Evaluation of Constructable Match Cost Measures for Stereo Correspondence Using Cluster Ranking

Daniel Neilson and Yee-Hong Yang
Computer Graphics Lab
Department of Computing Science
University of Alberta, Edmonton, Alberta, Canada
dneilson@ualberta.net, yang@cs.ualberta.ca

Abstract

Stereo correspondence research often involves the comparison of techniques to determine which are better under different circumstances. The methods of comparison employed often take the form of applying the techniques to a few stereo image pairs with the technique with the lowest error rate declared superior. However, the majority of these comparisons do not contain any discussion of statistical significance; making the declared superiority of a technique statistically unreliable. In this paper we present a new evaluation method called cluster ranking that yields a statistically significant comparison of the stereo techniques being compared. Cluster ranking leverages statistical inference techniques to first rank the performance of stereo techniques on a single stereo image pair and then combine the rankings from multiple stereo pairs into an over-all ranking; in both of these rankings, only stereo techniques that are statistically different are given different ranks. We demonstrate our framework with a comparison of constructable match cost measures (those that can be assembled from a base set of components) on a data set consisting of 30 synthetic stereo pairs, with varying amounts of noise, and 18 scenes from the 2005 and 2006 Middlebury data sets. Our analysis reveals match cost measures, and measure components, that are statistically superior to all other measures depending on amount of noise, illumination, or exposure time.

1. Introduction

Throughout the history of stereo correspondence research there has been one prevailing method for evaluating stereo techniques. Specifically, the techniques to be compared are employed to generate disparity maps on an, often small, set of stereo pairs; with usually only one disparity map generated per stereo pair. These disparity maps are then compared against ground truth disparity maps to arrive at an error rate for each technique on each stereo image; error rates are compared, and the technique with lower error rates is declared superior. However, there is rarely, if ever, any attempt to gauge the statistical significance of the comparison; resulting in techniques that may not generate statistically different results being declared different, and, even worse, techniques are declared superior when there is no statistical justification for such declaration.

The most popular, and arguably the de facto standard, method for evaluating stereo correspondence algorithms is to use the Middlebury online evaluation [10] supplied by Scharstein and Szeliski. To use this tool, a researcher submits a single disparity map calculated for each of four different stereo pairs. For each stereo pair the tool will then calculate the percentage of pixels, in the submitted disparity map, that differ from the ground truth by more than a threshold in three different evaluation regions. An over-all ranking is obtained from the average of all 12 ranks for each algorithm. Since its introduction as a standardized test set, the Middlebury online evaluation has helped foster stereo correspondence research by greatly simplifying the process of comparing a new technique to existing techniques.

A problem with this online ranking method, that is addressed by our proposed cluster ranking method, is that using only one disparity map per stereo pair does not allow one to determine whether there are two or more algorithms that produce statistically similar results on the stereo pair. Furthermore, the over-all ranking provided by the Middlebury online evaluation does not identify which, if any, algorithms have statistically similar performance over-all.

Our proposed cluster ranking evaluation method uses statistical significance tests combined with a greedy clustering algorithm to rank stereo algorithms such that only those that produce statistically dissimilar results, according to the statistical test employed, are assigned different ranks. When ranking algorithms by their results from a single stereo image pair, we use the error rates from both images combined

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

with an analysis of variance test (ANOVA) [11] to identify the algorithms that produce statistically similar results. When combining the rankings of algorithms from many different stereo pairs into a single over-all ranking, the Friedman test [11] is used instead of ANOVA.

We apply our cluster ranking method to the problem of comparing constructable match cost measures for stereo correspondence. A constructable match cost measure is one that can be assembled from a base set of components: a *channel function* (CF), *channel norm function* (CNF), *channel aggregate* (CA) in the case of measures for colour images, and an optional *spatial aggregate*. In this paper, 360 different measures from this class are compared. Their performance is evaluated using a stereo test set consisting of: 30 noise-free synthetic scenes, generated by a global illumination ray tracing method, with two differing amounts of noise added (resulting in 90 synthetic scenes in total); and 18 scenes from the Middlebury data set [6] (six from the 2005 data set, and 12 from the 2006 data set) that are captured with three different amounts of illumination, and with three different exposure times. Using our new cluster ranking evaluation method on the results allows us to identify which match cost measures, and match cost measure components, outperform others over-all and under differing levels of noise, illumination, and exposure time.

2. Significance Testing

Given $k \geq 2$ items the goal of a statistical significance test is to determine the probability, p, that the k items are the same. Furthermore, given a confidence level, $C \in (0, 1]$, we can say that the k items are similar with confidence C if $p \geq C$; conversely, if we are interested in whether the items are different, we can say that the items differ with confidence C if $p \leq (1 - C)$.

We use two different significance tests in our proposed cluster ranking algorithm: one-way ANOVA, and the Friedman test. Which one we use depends on what data we are using to perform the ranking.

ANOVA [11] is the main method employed by the medical community in their clinical trials to determine whether or not a given treatment has any effect. Given $k \geq 2$ groups of normally distributed sample data, a one-way ANOVA gives the probability that the true values of the k group means are equal; that is, it gives the probability that the methods that produced the k groups of data are similar.

In the case of comparing stereo correspondence techniques, we expect that the error rates in an evaluation region, for a single technique, between the left and right images of a stereo pair will be very similar, and, thus, will follow a normal distribution sufficiently to use ANOVA to compare them. It is important to note that, in general, we cannot combine the error rates from different stereo image pairs, or from different evaluation regions, as this would violate the requirement that the sample data be normally distributed; as evidenced in the Middlebury [10] evaluation, error rates from different stereo scenes and evaluation regions can differ by extremely large margins.

When comparing stereo algorithms using rankings obtained from many different stereo pairs we cannot use an ANOVA test to determine whether or not the algorithms are similar; ranking information does not generally follow a normal distribution. In this case, we use the Friedman test [11]. Given k "tests", the result of which can be quantified with an ordinal value, and N "subjects" who each perform all k tests, the Friedman test gives the probability that all k tests represent populations with the same median value. For our purposes, a "test" is a stereo algorithm being compared and a "subject" is one of the stereo image pairs being used for comparison.

3. Cluster Ranking

The central component of our proposed evaluation method is a novel ranking technique that we call *cluster ranking*. This ranking algorithm gives us a ranking of stereo algorithms such that statistically similar techniques are easily identified by being assigned the same rank.

Given k stereo algorithms to rank, our cluster ranking algorithm is as follows:

```
1: $C \leftarrow$ desired confidence level (e.g. $95\%$)
2: $L \leftarrow$ list of techniques sorted by some criteria
3: Assign each technique a rank equal to its position in $L$
4: Perform a greedy partitioning of $L$ such that all techniques in the same partition are not statistically dissimilar at confidence $C$
5: Reassign each technique a rank equal to the average of the ranks in its partition
```

The sorting criteria on line 2 and the significance test on line 4 that are used depends on what information we are calculating a ranking from; the precise contents of these two lines are discussed in sections 3.1 and 3.2.

The greedy partitioning algorithm that we have implemented creates the first partition starting at the first element of L. We then add consecutive elements of L to this partition as long as there is no statistical difference, at confidence level C, between the techniques in the partition. When adding a technique would break this requirement, we start a new partition starting at the technique.

On line 5, rather than assigning a rank to all techniques in a partition that is equal to the rank-order of that partition in L, we assign a rank to all techniques in the partition that is the average of the ranks assigned to the techniques on line 3. This method of assigning ranks is used extensively in non-parametric tests of significance in the presence of ties [11]; we use it so that rankings obtained by our method can be used in later non-parametric significance tests.

3.1. Ranking From Error Rates

When we are given at least two error rates of the results on a stereo image pair from each stereo technique, we can use one-way ANOVA to determine whether or not the results are statistically similar; note that the error rates must also be from the same evaluation region. Thus, in this case, the test performed on line 4 of the cluster ranking algorithm is the one-way ANOVA test.

Furthermore, since the one-way ANOVA test determines similarity by comparing group means, the sort on line 2 of the cluster ranking algorithm sorts the stereo techniques by increasing order of their mean error rate on the stereo pair.

3.2. Ranking From Rankings

To create partitions the stereo techniques are sorted, on line 2 of the cluster ranking algorithm, in increasing order of their median rank over all the stereo pairs with ties broken in increasing order of their mean rank; this is done because even though two stereo techniques may have the same median rank, they may not be statistically similar at the desired confidence level – so, the mean rank is used as an indicator of which stereo technique is better in these cases.

When combining the ranking results from many stereo image pairs into a single over-all ranking we cannot assume that the rankings for a stereo technique are normally distributed. Thus, we cannot use a one-way ANOVA to determine whether or not techniques are statistically similar. We can, however, use the Friedman test in this case. So, when calculating an over-all ranking from many rankings the similarity test on line 4 of the cluster ranking algorithm is performed using the Friedman test.

4. Constructable Match Cost Measures

Many of the match cost measures in use today can be constructed from a set of up to four basic components: a *channel function* (CF), *channel norm function* (CNF), *channel aggregate* (CA) in the case of colour images, and an optional *spatial aggregate*. This construction method is easy to implement and allows us to determine the relative performance of individual components.

The CF and CNF together define a gray scale match cost measure that can be applied to each colour channel of colour images and then combined into a colour match cost measure using a CA. These first components can be used to generate a disparity space image (DSI) for each image in a stereo pair. Once the DSI for each stereo image has been calculated, a *spatial aggregate* can be applied to each DSI to obtain a match cost for each pixel that takes a spatial neighbourhood of the pixel into account.

4.1. Channel Function

The CF is a function that is applied to a single colour channel of a pair of pixels to yield a non-negative measure of the difference between them. For our study, we have identified two CFs in common use; these functions can be found in table 1.

Function Name	**Abbr.**	**Function**
Difference	D	$g(c_1, c_2) = \|c_1 - c_2\|$
Birchfield & Tomasi	B	See [1]

Table 1: Channel functions.

Strictly speaking, the Birchfield & Tomasi measure [1] is presented as a gray scale match cost measure that should be re-derived for use on colour images. However, it is common practice [9] in the stereo correspondence community to use the gray scale derivation of this measure on each colour channel and then aggregate the three results into a single match cost measure. Since this is the common practice, we use the gray scale derivation of the Birchfield & Tomasi measure as a CF in our study.

4.2. Channel Norm Function

The CNF is applied to the results of the CF to alter the response of the match measure to differences in gray scale intensity. For this study, we have identified four different CNFs that have been used by the stereo correspondence research community (see table 2). We also use the truncated versions of the CNFs listed in this table in our study.

Function Name	**Abbr.**	**Function**
L1 Norm	A	$f(x) = \|x\|$
L2 Norm	S	$f(x) = x^2$
Generalized Gaussian	$\mathrm{G}_{\sigma,s}$	$f_{\sigma,s}(x) = \|x/\sigma\|^s$
Lorentzian	L_σ	$f_\sigma(x) = \ln(1 + \frac{1}{2}(\frac{x}{\sigma})^2)$

Table 2: Channel norm functions.

4.3. Channel Aggregate

When using colour images (for example, BGR images) the results of the CNF for each channel must be combined into a single value representing the quality of a match; this is done via a CA. For this study we have experimented with the five CAs found in table 3, as well as truncated versions of the *sum*, *weighted sum*, and *sum minus max* aggregates; truncated versions of the *median* and *max* aggregates are redundant given that we apply truncation to the channel norm functions. Truncation of a CA is considered [12] to make a match cost measure robust to noise and outliers.

Aggregate Name	Abbreviation	Aggregate
Sum	S	$c_S(p,q) = \sum_{c \in \{b,g,r\}} f(x_c)$
Weighted Sum	Sw_{w_b,w_g,w_r}	$c_{Sw}(p,q) = \sum_{c \in \{b,g,r\}} w_c f(x_c)$
Median	Me	$c_{Me}(p,q) = \text{median}\{f(x_c) : c \in \{b,g,r\}\}$
Max	Mx	$c_{Mx}(p,q) = \max\{f(x_c) : c \in \{b,g,r\}\}$
Sum Minus Max	Sx	$c_{Sx}(p,q) = c_S(p,q) - c_{Mx}(p,q)$

Table 3: Channel aggregates between pixels p and q. $f()$ denotes a channel norm function, and x_c result of a channel function.

We have observed the *sum*, *weighted sum*, and *max* CAs utilized by researchers in various match cost measures. At the start of our study we thought that the *median* and *sum minus max* CAs may be more robust to noise than either the *sum* or *max* aggregates, by virtue of them excluding the channel with the largest match penalty from the calculation of the aggregate, and thus included them in our study; our results show this thinking to be false in practice.

4.4. Spatial Aggregate

Once the DSIs for a stereo image pair have been calculated, a *spatial aggregate* can be applied to the DSIs to obtain a match cost for each pixel that takes the spatial neighbourhood of the pixel into account. Options for a spatial aggregate range from applying a simple $(2n+1) \times (2m+1)$ mean, median, or summation filter centered on each pixel, to the similarity-based adaptive neighbourhood presented by Patricio et al. [7], or Yoon and Kweon's [13] adaptive-weights aggregate.

Spatial aggregates are not a part of this study, due to the processing requirements of including them, but, their inclusion is a goal for our future work.

4.5. Putting it all Together

To make describing each constructable match cost measure less cumbersome we have defined an abbreviation for each value of each component, and derive an abbreviation for a measure by listing the abbreviations for its components in the order: spatial aggregate (when present), CA, CNF, then CF. If a component's results are being truncated we postfix the abbreviation for the component with the truncation value. For example, the abbreviation for the measure that uses a *sum minus max* CA, *L2 norm* CNF, and *Birchfield & Tomasi* CF is SxSB. The same measure with its CNF truncated by τ is SxS(τ)B.

5. Stereo Image Test Set

For our study of constructable match cost measures we have assembled a stereo image test suite consisting of 252 stereo image pairs from 48 different scenes; 90 synthetic stereo image pairs that we created, and 162 non-synthetic stereo image pairs from the Middlebury data set.

We created a set of 30 noise-free stereo image pairs using the PBRT [8] ray tracer so that we had a clean data set that could be used to test the effects of noise. Disparity maps were generated for each of these ray traced images from the output of an in-house developed plugin that generates a floating point depth map for each ray traced image. All images were generated with a path tracing algorithm configured to virtually eliminate noise.

The 30 synthetic scenes were rendered from three geometric configurations with ten random texture assignments on each configuration; an example of each of the geometric configurations used can be seen in figure 1. Our synthetic scenes were created in this manner so there would be high variability of what colours appear on either side of depth discontinuities in each geometric configuration.

For each of these 30 scenes we rendered a single stereo image pair with camera positions set so that the resulting images would be perfectly rectified at a baseline separation of 30 pixels. For each geometry, the camera positions used to render images from the ten scenes using the geometry were identical.

To simulate the Gaussian noise characteristics [5] of a real digital camera we used a Sony DCR-TRV230 NTSC video camcorder to capture 1022 images of a static scene under constant lighting. We then averaged all of these images to arrive at an estimate of the ground truth, and compared all 1022 images to it to determine the mean and variance of the noise for each intensity in each channel.

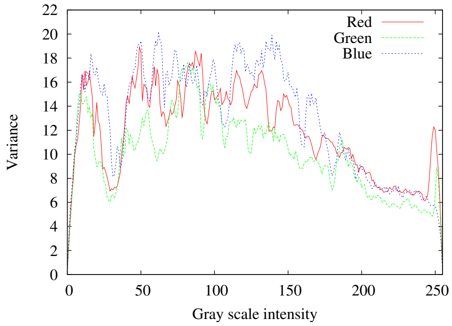
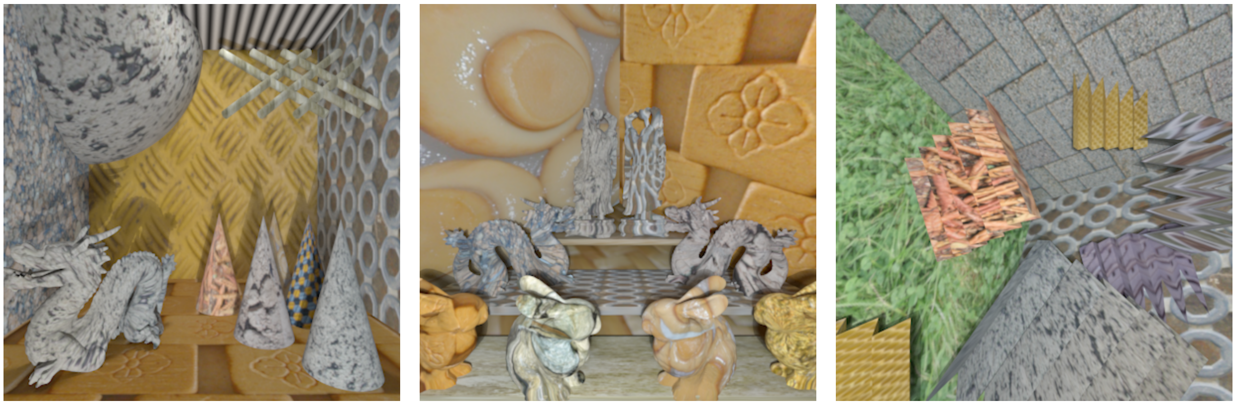


Figure 2: Gray scale intensity on each colour channel vs noise variance for the Sony DCR-TRV230 NTSC video camcorder.



(a) Geometry 1

(b) Geometry 2

(c) Geometry 3

Figure 1: Synthetic images from the three different geometries used in the synthetic portion of our test set.

The calculated variance curves (see figure 2) were then used to introduce synthetic zero mean Gaussian noise to our noise-free images in two levels. The first level (noise level 1) used the noise variance of the Sony camera, and the second (noise level 2) used four times the variance of the Sony camera. The result is 90 synthetic image pairs at three different noise levels.

To determine whether the amount of illumination or exposure-time used has any effect on the accuracy of the match cost measures in our study we used 18 of the $\frac{1}{3}$-sized scenes from the Middlebury 2005 and 2006 data sets. Specifically we used: Aloe, Art, Baby2, Baby3, Books, Bowling2, Dolls, Flowerpots, Lampshade2, Laundry, Midd1, Moebius, Monopoly, Plastic, Reindeer, Rocks1, and Wood1. For each scene in this data set there are nine stereo pairs taken under a combination of three different illumination levels and three different exposure levels; resulting in nine stereo pairs per scene, or 252 stereo pairs in total. Ground truth disparity maps are available for all of these stereo image pairs.

From the ground truth disparity maps, we automatically generated pixel masks for the evaluation regions: *all* (all pixels except a small border on the left, for the left image, or right, for the right image), *occluded* (all pixels that do not adhere to the weak consistency constraint [4] in a left-right consistency check), *discontinuous* (all occluded pixels plus all pixels within 10 pixels of either an occluded pixel or a disparity discontinuity greater than $\frac{1}{2}$ a disparity level), and *non-occluded* (all pixels in the *all* mask that are not in the *occluded* mask).

6. Hierarchical Belief Propagation

During the course of this study, the stereo correspondence algorithm we chose was run around 11.76 million times. Thus, we required an algorithm that was both very fast, in a CPU-only (non-GPU) implementation, and produces reasonably decent results. We decided to use a global stereo algorithm for our testing because most of the top performing algorithms use a global formulation. However, we did not consider segmentation-based algorithms because the choice of segmentation parameters adds another level of complexity to our evaluation that we did not have the computing resources to address. This restriction combined with the speed requirement removed many of the top rated algorithms from consideration. We chose a modification of Felzenszwalb and Huttenlocher's [3] hierarchical belief propagation algorithm that maximizes the posterior probability:

$$\rho(f|\mathcal{I}) \propto \prod_{p\in\mathbf{P}} \rho_p(f(p)) \prod_{\{p,q\}\in N} \rho_{p,q}(f(p), f(q)). \quad (1)$$

where f is the calculated disparity map, $\mathcal{I}$ is the set of input images, $\mathbf{P}$ is the set of pixels in the reference image, and N denotes the set of 4-connected pixels in $\mathbf{P}$. $\rho_p(\delta)$ is the data cost:

$$\rho_p(\delta) = e^{-c(p,\delta)} \quad (2)$$

where $c(p, \delta)$ denotes the match cost of pixel p at disparity level δ, and $\rho_{p,q}(\delta_1, \delta_2)$ is the smoothing term, for which we use the truncated linear model with parameters s and d:

$$\rho_{p,q}(\delta_1, \delta_2) = e^{-\min\{s|\delta_1-\delta_2|,d\}}. \quad (3)$$

Note that the algorithm presented by Felzenszwalb and Huttenlocher minimizes the negative log of this posterior probability while we modified the algorithm to maximize the posterior probability. Our implementation was also optimized using SIMD instructions where we could, and used floating point throughout to remove errors introduced by rounding to integer values; images are represented as floating point with each intensity value in the range $[0, 1]$.

7. Experimental Setup

In this study we compare the accuracy of disparity maps generated by the hierarchical belief propagation algorithm using 360 different constructable match cost measures, on all 252 image pairs in our data set (504 disparity maps in total). Our match cost measures are constructed using all of the components listed in tables 1, 2, and 3. Since the Generalized Gaussian and Lorentzian CNFs are parametrized, we conducted some preliminary testing to find some decent parameters for these two CNFs to use in our study; for the Generalized Gaussian we used values of $(\sigma, s) \in \{(0.5, 1.5), (1, 1.5), (1.5, 1.5)\}$, and for the Lorentzian we used values of $\sigma \in \{0.5, 1, 1.5, 2\}$. Furthermore, we included measures with and without truncation values on their CNF and CA subject to the restrictions in section 4.3.

For each of these 360 match cost measures there are up to two different truncation values, τ_1 (for the CA) and τ_2 (for the CNF), that must be used. Ideally, we would optimize on these values; however, given the size of this study and the sizable computation resources such an optimization would require this was not a route we could take. Instead, we decided to use all combinations of the truncation values $\tau_1 \in \{\frac{k}{255} : k = 6, 8, 10\}$ and $\tau_2 \in \{\frac{k}{255} : k = 2, 4, 6, 8, 10\}$ for each match cost measure. For the measures that use truncation values, we ran the measure with all of the truncation values and chose the disparity map for a stereo image that minimized the percentage of pixels that differed from the ground truth by more than one disparity level in the *non-occluded* evaluation region. The *non-occluded* evaluation region was chosen because the disparity values for occluded pixels will be highly influenced by the smoothing term due to lack of a pixel correspondence. Counting different truncation values, we have a total of 1944 different instances of match cost measures that we use to calculate disparity maps.

Furthermore, we observed early on in our study that no single parameter setting for the parameters in the smoothing term, $\rho_{p,q}$, work well for every match cost measure or even every stereo image pair. In fact, it seems as though for every match cost measure that a set of parameters works very well for on a stereo image pair there is at least one other match measure that they work very poorly for on the same stereo image pair. Again, ideally we would optimize on these parameters. However, optimizing on each of 252 stereo image pairs for each of the 1944 match cost measures is computationally intractable given our resources. So, instead we chose 12 different parameter settings for these smoothing parameters that preliminary testing indicated would be acceptable (see table 4). Of the 12 disparity maps calculated for a match cost measure on a single stereo image, we chose the disparity map in the same manner as when choosing between disparity maps resulting from different truncation values.

(s, d)	(s, d)	(s, d)	(s, d)
$(\frac{3}{255}, \frac{20}{255})$ $(\frac{1}{255}, \frac{10}{255})$ $(\frac{1}{255}, \frac{5}{255})$	$(\frac{5}{255}, \frac{20}{255})$ $(\frac{2}{255}, \frac{10}{255})$ $(\frac{2}{255}, \frac{5}{255})$	$(\frac{7}{255}, \frac{20}{255})$ $(\frac{3}{255}, \frac{10}{255})$ $(\frac{1}{255}, \frac{3}{255})$	$(\frac{9}{255}, \frac{20}{255})$ $(\frac{4}{255}, \frac{10}{255})$ $(\frac{1}{255}, \frac{2}{255})$

Table 4: Parameter settings used for the truncated linear smoothing term.

In all, our study required calculating $1944 \times 12 \times 252 \times 2 = 11,757,312$ disparity maps; taking approximately 4.8 compute-years on a single 3.0GHz Xeon processor. We used a shared Beowulf cluster, containing 1680 3.0GHz Xeon processors, to complete this computation in approximately 10 weeks.

8. Discussion of Results

To analyze our results, we applied our cluster ranking evaluation method with a confidence level of 95%. The amount of analysis that we performed is too voluminous to fully discuss in this short format, so we focus on some of the over-all ranking results here. Since there is so much data available from our study (disparity maps, tables and graphs of measure rankings, etc) we designed an interactive web site to make navigating this data easier; it is available at http://www-user.cs.ualberta.ca/stereo.

8.1. Match Cost Measure Analysis

To analyze the relative performance of the 360 match cost measures in our study we employed our new cluster ranking evaluation method in two stages. First, we used *cluster ranking from error rates* to calculate a ranking of the 360 measures for each of the 252 stereo pairs in our data set. Then, using these rankings, we used *cluster ranking from ranks* to calculate over-all rankings of the match cost measures; over-all rankings were calculated for each noise level (by only including the stereo pairs at the noise level), each illumination level, each exposure level, over all synthetic images, over all real (Middlebury) images, and over the entire test set.

Measure	**Rank**	**Median**	**Mean**
$SG_{0.5,1.5}(\tau_2)D$	1.5	30.5	46.93
$Sw_{0.33,0.33,0.33}AD$	1.5	31.5	39.10
$MxG_{0.5,1.5}D$	4.5	33.5	41.71
$MxAD$	4.5	32.0	40.26
$SG_{0.5,1.5}D$	4.5	32.5	42.74
$SG_{1,1.5}D$	4.5	32.5	41.03

Table 5: Top six, of 360, match cost measures over the entire test set. Shown: Over-all rank, median and mean rank over all 252 stereo pairs.

Table 5 shows the top six match cost measures from the ranking obtained by aggregating the rankings from the entire test set using cluster ranking. There are a few interesting things to note even from this small snapshot of the over-all rank table.

First, none of the top six match cost measures use the Birchfield & Tomasi CF; in fact, the best ranking match cost measure that uses a Birchfield & Tomasi CF is the $SA(\tau_2)B$ measure at rank 51. Furthermore, although the $MxAD$ measure ranks 4.5 over-all its Birchfield & Tomasi equivalent ($MxAB$) ranks 51^{st} with a median rank of 57. As our measure component analysis shows (see table 6), this trend occurs more often than not. This suggests that using the gray scale Birchfield & Tomasi measure as a component for a colour match cost measure is typically inferior to using the simpler *difference* CF.

Also of interest is that four of the top six (15 of the top 20, 22 of the top 30, and 30 of the top 42) match cost measures use the Generalized Gaussian CNF, with and without truncation, with varying parameter choices; the remainder of the top 42 all use the L1 norm CNF. In fact, the Lorentzian CNF, the only CNF not based on a Generalized Gaussian, does not appear in a match cost measure until $SL_{0.5}D$ at rank 51 with median rank 59.5.

8.2. Match Cost Measure Component Analysis

Channel Function	Rank	Median	Mean
Difference	1	1	1.01
Birchfield & Tomasi	2	2	1.99

Table 6: Over-all ranking of channel functions.

We also leveraged the constructive nature of the match cost measures in our study to analyze the typical performance of each of the match cost measure components used in our study. We performed this analysis using our *cluster ranking from ranks* method. For each component type (CF, CNF, or CA) cluster ranking was first applied to obtain a ranking of the component's functions on each of the 252 stereo pairs in our data set. We then used cluster ranking to obtain over-all rankings in each of the same categories as in our analysis of match cost measures.

Table 7 shows the top ten CNFs used in our study. There are a few interesting items of note that jump out when looking at this table. First, all of the CNFs are statistically dissimilar at 95% confidence (all rank values are unique). Second, eight of the top ten CNFs are based on the Generalized Gaussian. This suggests that a stereo algorithm that uses a Generalized Gaussian based match cost measure, and optimizes on the Generalized Gaussian parameters, would be worth further investigation; this has been done with some success by Cheng and Caelli [2]. Finally, with one exception, the truncated version of a CNF is always superior to the non-truncated version of the same.

Channel Norm Function	Rank	Median	Mean
A	1	2.5	2.92
$A(\tau)$	2	2.5	3.35
$G_{0.5,1.5}(\tau)$	3	3.0	3.29
$G_{0.5,1.5}$	4	3.5	3.31
$G_{1,1.5}(\tau)$	5	5.5	5.12
$G_{1,1.5}$	6	5.5	5.42
$G_{1.5,1.5}(\tau)$	7	7.5	7.16
$G_{1.5,1.5}$	8	7.5	7.49
$L_{0.5}(\tau)$	9	9.5	8.84
$L_{0.5}$	10	9.5	9.15

Table 7: Top 10, of 18, over-all channel norm functions.

Channel Aggregate	Rank	Median	Mean
S	1	1.5	1.33
$S(\tau)$	2	1.5	1.87
Mx	3	3.0	2.80
$Sw_{0.4,0.59,0.11}(\tau)$	4.5	5	5.17
$Sw_{0.4,0.59,0.11}$	4.5	5.0	5.27
$Sw_{0.33,0.33,0.33}(\tau)$	7	7.0	6.75
$Sw_{0.33,0.33,0.33}$	7	7.0	7.03
Sx	7	7.5	7.40
$Sx(\tau)$	9	7.5	7.48
Me	10	10.0	9.90

Table 8: Channel aggregate rankings on images with noise characteristics similar to the Sony DCR-TRV230 NTSC video camcorder.

In table 8 we present the over-all rankings of channel aggregates over the synthetic images with noise variance equal to that of the Sony DCR-TRV230 NTSC video camcorder. These results show that the practice of truncating a CA to yield a colour match cost measure that is thought to be robust to noise typically produces a colour match cost measure that is worse, or is not statistically significantly different, than its non-truncated counterpart on noisy images. As an aside, note that simply using the mean rank of these CAs as an indicator of performance would yield the erroneous conclusion that truncating sometimes yields a better measure even though there is no statistically significant evidence for such a conclusion.

One must be cautious when combining the results of the over-all rankings from this study. From these results, one might expect the commonly used SAD measure to be typically the best performer even though it is not; this measure does rank highly, over-all, with a rank of 7.5 but is not the top performer.

9. Conclusion and Future Work

In this paper we present a novel evaluation method for comparing stereo correspondence techniques that allows one to easily determine whether results are statistically significant. We also present the class of constructable match cost measures and use our evaluation method to compare 360 different match cost measures from this class. This comparison is performed using a hierarchical belief propagation algorithm and a test data set consisting of 252 stereo image pairs. For our stereo data set, we develop 90 synthetic stereo image pairs with varying amounts of noise that closely approximates the noise characteristics of a real digital camera.

We present some of the results of this study in this paper and have made the full analysis, with most of the raw data, available on an interactive website. Through this website we also make available our synthetic data set, and most of the program code used in our study. Furthermore, for further analysis, all data generated during our study are available upon request; that is, all 11+ million disparity maps (at $\frac{1}{2}$ scale), percent-error rates at five different threshold values, and RMS error rates.

In the past, the choice of match cost measure used in a stereo algorithm has been made with little, if any, justification; instead, researchers typically choose "tried and true" match cost measures (SSD, SAD, Birchfield & Tomasi, etc) with little apparent justification about its impact, if any, on their proposed algorithms. The results of this study show that the choice of match cost measure used in an algorithm can have a large impact on the accuracy of the disparity maps generated by the algorithm.

In analyzing the results of our study we have observed that the practice of creating a colour match cost measure by aggregating the gray scale Birchfield & Tomasi measure typically produces worse results than using a simple difference operator in its place. Furthermore, we have observed that the common practice of truncating a match cost measure to produce one robust to noise does not typically perform better than its non-truncated counterpart in the presence of noise.

In the immediate future, we plan to apply our evaluation method to reevaluate different stereo algorithms, from the simplest winner-take-all to more sophisticated global optimization algorithms such as dynamic programming and graph-cuts based algorithms. Additionally, we will investigate the impact of many of the commonly used, but seldom discussed, pre- and post-processing methods to the performance of stereo algorithms. We also plan to extend our analysis of match cost measures to include measures that incorporate spatial aggregation.

10. Acknowledgements

Computing resources for this work generously provided by the Western Canada Research Grid are gratefully acknowledged. The authors also want to thank the financial supports provided by NSERC, the University of Alberta and iCore.

References

[1] S. Birchfield and C. Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(4):401–406, April 1998.

[2] L. Cheng and T. Caelli. Bayesian stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 192 – 192, 2004.

[3] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–261 – I–268, 2004.

[4] M. Gong and Y.-H. Yang. Fast stereo matching using reliability-based dynamic programming and consistency constraints. In *International Conference on Computer Vision*, volume 1, pages 610 – 617, 2003.

[5] G. Healey and R. Kondepudy. Radiometric ccd camera calibration and noise estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(3):267 – 276, 1994.

[6] H. Hirschmüller and D. Scharstein. Evaluation of cost functions for stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[7] M. P. Patricio, F. Cabestaing, O. Colot, and P. Bonnet. A similarity-based adaptive neighborhood method for correlation-based stereo matching. In *International Conference on Image Processing*, volume 2, pages 1341 – 1344, 2004.

[8] M. Pharr and G. Humphreys. *Physically Based Rendering: From Theory To Implementation*. Morgan Kaufmann, 2004.

[9] D. Scharstein and C. Pal. Learning conditional random fields for stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[10] D. Scharstein and R. Szeliski. http://vision.middlebury.edu/stereo.

[11] D. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press, 2004.

[12] J. Sun, Y. Li, S.-B. Kang, and H.-Y. Shum. Symmetric stereo matching for occlusion handling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 399 – 406, 2005.

[13] K.-J. Yoon and I. S. Kweon. Adaptive support-weight approach for correspondence search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):650 – 656, 2006.