A decision-theoretic formulation for sparse stereo correspondence problems

Tom Botterill and Richard Green
Department of Computer Science
University of Canterbury, Christchurch, NZ
tom.botterill@canterbury.ac.nz

Steven Mills
Department of Computer Science
University of Otago
Dunedin, NZ

Abstract

Stereo reconstruction is challenging in scenes with many similar-looking objects, as matches between features are often ambiguous. Features matched incorrectly lead to an incorrect 3D reconstruction, whereas if correct matches are missed, the reconstruction will be incomplete. Previous systems for selecting a correspondence (set of matched features) select either a maximum likelihood correspondence, which may contain many incorrect matches, or use some heuristic for discarding ambiguous matches. In this paper we propose a new method for selecting a correspondence: we select the correspondence which minimises an expected loss function. Match probabilities are computed by Gibbs sampling, then the minimum expected loss correspondence is selected based on these probabilities. A parameter of the loss function controls the tradeoff between selecting incorrect matches versus missing correct matches.

The proposed correspondence selection method is evaluated in a model-based framework for reconstructing branching plants, and on simulated data. In both cases it outperforms alternative approaches in terms of precision and recall, giving more complete and accurate 3D models.

1. Introduction

Decades of research into 3D reconstruction from stereo image pairs has resulted in general purpose algorithms which can successfully reconstruct many different scenes, including road scenes [13], buildings [12], and aerial images [20]. There are, however, environments in which dense stereo algorithms perform poorly; these include environments containing many similar looking objects, and environments with many occlusions and depth discontinuities, where the assumptions built into these algorithms are not appropriate. Examples of these environments include branching plants imaged by agricultural robots [2], swarming flies imaged for biological research [22], and fingertips imaged in hand tracking applications [9]. In situations like these, model-based schemes which use domain knowledge to match a sparse set of features between views are often more suitable.

Previous methods for matching features between views resolve ambiguous matches by either finding a maximum likelihood correspondence (set of matched features), by constraining the order in which matched features can occur in the images, or by detecting and discarding ambiguous matches. These strategies are not necessarily the most appropriate for complete and accurate 3D reconstruction, as they frequently result in incorrectly matched features, or fail to select features which could be matched. In this paper we propose a new algorithm for selecting a correspondence between features detected in the two views from a calibrated stereo camera. We treat the problem as a problem in decision theory: of all possible correspondences, which is the best to choose to give as complete and accurate a 3D reconstruction as possible? To solve this problem, we compute probabilities that feature matches are correct by Gibbs sampling, then define a loss function which quantifies the benefits of correctly matching features and the implications of incorrectly matching features. We then derive a decision rule for selecting the correspondence which minimises the expected loss. A parameter α controls the tradeoff between selecting incorrect matches versus missing correct matches.

The proposed scheme is ideal for the situation where features have multiple match candidates, and these matching ambiguities cannot be resolved by regularising depths or by imposing ordering constraints. It is appropriate when matches are ambiguous *despite* knowing the relative camera poses, unlike the situation where matches are ambiguous only because the camera pose is unknown, and the matches and camera pose can be estimated jointly by RANSAC [11].

We evaluate our proposed scheme on a challenging 3D reconstruction problem where branching vines are reconstructed by a pruning robot. The proposed system outperforms alternative approaches in terms of precision and recall, leading to more complete and accurate 3D models.

2. The sparse stereo correspondence problem

A calibrated stereo camera rig captures two images of a scene. A feature detector is run on each image, and detects a set of features R in the one image and S in the other. A pair of features $(r, s) \in R \times S$ is *correctly matched* if r and s each correspond to the same object in the scene. The aim of this research is to use attributes of the features, such as their appearance and position in the image, to find as many correct feature matches as possible, while avoiding incorrectly matched features. Matched features can then be reconstructed in 3D. We refer to a set of matched features $C \subset R \times S$ as a *correspondence* between the two images.

In general, one feature from R can be correctly matched to at most one feature from S, and vice versa. This constraint is known as the *one-to-one matching constraint*. We say a feature match (r, s) is *compatible* with a correspondence C if it could be added to C without violating this constraint.

2.1. Matching features between images

To find correspondences between two images, we use attributes of the features including their appearance and their position in the image. The appearance (i.e. colour and shape) of pairs of features are compared, and only features which appear similar should be matched. In some environments, an image feature may appear similar to many different features in the other image, but at most one of these matches will be correct. In this situation more information, such as position, helps to resolve which is the correct match.

The main constraint on the positions of matched features comes from the epipolar geometry of the two views [16]. The epipolar geometry constrains any correct match to a point feature in one image to lie on the epipolar line corresponding to that feature's location in the other image. The depth to the corresponding point in 3D is determined by the matching feature's position on the epipolar line, so the range of feasible depths for the object bounds the segment of the epipolar line on which the correctly matched feature must lie. In practice, errors in localising features in the image mean that features could be correctly matched to points anywhere within a few pixels of the corresponding epipolar line.

For image features which do not have point locations, such as linear features without well-defined endpoints, or objects with boundaries occluded by other objects, the epipolar constraint provides only a weak constraint on the position of matching features. For the example of images of branching plants, the endpoints of detected branches are often at either the image edge or at occlusion boundaries with other branches. Features such as these often have multiple possible match candidates.

3. Previous solutions to ambiguous stereo correspondence problems

Matching sets of features between views is a common problem in computer vision. This section reviews three different situations where correspondence problems are solved. The first methods considered are feature-based methods, where ambiguous matches are generally detected and discarded. The second class of methods are dense stereo algorithms, which use assumptions about the scene's 3D structure to resolve ambiguities. The third class of methods are application-specific 3D reconstruction schemes, which combine domain knowledge with a variety of heuristics for resolving ambiguities.

Firstly, feature based methods are widely used for reconstructing scenes where relative camera poses are initially unknown. A set of feature descriptors is extracted from each view, and similar-looking features are matched between views. Feature descriptors such as SIFT (Scale Invariant Feature Transform; [17]) assign a vector descriptor to a small region of an image; if the Euclidean distance between two SIFT descriptors is small then the regions appear similar. Typically each feature descriptor from one image is matched to its nearest neighbour in the other image [18], to provide a candidate match. From a set of candidate matches, a set of (mostly) correct matches, together with the epipolar geometry is estimated jointly, often using RANSAC (Random Sample Consensus; [11]).

A common heuristic used to identify ambiguous matches is to match feature descriptors only when the distance to the second nearest neighbour is greater than the distance to the nearest neighbour by some threshold [17, 18], however this eliminates many correct matches. For example, Dey et al. [8] used a feature-based approach to reconstruct a 3D model of vines, but the reconstructions of branches and wires are incomplete, because too few unambiguous matches are found. Extensions to RANSAC have been proposed to make use of multiple candidate matches to each feature [21, 4], but RANSAC still relies on initially-ambiguous matches being resolved once the epipolar geometry is known. By contrast, our proposed approach is designed for the situation where ambiguities exist despite knowing the relative camera pose.

One situation where matches are ambiguous despite knowing the camera geometry is line-segment based stereo matching: straight line segments are extracted from two images, and a correspondence between these sets of line segments is used to reconstruct the scene. Christmas et al. [7] find the maximum likelihood correspondence between line sections by probabilistic relaxation labelling, an algorithm which has been applied to a variety of matching problems in computer vision. The matching found is a maximum likelihood estimate given assumptions about distances between matched lines, which are equivalent to a constraint on the ordering of matches. Features are not matched to anything if this is the most likely outcome, however no one-to-one constraint on matches is enforced. The maximum likelihood matching results in some lines being incorrectly matched.

The second situation where a correspondence between two views is found is dense stereo matching. Dense stereo algorithms use a pair of images from cameras with known relative pose to find a disparity map mapping each pixel in one image to its matching pixel in the other. Some optimisation algorithm is used to estimate a disparity map minimising an objective function. The objective function measures both the similarity in appearance between matching pixels, and how well the disparity map matches assumptions about the structure of the world. Objective functions often constrain the depth map to be piecewise planar or piecewise continuous [5, 19]. Approaches using dynamic programming for the optimisation also impose an ordering constraint on matched pixels—along two epipolar lines, the order in which features appear must be the same [10, 15]. Other approaches implicitly prefer reconstructions where ordering is preserved between images by minimising objective functions which assign lower costs to smaller depth discontinuities [5, 19].

If the objective function is minimised exactly, then the depth map found is a maximum likelihood estimate, given the assumptions represented by the objective function, although in practice, optimisers which find approximate solutions are generally used. Modern dense stereo algorithms allow pixels to be unmatched if they are occluded (often near to depth discontinuities), however penalty terms ensure that as many pixels as possible are matched [15, 19].

In summary, dense stereo algorithms find a correspondence between views by using assumptions about the structure of the scene to resolve ambiguities. These algorithms perform well in many environments, but can perform poorly in environments where ordering and depth continuity assumptions are inappropriate. In these situations, model-based approaches which use knowledge of the scene to resolve ambiguities may be more appropriate. A model based approach is used to find a correspondence between fingertips for a hand-tracking application by Dorfmuller and Schmalstieg [9], who use the epipolar constraint together with a model of the dimensions of a hand to constrain matches. Chen et al. [6] match networks of arteries between views of the retina by first identifying a minimal set of junctions which can be matched unambiguously, then using these matches, together with the shape of the retina, to disambiguate remaining matches. Zou et al. [22] estimate the trajectories of a swarm of fruit flies by estimating the correspondence between detected flies which minimises a cost function. The cost function combines the epipolar error, differences in measured sizes, and velocity information from tracking. The optimal correspondence is selected using dynamic programming, subject to a one-to-one matching constraint. Detected flies are always matched unless there is no suitable match candidate.

Of these contemporary solutions to correspondence problems, dense stereo and model-based approaches tend to match as many features as possible, even when there is a high risk of incorrect matches. However, incorrect 3D reconstructions can cause significant problems, so it may be more appropriate to choose not to match features, rather than risk reconstructing objects incorrectly. By contrast, feature based approaches often discard large numbers of potentially ambiguous matches, at least until a subset of good correspondences is found. In this paper we propose a novel framework for selecting which features to match, which balances the risk of incorrectly matching features with the benefits of obtaining a more complete 3D reconstruction. Unlike a maximum likelihood approach, the new framework will leave features unmatched if the most probable match is likely to be incorrect, and unlike contemporary stereo algorithms, we avoid making assumptions about the 3D structure of the scene.

4. Proposed minimum expected loss matching framework

This section describes our proposed system for selecting a correspondence from a set of ambiguous feature matches. Our proposed system is designed to give as complete and accurate a 3D reconstruction as possible. The method depends on estimating the probability that each feature match is correct, so we first describe how feature match probabilities, conditional on other matches, are computed. Secondly, we describe how to estimate marginal probabilities for each feature match by Gibbs sampling. Thirdly, we derive a decision rule for selecting the minimum expected loss correspondence, based on these marginal probabilities.

4.1. Estimating feature match probabilities

This section describes how to estimate the probability of a pair of features (r, s) from $R \times S$ being correctly matched, given attributes of the features. Estimating candidate feature match probabilities directly is challenging, because different matches to a particular feature are mutually incompatible, and hence the probabilities are all interdependent. However, estimating the probability of a feature being correctly matched conditional on the state of every other feature match is considerably easier, as each match can be considered in isolation. We write $P((r, s))$ for the probability that r and s are correctly matched. Assuming a correspondence C consists of correctly matched features, we compute $P((r, s) \mid C_{-(r,s)})$, where $C_{-(r,s)}$ is the correspondence C excluding the match (r, s). If either of r or s are matched in $C_{-(r,s)}$ then (r, s) cannot also be correctly matched, so $P((r,s) \mid C_{-(r,s)}) = 0$. If neither r nor s are matched, then we compute the probability $P((r,s) \mid C_{-(r,s)})$ from a vector of attributes of the feature match, $\mathbf{x}_{(r,s)}$. $\mathbf{x}_{(r,s)}$ includes knowledge about the match including errors in the epipolar constraint, and differences in the appearance of the two features. For correctly matched features, $\mathbf{x}_{(r,s)}$ is sampled from a distribution with probability density function (PDF) $p_c(\mathbf{x}_{(r,s)})$. For example, if components of $\mathbf{x}_{(r,s)}$ represent differences in pixel colour values, and we use a normal distribution to model the difference in colour between correctly matched features, we would choose $p_c(\mathbf{x}_{(r,s)}) = \phi_{\mu,\Sigma}(\mathbf{x}_{(r,s)})$, where $\phi_{\mu,\Sigma}$ is the normal PDF with mean μ and covariance Σ, and μ and Σ can be estimated by fitting a model to training data. Similarly, we fit a model with PDF $p_{\bar{c}}(\mathbf{x}_{(r,s)})$ to the attributes of feature pairs which are not correctly matched. From these distributions, and given that the prior probability of r or s having any match is π, Bayes rule gives us the probability:

$$P((r,s) \mid C_{-(r,s)}) = \begin{cases} \frac{\pi p_c(\mathbf{x}_{(r,s)})}{\pi p_c(\mathbf{x}_{(r,s)}) + (1-\pi) p_{\bar{c}}(\mathbf{x}_{(r,s)})}, & \text{if } (r,s) \text{ compatible with } C_{-(r,s)} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The same model and formula are used by Christmas et al. [7] to compute labelling probabilities given feature attributes. In practice, most pairs $(r,s) \in R \times S$ can be considered infeasible because $p_c(\mathbf{x}_{(r,s)})$, and hence $P((r,s) \mid C_{-(r,s)})$, is negligibly small, for example because the match violates the epipolar constraint by a large margin. We are only interested in selecting matches from the remaining set of feasible matches F.

4.2. Estimating marginal probabilities by Gibbs sampling

Gibbs sampling [14] is an algorithm for sampling from a multivariate distribution where the distribution of each component, conditional on every other component, is known. This is precisely the case for feasible correspondences. We use Gibbs sampling to sample mutually compatible sets of feasible matches $C \subset F$, given the conditional probabilities of each match being correct.

To represent the set C as a random vector, we encode it as a vector $\mathbf{f}^C$ of length $|F|$ of Bernoulli (true/false) random variables. Each component f_i^C of $\mathbf{f}^C$ indicates whether the match $(r,s)_i$ is in C. Each element f_i^C has probability $P((r,s)_i \mid C_{-(r,s)_i})$ given by Equation 1. We then use Gibbs sampling to sample feasible correspondences. Each iteration of Gibbs sampling computes a sample C_t from the previous C_{t-1} by initialising $C_t = C_{t-1}$ and then sampling each $f_i^{C_t}$ from *Bernoulli*$\big(P((r,s)_i \mid C_{t-(r,s)_i})\big)$ in turn. We use the resulting sample to estimate the marginal probability of each feature match being correct. Asymptotically:

$$P((r,s)_i) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} f_i^{C_t}. \quad (2)$$

After a fixed number of iterations T, we use the estimate:

$$P((r,s)_i) \approx \frac{1}{T} \sum_{t=1}^{T} f_i^{C_t}. \quad (3)$$

The Gibbs sampler cannot sample incompatible correspondences, e.g. containing both (r,s_1) and (r,s_2), so each sample contains at most one match to each feature, and hence the total probability of a match to each feature is at most 1. For each r, $1 - \sum_{s\in S} P((r,s))$ is the probability that r has no match in S.

4.3. Selecting the minimum expected loss correspondence

We formulate the problem of choosing the best correspondence C_Δ as a problem in decision theory [1]. We define a loss function $L(A,C)$, which represents the cost of choosing the correspondence C if the true correspondence is the set A. The choice of loss function is application-dependent; for 3D reconstruction we should choose a cost function which penalises incorrect matches, as these lead to incorrect 3D models, but which also penalises matches which are missing from C. Our proposed loss function is:

$$L(A,C) = |C \setminus A| + \alpha |A \setminus C|, \quad (4)$$

where '$\setminus$' is the 'set minus' operator, $|C \setminus A|$ is the number of incorrect matches in C, $|A \setminus C|$ is the number of correct feature matches which were not selected, and α is a penalty for not identifying matches, relative to the cost of incorrectly matching a feature. Normally $0 < \alpha \leq 1$, indicating that missing matches are less of a problem than incorrect matches. We then select the correspondence set C_Δ which minimises the expected loss:

$$C_\Delta = \operatorname*{arg\,min}_{C \subset F} \mathrm{E}(L(A,C)). \quad (5)$$

The expected loss from choosing correspondence C is defined in terms of the probabilities $P(A)$ of each possible correspondence A being correct:

$$\mathrm{E}(L(A,C)) = \sum_{A \subset F} P(A) L(A,C). \quad (6)$$

For our proposed loss function, losses from different feasible matches are independent, so the loss function can be decomposed into a sum of losses due to different feasible matches:

$$L(A,C) = \sum_{(r,s)\in F} \Big(I\big((r,s) \in C \setminus A\big) + \alpha I\big((r,s) \in A \setminus C\big) \Big), \quad (7)$$

where I is the indicator function. Substituting Equation 7 in Equation 6 gives

$$\begin{aligned}\mathrm{E}(L(A,C)) &= \sum_{A\subset F}\Big[P(A)\sum_{(r,s)\in F}\Big(I\big((r,s)\in C\setminus A\big) \\ &\qquad + \alpha I\big((r,s)\in A\setminus C\big)\Big)\Big] \quad (8)\\ &= \sum_{(r,s)\in F}\sum_{A\subset F}P(A)\Big(I\big((r,s)\in C\setminus A\big) \\ &\qquad + \alpha I\big((r,s)\in A\setminus C\big)\Big). \quad (9)\end{aligned}$$

Each marginal probability $P((r,s))$ is the probability that A contains (r,s), so:

$$\sum_{A\subset F}P(A)I\big((r,s)\in A\big) = P((r,s)). \qquad (10)$$

By substituting these marginal probabilities in Equation 9, we can write the expected loss as a sum of the losses due to each feasible match:

$$\begin{aligned}\mathrm{E}(L(A,C)) &= \sum_{(r,s)\in F}\Big((1-P((r,s)))I\big((r,s)\in C\big) \\ &\qquad + \alpha P((r,s))I\big((r,s)\notin C\big)\Big). \quad (11)\end{aligned}$$

Including each (r,s) in C results in an expected loss of $1-P((r,s))$, whereas not including (r,s) in C results in an expected loss of $\alpha P((r,s))$. The loss for each match is minimised when we select each (r,s) where $P((r,s)) > 1/(1+\alpha)$, therefore the total expected loss is minimised when we select:

$$C_\Delta = \{(r,s)\in F : P((r,s)) > 1/(1+\alpha)\}. \qquad (12)$$

If $\alpha \leq 1$, this decision strategy will only select matches (r,s) with $P((r,s)) > 1/(1+\alpha) \geq 1/2$. As the total probability of matches to either r or s cannot exceed 1, this guarantees that at most one match to each feature will be selected. This ensures that C_Δ satisfies the one-to-one matching constraint, and also that if a match is selected, it is the most probable match to both r and s.

Note that in general we cannot compute C_Δ by evaluating the loss for every C in our Gibbs sample using Equation 6, because C_Δ may be unlikely to ever actually occur—C_Δ is selected by a decision strategy designed to give a 3D reconstruction with a low error rate, it is not an estimate of the actual correct correspondence.

The proposed decision strategy can also be used to select correspondences when $\alpha > 1$, i.e. when the cost of missing a match is higher than the cost of incorrectly matching a feature. In this situation multiple matches to each feature may be found. Applications where this is the case include robot path planning, where failure to reconstruct an object correctly could result in a collision. In this case it may be appropriate to select multiple candidate matches to each feature.

The Gibbs sampler is also used to estimate the maximum likelihood (ML) correspondence—this is simply the correspondence which is sampled most often. The ML correspondence is useful for evaluating our proposed strategy, however finding it by Gibbs sampling can be slow for large feature sets, as a large number of samples are needed to find one which clearly occurs more often.

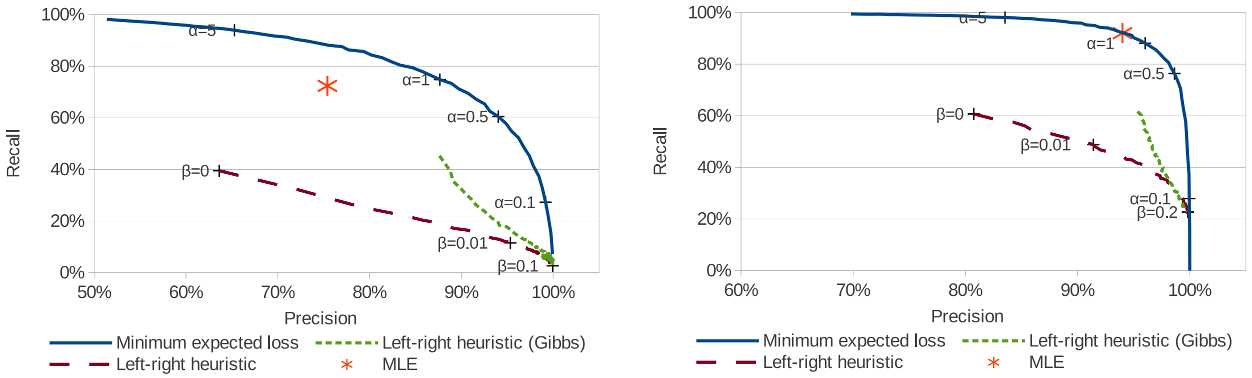
In contrast to the ML correspondence, the proposed method will often choose not to match a feature which has multiple feasible match candidates, even when the event that the feature has no match is unlikely. In contrast to heuristics based on comparing the most likely match to the second most likely, the proposed method selects matches when the absolute risk of each match being incorrect is low, rather than only considering only the risk of the second most likely candidate being the true match.

5. Experimental results

We first evaluate the proposed method using simulated data. We simulate matches between two images, with 10 or 100 features in each image, each of which could be matched to any feature in the other image. Attributes are simulated for each feature match, with correct matches having attributes sampled from $\mathcal{N}(0,1^2)$ (so $p_c = \phi_{0,1^2}$), and incorrect matches having attributes sampled from $\mathcal{N}(0,20^2)$ or $\mathcal{N}(0,100^2)$ respectively (so $p_{\bar{c}} = \phi_{0,20^2}$ or $p_{\bar{c}} = \phi_{0,100^2}$). For real data, these attributes might be the epipolar error in pixels. Equation 1, with prior probability $\pi = 0.5$ then gives the conditional probability of each match being correct.

When 10 features per image are simulated, each has an average of 2.2 feasible match candidates; and for each feature there is an incorrect match with higher conditional probability than the correct match with probability 0.19. This is a relatively low level of ambiguity, and a high level of accuracy can be obtained. When 100 features are simulated, each has an average of 4.3 feasible match candidates; and for each feature there is an incorrect matches with higher conditional probability than the correct match with probability 0.36, providing a greater level of ambiguity and making accurate matching challenging. All results given are averages over at least 500 runs.

We compared the proposed approach with the maximum likelihood correspondence (from Gibbs sampling), and with a heuristic which selects the most probable match to each feature, except when there is another candidate match to either with probability within a threshold β of the most probable match. This heuristic uses either the probabilities computed from Equation 1 (labelled 'Left-Right heuristic'), or marginal probabilities from Gibbs sampling (labelled



(a) 100 features per image with an average 4.3 feasible matches per feature. (b) 10 features per image with an average 2.2 feasible matches per feature.

Figure 1. Precision-recall for different correspondence selection criteria.

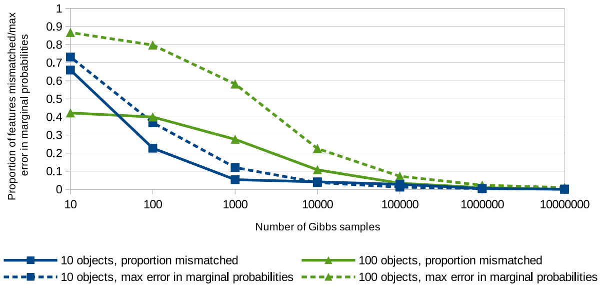


Figure 2. Mixing times for Gibbs sampler: as the number of samples (T) increases, the maximum absolute error in marginal probabilities falls, and the probability that a feature's match is selected or not selected incorrectly (according to Equation 12) falls.

'Left-Right heuristic (Gibbs)'). The 'Left-Right heuristic' is based on common heuristics used for discarding ambiguous matches between features [2, 17, 18].

We evaluated each of these approaches with a range of values for α and β; results are shown in Figure 1. These results show that by varying α, the tradeoff between precision and recall of the proposed approach can be adjusted. α can be chosen so that the proposed approach outperforms any other method on precision or recall.

Next, we test how many samples are needed before the Gibbs sampler converges to a useful level of accuracy. Figure 2 shows the mean proportion of incorrect decisions (the proportion of matches which were incorrectly selected or incorrectly omitted, according to Equation 12) when the sampling is terminated after different values of T. The experiment with 100 features has more interdependencies between matches, and is slower to converge than the experiment with 10 features. With 10000 samples and the 100-feature simulation, the sampler makes an incorrect decision 11% of the time, whereas for the 10 feature simulation, only 4% of decisions are incorrect. These errors have a surprisingly small effect on the precision or recall of the proposed method: with $\alpha = 0.5$ and 100 features, using 100000 samples rather than 5000 samples increases precision by just 0.6%, with the same level of recall. The explanation for the good performance given the significant numbers of incorrectly selected matches is that these matches are close to the decision boundary, so lead to only a small increase in the expected loss when selected incorrectly. All other experiments are carried out with 5000 samples, which enables a correspondence to be selected in 19ms for the 100-feature simulation, and 1ms for the 10-feature simulation (using C++ code compiled with clang and running on a single core of an Intel i7 2.93GHz processor). These results indicates that the proposed approach is suitable for real-time matching applications.

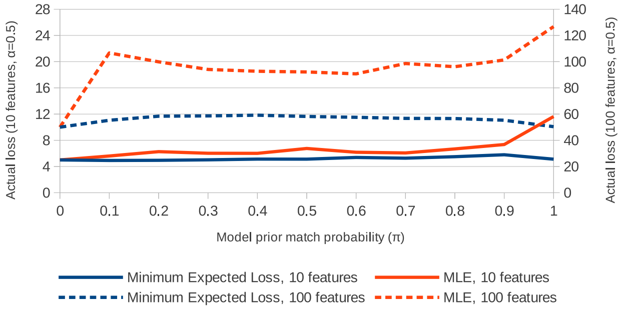


Figure 3. Effects of different choices for the prior inlier probability π. Features have a correct match with probability 0.5.

One parameter for any system computing match probabilities is the prior probability, π, of a feature having any correct match. Figure 3 shows that π does not strongly affect the accuracy of minimum expected loss or MLE selection criteria for selecting a correspondence set.

5.1. Application to model-based 3D reconstruction

Our intended application is the 3D reconstruction of vines for a vine pruning robot. A 3D reconstruction which is

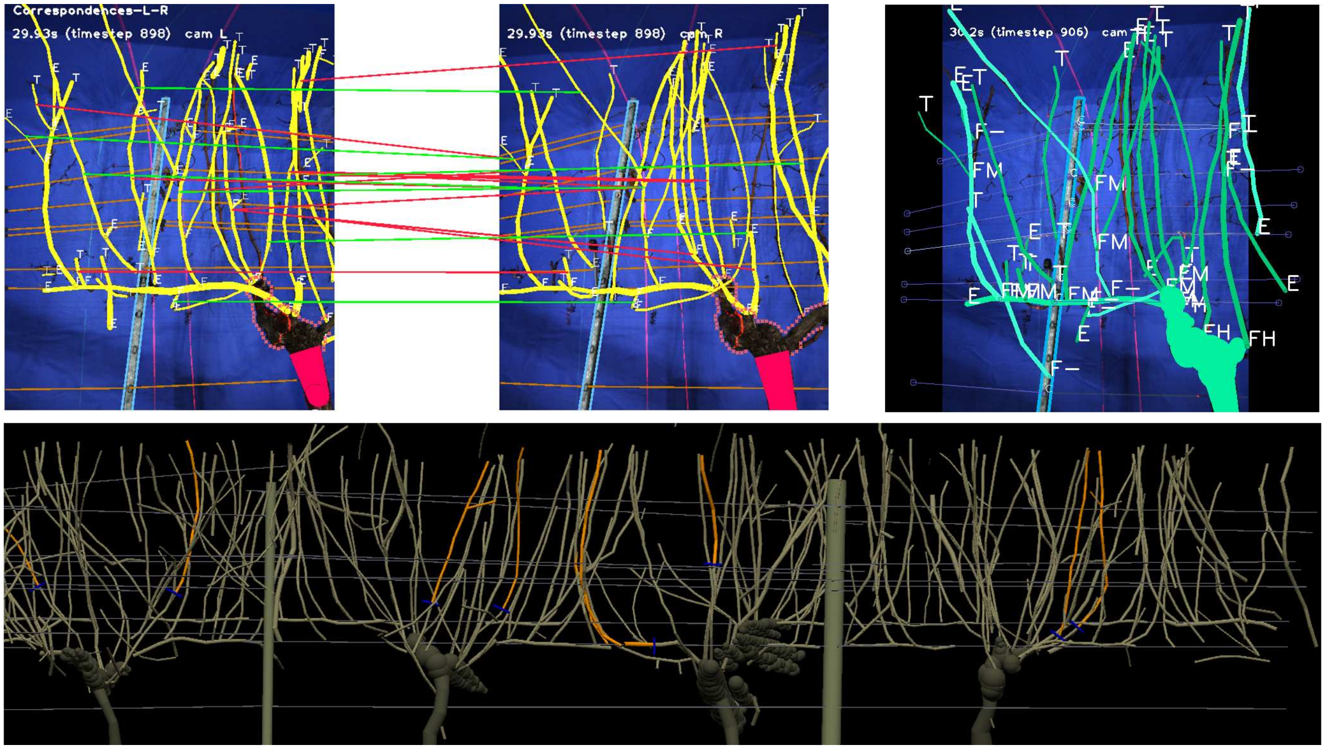


Figure 4. 3D vine reconstruction pipeline. Top-left: vines (yellow) and wires (orange) are detected in each frame. Candidate matches between vines are shown, with matches selected by the proposed minimum expected loss corresponder in green. Top-right: 3D model of vines and trellis backprojected onto a video frame. Bottom: 3D model built up incrementally over many frames, with cutpoints selected (cut vines highlighted in orange).

as complete and structurally correct as possible is required, so that a robot arm can prune the vines. The trellis over which the vines grow is also reconstructed so that wires are not accidentally cut. The robot images the vines with colour cameras, and vines, wires and posts are detected in each 2D image by applying standard computer vision methods (Figure 4; [2, 3]). To build a 3D model, a correspondence between pairs of 2D vines and a correspondence between pairs of 2D wires is found. The system attempts to reconstruct every pair of vines or wires, then the conditional likelihood of each match being correct is computed from measurements including thicknesses, curvature, and (where available) the reprojection error.

We tested the proposed minimum expected loss correspondence method on 12 sets of stereo images of 4 different plants with ground-truth. In these images the positions of 1127 2D vines and 486 wires, and all correspondences between them, have been hand-labelled. Figure 5 shows the precision and recall of each of the correspondence methods.

The 2D vines are challenging to match because occlusions lead to many 2D detections being incomplete or missing (the 2D vine detector has a precision of 73% at a recall of 75%, but most are incompletely-detected). 423 pairs of 2D vines can be reconstructed; of these 226 are correctly matched. Figure 5 shows that the proposed minimum expected loss corresponder gives a small improvement in precision over the MLE corresponder (77% versus 74% with $\alpha = 1$) at the same level of recall. Performance is similar to the heuristics for removing ambiguous matches.

The 2D wires are easier to detect than the vines (with a precision of 87% at a recall of 93%, with most detections being complete), but are challenging to match because they all appear similar, and because the epipolar constraint can eliminate few of the possible matches. Each matched wire part has feasible matches to an average of 2.1 wires in the other image, and just 99 of the 283 feasible matches found are correct. Figure 5 shows that the proposed minimum expected loss corresponder gives a substantial improvement in precision over the MLE corresponder—64% versus 56% at same level of recall. The biggest advantage of the proposed corresponder however is that α can be chosen to give a much higher level of precision, e.g. 83% at 44% recall.

We integrated the minimum expected loss corresponder into the 3D vine reconstruction pipeline. As the pruning machine moves, detected 2D vines are either assigned to existing 3D vine models, or are corresponded to generate new 3D models. The 3D vines are connected into a complete model and are optimised in an incremental bundle ad-

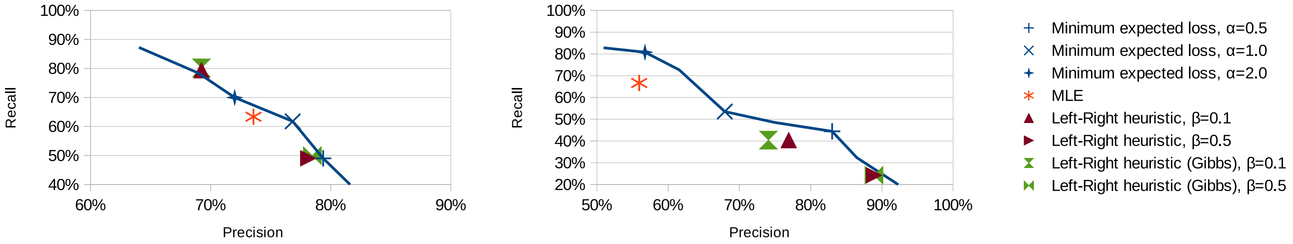


Figure 5. Precision-recall for the correspondence problem involving vines (left) and wires (right).

justment framework. Figure 4 shows 2D detected features, feature correspondences, and a 3D vine model. Finding correspondences takes less than 4ms of the 500ms required per frame; this framerate is sufficient for a speed of 0.1m/s.

6. Discussion

This paper described a novel system for selecting a correspondence (set of feature matches) between features detected in two stereo images. The method is designed to select the correspondence which gives as complete and accurate a 3D reconstruction as possible in the presence of ambiguous matches between features. The method uses a decision rule to select the minimum expected loss correspondence, where the proposed loss function quantifies the risk of selecting matches incorrectly, and of missing correct matches. A parameter of the loss function controls the tradeoff between precision and recall. The proposed approach outperforms alternative selection criteria in terms of precision and recall on real and simulated stereo matching problems. For a model-based sparse 3D reconstruction problem, the 3D model is more accurate and complete than with other correspondence selection criteria. The proposed method can be applied to other sparse stereo matching problems where matches are ambiguous.

References

[1] J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Verlag, 1985. 4

[2] T. Botterill, R. Green, and S. Mills. Reconstructing partially visible models using stereo vision, structured light, and the g2o framework. In *Proc. IVCNZ*, 2012. 1, 6, 7

[3] T. Botterill, R. Green, and S. Mills. Finding a vine's structure by bottom-up parsing of cane edges. In *Proc. IVCNZ*, pages 1–6, 2013. 7

[4] T. Botterill, S. Mills, and R. Green. New conditional sampling strategies for speeded-up RANSAC. In *Proc. British Machine Vision Conference*, 2009. 2

[5] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *Trans. PAMI*, 23(11):1222–1239, 2001. 3

[6] L. Chen, Y. Xiang, Y. Chen, and X. Zhang. Retinal image registration using bifurcation structures. In *International Conference on Image Processing*, pages 2169–2172, 2011. 3

[7] W. J. Christmas, J. Kittler, and M. Petrou. Structural matching in computer vision using probabilistic relaxation. *Trans. PAMI*, 17(8):749–764, 1995. 2, 4

[8] D. Dey, L. Mummert, and R. Sukthankar. Classification of plant structures from uncalibrated image sequences. In *Workshop on the Applications of Computer Vision*, 2012. 2

[9] K. Dorfmuller-Ulhaas and D. Schmalstieg. Finger tracking for interaction in augmented environments. In *Proc. Int. Symp. Augmented Reality*, pages 55–64, 2001. 1, 3

[10] G. Fielding and M. Kam. Weighted matchings for dense stereo correspondence. *Pattern Recognition*, 33(9):1511–1524, 2000. 3

[11] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1, 2

[12] D. Gallup, J.-M. Frahm, and M. Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *CVPR*, pages 1418–1425, 2010. 1

[13] S. K. Gehrig, F. Eberli, and T. Meyer. A real-time low-power stereo vision engine using semi-global matching. In *Computer Vision Systems*, pages 134–143. 2009. 1

[14] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Trans. PAMI*, (6):721–741, 1984. 4

[15] G. Gimel'farb. Probabilistic regularisation and symmetry in binocular dynamic programming stereo. *Pattern Recognition Letters*, 23(4):431–442, 2002. 3

[16] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. CUP, second edition, 2003. 2

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2, 6

[18] R. Szeliski. *Computer vision: algorithms and applications*. Springer, 2010. 2, 6

[19] O. Woodford, P. Torr, I. Reid, and A. Fitzgibbon. Global stereo reconstruction under second-order smoothness priors. *Trans. PAMI*, 31(12):2115–2128, 2009. 3

[20] L. Zebedin, A. Klaus, B. Gruber-Geymayer, and K. Karner. Towards 3D map generation from digital aerial images. *J. Photogrammetry and Remote Sensing*, 60(6):413, 2006. 1

[21] W. Zhang and J. Kosecka. Generalized RANSAC framework for relaxed correspondence problems. In *Int. Symp. 3D Data Processing, Visualization, and Transmission*, 2006. 2

[22] D. Zou, Q. Zhao, H. S. Wu, and Y. Q. Chen. Reconstructing 3D motion trajectories of particle swarms by global correspondence selection. In *ICCV*, pages 1578–85, 2009. 1, 3