Maximum Likelihood Stereo Matching

Nicu Sebe Michael S. Lew
Leiden Institute of Advanced Computer Science,
Niels Bohrweg 1, 2333 CA, Leiden, The Netherlands
{nicu mlew}@liacs.nl

Abstract

In the research literature, maximum likelihood principles were applied to stereo matching by altering the stereo pair so that the difference would have a Gaussian distribution. Here in this paper we present a novel method of applying maximum likelihood to stereo matching. In our approach, we measure the real noise distribution from a training set, and then construct a new metric which we denote the maximum likelihood metric for comparing the stereo pair. The maximum likelihood metric is optimal in the sense that it maximizes the probability of similarity. In our experiments and discussion, we compared the maximum likelihood metric to other promising algorithms from the research literature using international stereo data sets. Furthermore, we showed that the algorithms from the research literature could be improved by using the maximum likelihood metric instead of the sum of squared differences.

1. Introduction

The common assumption is that the real noise distribution should fit either the Gaussian or the Exponential, but what if this assumption is invalid? What if there is another distribution which fits the real noise distribution better than the Gaussian or the Exponential? It is precisely this question which we examined in this paper. Toward answering this question, we have endeavored to use international test sets and promising algorithms from the research literature. Furthermore, one of the canonical measures of similarity from the field of information theory, the Kullback relative information, was also implemented and compared to the metrics based on maximum likelihood.

Stereo matching implies finding correspondences between two or more images. If these correspondences can be found accurately and the camera geometry is known, then a 3D model of the environment can be reconstructed [2]. Several algorithms have been developed to compute the disparity between images, e.g. the correlation methods [10] or correspondence methods [6]. In [5], pixel correspondences are found by adaptive, multi-window template matching. The templates are compared using the SSD. Recent research by [3] concluded that the SSD is sensitive to outliers and therefore robust M-estimators should be used regarding stereo matching. However, the authors [3] did not consider metrics based on similarity distributions. They considered ordinal metrics, where an ordinal metric is based on relative ordering of intensity values in windows - rank permutations. Cox, et al. [4] presented a stereo algorithm that optimizes a maximum likelihood cost function. This function assumes that corresponding features in the left and right images are normally distributed about a common true value. However, the authors [4] noticed the normal distribution assumption used to compare corresponding intensity values is violated for some of their test sets. They altered the stereo pair so that the noise distribution would be closer to a Gaussian. In our approach, we attempt to find a better model for the real noise distribution instead of altering the stereo pair.

Section 2 describes the mathematical support for the maximum likelihood approach. The setup of our experiments is given in Section 3. In Section 4 we present and discuss our experiments using the maximum likelihood metric in stereo matching applications. Conclusions are given in Section 5.

2. Maximum Likelihood Approach

Consider M image pairs (or more generally, feature vectors) from the database (D): $(x_i, y_i) \in D$, with $i = 1, \cdots, M$ which according to the ground truth (G) are similar: $x_i \equiv y_i$. Considering n_i as the "noise" image obtained as the difference between the other two images (x_i and y_i), the similarity probability can be defined:

$$P(G) = \prod_{i=1}^{M} \{\exp[-\rho(n_i)]\} \tag{1}$$

where function ρ is the negative logarithm of the probability density of the noise. According to (1) we have to find the probability density function of the noise that maximizes the similarity probability: *maximum likelihood* estimate for the noise distribution [7]. Taking the logarithm of (1) it can be shown (due to space limitations, we omitted the full proof) that we have to minimize the expression:

$$\sum_{i=1}^{M} \rho(n_i) \tag{2}$$

Note that when the Exponential and Gaussian distributions are used in equation (2), we arrive at the L_1 and L_2 metrics, respectively. A distribution with more extensive tails is the Cauchy distribution, and the corresponding metric L_c is given by the expression:

$$L_c(X,Y) = \sum_{i=1}^{M} \log(\mathbf{a}^2 + (x_i - y_i)^2) \quad (3)$$

where **a** is a parameter which determines the height and the tails of the distribution.

For a general noise distribution, considering ρ as the negative logarithm of the probability density of the noise, the corresponding metric is given by equation (2). In practice, the probability density of the noise can be estimated from the normalized histogram of the absolute differences.

3. Experimental Setup

The setup of our experiments was the following. First, we assumed that representative ground truth was provided. The ground truth was split into two non-overlapping sets: the training set and the test set. Second, the training set was converted to a histogram which was then normalized to what we denoted the real noise distribution. The Gaussian, Exponential, and Cauchy distributions were fitted to the real distribution.

The Chi-square test was used to find the fit between each of the model distributions and the real distribution. Why was the Chi-square test used to find the fit between the model distributions and the real distributions? We could not use a maximum likelihood distance measure between the distributions because the training set data was not sufficient. We would need to accumulate training data over thousands of applications instead of over thousands of examples within an application.

We selected the model distribution which had the best fit and its corresponding metric (L_d) was used in ranking. Note that L_d is a notation for all possible metrics that can be used, e.g. L_1, L_2, L_c. The ranking is done using only the test set. For benchmarking purposes in all of the experiments we compared our results with the ones obtained using the Kullback relative information (K) [8]. We chose the Kullback relative information because it is the most frequently used information theoretic similarity measure. Furthermore, Rissanen [11] showed that it serves as the foundation for other minimum description length measures such as the Akaike's [1] information criterion. Regarding the relationship between the Kullback relative information and the maximum likelihood approach, Akaike [1] showed that maximizing the expected log likelihood ratio in maximum likelihood estimation is equivalent to maximizing the Kullback relative information.

It is important to note that for real applications, the parameter in the Cauchy distribution is found when fitting this distribution to the real distribution from the training set. This parameter setting would be used for the test set and any future comparisons in that application.

In stereo matching, the ground truth is typically generated manually. A set of reference points are defined in the images and then a person finds the correspondences for the stereo pair.

As noted in the previous section, it is also possible to create a metric using the real noise distribution based on maximum likelihood principles. Consequently, we denoted the maximum likelihood (ML) metric as equation (2) where ρ is the negative logarithm of the normalized histogram of the absolute differences from the training set. Note that the histogram of the absolute differences was normalized to have area equal to one by dividing the histogram by the total number of examples in the training set. This normalized histogram was our approximation for the probability density function.

4. Maximum Likelihood Stereo Matching

Stereo matching is the process of finding correspondences between entities in images with overlapping scene content. The images are typically taken from cameras at different viewpoints which implies that the intensity of corresponding pixels may not be the same.

In the first experiments we used two standard stereo data sets (Castle set and Tower set) provided by Carnegie Mellon University. These datasets contain multiple images of static scenes with accurate information about object locations in 3D. The 3D locations are given in X-Y-Z coordinates with a simple text description (at best accurate to 0.3 mm) and the corresponding image coordinates (the ground truth) are provided for all eleven images taken for each scene. For each image there are 28 points as ground truth in the Castle set and 18 points in the Tower set. An example of two stereo images from the Castle data set is given in Figure 1.

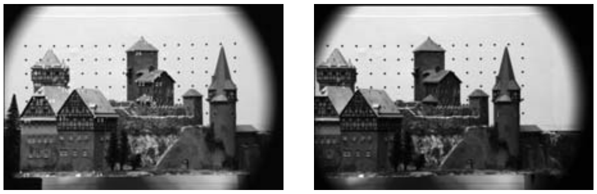
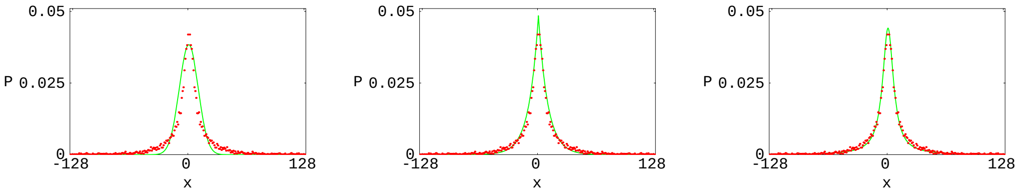


Figure 1. A stereo image pair from the Castle data set

Let I_1 and I_2 represent intensities in two templates i.e. there exist n tuples $(I_1^1, I_2^1), \cdots, (I_1^n, I_2^n)$, n depending on the size of the template used. The quantity $SSD = \sum_{i=1}^{n}(I_1^i - I_2^i)^2$ measures the squared Euclidean distance between (I_1, I_2) and a value close to zero indicates a strong match. The other metrics L_1 and L_c can be defined similarly.

In each image we considered the templates around points which were given by the ground truth. We wanted to find the model for the real noise distribution which gave the best accuracy in finding the corresponding templates in the other image. As a measure of performance we computed the accuracy of finding the corresponding points in the neighborhood of one pixel around the points provided by the test set.



(a) Gauss (b) Exponential (c) Cauchy (**a**=7.47)

Figure 2. Noise distribution in the stereo matcher using Castle data set

In searching for the corresponding pixel, we examined a band of height 7 pixels and width equal to the image dimension centered at the row coordinate of the pixel provided by the test set.

In this application we used a template size of n=25, i.e. a 5x5 window around the central point. For the training sets, we placed templates around 10 points which were obtained from the ground truth.

We present the real noise distribution in Figure 2. As one can see from Table 1 the Cauchy distribution had the best fit to the measured distribution relative to L_1 and L_2. Therefore, one expects L_c to have greater accuracy (Table 2). In all cases the results obtained with L_2 are the worst.

In addition, we investigated the influence of similarity noise using two stereo algorithms from the research literature. The first algorithm [5] is an adaptive, multi-window scheme using left-right consistency to compute disparity. For each pixel the correlation with nine different windows (Figure 3) is performed and the disparity with the smallest SSD (L_2) error value is retained. The authors conclude that the adaptive, multi-window scheme clearly outperforms fixed window schemes. Moreover, the left-right consistency check proves to be effective in eliminating false matches and identifying occluded regions.

Set	Gauss	Exponential	Cauchy
Castle	0.0486	0.0286	0.0246
Tower	0.049	0.045	0.043

Table 1. The approximation error for the corresponding point noise distribution in stereo matching

Set	L_2	L_1	K	L_c	ML
Castle	91.05	92.43	92.12	93.71 **a**=7.47	94.52
Tower	91.11	93.32	92.84	94.26 **a**=5.23	95.07

Table 2. The accuracy of the template stereo matcher (%)

The second algorithm we implemented and tested was introduced by Cox, et al. [4]. Their algorithm optimizes a maximum likelihood cost function. This function assumes that corresponding features in the left and right images are normally distributed about a common true value and consists of a weighted squared error term if two features are matched or a (fixed) cost if a feature is determined to be occluded. Their interesting idea was to perform matching on the individual pixel intensity, instead of using an adaptive window as in the area-based correlation methods.

In order to evaluate the performance of the stereo matching algorithms under difficult matching conditions we also used the Robots stereo pair [9]. This stereo pair is more difficult due to varying levels of depth and occlusions (Figure 4). For this stereo pair, the ground truth consists of 1276 point pairs, given with one pixel accuracy.

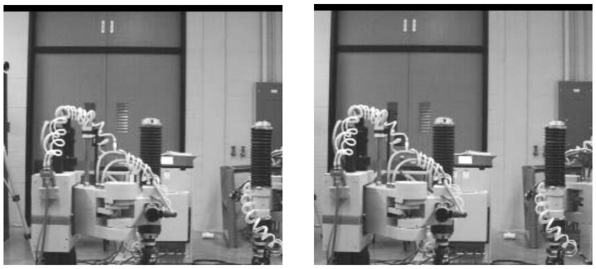


Figure 4. Robots stereo pair

In addition, we also used the two stereo datasets which contain an aerial view of a suburban region. These were taken from Stuttgart ISPRS Image Understanding datasets and are referred to as the Flat and Suburb stereo data sets. In Tables 3 and 4 the results using different distance measures are presented. For all of the stereo sets, ML had the highest accuracy. For the multiple window stereo algorithm, the ML beat L_2 by 2 to 7 percent. Even with the normalization in Cox, et al. [4], the ML metric had improved accuracy over the L_2 metric of approximately 3 to 9 percent.

5. Conclusions and Discussion

We implemented a template matching algorithm, an adaptive, multi-window algorithm by Fusiello, et al. [5], and a maximum likelihood method using pixel intensities by Cox, et al. [4]. Note that the SSD was used in the paper by Fusiello, et al. [5] and in the work by Cox, et al. [4]. Furthermore, we used international stereo data sets from Carnegie Mellon University(Castle and Tower), University of Illinois at Urbana-Champaign (Robots) and University of Stuttgart (Flat and Suburb).

From our experiments, it was clear that choosing the correct metric had significant impact on the accuracy. Specifi-

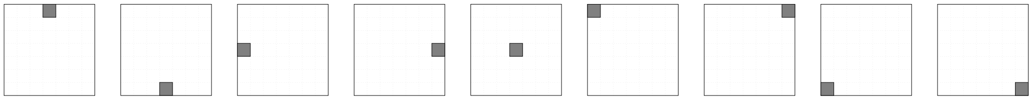


Figure 3. The nine asymmetric correlation windows in Fusiello's algorithm

Set	L_2	L_1	K	L_c	ML
Castle	92.27	92.92	92.76	94.82 **a**=7.47	95.73
Tower	91.79	93.67	93.14	95.28 **a**=5.23	96.05
Robots	72.15	73.74	75.87	77.69 **a**=26.2	79.54
Flat	78.43	77.92	77.76	76.82 **a**=17.17	80.69
Suburb	80.14	79.67	79.14	78.28 **a**=15.66	82.15

Table 3. The accuracy of Fusiello's multiple window stereo algorithm

Set	L_2	L_1	K	L_c	ML
Castle	93.45	94.72	94.53	95.72 **a**=7.47	96.37
Tower	93.18	95.07	94.74	96.18 **a**=5.23	97.04
Robots	74.81	76.76	78.15	82.51 **a**=26.2	84.38
Flat	81.19	80.67	80.15	79.23 **a**=17.17	84.07
Suburb	82.07	81.53	80.97	80.01 **a**=15.66	86.18

Table 4. The accuracy of Cox's maximum likelihood stereo algorithm (with images normalized)

cally, among the L_2, L_1, Cauchy, and Kullback metrics, the accuracy varied up to 7%.

One paradigm for choosing the correct metric is based on maximum likelihood theory. Assuming that the probability of similarity is based on the distribution of the difference between image elements, then it can be proven that a Gaussian distribution results in the L_2 metric, an Exponential distribution results in the L_1 metric, and a Cauchy distribution results in the L_c metric. Therefore, it is also provable that minimizing the L_2 metric maximizes the probability of similarity when the distribution is Gaussian. However, the distribution may not be Gaussian. In fact, none of the international stereo data sets displayed Gaussian nor Exponential distributions, but more Cauchy. However, even the Cauchy distribution proved to be not a very good approximation. Consequently, we introduced a novel metric which is based directly on the real noise distribution, which we denoted the maximum likelihood metric.

For the stereo pairs and the algorithms in our experiments, the maximum likelihood metric consistently outperformed all of the other metrics. Furthermore, it is optimal with respect to maximizing the probability of similarity. The breaking points occur when there is no ground truth, or when the ground truth is not representative.

There appear to be two methods of applying maximum likelihood toward improving the accuracy of matching algorithms. The first method recommends altering the images so that the measured noise distribution is closer to the Gaussian and then using the SSD. The second method is to find a metric which has a distribution which is close to the real noise distribution. Regarding the method of altering the images, our experiments indicate that this technique may have varying accuracy depending on whether the resulting noise distribution is Gaussian.

Our main contribution was showing how to create a maximum likelihood metric based on the real noise distribution. Furthermore, our experiments suggested that for any stereo matching algorithm which uses the SSD or SAD, it is possible (but not necessary) to increase the accuracy by using the maximum likelihood metric if the real noise distribution is neither Gaussian nor Exponential. In the case where there is minimal ground truth and an analytic metric must be used, our experiments suggest that the method of altering the images and using the SSD gives good results. It is noteworthy that in typical photogrammetry applications, extensive ground truth is usually taken.

In future work we intend to examine the influence of multi-parameter distributions towards achieving a better fit to the real distribution.

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory, Armenia*, 1971.

[2] S. Barnard and M. Fischler. Computational stereo, comp. survey. *Science*, 194:283–287, 1976.

[3] D. Bhat and S. Nayar. Ordinal measures for image correspondence. *IEEE Trans. on PAMI*, 20(4):415–423, 1998.

[4] I. Cox, S. Hingorani, and S. Rao. A maximum likelihood stereo algorithm. *CVIU*, 63(3):542–567, 1996.

[5] A. Fusiello, V. Roberto, and E. Trucco. Efficient stereo with multiple windowing. *CVPR*, pages 858–863, 1997.

[6] W. Grimson. Computational experiments with a feature spaced stereo algorithm. *IEEE Trans. on PAMI*, 7:17–34, 1985.

[7] P. Huber. *Robust Statistic*. NewYork: Wiley, 1981.

[8] S. Kullback. *Information theory and statistics*. Dover Publications, 1968.

[9] M. Lew, T. Huang, and K. Wong. Learning and feature selection in stereo matching. *IEEE Trans. on PAMI*, 16(9):869–882, 1994.

[10] W. Luo and H. Maitre. Using surface model to correct and fit disparity data in stereo vision. *IEEE Conf. on Pattern Recognition*, 1:60–64, 1990.

[11] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.