reduction

As operating in high dimensional spaces is not desirable, we wish to find a lower dimensional representation U of the initial visual space X. As previously shown [7], dimensionality reduction can be done using Mixture of Factor Analyzers (MoFA), which can be expressed probabilistically as follows:

$$P(X, U) = \sum_{j=1}^{K} P(X|U, C = j)P(U|C = j)P(C = j) \quad (3)$$

in which it is assumed that $\{X|U, C = j\} \sim \mathcal{N}(\Lambda_j U + \eta_j, \Psi_j)$ and $U \sim \mathcal{N}(\mu_j, \Sigma_j)$. In other words, the joint probability of X and U is assumed to be modeled as a mixture of K local linear projections, or factors (see Equation (2)) [7], [17]. In this paper we assume that U are latent variables. This is a more general case than both [7] and [17]. After introducing auxiliary latent variables L_{ij}, as above, we can write Equation (3) in the following way (which corresponds to the graphical model in Figure 2, middle):

$$P(X, U, L|\Theta_0) = P(X|U, L, \Theta_0)P(U|L, \Theta_0)P(L|\Theta_0),$$

where $\Theta_0 = \{\mu_j, \Sigma_j, \Lambda_j, \eta_j, \Psi_j, \pi_j\}_{j=1}^{K}$ contains the unknown parameters of the model. Because of the particular assumptions about the model, made in Equation (2), the

probability of a data point $\mathbf{x}_i$ belonging to a terrain class j, given a latent representation $\mathbf{u}_i$, and the probability of the latent representation $\mathbf{u}_i$, given the class j, are expressed as:

$$P(\mathbf{x}_i|\mathbf{u}_i, L_{ij} = 1) = \frac{e^{-\frac{1}{2}(\mathbf{x}_i - \Lambda_j \mathbf{u}_i - \eta_j)^T \Psi_j^{-1}(\mathbf{x}_i - \Lambda_j \mathbf{u}_i - \eta_j)}}{(2\pi)^{D/2}|\Psi_j|^{1/2}}$$

$$P(\mathbf{u}_i|L_{ij} = 1) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} e^{-\frac{1}{2}(\mathbf{u}_i - \mu_j)^T \Sigma_j^{-1}(\mathbf{u}_i - \mu_j)},$$

where D and d are the dimensionalities of the initial visual space and the projected representation, respectively. Those distributions are modeled, so that a tractable solution to the maximum likelihood estimation problem is achieved.

C. Automatically supervised dimensionality reduction

Previous approaches have assumed the projections U of the data are known [12], [17] or have obtained them by unsupervised learning [7]. In this work we wish to have the automatic supervision influence which projections are chosen to best represent and consequently discriminate the visual classes. For that purpose we introduce supervision into the whole maximum likelihood framework, thus solving the initial problem in Equation (1), considering all the data available to the system. That is, the ambiguous mechanical supervision also takes part in the maximum likelihood decision.

In particular, we have two parts, a vision part, in which dimensionality reduction is done, and a mechanical behavior part, in which the slip measurements act as supervision. They are linked through the fact that they refer to the same terrain type, so they both give some information about this terrain. In other words, during learning, we can use visual information to learn something about the nonlinear mechanical models, and conversely, the mechanical feedback to supervise the vision based dimensionality reduction and terrain classification. Our goal is to make those two different sets of information interact.

The main problem is that the decision about the terrain types and learning of their mechanical behavior are not directly related (i.e. they are done in different, decoupled spaces) but they do refer to the same terrains. We can do that decoupling by using again the hidden variables L which define the class-membership of each training example (here $L_{ij} = 1$ if the ith training example $(\mathbf{x}_i, \mathbf{y}_i, z_i)$ has been generated by the jth nonlinear slip model and belongs to the jth terrain class). As

Input: Training data $\{\mathbf{x}_i, \mathbf{y}_i, z_i\}_{i=1}^{N}$, where $\mathbf{x}_i$ are the vision domain data, $\mathbf{y}_i$ are the geometry domain data, z_i are the mechanical supervision measurements. **Output:** Estimated parameters Θ of the system.

Algorithm: Initialize the unknown parameters Θ^0. Set $t = 0$. Repeat until convergence:

1. (E-step) Estimate the expected values of L_{ij}, $\mathbf{u}_{ij}$ (we denote $\mathbf{u}_{ij} = E(\mathbf{u}|\mathbf{x}_i, L_{ij} = 1)$):

$$L_{ij}^{t+1} = \frac{P(\mathbf{x}_i|L_{ij}=1,\Theta^t)P(\mathbf{y}_i,z_i|L_{ij}=1,\Theta^t)\pi_j^t}{\sum_{k=1}^{K} P(\mathbf{x}_i|L_{ik}=1,\Theta^t)P(\mathbf{y}_i,z_i|L_{ik}=1,\Theta^t)\pi_k^t}, \text{ where } \mathbf{x}_i \sim \mathcal{N}(\Lambda_j^t\mu_j^t + \eta_j^t, \Psi_j^t + \Lambda_j^t\Sigma_j^t(\Lambda_j^t)')$$

$$\mathbf{u}_{ij}^{t+1} = \Upsilon[(\Lambda_j^t)'(\Psi_j^t)^{-1}(\mathbf{x}_i - \eta_j^t) + (\Sigma_j^t)^{-1}\mu_j^t], \text{ where } \Upsilon = [(\Sigma_j^t)^{-1} + (\Lambda_j^t)'(\Psi_j^t)^{-1}\Lambda_j^t]^{-1}.$$

2. (M-step) Select the parameters Θ^{t+1} to maximize $CL(X,U,Y,Z,L|\Theta^t)$. Let $l_{ij}^{t+1} = L_{ij}^{t+1}/\sum_{r=1}^{N} L_{rj}^{t+1}$:

$$\mu_j^{t+1} = \sum_{i=1}^{N} l_{ij}^{t+1}\mathbf{u}_{ij}^{t+1}; \quad \Sigma_j^{t+1} = \sum_{i=1}^{N} l_{ij}^{t+1}\mathbf{u}_{ij}^{t+1}(\mathbf{u}_{ij}^{t+1})' - \mu_j^{t+1}(\mu_j^{t+1})' + \Upsilon; \quad \eta_j^{t+1} = \sum_{i=1}^{N} l_{ij}^{t+1}(\mathbf{x}_i - \Lambda_j^t\mathbf{u}_{ij}^{t+1})$$

$$\Lambda_j^{t+1} = [\sum_{i=1}^{N} L_{ij}^{t+1}(\mathbf{x}_i - \eta_j^t)(\mathbf{u}_{ij}^{t+1})'][\sum_{i=1}^{N} L_{ij}^{t+1}(\mathbf{u}_{ij}^{t+1}(\mathbf{u}_{ij}^{t+1})' + \Upsilon)]^{-1}; \quad \Psi_j^{t+1} = \sum_{i=1}^{N} l_{ij}^{t+1}(\mathbf{x}_i - \eta_j^{t+1} - \Lambda_j^{t+1}\mathbf{u}_{ij}^{t+1})(\mathbf{x}_i - \eta_j^{t+1})'$$

$$\theta_j^{t+1} = (G'L_j^{t+1}G)^{-1}G'L_j^{t+1}Z; \quad (\sigma_j^2)^{t+1} = \sum_{i=1}^{N} l_{ij}^{t+1}(z_i - G(\mathbf{y}_i, \theta_j^{t+1}))^2; \quad \pi_j^{t+1} = \sum_{i=1}^{N} L_{ij}^{t+1}/N.$$

3. $t = t + 1$

Fig. 3.   EM algorithm updates (see [1] for details).

an additional step, a dimensionality reduction of the visual part of the data is done, so now the supervision can affect the parameters related to the dimensionality reduction too. This essentially means preferring projections which fit the data, and therefore also the supervision, well. Now, given the labeling of an example is known, the slip supervision measurements and the visual information are independent. So, the complete likelihood factors as follows:

$$P(X,U,Y,Z,L|\Theta) =$$
$$\underbrace{P(X|U,L,\Theta)P(U|L,\Theta)}_{\text{Vision part, dim. red.}} \ \underbrace{P(Y,Z|L,\Theta)}_{\text{Autom. supervision}} \ \underbrace{P(L|\Theta)}_{\text{Prior}}$$

where $\Theta = \{\mu_j, \Sigma_j, \Lambda_j, \eta_j, \Psi_j, \theta_j, \sigma_j, \pi_j\}_{j=1}^{K}$ contains all the parameters that need to be estimated in the system. θ_j are the parameters of the nonlinear fit of the slip data and σ_j are their covariances (here they are the standard deviations, as the final measurement is one dimensional). The graphical model corresponding to this case is shown in Figure 2, right. This model allows the automatically obtained mechanical supervision to affect both the dimensionality reduction and the clustering process, thus improving a purely unsupervised learning for the purposes of the task at hand. Note that here the lower dimensional representation is hidden and that the supervision part can influence the visual learning and the dimensionality reduction through the latent variables L_{ij}.

The supervision part is as follows. The mechanical measurement data are assumed to have come from a nonlinear fit, which is modeled as a General Linear Regression (GLR) [18]. GLR is appropriate for expressing nonlinear behavior and is convenient for computation because it is linear in terms of the parameters to be estimated. For each terrain type j, the regression function $\tilde{Z}(Y) = E(Z|Y)$ is assumed to have come from a GLR with Gaussian noise: $f_j(Y) \equiv Z(Y) = \tilde{Z}(Y) + \epsilon_j$, where $\tilde{Z}(Y) = \theta_j^0 + \sum_{r=1}^{R}\theta_j^r g_r(Y)$, $\epsilon_j \sim \mathcal{N}(0, \sigma_j)$, and g_r are several nonlinear functions selected before the learning has started. Some example nonlinear functions to be

used as building blocks for slip approximation are: x, x^2, e^x, $\log x$, $\tanh x$ (those functions are used later on in our experiments with the difference that the input parameter is scaled first). The parameters $\theta_j^0, ..., \theta_j^R, \sigma_j$ are to be learned for each model j. We assume the following probability model for z_i belonging to the jth nonlinear model conditioned on $\mathbf{y}_i$:

$$P(z_i|\mathbf{y}_i, L_{ij} = 1, \theta_j, \sigma_j) = \frac{1}{(2\pi)^{1/2}\sigma_j}e^{-\frac{1}{2\sigma_j^2}(z_i - G(\mathbf{y}_i, \theta_j))^2},$$

where $G(\mathbf{y}, \theta_j) = \theta_j^0 + \sum_{r=1}^{R}\theta_j^r g_r(\mathbf{y})$ and $\theta_j = (\theta_j^0, \theta_j^1, ..., \theta_j^R)$. $P(\mathbf{y}_i)$ is given an uninformative prior (here, uniform over a range of slopes).

With the help of the hidden variables L, the complete log likelihood function (CL) can be written as:

$$CL(X,U,Y,Z,L|\Theta) = \sum_{i=1}^{N}\sum_{j=1}^{K} L_{ij}[\log P(\mathbf{x}_i|\mathbf{u}_i, L_{ij} = 1, \Lambda_j, \eta_j, \Psi_j) +$$
$$\log P(\mathbf{u}_i|L_{ij} = 1, \mu_j, \Sigma_j) + \log P(\mathbf{y}_i, z_i|L_{ij} = 1, \theta_j, \sigma_j) + \log \pi_j]$$

The introduction of the hidden variables L is crucial to simplifying the problem and allows for it to be solved efficiently with the Expectation Maximization (EM) algorithm [5], which tries to maximize the complete log likelihood (CL). The EM algorithm updates applied to our formulation of the problem are shown in Figure 3 (the detailed derivations of the updates are provided in [1]). In brief, the algorithm performs the following steps until convergence. In the E-step, the expected values of the unobserved variables $\mathbf{u}_{ij}$ and label assignments L_{ij} are estimated. In the M-step, the parameters for both the vision and the mechanical supervision side are selected, so as to maximize the complete log likelihood. In other words, at each iteration better parameters Θ are selected, in a sense that they increase the likelihood of the available data. As the two views are conditionally independent, the parameters for the vision and the mechanical side are updated independently of

one another in the M-step. Note that it is through the variable L that the visual data and the mechanical supervision interact and that the automatic supervision can affect the local projections defining the dimensionality reduction through the variable U. The interaction happens in the E-step of each iteration, by updating the expected values of L and U which depend on *both* the visual data and the supervision. The new variables introduced in Figure 3 are defined as follows: L_j^t is a diagonal NxN matrix which has $L_{1j}^t, ... L_{Nj}^t$ on its diagonal, G is a Nx$(R+1)$ matrix such that $G_{ir} = g_r(\mathbf{y}_i)$, $G_{i(R+1)} = 1$, and Z is a Nx1 vector containing the measurements z_i [1].

D. Discussion

The main difference from previous approaches [7], [12], [17] is that we have incorporated automatic supervision into the framework, which directly affects the lower dimensionality projections and the terrain classification. Furthermore, the variables U corresponding to the low dimensional representation are *latent* (unlike [12], where they are known and obtained from Isomap, prior to learning) and can have arbitrary means and covariances which are learned (unlike [7], where they are assumed to be zero mean and unit variance). This is an important point, because it is through the latent variables U that the supervision can influence the dimensionality reduction process during learning.

The proposed maximum likelihood approach solves the abovementioned combinatorial enumeration problem [10] approximately by producing a solution which is guaranteed to be a local maximum only. Indeed, the EM solution is prone to getting stuck in a local maximum. For example, one can imagine creating adversarial mechanical models to contradict the clustering in visual space. In practice, for the autonomous navigation problem we are addressing, our intuition is that the mechanical measurements are correlated to a large extent with the vision input and will be only improving the vision based classification. This is seen later in the experiments.

V. Experimental evaluation

In this section we apply the proposed automatically supervised dimensionality reduction algorithm to vision-based learning of different terrain types, using slip supervision obtained by the robot.

The learning setup is as follows. The robot collects data by building a map of the environment and obtaining geometry and appearance information for each map cell. When a particular cell is traversed, the robot measures the amount of slippage occurring and saves a training example composed of a visual feature vector (corresponding to a terrain patch), geometry feature vector (here only the slope angle), and the corresponding slip. The collected training examples are used for learning of the mapping between the input visual and geometric features and the output slip. This strategy is commonly applied to learning traversability or other terrain properties from vision [2], [11], [24]. VO [15] is used for robot localization.
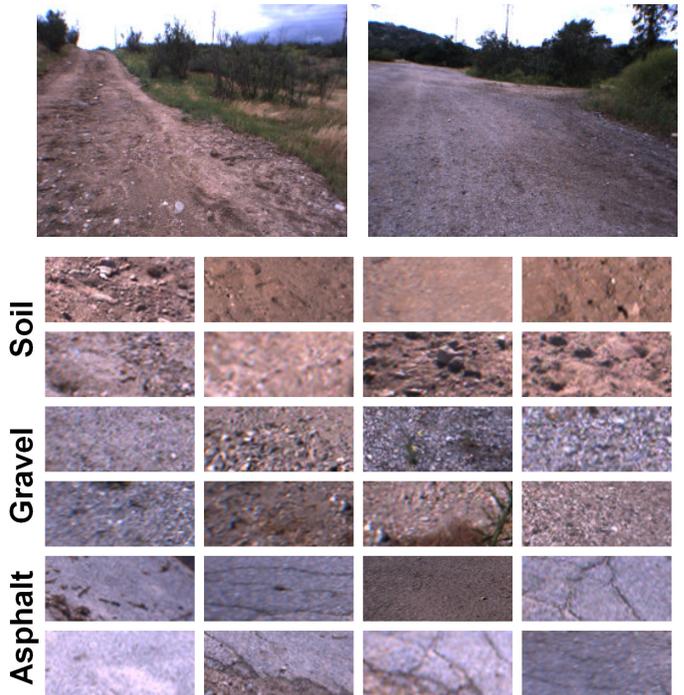

Fig. 4. Top: Example frames from driving on soil (left) and on gravel (right). Bottom: Patches from the classes in our dataset. The variability in texture appearance is one of the challenges present in our application domain. The dataset is collected under various weather conditions.

A. Dataset

The dataset has been collected by an autonomous LAGR[1] robot while driving on three terrains with different mobility in a natural park: soil, gravel and asphalt. Figure 4 shows example patches from the terrains and Figure 1 shows the collected slip measurements in the dataset. It is not known to the algorithm which terrain classes the input examples belong to: the slip and slope measurements (Figure 1) are the only information to be used for automatic supervision. The dataset is quite challenging as it is obtained in outdoor, off-road environments. In particular, a lot of intra-class variability can be observed in the appearance of the terrain patches and the mechanical slip measurements are very noisy.

B. Visual representation

Each terrain patch is represented as the frequency of occurrence (i.e. a histogram) of visual features, called textons, within a patch [23]. The textons are collected by using k-means of 5x5 pixel neighborhoods extracted at random from a pool of training images coming from all the classes (see [23] for details). In this case, 5 textons are selected for each terrain class in the data, constructing a 15-dimensional input feature vector. This representation, based on both color and texture, has been shown to achieve satisfactory classification results for generic textures [23], as well as for natural off-road terrains [2].

C. Mechanical supervision

Robot slip is defined as the difference between the commanded velocity of the robot, obtained from its kinematics model and wheel encoder sensors, and its actual velocity between two consecutive steps [2]. The VO algorithm [15], running onboard the robot, is used to compute its actual velocity. Thus, the slip-based supervision is measured fully automatically by the robot. In these experiments we focus on slip in the forward motion direction as dependent on the longitudinal slope. The terrain slope is retrieved by performing a least-mean-squares plane fit on the average elevations of the map cells in a 4x4 cell neighborhood.

D. Experimental results

In this section we present experimental results of the dimensionality reduction with automatic supervision. We quantitatively evaluate the performance of the proposed algorithm for automatically supervised learning (Figure 2, right) compared to both unsupervised learning (Figure 2, middle [7]) and human supervised learning. While testing, terrain classification is performed first to find the most likely class index j^* given the input data X (let us denote $P(L_j) = P(C = j)$):

$$j^* = argmax_j P(C = j|X) \propto P(X|C = j)P(C = j) =$$

$$\int_u P(X|u, L_j)P(u|L_j)du P(L_j) \approx P(X|U_{ML}, L_j)P(L_j),$$

in which we approximate the integral by using the maximum likelihood lower dimensional projection (U_{ML}). Note that only the visual input is used to make this decision. Then, the expected slip is predicted by evaluating the j^*-th learned slip model $f_{j^*}(Y) = \theta_{j^*}^0 + \sum_{r=1}^R \theta_{j^*}^r g_r(Y)$ for the given slope Y.

The average terrain classification and slip prediction errors and their standard deviations across 50 independent runs are shown in Figure 5. We compare learning and dimensionality reduction without supervision, with automatic supervision, and with human supervision. We have about ∼1000 examples which are split randomly into 70% training and 30% test sets in each run. As the correct slip models are not known, the ultimate test of performance is by comparing the predicted slip to the actual measured slip on a test set (not used in training). Slip prediction error is computed as: Err=$\sum_{i=1}^N |F(\mathbf{x}_i, \mathbf{y}_i) - z_i|/N$, where $F(\mathbf{x}_i, \mathbf{y}_i)$ is the predicted and z_i is the target slip for a test example $(\mathbf{x}_i, \mathbf{y}_i)$. The terrain classification results are evaluated by comparing to human labeled terrains. When using human supervision, the class-membership of each example is known, but the parameters of each class need to be estimated. The latter is equivalent to doing Factor Analysis in each class independently. Due to some overlap between the classes in the original visual space, the classification with human supervision can still incur some nonzero test error in terrain classification. To reflect the monotonic nature of slip, an additional constraint ($\theta_j \geq 0$) is imposed (see [1] for details).

As seen in Figure 5, learning with automatically supervised dimensionality reduction outperforms the unsupervised learning method and decreases the gap to learning with human
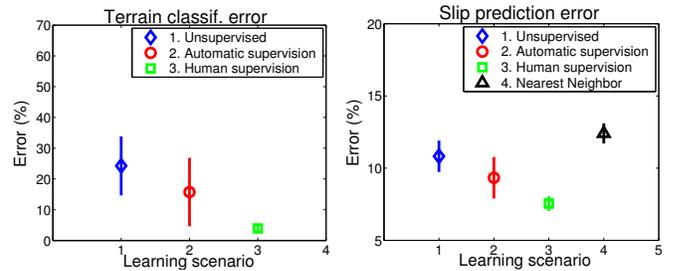

Fig. 5. Average test results for terrain recognition (left) and slip prediction (right). Comparison to a baseline nonlinear regression method is also shown.
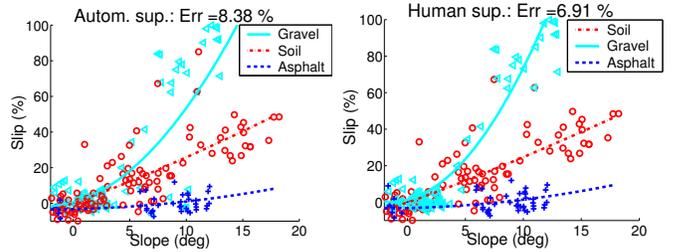

Fig. 6. The learned slip models and the classification of the test examples when learning with automatic supervision (left) and learning with human superivsion (right). The examples are marked according to their predicted terrain class labels (the colors and markers are consistent with Figure 1).

supervision. More precisely, learning with automatic supervision achieves about 42% and 45% of the possible margin for improvement between the unsupervised and the human supervised learning for terrain classification and slip prediction, respectively. Naturally, for the type of supervision used in these experiments (Figure 1), we cannot expect to fully close the gap to human supervision, because the supervision signals are not sufficiently well separable. The improved performance of the supervised dimensionality reduction compared to the unsupervised one is due to selecting more appropriate low dimensional visual representations which provide for better discrimination among the terrain classes and respectively for learning of more accurate slip models for each terrain. Comparing the results to [3] we can see that working with more descriptive high dimensional representations is instrumental to achieving better performance. At the same time, as the representation is more powerful, there is a smaller margin for improvement between the unsupervised and the human supervised learning. We also compared the results to a baseline nonlinear regression method, k-Nearest Neighbor, which learns directly the mapping from the inputs (visual features $\mathbf{x}$ and slope $\mathbf{y}$) to the output (slip z) and does not apply dimensionality reduction as an intermediate step. Note that directly learning the desired outputs, as is with k-Nearest Neighbor, important information about the structure of the problem, namely that there are several underlying terrain types on which potentially different slip behaviors occur, is ignored. As seen in Figure 5, the k-Nearest Neighbor is outperformed by the other three methods.

The learned nonlinear models for one of the runs are shown

in Figure 6. The resultant slip models when learning with automatic supervision are very similar to the ones generated by human supervision, which is due to having learned the correct terrain classification in the visual space. Note that, although the correct slip models have been learned, there are still examples which are misclassified for both learning scenarios because only the visual information is used during testing. The slip model used here has less inputs than in [2] and its main purpose is to act as supervision rather than achieve a good approximation of the slip signal. Now, given that the robot has *automatically* learned how to visually discriminate terrains by using the slip signals as supervision, the final slip prediction results can be further improved by applying a more advanced slip learning algorithm, e.g. by taking into consideration more inputs [2].

Our results show that using additional, automatically obtained, signals as supervision is worthwhile: it outperforms purely unsupervised vision-based learning and has the potential to substitute the expensive, tedious, and inefficient human labeling in applications related to autonomous navigation. Secondly, as more descriptive high dimensional feature representations are crucial to achieving better recognition performance, performing dimensionality reduction and utilizing the automatic supervision in the process is more advantageous than working with simpler lower dimensional representations.

VI. Conclusions and future work

We have proposed a novel probabilistic framework for dimensionality reduction which takes advantage of ambiguous and noisy supervision obtained automatically from the robot's onboard sensors. As a result, simultaneous learning of the lower dimensional representation, the terrain classification, and the nonlinear slip behavior on each terrain is done by using only automatically obtained measurements. The proposed method stands in between reasoning under uncertainty using probabilistic models and retrieving the underlying structure of the data (i.e. dimensionality reduction). The impact of the proposed method of automatically supervised dimensionality reduction is that: 1) a better visual representation can be created by utilizing the supervision from the robot, or the task at hand; 2) the robot can learn about terrains and their visual representation by using its own sensors as supervision; 3) after the learning has completed, the expected mobility behavior on different terrains can be predicted remotely.

We have shown experiments on a dataset collected while driving in the field, in which different terrain types are learned better from both vision and slip supervision than from vision alone and unsupervised dimensionality reduction. Significant improvements, currently under investigation, can be done by introducing temporal/spatial continuity to the consecutive/neighboring terrain measurements. Extending the method to online learning is an important future direction, in which the main challenges are determining which examples to keep in memory and estimating the number of terrains.

References

[1] A. Angelova. EM algorithm updates for dimensionality reduction using automatic supervision. *Technical report*, 2007. http://www.vision.caltech.edu/anelia/publications/DimRedTR.pdf.

[2] A. Angelova, L. Matthies, D. Helmick, and P. Perona. Slip prediction using visual information. *Robotics: Science and Systems Conf.*, 2006.

[3] A. Angelova, L. Matthies, D. Helmick, and P. Perona. Learning slip behavior using automatic mechanical supervision. *International Conference on Robotics and Automation*, 2007.

[4] H. Dahlkamp, A. Kaehler, D. Stavens, S. Thrun, and G. Bradski. Self-supervised monocular road detection in desert terrain. *Robotics: Science and Systems Conference*, 2006.

[5] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–37, 1977.

[6] E. DuPont, R. Roberts, C. Moore, M. Selekwa, and E. Collins. On-line terrain classification for mobile robots. *International Mechanical Engineering Congress and Exposition Conference*, 2005.

[7] Z. Ghahramani and G. Hinton. The EM algorithm for mixtures of factor analyzers. *Tech. Report CRG-TR-96-1, Department of Computer Science, University of Toronto*, 1997.

[8] D. Grollman, O. Jenkins, and F. Wood. Discovering natural kinds of robot sensory experiences in unstructured environments. *Journal of field robotics*, 2006.

[9] M. Happold, M. Ollis, and N. Johnson. Enhancing supervised terrain classification with predictive unsupervised learning. *Robotics: Science and Systems Conference*, 2006.

[10] A. Julosky, S. Weiland, and W. Heemels. A Bayesian approach to identification of hybrid systems. *IEEE Trans. on Automatic Control*, 50(10):1520–1533, 2005.

[11] D. Kim, J. Sun, S. Oh, J. Rehg, and A. Bobick. Traversability classification using unsupervised on-line visual learning for outdoor robot navigation. *Int. Conference on Robotics and Automation*, 2006.

[12] S. Kumar, F. Ramos, B. Upcroft, and H. Durrant-Whyte. A statistical framework for natural feature representation. *International Conference on Intelligent Robots and Systems*, 2005.

[13] D. Lieb, A. Lookingbill, and S. Thrun. Adaptive road following using self-supervised learning and reverse optical flow. *Robotics: Science and Systems Conference*, 2005.

[14] R. Lindemann and C. Voorhees. Mars Exploration Rover mobility assembly design, test and performance. *IEEE International Conference on Systems, Man and Cybernetics*, 2005.

[15] L. Matthies and S. Schafer. Error modeling in stereo navigation. *IEEE Journal of Robotics and Automation*, RA-3(3), June 1987.

[16] L. Matthies, M. Turmon, A. Howard, A. Angelova, B. Tang, and E. Mjolsness. Learning for autonomous navigation: Extrapolating from underfoot to the far field. *NIPS, Workshop on Machine Learning Based Robotics in Unstructured Environments*, 2005.

[17] L. Saul and S. Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.

[18] G. Seber and C. Wild. *Nonlinear Regression*. John Wiley & Sons, New York, 1989.

[19] B. Sofman, E. Lin, J. Bagnell, N. Vandapel, and A. Stentz. Improving robot navigation through self-supervised online learning. *Robotics: Science and Systems Conference*, 2006.

[20] M. Sugiyama. Local Fisher Discriminant Analysis for supervised dimensionality reduction. *Int. Conference on Machine learning*, 2006.

[21] J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000.

[22] B. Upcroft et al. Multi-level state estimation in an outdoor decentralised sensor network. *ISER*, 2006.

[23] M. Varma and A. Zisserman. Texture classification: Are filter banks necessary? *Conf. on Computer Vision and Pattern Recognition*, 2003.

[24] C. Wellington and A. Stentz. Online adaptive rough-terrain navigation in vegetation. *Int. Conference on Robotics and Automation*, 2004.

[25] X. Yang, H. Fu, H. Zha, and J. Barlow. Semi-supervised nonlinear dimensionality reduction. *Int. Conference on Machine learning*, 2006.