

# Learning the Probability of Correspondences Without Ground Truth

Qingxiong Yang R. Matt Steele David Nistér Christopher Jaynes  
Center for Visualization and Virtual Environments  
Department of Computer Science  
University of Kentucky, Lexington, USA

## Abstract

We present a quality assessment procedure for correspondence estimation based on geometric coherence rather than ground truth. The procedure can be used for performance evaluation of correspondence extraction schemes developed by researchers, as well as for online learning and adaptation aimed at better system performance.

A very important aspect of the proposed procedure is that it considers uncertainty in the correspondence extraction, and encourages the evaluated methods to deal correctly with uncertainty.

Other important strengths of the procedure are that it does not use any manual work, and that it does not put any strong constraints on the scene, but rather relies on geometric coherence in the motion. Thanks to these strengths, it can therefore be used with large amounts of real, potentially application specific data, or even data acquired during system operation.

In the evaluation the correspondence extractor is handled as a black box producing a probability distribution for the local motion vector between a pair of image patches. The procedure is therefore quite general. We are making the evaluation procedure available for public use.

## 1. Introduction

Estimating correspondences is a fundamental task in computer vision. It is a natural part of tasks as diverse as structure from motion, tracking, recognition, and multi-view calibration. It is unsurprising, then, that correspondence estimation has received significant attention and a number of diverse approaches have been proposed. As the number and theoretical depth of these algorithms continues to grow, empirical analysis is important and in response to this need a healthy trend towards thorough performance evaluation has emerged [10, 5, 4, 6, 1, 3, 11, 12].

In this work, we introduce a quality assessment procedure for correspondence estimation that does not rely on ground truth or scene content and incorporates assessment of uncertainty produced by the low-level correspondence extractor. The method was developed as a tool for char-

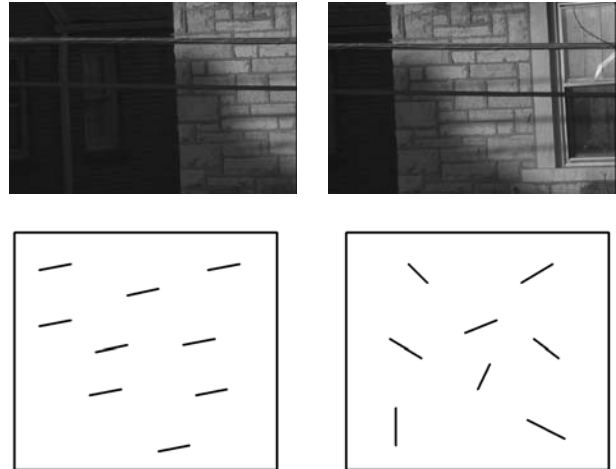


Figure 1: We present a quality assessment procedure for correspondence estimation based on geometric coherence rather than ground truth. A very important aspect of the proposed procedure is that it considers uncertainty in the correspondence extraction, and encourages the evaluated methods to deal correctly with uncertainty. Lower left: A coherent, and therefore possible set of correspondences. Lower right: An incoherent, and therefore incorrect set.

acterizing the relative performance of correspondence algorithms on realistic data sets. However, the procedure also suggests an online learning approach to correspondence extraction that can adapt to a variety of imaging conditions.

The approach is inspired by similar techniques that utilize geometry. For example, Baker et. al. evaluate edge detection using carefully designed scenes that contain parallel lines and gauge performance as complex averages of the variances of projective quantities [1]. In contrast to this approach, we leverage constraints on camera motion rather than scene constraints.

Other techniques have explored the advantages of avoiding constraints on scene content by utilizing motion models in the evaluation process. For example Schmid et. al. use

a known homography to describe inter-frame motions and measure feature detection repeatability [6] as well as feature descriptor invariance [4] using this model. The work here differs from these approaches in three important respects. Firstly, we do not rely on manual supervision to acquire the motion constraints. This is accomplished by marginalizing over all possible motions. Secondly, we achieve a higher level of generality by treating the whole correspondence extraction process as a black box, not just feature detection or region description. Hence, we are not restricted to feature-based methods. Finally, uncertainty is directly incorporated into the evaluation framework by treating the output of the low-level correspondence extractor as a local probability distribution. This is in accordance with a widespread understanding that low-level tasks such as correspondence extraction or segmentation should not be performed in isolation [8, 7]. It is hard or perhaps even impossible to take final decisions such as thresholding or discrete labeling locally without integration into the appropriate high-level task. Essentially, there is not enough information locally, and by thresholding prematurely, the local information is impaired before it reaches the high-level task. The remedy is to propagate a probability distribution from the low-level process. This has the advantage of not destructing information, while maintaining some degree of decoupling between the low-level and the high-level process, which is attractive from an engineering standpoint. Our approach is thus partly motivated by the trend that low-level vision algorithms must be capable of producing a reasonable measure of uncertainty of their own estimates.

The rest of the paper is organized as follows. Section 2 presents some motivation. In Section 3 we outline our approach. The mathematical theory behind our approach is presented in Section 4. Section 5 presents experimental results and validation of the framework. Section 6 concludes.

## 2. Motivation

By measuring the coherency of correspondence extraction, an algorithm’s performance is measured in terms of its ability to correctly generate correspondences that globally agree on an underlying motion model. Although this approach constrains the motion between frames to some known model, it is not directly related to assumptions about scene content that some previous methods required. This has the advantage that evaluation metrics are decoupled from scene content and do not bias the performance measures towards methods that are particularly good in the presence of certain scenes. Perhaps more importantly, any video data that conforms to some known motion model can be utilized, encouraging the use of data sets that are not amenable to hand labeling of features but are perhaps more representative of real-world situations.

In conceiving our approach, we had alignment of surveillance video in mind, but the methodology is generally applicable. The typical state of the art alignment algorithm will in some fashion maximize the image similarity after alignment subject to some motion model, where the measure of similarity varies between algorithms. The measure is most often some type of accumulation of local similarities across the whole image. In this case, the task can be equivalently reformulated as maximizing a joint likelihood that is a product of the local likelihoods arising from observing local pairs of patches in the two images that are to be aligned. The local likelihood is parameterized by the shift between the images for that pair of patches.

Even in low texture areas of imagery, there is sufficient information to perform correspondence estimation. However, in order to utilize this information effectively, uncertainty has to be correctly handled, and robustness is necessary in the estimation. This framework directly addresses uncertainty by representing displacement hypotheses, generated by the similarity measure, as a probability distribution in the image. The mutual consistency of these likelihoods is measured on the basis of geometric coherence.

It is standard to constrain the correspondences using the motion model, which leads to more correct correspondences between the current pair of video frames. However, what we are aiming for here is to use the motion model as a measure of the quality of a particular registration algorithm.

It should be noted that this approach is a first step towards online improvement of the correspondence quality between *future* video frames, where the motion model may no longer apply. This is accomplished by using the motion model to select the estimator of the local likelihoods that is most appropriate for the data. If we can accomplish this, the performance of the overall system will also be improved. This approach is in tune with a growing interest in applying machine learning techniques to real image data [2, 9]. Note however that the approach is quite different from what is normally called unsupervised learning in pattern recognition terms, which usually intends clustering of data.

Aside from the method’s potential for learning correspondence extraction, two immediate applications of our approach are:

- A uniform performance comparison of existing correspondence extraction schemes.
- Performance analysis as an online tool to aid in the development of new correspondence extraction schemes.

## 3. Approach

In this section we give an intuitive outline of our approach. The mathematical foundation will be given in the following section.

The underlying idea of our approach is that if we feed a correspondence extractor pairs of patches of local image data, without contextual information such as which image or where in the image the patches came from, the only coherent information that the correspondence extractor can deliver is information which is common between several patch pairs. If there is a global motion model, the motion vectors constitute common information. Other global effects such as overall lighting level, color distribution, contrast or frequency spectrum can also constitute common information. Such cues, or even intrinsic bias in the correspondence extractor, can lead the algorithm to exhibit false coherence, agreeing consistently with a global motion that is nevertheless far from the true motion.

However, we can distinguish between the motion vectors and other common information by exploiting that the motion vectors constitute shift-covariant information. If we synthetically shift one of the patches from a pair of local patches, the motion vector extracted from that pair should ideally also shift by the same amount, and this holds for all small shifts. If we use small known random shifts on all of the patches from an image pair that follow an unknown global motion, we know that the correspondence vectors minus the shifts should be a coherent motion. Combining the global real motion with synthetic random shifts makes it virtually impossible for the correspondence extractor to produce such a signal without actually performing successful correspondence extraction. Hence we obtain a basis for performance evaluation.

## 4. Theory

We assume that there is some unknown data model, embodied by the probability  $p(I, J|x)$  of observing the patch patterns  $I, J$  given the correspondence vector  $x$  for that pair of patches. This conditional probability is local and can not take global image information or image location into account, but it can cater to a higher level process that does.

We also assume that there is some global geometric relation  $g$ , such as a homography or epipolar geometry, that restricts the possible correspondences. For simplicity, we assume that this relation is completely embedded in the probability  $p(X)$  of the correspondences, where  $X$  denotes the joint correspondence set. That is, certain sets of correspondences that do not jointly obey the geometric relation cannot occur. Note that this means that we assume that  $p(X)$  takes into account probabilities of correspondences including parallax for the case of the fundamental matrix. Note also that the correspondences  $X$  are ideal and that image noise is taken into account by  $p(I, J|x)$ . Let  $i$  index over all the image patches we wish to consider, possibly taken from multiple images. Assuming conditional independence between the generative models for the patches given the cor-

respondences, Bayes' rule yields

$$p(X|\mathbf{I}, \mathbf{J}) \sim \prod_i p(I_i, J_i|x_i)p(X). \quad (1)$$

The standard correspondence estimation task is to compute the posterior distribution on the left hand side given all the image data  $\mathbf{I}, \mathbf{J}$  and the models  $p(I, J|x)$  and  $p(X)$ . However, what we wish to do in this paper is *estimate*  $p(I, J|x)$ , not assume that it is known. To formalize that it is unknown we define the unknown function  $f(I, J, x) = p(I, J|x)$  and assign it some prior distribution  $p(f)$ . To make this more explicit, we extend Equation (1) to

$$p(f, X|\mathbf{I}, \mathbf{J}) \sim \prod_i f(I_i, J_i, x_i)p(f)p(X), \quad (2)$$

where the independence  $p(f, X) = p(f)p(X)$  is assumed. Marginalizing over the set of possible correspondences leads to

$$p(f|\mathbf{I}, \mathbf{J}) \sim p(f) \int_X \prod_i f(I_i, J_i, x_i)p(X)dX, \quad (3)$$

which is our most important equation. It is the posterior distribution for the local generative model  $f$  given the image data. We consider local correspondence extraction schemes to be proposals for  $f$ , and Equation (3) is our basis upon which to prefer one proposal over the other. We can also try to find the maximum a posteriori estimate of  $f$  over a continuous set of proposals. Once  $f$  is estimated or selected, it can be used in Equation (1) to predict correspondences in new images.

A local correspondence extraction scheme should ideally return a representation of the distribution

$$p(x_i|I_i, J_i) \sim f(I_i, J_i, x_i)p(x_i). \quad (4)$$

Hence a correspondence extraction scheme's proposal for  $f$  would be

$$f(I_i, J_i, x_i) \sim \frac{p(x_i|I_i, J_i)}{p(x_i)}. \quad (5)$$

Note that  $p(x_i)$  is just the prior for the correspondence induced by the patch motion, i.e. the prior belief on motions, before the patch has been observed. If it is explicit for a correspondence extraction scheme, the right hand side can readily be computed. If not (which we consider a drawback of the correspondence scheme), one can assume a uniform prior in order to interpret the scheme as a proposal for  $f$ .

To fairly compare different correspondence extraction schemes, the prior probability  $p(f)$  should employ as few ad-hoc assumptions as possible. The prior could potentially enforce properties of  $f$  that can be universally agreed upon, such as for example shift-covariance, i.e.

$$f(I, J, x) = f(I, J + t, x + t), \quad (6)$$

where  $J + t$  denotes the patch shifted by  $t$ . However, we set the prior to uniform in order to completely avoid bias towards any method. This also avoids any ad-hoc weights on a penalty for not obeying shift-covariance. It is also possible to require rotation invariance, but it could be argued that the evaluation scheme should not choose or impose a certain way of handling resampling effects connected with rotating a discretized patch, which is required for rotation, while integer pixel shifts require no such choices.

## 5. Experiments

In this section we present experimental results. We use two ways to validate our performance assessment method. The first is to show that the score deteriorates as expected when we deteriorate a correspondence extraction algorithm on purpose by introducing noise to the image patches before the algorithm is run. The second is to show that the results correlate appropriately with evaluation using ground truth.

In the former case we go backwards from a known algorithm in a way that should be intuitively clear will deteriorate it, rather than trying to go forward from a known algorithm and show that the score improves. Had we done the latter, it would not have been clear whether the algorithm actually improved and the score correctly showed it, or the score improved but the algorithm actually did not.

For clarity, we use simple and widely known correspondence extraction measures. Our goal in this paper is to validate the performance measure, not to use the performance measure to find algorithms better than what is currently known.

The measures we use for patch similarity are

- Sum of Squared Differences (SSD):

$$SSD(I, J, t) = \sum_i (I(x_i) - J(x_i + t))^2 \quad (7)$$

- Sum of Absolute Differences (SAD):

$$SAD(I, J, t) = \sum_i |I(x_i) - J(x_i + t)| \quad (8)$$

- Normalized Correlation (NC):

$$NC(I, J, t) = \sum_i I(x_i)J(x_i + t), \quad (9)$$

where it is assumed that the overlapping portions of  $I$  and  $J$  have already been normalized to sum to zero and square-sum to unity, i.e.  $\sum_i I(x_i) = 0$ ,  $\sum_i I(x_i)^2 = 1$  and likewise for  $J$ .

Note that there are some intricacies here involved with the non-overlapping regions arising when shifting one of

the patches in a pair of patches. We have tried two different ways of handling this, with very similar results. The first is to compute the similarity measure over just the overlapping area and normalize the SSD and SAD according to the overlapping area (normalized correlation is unaffected). The second is to use a smaller patch in the second image so that the larger window from the first image can fit the true motion plus the random shift and still cover the small window.

We use indoor and outdoor video with simple approximately translational motion. This type of motion typically occurs in surveillance video when the camera is highly zoomed in and is panning. The two-dimensional motion space makes it an easy task to evaluate the integral in Equation (3) explicitly as the sum

$$p(f|\mathbf{I}, \mathbf{J}) \sim \sum_t \prod_i f(I_i, J_i, t), \quad (10)$$

where we have set the motion model  $X$  to a translational shift  $t$ . We have also set the generative model prior  $p(f)$  and the motion prior  $p(t)$  to uniform (within a disparity limit). Although the motion is simple, it does get us realistic noise, both temporal sensor noise, and variability between different pixel responses in the sensor. It can also cover effects such as motion blur.

A scheme that only returns a single correspondence hypothesis rather than a probability distribution could be considered as a function

$$\delta(I, J)(t) \quad (11)$$

taking a pair of image patches  $I, J$  and returning a delta-function in the correspondence vector  $t$  as its posterior. Used blindly, such a scheme would not perform well, since it is not 'hedging its bets' appropriately. The delta-function would indicate that a mismatch is impossible. However, mismatches are bound to occur and will then hurt the score immensely, since the scheme is incorrectly claiming that a mismatch is impossible. Thus appropriate representation of uncertainty is enforced by the raw score.

We handle this by searching a space of possible ways to hedge the bets of any scheme that returns matches without uncertainty information. More precisely we use a contaminated Gaussian

$$f(t) = \frac{\alpha}{N} + (1 - \alpha)k e^{-\frac{(t-c)^2}{2\sigma^2}}, \quad (12)$$

where  $N$  is the number of pixels in the pdf-representation,  $\alpha$  is a contamination fraction and  $\sigma$  is a standard deviation given in pixels of the pdf-representation. The contaminated Gaussian is centered on the correspondence  $c$  given by the correspondence extractor. The constant  $k$  is set by requiring that the clean shifted Gaussian  $k e^{-\frac{(t-c)^2}{2\sigma^2}}$  sums to one in the pdf-representation.

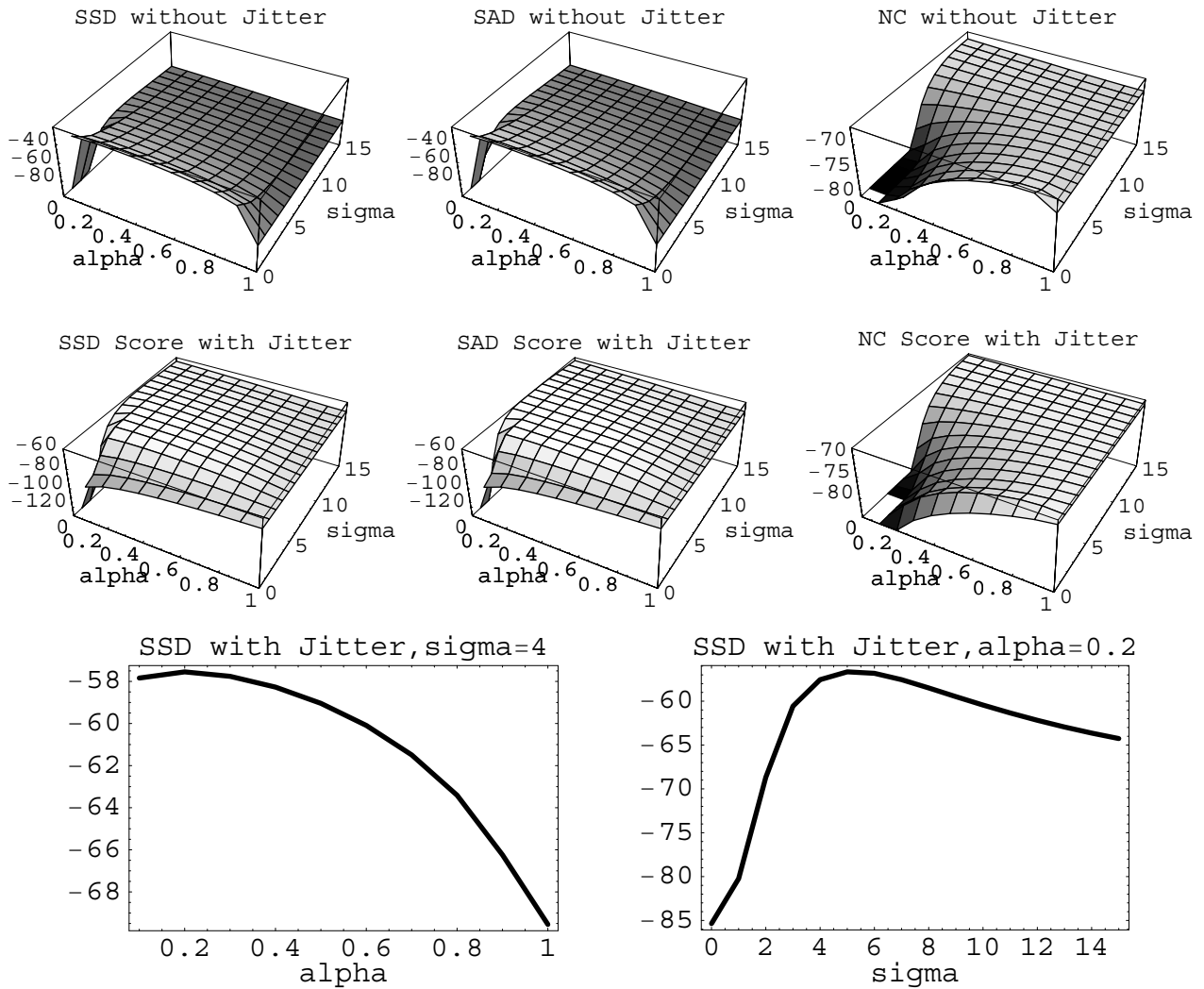


Figure 2: An example on indoor video. The second row shows the matrix of coherence scores with varying contamination  $\alpha$  and standard deviation  $\sigma$  for SSD, SAD and NC, respectively. The second row uses performance scores without enforcing translation covariance. The third row shows the same scores when translation covariance is enforced by introducing random jitter to the image patches. Note that appropriate choices of  $\alpha$  and  $\sigma$  lead to optimal scores, illustrating how the score encourages faithful uncertainty representation. Note that NC is not significantly affected by the jitter. However, SSD and SAD exhibit some false coherence that is suppressed when enforcing translation covariance. The bottom row shows the  $\sigma = 4$  and  $\alpha = 0.2$  slices through the surface plots for SSD with jitter, respectively.

The contaminated Gaussian is motivated by the relatively popular way of first extracting correspondences and then minimizing a robust cost function on the correspondences. We search over a whole matrix of settings for  $\alpha$  and  $\sigma$  and compute the scores for each setting. The plots are shown in Figure 2. Throughout this section, we use the log-likelihood of the score from Equation (10).

Note that the choices for  $\alpha$  and  $\sigma$  significantly impact the resulting overall coherence score. As expected there is a choice of  $\alpha$  and  $\sigma$  that results in an optimal score. We will use this optimal score later as the overall measure of a method’s performance.

The example illustrated in Figure 2 also shows the importance of enforcing translational covariance. Without introducing random jitter to the tiles, SAD and SSD seem to dramatically outperform NC. This is even more evident in Figure 3, where the consistency scores are plotted with respect to increasing jitter magnitude. The cause of this is that a lighting gradient in the first image tends to drive the SAD and SSD matchers to estimate a shift placing each uniformly bright tile on top of the bright corner of its mate. Since the majority of tiles are dominated by these lighting effects, resulting in motion vectors pointing at the corners, the motion vectors exhibit false coherence that does not correspond to the true motion. As seen in Figure 2, without jitter, the motion vectors of these tiles appear coherent enough that only small values of  $\alpha$  and  $\sigma$  are necessary to achieve maximum score. When jitter is brought in, the true performance comes out, which is manifested in Figure 2 by a lower overall coherence score and more sensible values of  $\alpha$  and  $\sigma$  leading to maximum score. It is also clearly shown in Figure 3, where the coherence scores are brought down below the scores for NC, as also indicated by the ground truth.

In Figure 4 an experiment is illustrated that validates the coherence score from Equation (10). Image noise with increasing standard deviation is deliberately introduced before computing the patch similarities. For all methods and in all cases the score for the best settings of  $\alpha$  and  $\sigma$  is used. As desired, the score deteriorates accordingly in a monotonic fashion. To further validate the score, ground truth performance evaluation is also conducted with two different scores, resulting in similar performance fall-off. The first, probabilistic ground truth score, is obtained by incorporating the impulse function aligned with ground truth as the prior in Equation (3). This yields a concentration of the sum from Equation (10) to the true motion, thereby requiring not only coherence, but coincidence with the truth. The final score is the negated geometric distance between the extracted correspondences and the true motion vectors. This score was computed by averaging the pixel distance between the ground truth motion and the estimated displacement for each pair. Note that the coherence score correlates well with the other scores, as desired.

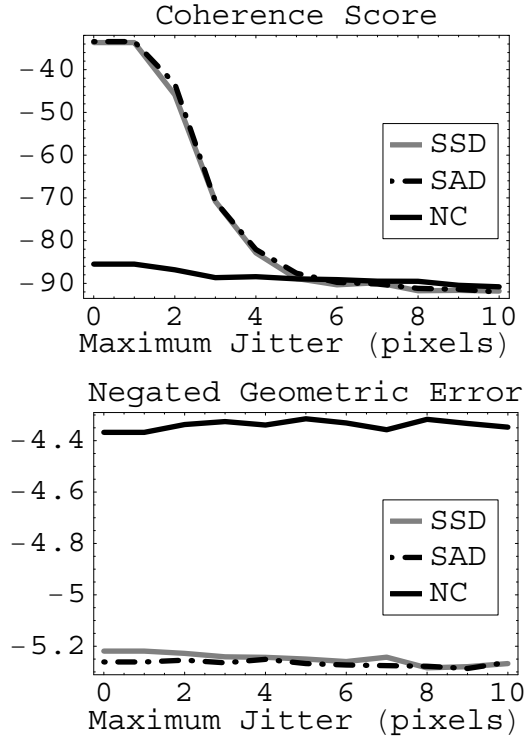


Figure 3: Top: A plot of how the coherence scores decrease for SSD, SAD and NC as translation covariance is more stringently enforced by using increasing magnitudes (x-axis) of random jitter drawn from a uniform distribution. In all three cases,  $\alpha = 0.1$  and  $\sigma = 1.0$  were used. Each datapoint represents an average of 20 trials. Bottom: Plot of the ground truth geometric error as Euclidean pixel distance averaged over each trial and patch pair, negated so that higher values are closer to ground truth. Note how sufficient amounts of jitter brings SSD and SAD down below NC, suppressing the false coherence and resulting in a score correctly reflecting the ground truth performance.

The optimal settings for  $\alpha$  and  $\sigma$  with the first two scores are depicted in Figure 5, showing that the optimal setting varies with noise level, meaning that evaluation of this type is important in order to obtain the best performance for a particular application. Figure 6 shows the performance versus noise of several settings plotted into the same figure. This illustrates how the various setting win at different noise levels, jointly producing the curves in Figure 4.

## 6. Conclusions

We have presented a method for performance evaluation of correspondence extraction. The method does not rely on scene content and does not require ground truth. We have

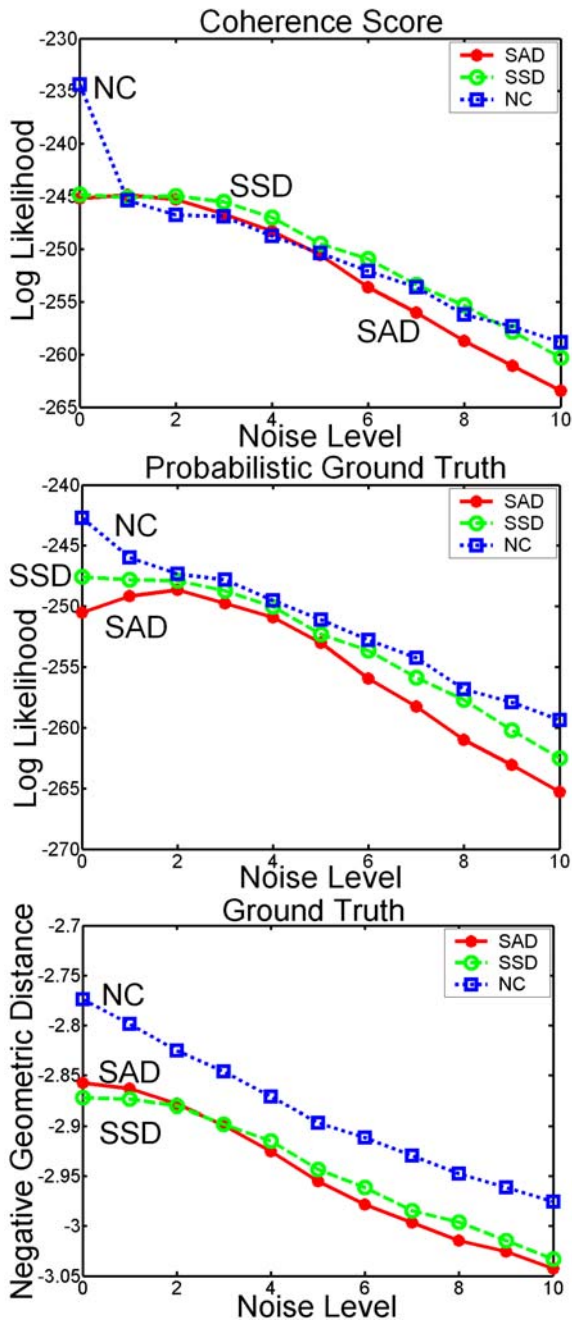


Figure 4: Top: This figure shows that the coherence score decreases monotonically when noise is introduced on purpose before applying the correspondence extractors. This proves that the coherence score correctly signals performance according to what is intuitively expected, and makes it plausible that it would correctly signal better correspondence extractors, including ones that are not yet known. Middle and Bottom: Here we also show two ground truth evaluation scores, the first is computed by introducing a prior indicating the true motion into Equation (3). The second is computed as the geometric distance between the extracted correspondences and the true motion (distance negated to allow direct comparison with above graphs). Note that the scores correlate well.

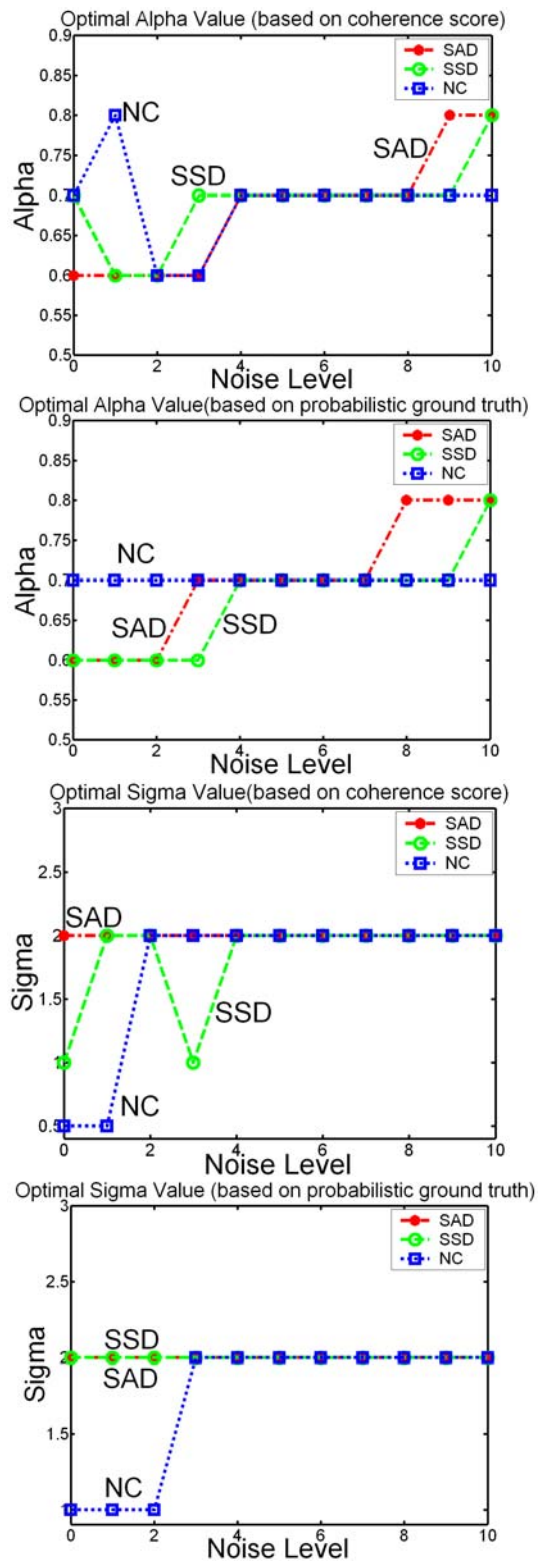


Figure 5: First and Second: Optimal settings for  $\alpha$  according to coherence score and probabilistic ground truth, respectively. Third and Fourth: Optimal setting for  $\sigma$  according to coherence score and probabilistic ground truth, respectively. Note the desired correlation within the figure pairs.

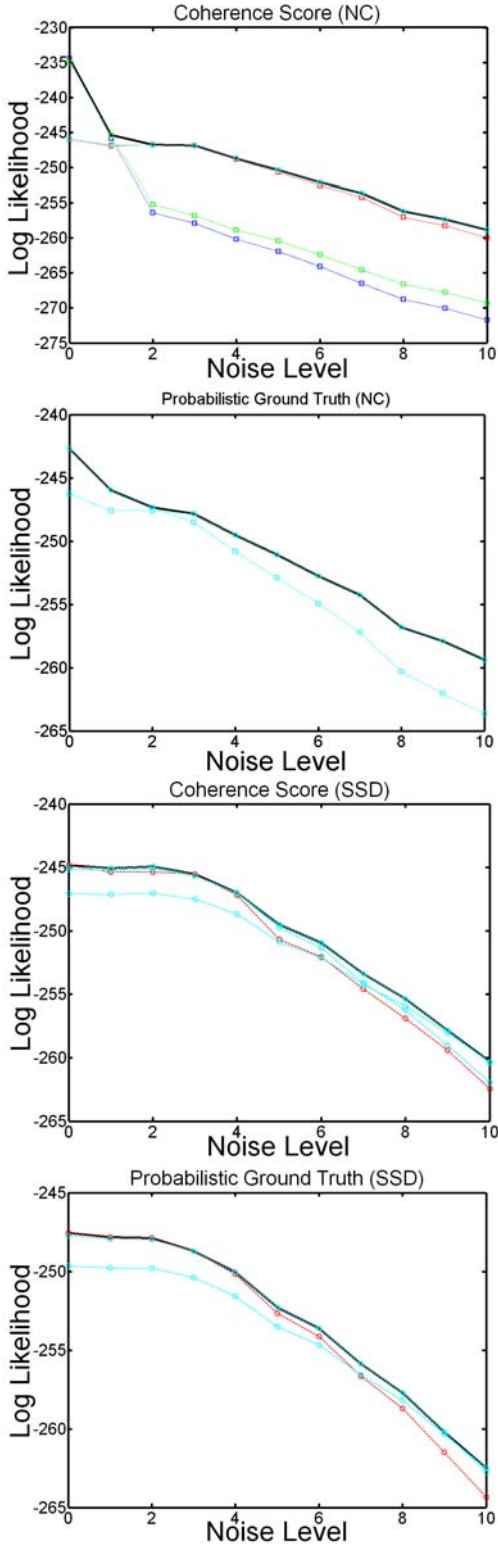


Figure 6: First and Second: Score versus noise for various settings of  $\alpha$  and  $\sigma$  of NC, for coherence score and probabilistic ground truth, respectively. Third and Fourth: Score versus noise for various settings of  $\alpha$  and  $\sigma$  of SSD, for coherence score and probabilistic ground truth, respectively. Again, note the desired correlation within the figure pairs.

validated the method in two ways. First, we have shown that the evaluation results correlate with evaluation using ground truth, by comparing to ground truth evaluation for examples where it is available. Second, we have shown that the evaluation scores correlate correctly with what is expected when we gradually deteriorate a correspondence algorithm on purpose. This makes it plausible that the evaluation would correctly signal algorithms that are better than the known state of the art, but have yet to be found. We have argued that performance evaluation without reliance on constrained scene content or ground truth is a powerful and important tool, since it allows application specific evaluation with large amounts of realistic data. It also suggests a vehicle for online learning of correspondence extraction.

## References

- [1] S. Baker and Shree K. Nayar, Global Measures of Coherence for Edge Detector Evaluation, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Volume 2, Pages 373-379, Fort Collins, Colorado, June, 1999.
- [2] D. Claus and A. Fitzgibbon, Reliable Fiducial Detection in Natural Scenes *European Conference on Computer Vision*, Prague, Czech Republic, 2004.
- [3] A. Heyden, K. Rohr, Evaluation of Corner Extraction Schemes Using Invariance Methods, *ICPR96*, A94.3.
- [4] K. Mikolajczyk and C. Schmid, A performance evaluation of local descriptors. *IEEE Conference on Computer Vision and Pattern Recognition*, June 2003.
- [5] D. Scharstein and R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, *International Journal of Computer Vision*, 47(1):7-42, May 2002.
- [6] C. Schmid, R. Mohr and C. Bauckhage, Evaluation of Interest Point Detectors, *In International Journal of Computer Vision*, 37(2), 151-172, 2000.
- [7] Z. Tu, X. Chen, A.L. Yuille and S.C. Zhu. Image Parsing: Unifying Segmentation, Detection, and Recognition. *International Conference on Computer Vision*, 2003.
- [8] J.M. Coughlan and A.L. Yuille, Manhattan World: Compass Direction from a Single Image by Bayesian Inference. *International Conference on Computer Vision*, Corfu, Greece, 1999.
- [9] Y. Wexler, A. Fitzgibbon and A. Zisserman, Learning Epipolar Geometry from Image Sequences, *IEEE Conference on Computer Vision and Pattern Recognition*, Volume 2, pp. 209-216, 2003.
- [10] W. Förstner, 10 Pros and Cons Against Performance Characterization of Vision Algorithms, *ECCV Workshop on Performance Characteristics of Vision Algorithms*, 1996.
- [11] P. Rosin and E. Ioannidis, Evaluation of global image thresholding for change detection, *Pattern Recogn. Lett.*, Elsevier Science Inc. 24(14):2345-2356, 2003.
- [12] K. Bowyer and P. J. Phillips, *Empirical Evaluation Techniques in Computer Vision*, IEEE Computer Society Press, Los Alamitos, CA, 1998.