
# Evaluation of Computer Vision Stereo Algorithms for Surgical Applications

**Sepehr Vosoughi, Elle Ameli and Richard P. Wildes**

Technical Report EECS-2014-01

February 10 2014

Department of Electrical Engineering and Computer Science
4700 Keele Street, Toronto, Ontario M3J 1P3 Canada

# Evaluation of Computer Vision Stereo Algorithms for Surgical Applications

Sepehr Vosoughi, Elle Ameli and Richard P. Wildes
Department of Electrical Engineering and Computer Science
York University
Toronto, Ontario

**Abstract**

This project contributes to the application of computerized stereo vision in the medical domain. A dataset that is representative of medical surgery has been obtained that contains twenty-five left-right stereo image pairs along with dense groundtruth disparity maps. Furthermore, a representative set of four state-of-the-art stereo algorithms are run on the datasets and resulting disparity maps are presented. These results are qualitatively and quantitatively analyzed and performance of each stereo algorithm is compared. Key performance issues for each algorithm and for each scene are pointed out and suggestion for future improvements to the stereo algorithms are made. Finally, suggestions and plans for future work are presented.

# 1 Introduction

## 1.1 Motivation

The term "computer vision" is straightforward enough for most people to realize what it encompasses: to give vision to computers. However, many still wonder what applications this field has. Indeed, the possible applications of computer vision are vast and one that quickly comes to mind is the medical domain. Along these lines, one interesting and ambitious goal is the realization of a robot surgical assistant. For this robot assistant to work usefully it will be advantageous for it to perceive the world's 3D structure from acquired imagery, analogous to human capabilities.

Intensive research in computer vision has yielded a range of stereo vision algorithms that are capable of reconstructing the 3D layout of a scene from binocular imagery [4]. In relatively benign conditions (e.g., laboratory set-ups with simple lighting, surface reflectances and surface layout), some algorithms are able to provide 3D estimates with a high degree of precision and accuracy [18]. Performance in the presence of less controlled, real-world conditions yields much less reliable performance [22]. Further, it is found that algorithm performance can be scenario and application dependent [29].

In the light of the above observations, it becomes important to evaluate stereo vision algorithms on image data that is representative of any specific application domain of interest. Correspondingly, there has been much research evaluating the performance of different stereo algorithms in different scenarios [18, 11, 2, 9, 27]. There has not, however, been a systematic study evaluat-

ing the performance of current state-of-the-art stereo algorithms on medical imagery. Further, nor has there emerged a database of stereo imagery with associated groundtruth disparity that is representative of medical applications. Significantly, without such a database, quantitative evaluation of stereo vision algorithms in the medical domain is not a possibility.

The research project documented in this report responds to both of the shortcomings mentioned in the previous paragraph. A representative set of medically relevant images along with associated groundtruth disparity measurements are attained in a lab environment. Furthermore, the performance of four of the most widely used stereo algorithms (i.e. Normalized Cross Correlation (NCC) [4], Adaptive Coarse-to-Fine NCC (CTF) [20], Semi-Global Block Matching (SGM) [10] and Graph Cuts [3]) [18] have been qualitatively and quantitatively evaluated on the database.

## 1.2 Related Research

There has been some previous work on recovery of groundtruth and evaluation/use of stereo algorithms or other computer vision tools to recover 3D representations of a surgical scene or even model a specific organ. However, it appears that no previous work has presented a database with dense groundtruth disparity of real-world medical imagery nor has there appeared a detailed quantitative empirical comparison of computer vision stereo algorithms on such data. Nevertheless, the remainder of this section provides a summary of the most relevant previous research.

Some previous research only performs qualitative evaluation on their pro-

posed algorithms or other algorithms they are comparing against. For example, one investigation uses the Normalized Cross Correlation (NCC) algorithm while enforcing strict matching criteria to achieve reconstruction of the 3D scene using stereo images in invasive surgery [1]. The proposed approach introduces three constraints on matching points (which are the matched patch should resemble the expected spatial structure within some error range, match confidence must exceed some specified value and identical left-to-right and right-to-left matching), which limit the number of qualifying match points to potentially a very small number. The presented results are only compared with two other contemporary algorithms [16, 25] but this comparison is only done qualitatively. Furthermore, the research does not attempt to acquire or even mention quantitative analysis with groundtruth.

Several research efforts perform quantitative evaluation on phantoms and only on handpicked points from sparse groundtruth (e.g., provided by a CT scan). For example, in one such effort a GPU based method was used for 3D reconstruction for intraoperative navigation [16]. The research uses a video based stereo algorithm called Hybrid Recursive Matching modified to support sub-pixel matching and to run on a GPU. However, for dense quantitative analysis, this research only uses virtual (i.e., computer graphics generated) image sequences with known groundtruth. This research also performs quantitative analysis on images obtained from a phantom, but limited to sparse groundtruth points acquired by CT (only 20 points are provided). Another effort uses a factorization method to reconstruct 3D points from 2D endoscopic images [30]. Using this approach the authors were able to show better results compared to

more conventional methods. Full dense evaluation was only done on synthetic data. Similar to the previous approach, real data evaluation was done only against sparse groundtruth provided by a CT scanner with matching points. Other research also presented a method to match stereo images using an initial sparse set of 3D matched feature points that subsequently are propagated throughout the scene to create a semi-dense disparity map [25]. However, the sparse groundtruth for phantom-based evaluation was obtained using a CT scanner and the evaluation on a real data set was performed qualitatively only. Therefore, no real world groundtruth acquisition or complete-dense 3D recovery is used. In still other research, the NCC match measure between stereo images from calibrated stereo endoscopic cameras was used to recover reasonable disparity maps [24]. This was done using both stereo and temporal correspondences. The evaluation of the algorithm was done using groundtruth obtained by a CT scan of a phantom. The real data evaluation was done only qualitatively on an in vivo stereoscopic sequence. Moreover, no other match measures or matching algorithms were evaluated or discussed. Other research was able to recover liver structure and motion using calibrated rigid laparoscopic cameras [23]. However, this approach used handpicked stereo correspondence points, as it was found that established automated methods failed in this scenario. Nevertheless, the results obtained with this experiment only contains the 3D reconstruction of selected points of interest (which were manually selected) and no comparison to dense groundtruth is presented.

There has been some research that has looked at alternative approaches in 3D recovery for medical application. For example, one research effort pro-

poses Time of Flight (ToF) sensors and measurement on a single endoscope to acquire real-time per pixel distance information [14]. The recovered 3D groundtruth is impressive. This approach would require installation of a ToF sensor in the operation, including near infrared illumination. This has a cost ($7000) and may not be allowed in all circumstances since the actual endoscope optics need to be altered. Furthermore, no evaluation of any multiview stereo algorithms is provided. In another example, traditional SFM (structure from motion) methods with some pre-processing and constraints to deal with missing data and outliers were used to produce 3D reconstructions from temporal sequences of single 2D endoscopic images [12]. Some of the groundtruth acquisition methods used in this research for data evaluation is similar to the method used in the work described in the present project (i.e. using a stereo camera and a light projector). However, there was no groundtruth acquisition done on real internal organs and only "leave-n-image-out cross-validation" was used to evaluate consistency of the proposed method. Further, no evaluation against other 3D reconstruction algorithms was performed.

Finally, a group of research efforts have considered the particular application of 3D registration of a model organ to disparity-based 3D surface estimates. This research mostly performs its evaluations based on the correctness of the registration itself. For example, in one such effort a tracking algorithm using images from a stereo endoscopic system was proposed [13]. Stereo matching was done using the zero-mean sum of squared differences [15]. Using this technique, the authors were able to recover the coherent motion of the heart. The acquired results are not quantitatively compared with groundtruth, but

the acquired disparity is used to register the medical images and further calculate the heart rate frequency. Furthermore, in other research, a multiple step method is employed to add timed 3D information to augmented reality to assist possible robot surgeons [5]. This approach uses a timed model of a heart obtained by CT or MRI scans. Then a 3D reconstructed operation surface using a correlation-based stereo algorithm [7] is registered with segments in the CT/MRI model. Finally, the location of coronary arteries is super-imposed on the recovered 3D structure. Unfortunately, there is no documented evaluation of the proposed method. Yet another study considered a dynamic programming optimization based stereo algorithm to calculate 3D disparity using calibrated stereo images [28]. Subsequently, another method is proposed to match the dense disparity maps to a 3D model of the surface of interest. However, the results were once again only qualitatively evaluated using registration accuracy from in vivo animal and patient data.

Overall, previous research suggests the potential applicability of computer stereo vision to medical surgery. However, it is limited by the lack of a representative database of groundtruthed imagery and consideration of a representative range of contemporary algorithms. As can be seen from previous research, no single effort, nor all previous work combined duplicates the dense groundtruth acquisition and coupled evaluation of the suite of stereo algorithms performed in this project.

## 1.3 Contributions

The significant contributions of this project to the state-of-the-art in computer vision are twofold. First, a significant and apparently unique database of stereo imagery with groundtruth relevant to medical surgical scenarios has been constructed. This database contains interesting and plausible scenes of actual organs and anatomical models. Second, and perhaps more significantly, the performance of a representative set of contemporary computer vision stereo algorithms has been evaluated on this dataset. These results are further analyzed to determine the deficiencies and strengths of each algorithm and also to propose new possible enhancements to tailor each algorithm to medical scenarios.

## 1.4 Outline

This report unfolds in the following fashion. This first section has motivated the study of the application of computer stereo vision to the medical domain and reviewed previous research. The second section explains the data acquisition methodology and the resulting image database. Section three documents the algorithms that were evaluated on the database. Following the database and algorithm documentation, the results of the empirical evaluation are presented. Finally, a summary is provided.

# 2 Data Acquisition

## 2.1 Methodology

Key to the recovery of our 3D groundtruth is the initial recovery of disparity maps for the acquired stereo image pairs. Here, a well known structured light approach [17] is employed, which has been used previously for groundtruth acquisition in other situations. In comparison to approaches that make use of laser scanned objects that are subsequently acquired in visible images, the use of structured lighting more readily provides visible imagery and groundtruth 3D information in common coordinate frames.

To enable groundtruth construction, binocular images are acquired for each scene. Groundtruth disparity is constructed for the images as follows. In addition to ambient illumination, separate images are acquired with structured lighting. The particular structured lighting that is used consists of binary (black/white) striped patterns projected by an LCD projector onto the scene. To distinguish N image positions, it is necessary to project $log_2$ N patterns, i.e., in such a manner a unique grey code is assigned to each pixel by concatenating binary values corresponding to whether the pixel is illuminated or not by each pattern.

With grey codes assigned to left and right images, disparity can be computed redundantly. First, left-to-right and right-to-left disparity can be computed through simple search for the unique match in the other image (with allowances made for nearly identical matches that can occur in practice). Second, disparity can be computed relative to each image/illumination combina-

tion (analogous to traditional structured lighting, but without requiring separate illumination calibration, as illumination projection matrices are determined from the pixel/illumination correspondences). Overall, $2N+2$ disparity maps result: N in relating each illumination to the left camera, N in relating each illumination to the right camera, 1 left-to-right camera correspondence and 1 right-to-left camera correspondence. Measurements at each pixel are combined using a robust approach to reject loci that do not yield consistent results and thus ensure high quality groundtruth. The result is a dense map of disparity with single pixel precision referenced to either camera view (left or right). Further details of this approach are well documented elsewhere, e.g., [17] and references therein.

The system was initially installed and working from a tower server machine along with an optical bench and multiple light projectors. Therefore this system was bulky and not portable, as shown in Figure 1. As part of this project it was necessary to make the system portable so image and data acquisition could be done in actual hospital environments. A considerable engineering effort ensued to reduce the system to comprise merely a calibrated pair of machine vision cameras and a single projector (both mounted on their own tripods) for control from a laptop computer. The whole system of image and groundtruth acquisition has been implemented as a single, self-contained program in our lab using C++ and OpenCV libraries. The final system setup that was used to acquire imagery and groundtruth in our lab and also outside of the lab is presented in Figure 2.

The stereo images are acquired with a 6 CM baseline and at a resolution

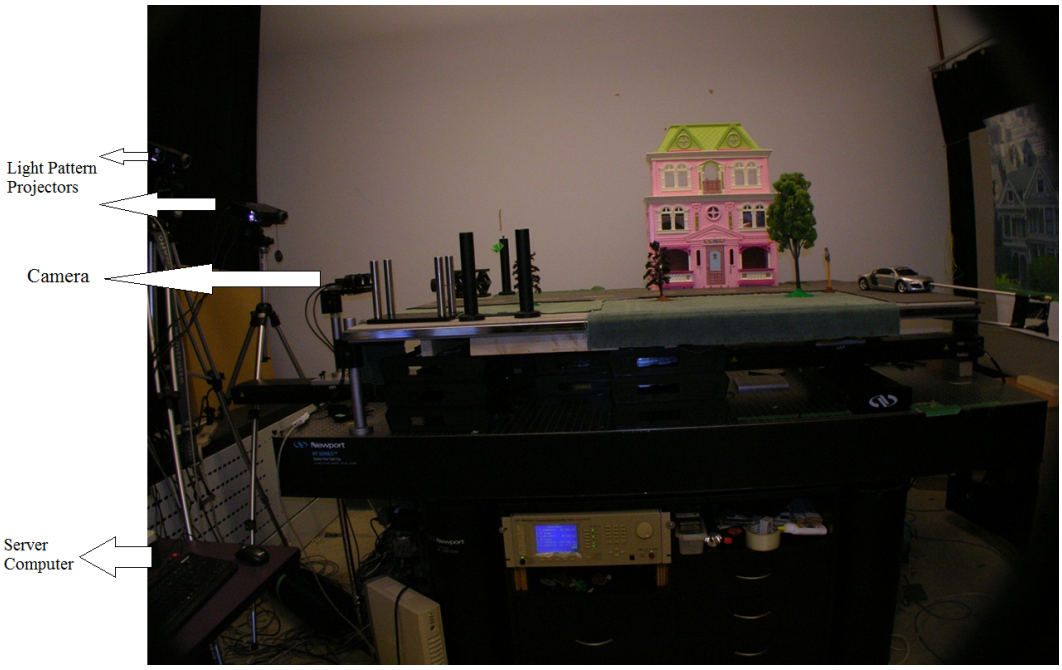


Figure 1: Original Stereo and Groundtruth Acquisition System.

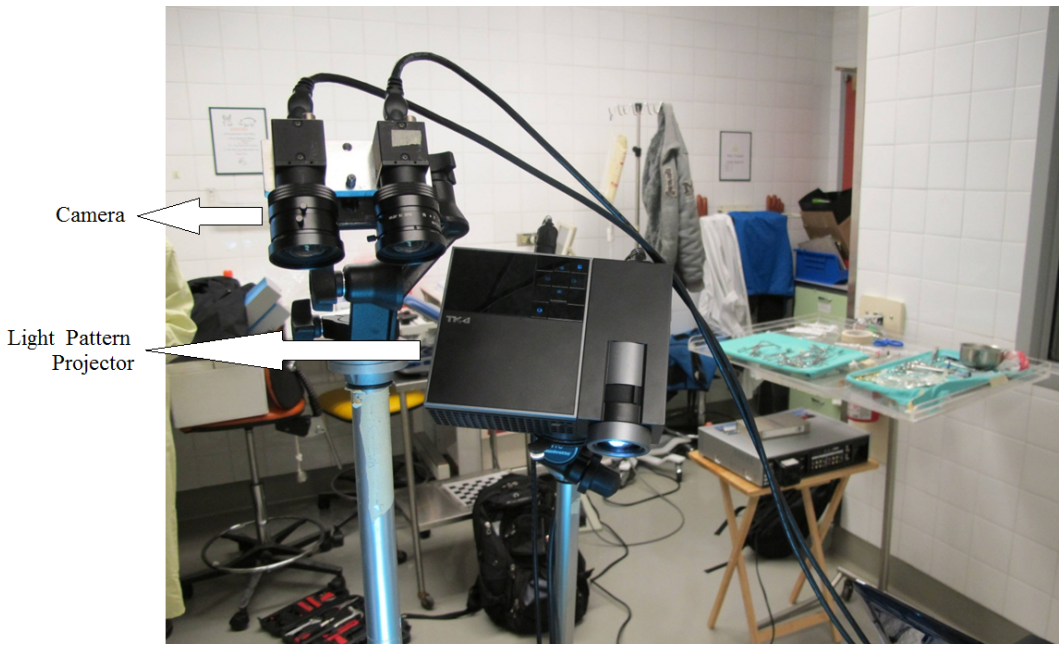


Figure 2: Modified and Portable Stereo and Groundtruth Acquisition System.

of $1024 \times 768$ pixels. 75 degree horizontal field of view lenses were used to represent the wide angle view usually used in medical surgery imagery. The time required to acquire a left-right stereo pair with groundtruth is approximately 90 seconds.

It is interesting to note that because the field of view of the camera and the light projector are not exactly the same, some points that are mutually visible to the cameras are occluded from the projector's view and hence do not produce groundtruth disparity. Significantly, our industry partner (MDA) has recently constructed a small, ultra-portable system that puts the camera and the projector in almost the same field of view. Furthermore, this new technology could potentially result in higher resolution imagery (up to $1600 \times 1200$) and is also faster in acquiring groundtruth. Our control software has been upgraded to work with this system; it has been evaluated and tested in the York Vision Lab and is ready to be used in the next stages of this project.

## 2.2 Datasets

The image database was acquired in two different settings. First, laboratory images were acquired at York. Second, images were acquired at the Hospital for Sick Children, Toronto (SickKids). All scenes in the acquired database were captured as a pair of calibrated binocular (left-right) images in 8 bit monochrome at $1024 \times 768$ spatial resolution. Associated with each pair is a pixel precision groundtruth disparity map, recovered according to the methodology described in Section 2.1. (Note that while projected light patterns were used to construct the groundtruth, the left-right database images are acquired

without light patterns.)

### 2.2.1 Lab Dataset

The lab images were acquired in the Vision Lab at York University using meat products acquired from a butcher shop. Meat products, indeed individual organs, were selected to construct the scenes as their shapes, surface reflectances and textures would be reasonably representative of what would be encountered in actual surgical scenarios. Organs acquired included lungs, heart, liver, kidneys and intestines.

The organs were arrayed in the following fashion. First, they were assembled into an anatomical model of the chest and abdomen region of a mammal; see top row of Figure 3. This set-up was selected to mimic what might be imaged during an actual surgery. Second, the heart, lungs, liver and kidneys were imaged separately. The heart was imaged both open and closed to reveal both its fine interior detail and smooth exterior. Individual organs were imaged to facilitate understanding of how the evaluated algorithms would perform on surgery focused on a particular organ. Thus, six different arrangements were considered (anatomical model, heart open, heart closed, lungs, liver and kidneys). In addition to the meat product organs, images of a liver phantom also were acquired, as they are often employed in conjunction with medical imaging studies.

All scenes were captured at three different distances (33 cm, 41 cm and 48 cm). Distance was measured from the stereo camera baseline to the nearest point in the viewed scene. The employed distances were selected to be rep-

resentative of possible placements of a camera in an actual operating room. Overall, 21 scenes were captured (6 organ arrangements, plus the phantom, all viewed at 3 distances) as specified in Table 1. Ambient overhead illumination in the lab was the only light source, beyond the light pattern projector, which was employed only during groundtruth acquisition. Figure 3 shows the dataset for the anatomical model at 33 cm. All datasets are shown in Appendix A.

Table 1: Lab Dataset

| Case # | Description | Distance |
|---|---|---|
| 1 | Full Anatomical Model | 33 cm |
| 2 | Full Anatomical Model | 41 cm |
| 3 | Full Anatomical Model | 48 cm |
| 4 | Heart Closed | 33 cm |
| 5 | Heart Closed | 41 cm |
| 6 | Heart Closed | 48 cm |
| 7 | Heart Open | 33 cm |
| 8 | Heart Open | 41 cm |
| 9 | Heart Open | 48 cm |
| 10 | Kidney | 33 cm |
| 11 | Kidney | 41 cm |
| 12 | Kidney | 48 cm |
| 13 | Liver | 33 cm |
| 14 | Liver | 41 cm |
| 15 | Liver | 48 cm |
| 16 | Lungs | 33 cm |
| 17 | Lungs | 41 cm |
| 18 | Lungs | 48 cm |
| 19 | Phantom Liver | 48 cm |
| 20 | Phantom Liver | 48 cm |
| 21 | Phantom Liver | 48 cm |

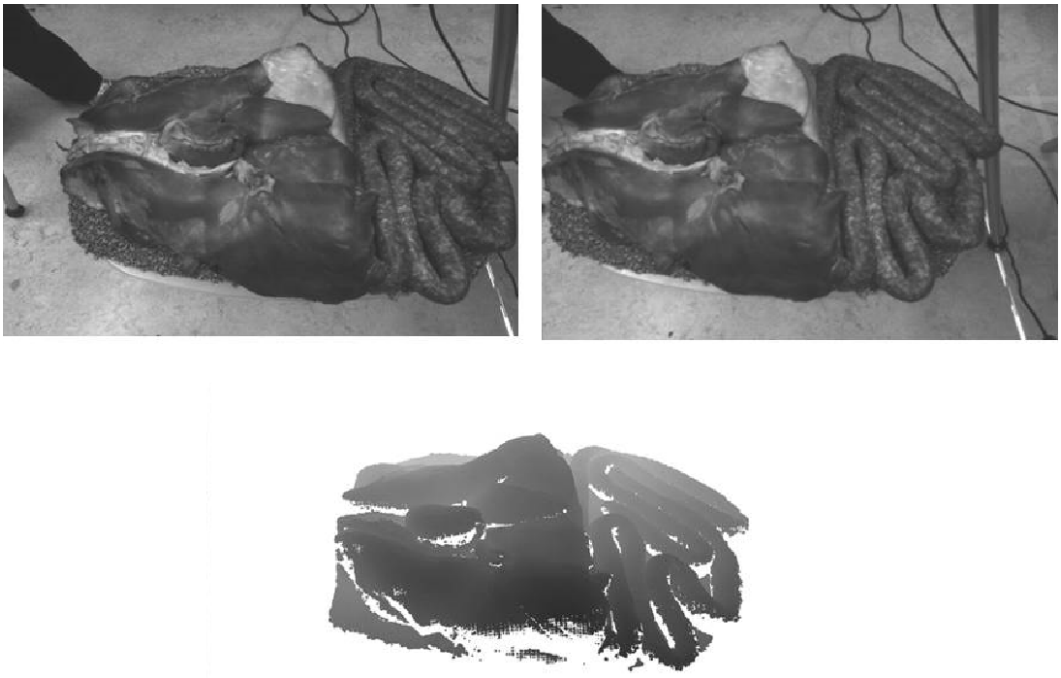

Figure 3: Anatomical Model at 33 cm Acquired in the Vision Lab at York University. The top row shows left and right camera views, the second row shows the left-based groundtruth disparity, with darker intensities depicting closer distances.

### 2.2.2 Hospital Dataset

Images of ex vivo porcine samples were acquired at SickKids hospital. In all cases, the samples were made available after the animal had been sacrificed for unrelated experiments. The datasets acquired consisted of a front anatomical view of the abdominal region, a back anatomical view of the abdominal region, a front view of the heart and lungs and a back view of the heart and lungs. Therefore, in total 4 datasets were acquired. All these datasets were acquired from a baseline to subject distance of 41 cm.

Figure 4 shows the dataset for the front anatomical model at 41 cm. All the datasets are shown in Appendix A.

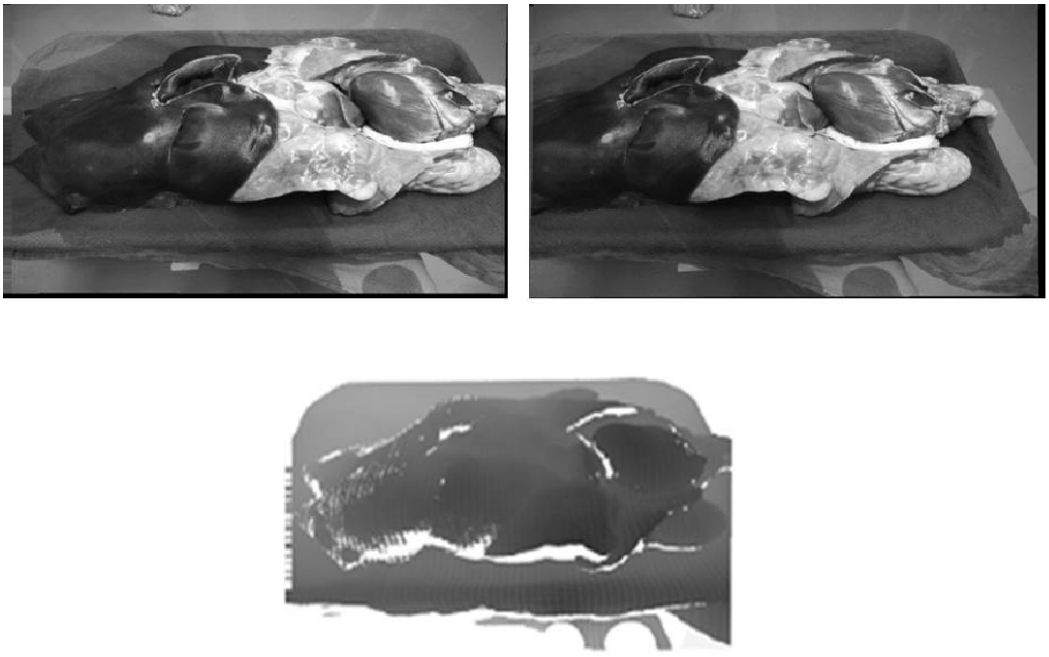

Figure 4: Anatomical Model of Abdominal Region at 41 cm Acquired in the Toronto SickKids Hospital. The top row shows left and right camera views, the second row shows the left-based groundtruth disparity, with darker intensities depicting closer distances.

### 2.2.3 Discussion

Visual inspection of both the lab and hospital portions of the acquired database shows that the the quality of the binocular images and groundtruth are of high quality. The binocular image pairs are in focus, with good dynamic range and the objects of interest are well framed. These observations hold across the range of viewing distances considered. The resulting groundtruth is quiet dense with good visual presentation of depth variation. In preliminary studies, the expected pixel precision of the disparity groundtruth was verified by having a human operator visually select corresponding left-right points sampled across

the images. Significantly, the depth variations are captured consistently across the range of viewing distances considered. More specifically, for lungs and livers the continuous surfaces are apparent in the groundtruth. In contrast, in the cases of heart open and intestines frequent and sometimes sudden depth variations are easily apparent and distinguishable with the naked eye. Further, the amount of occluded areas is kept to a minimum, with the current occluded areas being unavoidable (as explained in the Methodology section).

The quality of the groundtruth also seems to remain acceptable with different textures. For example, in the heart closed case an excessive amount of fat is apparent, which has a very bright surface that has been known to cause issues in other groundtruth acquisitions (e.g., because the image of the surface without the projector operating already is so bright that when the pattern is projected it is not readily discernible). However, here the recovered groundtruth remains smooth and acceptable even in the presence of the high reflectivity of the surface.

# 3 Stereo Algorithms

As discussed above, through years of research in computer science, there have been many stereo algorithms developed and there have been many papers comparing these algorithms in different scenarios. Standard taxonomies characterize these algorithms as either local or global in operation and thereby complexity, which ultimately impacts their speed and accuracy [18].

It was important for this project to consider a representative set of different

stereo algorithms. Correspondingly, a classic local block-matching algorithm [4] is considered and contrasted with a standard (arguably the best [26]) global matcher, graph cuts [3]. Still, this bimodal taxonomy does not reasonably capture other useful algorithmic instantiations. In particular, two additional considerations that have played a significant role in the design of stereo matchers that should be captured include the combination of local and global matching and the use of multiresolution and coarse-to-fine processing. Correspondingly, two more exemplars are included. First, the semiglobal stereo matcher is considered [10]. As its name suggests, this matcher can be seen as a blend of local and global approaches. Second, a coarse-to-fine matcher is considered [20]. As with all coarse-to-fine matchers, this algorithm makes use of initial coarse spatial resolution matching to guide subsequent finer resolution refinement; additionally, it makes use of adaptive windowing to ameliorate poor resolution of 3D boundaries, a standard shortcoming of multiresolution matching. In summary, four algorithms have been considered, as summarized in Table 2.

Table 2: Stereo Algorithms Evaluated

| Algorithm | Description |
|---|---|
| *NCC* | dense block matching [18] |
| *CFT* | coarse-to-fine adaptive block matching [19] |
| *SGM* | semiglobal matching [10] |
| *GC* | graph cuts stereo [3] |

The current study does not investigate the performance of different pointwise and area-based match metrics, as considerable previous investigations have concluded that real data requires normalization or rank-based measurements to get reliable results [11, 2]. Thus, all 4 algorithms rely on normalized

cross-correlation as their match measure computed over a $5 \times 5$ window, except NCC, which relies on $9 \times 9$ windows to obtain adequate match aggregation.

# 4 Empirical evaluation

## 4.1 Evaluation Methodology

Disparity recovered by the stereo algorithms is qualitatively and quantitatively compared to the groundtruth disparity. Here, evaluation is in terms of disparity (the spatial coordinate difference between matched binocular points), as it is the measurement that is recovered directly by stereo algorithms.

Qualitative evaluation comes in terms of two complementary visualizations. First, the recovered disparity maps will be presented; see Figure 5. This visualization allows for direct comparison between the recovered and groundtruth disparities. Second, difference maps between the recovered disparity and groundtruth will be displayed; see Figure 6. Difference maps help to isolate which portions of the acquired imagery are challenging the algorithms and thereby guide refinement efforts.

Quantification of performance will come in terms of three complementary measures. First, cumulative error distributions are calculated; see Figure 7. These statistics capture the proportion of points that lie within incremental error tolerances and thus are important in comparing algorithms according to the precision at which they can provide reliable estimates. Second, box plots will be used as a non-parametric way to characterize errors; see Figure 8. These plots allow for algorithms to be compared in terms of their overall error

distributions. Third, density plots will be used to characterize the proportion of points where valid estimates are returned; see Figure 9. Consideration of density is an important complement to the other statistics, as consideration of accuracy alone can become biased to algorithms that recover too few points to be of practical use. The full sets of qualitative visualizations and quantitative graphs for all datasets are shown in Appendix A.

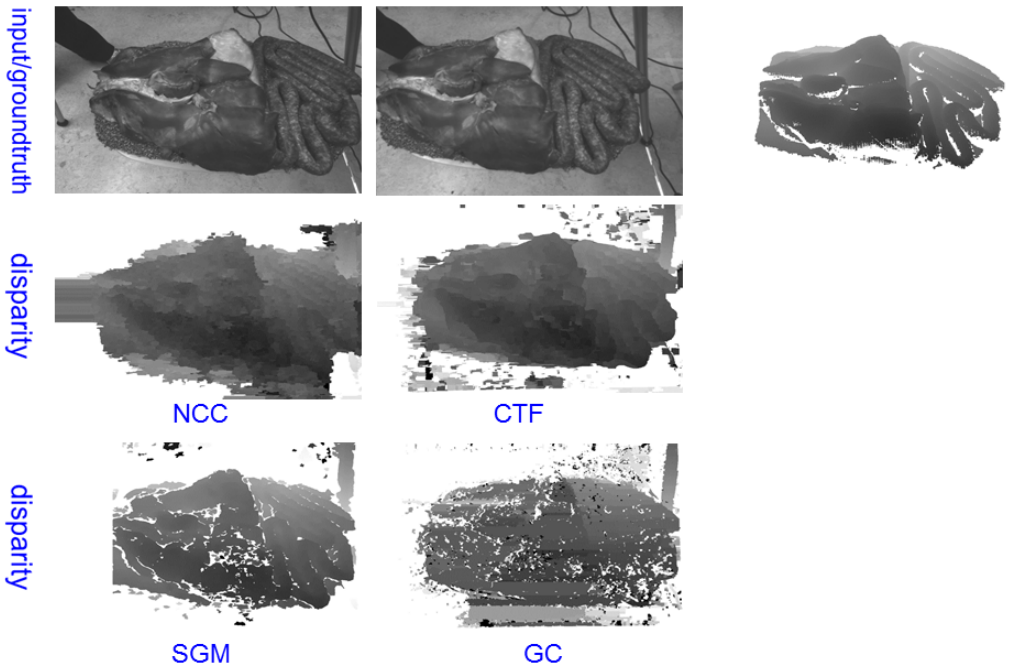


Figure 5: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Full Anatomical 33 cm Data Set.

In addition to producing quantitative plots for the anatomical model, individual organs and phantom at the 3 considered viewing distances, the data is also presented in 11 collapsed forms. These include 7 sets of graphs for

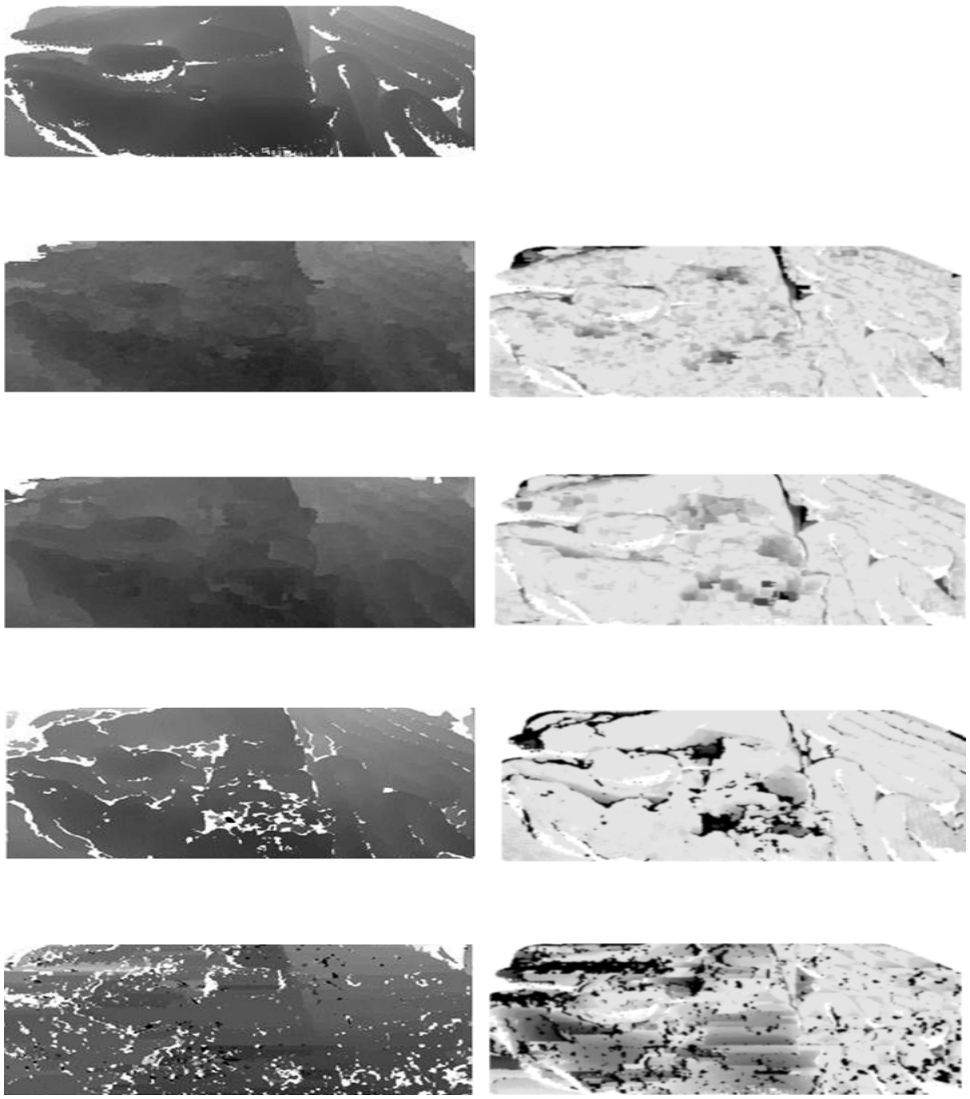

Figure 6: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Anatomical 33 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

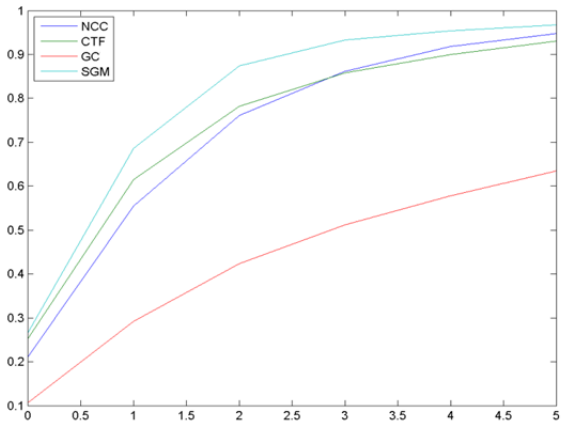


Figure 7: Cumulative Error Graph of Lab Anatomical 33 cm Data Set. The abscissa is the absolute pixel disparity error between the recovered disparity and the groundtruth disparity. The ordinate is the proportion of image points within each absolute disparity error.

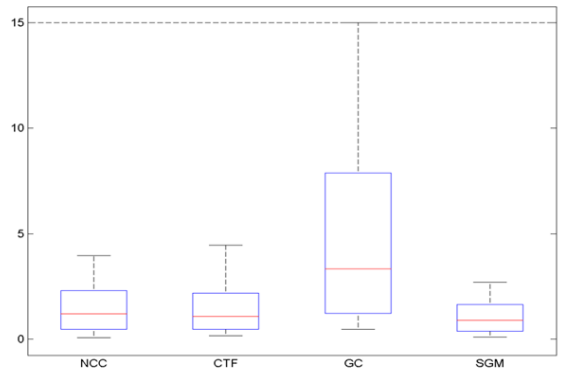


Figure 8: Box Plot of Lab Anatomical 33 cm Data Set. The ordinate is the absolute pixel disparity error between the recovered disparity and the groundtruth disparity. The bottom and top of the boxes show the 25th and 75th percentiles (resp.) and the red line shows the 50th percentile of image points in terms of absolute disparity error. The whiskers extend to the 10th and 90th percentiles below and above, respectively.

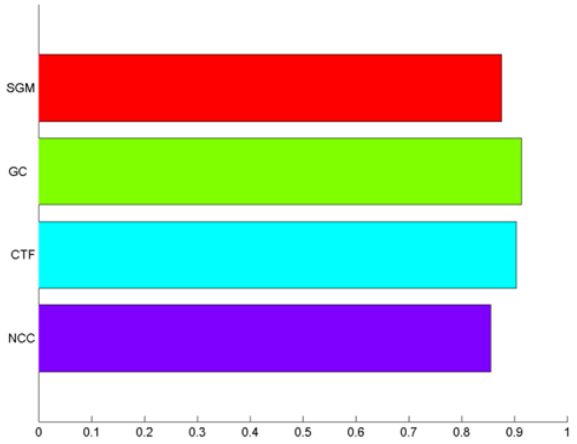


Figure 9: Density Plot of Lab Anatomical 33 cm Data Set. The abscissa is the proportion of image points where the stereo algorithm recovered a valid disparity value.

results collapsed across the 3 distances (i.e. one set for each of anatomical model, heart closed, heart open, kidney, liver, lungs and the phantom), 3 sets of graphs collapsed across organs excluding the phantom (i.e. one set for each camera baseline to subject distance of 33 cm, 41 cm and 48 cm) and a final graph collapsed across all the data sets excluding the phantom. The phantom data is not collapsed with the real organ data as it is not actual animal tissue. These results are presented in Appendix B.

## 4.2 Results

Figures 10-18 and 77 show results for the anatomical model captured at York. As can be seen from the overall results (Figure 77), SGM has the best performance in this scene followed closely by CTF and NCC. However, in terms of estimation density, CTF is the best in all 3 distances and SGM is one of the

worst. The full anatomical model consists of the intestine area, which has a non-specular surface that is rich in texture and is also mostly smooth. As can be seen from the disparity maps (Figures 10, 13, 16) and also the difference maps (Figures 11, 14, 17), all the algorithms perform well in this area at all three distances. Even GC, which has a poor overall performance, performs well in the intestine area. On the other hand, it is seen from the difference maps (Figures 11, 14, 17) that all the algorithms have issues with the middle area of the model. This area is where the liver is placed. The liver has a very highly specular surface and it also has a very weakly textured surface. These two properties together have resulted in poor performance of all three algorithms in this area. In particular, by looking at the anatomical model 33 cm results (Figures 10, 11) it is seen that even though SGM has a very strong performance overall, it has a poor performance in the liver area. The lungs area in the full anatomical model seems to have mediocre to good results with patches of disparity errors spread across the area. It is seen from the difference maps (Figures 11, 14, 17) that the only algorithm that performs particularly poorly in the lungs area is the GC algorithm. Finally, by looking at the box plots (Figures 12, 15, 18), it can be seen that all the algorithms perform better at a further baseline to subject distance. This result can be attributed to the fact that disparity scales inversely with distance; so, larger disparities and larger errors are present at closer distances.

Figures 19-27 and 78 show results for the heart closed captured at York. As can be seen from the overall results (Figure 78), the performance of all the algorithms is somewhat comparable in this scene. Nevertheless, SGM

still has the strongest performance at all 3 distances and GC has the worst performance. However, the estimation density of SGM is particularly low at all 3 distances compared to the other algorithms. By looking at the difference maps for the heart closed case at all 3 distances (Figures 20, 23, 26), it can be seen that all the algorithms have problems in the bottom area of the heart. By looking at the left and right images (Figures 19, 22, 25) it is seen that this area is the most specularly reflective area of the heart. In fact, it can be seen that the left and right images look differently in that area of the heart because of different reflection of light. The difference maps (Figures 20, 23, 26) show that in other areas of the heart the algorithms perform much better, even in the fatty regions near the top. The fat area has a rich texture and despite being bright, does not appear to be very specular; therefore, it does not seem to be a particular problem area for any of the algorithms. Finally, it is interesting that the performance of the algorithms does not differ much between the three distances for this scene.

Figures 28-36 and 79 show results for the heart open captured at York. As can be seen from the overall results (Figure 79) and also individually at each distance (Figures 30, 33, 36), once again SGM has the best performance in this scene. All the other 3 algorithms also perform strongly and have very similar performance. This is particularly interesting for GC, which so far has not performed up to par with CTF or even NCC. The SGM algorithm once again has the poorest estimation density among all the algorithms followed closely by NCC. Heart open has a mostly non-specular surface and also has a non-smooth surface. It is also very rich in texture (because of the exposed

arteries). This all has resulted in strong performance by all algorithms in this scene. By looking at the difference maps (Figures 29, 32, 35) it is seen that no particular problem areas exist in this scene. Finally, the performance of the algorithms do improve slightly as the baseline to subject distance increases, but not by much.

Figures 37-45 and 80 show results for the kidney captured at York. As can be seen from the overall results (Figure 80) and also for each of the three distances (Figures 39, 42, 45), all the algorithms perform very poorly at all three distances. SGM still has the best performance, but the difference is small. However, SGM once again has the worst estimation density followed closely by NCC. As can be seen from the actual images of the scene (Figures 37, 40, 43), the kidney has a highly specular surface. Moreover, the surface area of the kidney is very weakly textured. These two properties of the surface make it very difficult for any of the algorithms to perform well. It can be seen from the error plots (Figures 39, 42, 45) that the performance of the algorithms improve as the baseline to subject distance increases. However, as previously mentioned, this improvement is only because disparity scales inversely with distance; so, larger disparities and larger errors are present at closer distances. In general, if one looks at the difference maps for all three distances (Figures 38, 41, 44) it is seen that the algorithms have problems with the entire kidney area.

Figures 46-54 and 81 show results for the liver captured at York. Similar to the kidney, all the algorithms perform poorly in the liver case, except for SGM (Figure 81). However, the relatively strong performance of SGM comes

at the expense of very low estimation density across all three distances. The low density is shown not only in the density plot, but also the disparity maps (Figures 46, 49, 52) where the loci of unresolved disparities are indicated with white pixels. The areas where specular reflections are dominant can be seen in the raw left and right images (Figures 46, 49, 52); reference to the difference maps (Figures 47, 50, 53) shows that these areas are exactly those where the algorithms have most difficulties. For example, at the 33 cm distance (Figure 46), it is seen that the light is mostly reflected in the upper, left area of the liver. In the difference map (Figure 47) it is also seen this is the area where the bulk of the error is, while the lower-right area of the liver seems to be better estimated. Just like the kidney, the liver also has a very weakly textured surface, which adds to the difficulties, as the algorithms have little pattern structure on which to base a correct match. Finally, the performance of the algorithms improves once again as the baseline to subject distance increases. This improvement can again be attributed to the fact that at closer distances the disparity and its error are magnified.

Figures 55-63 and 82 show results for the lungs captured at York. It can be seen from the overall results (Figure 82) that SGM has the best performance followed by CTF. SGM also has the best performance at all the three distances as well, followed by CTF (Figures 57, 60, 63). The lungs surface is not as specular as the kidney or liver. Also, the lungs surface is mostly well textured. These physical properties of the lungs result in very good overall performance by all three algorithms for this organ. SGM once again has the poorest estimation density results (Figure 82), but the difference with the

other algorithms in this scene is slight. The one exception for the lung surface is the 48 cm distance, where the performance of the algorithms as well as the estimation density (especially for SGM) drops off (Figure 63). By looking at the actual images in the 48 cm distance (Figure 61), it is seen that there is actually more light specularly reflected off the surface of the lungs than the other 2 distances. By looking at the difference maps (Figure 62), it is seen that around the area where the light is specularly reflected (i.e. right side of the lungs) the errors are most concentrated. This state of affairs is especially noticeable in the NCC and SGM difference maps.

Figures 64-72 and 83 show results for the phantom liver captured at York. Here, it is seen that the overall results (Figure 83) indicate that SGM still has by far the best performance in this scene. This is also true for each distance (Figures 66, 69, 72). However, as consistently seen, SGM has the poorest estimation density followed closely by NCC. In general, all the algorithms have below average performance on the phantom. These results can be attributed to the fact that the phantom liver seems to have a very specularly reflective surface. In fact, if one looks at the left and right images in each case (Figures 64, 67, 70) the light specularly reflected off of the phantom is very evident. By looking at the difference maps (Figures 65, 68, 71) it can be seen that the problem areas arise in exactly the places where the light is most strongly specularly reflected. For example, in the 33 cm scene (Figure 64), it is seen that the upper portion of the phantom liver specularly reflects much light. By looking at the corresponding difference maps (Figure 65) it is seen that this is where the algorithms produced the most estimation error. Finally, the overall

performance of the algorithms seems to improve as the baseline to subject distance increases.

Figures 73-76 show results for the anatomical model and heart/lungs captured at the SickKids hospital. Owing to an error in calibration at the time of data acquisition at SickKids, these are only evaluated qualitatively. Nevertheless, it can be seen that like the lab datasets, in all hospital datasets both SGM and CTF seem to have the best performance. However, in the full anatomical model (both front and back) (Figures 73, 74), the SGM disparity results seems to have an inordinate number of drop outs (i.e. low estimation density), which could be attributed to the high specular reflectivity of the model, as consistent with the observations in the lab datasets. Furthermore, as can be seen from the left and right images on back and front views of heart and lung model (Figures 75, 76), the scene is much less specuarly reflective than the full anatomical model and has much more texture as well. These properties combine to result in improved performance and improved estimation density of all the algorithms in the heart and lung model, which is consistent with what was seen in the lab dataset results.

## 4.3 Discussion

As can be seen from the results collapsed across both organ and distance (Figure 87), SGM has the best overall performance followed by CTF. NCC and GC have the worst performance in these overall results.

NCC is the most basic algorithm evaluated in this project. The overall performance of NCC is also either the worst or the second worst in all datasets.

Despite its simplicity NCC also provided poor disparity density in all cases as well. NCC performs particularly poorly in the kidney and liver datasets. NCC generally fails in resolving 3D boundaries and this problem is particularly noticeable in the kidney at 33 cm, 41 cm and 48 cm (Figures 37-45). NCC does show some promising results in the full anatomical model case (Figures 10-18 and 77). As has been shown previously, NCC is capable in resolving medium-sized fairly textured objects [22].

The CTF algorithm evaluated here uses adaptive windowing for matching in the vicinity of 3D object boundaries. It has also been implemented to run in real-time [20, 19]. The disparity results produced by the CTF algorithm is shown to be very dense. Overall, CTF performs well in most cases. In particular, CTF is a strong performer in the full anatomical model at all three distances (Figures 12, 15, 18). The baseline to subject distance seems to have small effects on CTF and CTF performs to almost the same level at all three distances. However, CTF has some problem resolving low texture and specularly reflective areas. For example, CTF performs particularly poorly for the kidney and liver cases (Figures 80, 81) where the objects of interest have low texture and very specularly reflective surfaces. Notably, however, CTF is not alone in showing these limitations and all the algorithms perform poorly in these scenarios.

The SGM algorithm is another strong performer. The results produced by SGM are among the best in all cases. In particular, SGM is strong in the kidney and liver cases (Figures 80, 81) compared to the other algorithms. This observation suggests that SGM is a strong candidate to be used in cases

where high specular reflectivity is expected. However, the disparity estimation density is low for SGM and it is among the lowest in most cases. SGM has especially low density for kidney and liver, which suggests that the SGM algorithm only returns high confidence matches. Of course, in real medical scenarios high estimation density is important, but it is arguable that accuracy is more important than estimation density. That is, it is better to have no information about the disparity of a location than have wrong information. Therefore, the lower density higher accuracy approach of SGM is more desirable than low accuracy and high density results. This result, along with the fact that the algorithm has very low complexity and some real time implementations have appeared [6, 8], make SGM a very good algorithm choice to be further investigated/improved for medical stereo.

The GC algorithm is a global optimizer and it has been claimed by some to be the best general purpose performer [18, 4, 26]. However, in this evaluation GC is one of the worst performer in all the scenarios. It is evident that GC is hurt by its attempt to over-smooth. Alternatively, setting the GC parameters to smooth less yielded even worse results in preliminary evaluations. This state of affairs is particularly noticeable in the full anatomical model (Figures 10-18 and 77) where GC has poor performance. It is important to note that the optimization of the GC parameters is utterly important for each individual scene. This need for careful tuning, along with the fact that no real time implementation of GC exists, make GC the least desirable algorithm. It should be noted, however, that GC produced one of the best estimation densities in all cases, and the best overall average density of all the evaluated algorithms

(Figure 87).

It is evident that the kidney and liver datasets are the most problematic. All four algorithms fail to perform at a reasonable level in both these cases across all distances (Figures 80, 81). Looking at the left and right images for kidney (Figures 37, 40, 43) and liver (Figures 46, 49, 52) at all 3 distances it is obvious that both have very high specularly reflective surfaces and all stereo algorithms are known to have issues with such surfaces. Unfortunately, high specular reflectivity is expected in almost all medical imaging scenarios. Therefore, it is important that this issue is addresses by either a new algorithm or modification to existing algorithms. Interestingly, the estimation density also suffers the most in the kidney and liver cases. This is again most likely attributed to the high specular reflectivity of the surface in these scenarios, which limits the ability of the algorithms to establish any matches at all. Moreover, these surfaces also have weak texture; therefore, little information is available to define correct matches.

On the other hand, all algorithms perform well in the heart closed, heart open and lungs scenarios (Figures 78, 79, 82). These datasets consist of highly textured scenes with presence of lots of fat. This confirms that stereo algorithms perform better in high textured scenarios, as the local pattern structure defines correct matches. It also suggests that the algorithms have little problem in the presence of fat which obviously is promising as fat is expected in most medical scenarios.

The performance of the algorithms in the full anatomical model (Figure 77) is some where in the middle. This is expected as the performance suffers

in some organs (i.e., highly specular reflective surfaces such as the liver) and is good in other areas (i.e., low specularly reflective, high texture areas such as intestines). Looking at the difference maps for the anatomical model at all 3 distances (Figures 11, 14, 17) confirms these observations, as it is apparent that the intestine area has low error, while the liver and heart areas have higher errors.

By looking at the plots collapsed across the organs for all 3 distances (Figures 84-86), it is seen that the relative performance of the algorithms compared to each other remains mostly the same between difference baseline to subject distances of the same scene. However, it is evident that all the algorithms generally perform better at a further baseline to subject distance. As noted previously, this result can be attributed to the fact that disparity scales inversely with distance; so, larger disparities and larger errors are present at closer distances. In real medical scenarios the algorithms need to cover adequately all relevant distances; therefore, this issue needs to be addressed in any new algorithm or improvements to current algorithms.

From the above discussion, it can be suggested that the first thing that needs to be addressed in future work is better and denser disparity estimation in the presence of highly specular reflective surfaces. This problem, as magnified in the kidney and liver cases, is likely a contributor to below ideal performance of all algorithms on all datasets. Second, while most medical scenarios involve reasonably textured scenes, the possibility of low textured scene (such as the liver) is present and therefore needs to be addressed as well. Finally, to be of any practical use in real medical surgery scenarios, any al-

gorithm modifications proposed need to produce accurate disparity across the entire range of representative distances.

# 5 Summary

In summary, this report has helped advance the field of computer vision in application to medical surgery. This advance has been achieved by acquiring a database of stereo image pairs and 3D (disparity) groundtruth, representative of medical surgery, and further evaluating a representative range of stereo algorithms with respect to the acquired imagery. The representative imagery and associated groundtruth were obtained in the York Vision Lab and during visits to SickKids hospital. The stereo algorithms tested (NCC, CTF, SGM and GC) were evaluated both qualitatively and quantitatively on the acquired database.

In general, the results suggest that computer stereo vision technology has potential for application to medical scenarios; however, advances are required to realize this potential. More specifically, the results indicate that SGM and CTF generally outperform both NCC and GC. Indeed, in the presence of well textured, relatively matte surfaces all the algorithms recover estimates that are in reasonable agreement with groundtruth. However, the performance of all the algorithms was subpar in areas of specular reflection on all datasets. Furthermore, all algorithms generally performed better at a further camera baseline to subject distance. Finally, as seen in previous evaluations (e.g., [18]), algorithms performed better in scenes with rich texture as opposed to

those containing little surface detail. It can be concluded that current stereo algorithms are insufficient to deal with circumstances and properties that are common in surgical scenes. Significantly, however, this reported research has been able to narrow down the potential shortcomings of extant algorithms; and thereby provides a good starting point for further enhancement of these algorithms to better handle stereo medical imagery. For example, recent advancements in stereo video processing have shown promise in recovery of surface shape in the presence of specular reflections [21] and should be considered for incorporation into future developments of stereo for surgery.

## Acknowledgements

Thanks to S. Himidan, P. Jasiobedski, B. Ma and A. Simpson for valuable comments on the research that is documented in this report. The research was supported by OCE under TPS and ORF under CIVDDD.

# References


[1] Bernhardt, S., Abi-Nahed, J. and Abugharbieh, R.: Robust dense endoscopic stereo reconstruction for minimally invasive surgery. In Proceedings of Workshop on Medical Computer Vision (2012).

[2] Bleyer, M., Chambon, S.: Does color really help in dense stereo matching? In Proceedings of International symposium 3D data processing, visualization and transmission (2010).

[3] Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (11) (2001) 1222-1239.

[4] Brown, M.Z., Burschka, D., Hager, G.D.: Advances in computational stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (8) (2003) 993-1008.

[5] Devernay, F., Mourgues, F., Coste-Manier, E.: Towards endoscopic augmented reality for robotically assisted minimally invasive cardiac surgery. In Proceedings of Medical Imaging and Augmented Reality Workshop (2001).

[6] Ernst, I., Hirschmller, H.: Mutual information based semi-global stereo matching on the GPU. Advances in Visual Computing. Springer Berlin Heidelberg (2008) 228-239.

[7] Faugeras, O., Hotz, B., Mathieu, H., Vieville, T., Zhang, Z., Fua, P., Theron, E., Moll, L., Berry, G., Vuillemin, J., Bertin, P., Proy, C.: Real time correlation based stereo: algorithm implementations and applications. Technical Report 2013, INRIA Sophia-Antipolis, France (1993).

[8] Gehrig, S.K., Rabe, C.: Real-time semi-global matching on the CPU. In Proceedings of the IEEE Computer Vision and Pattern Recognition Workshops (CVPRW) (2010).

[9] Gong, M., Yang, R.,Wang, L., Gong, M.: A performance study on different cost aggregation approaches used in real-time stereo matching. International Journal of Computer Vision 75 (2) (2007) 283-296.

[10] Hirschmuller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2005).

[11] Hirschmuller, H., Scharstein, D.: Evaluation of stereo matching costs on images with radio-metric differences. IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (9) (2009) 1582-1599.

[12] Hu, M., Penney, G., Figl, M., Edwards, P., Bello, F., Casula, R., Rueckert, D., Hawkes, D.: Reconstruction of a 3D surface from video that is robust to missing data and outliers: Application to minimally invasive surgery using stereo and mono endoscopes. Medical Image Analysis 16 (3) (2011) 597-611.

[13] Lau, W., Ramey, N., Corso, J., Thakor, N., Hager, G.: Stereo-based endoscopic tracking of cardiac surface deformation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (2004) 494-501.

[14] Penne, J., Holler, K., Sturmer, M., Schrauder, T., Schneider, A., Engelbrecht, R., Feuner, H., Schmauss, B., Hornegger, J.: Time-of-flight 3-D endoscopy. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention Society (2009) 467-474.

[15] Ramey, N.A., Corso, J.J., Lau, W.W., Burschka, D., Hager, G.D.: Real time 3D surface tracking and its applications. In Proceedings of the IEEE Computer Vision and Pattern Recognition Workshop (2004).

[16] Rhl, S., Bodenstedt, S., Suwelack, S., Kenngott, H., Mller-Stich, B.P., Dillmann, R., Speidel, S: Dense GPU-enhanced surface reconstruction from stereo endoscopic images for intraoperative registration. The International Journal of Medical Physics Research and Practice 39 (3) (2012) 1632-1645.

[17] Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2003).

[18] Scharstein, D., Szeliski, R.: Taxonomy and evaluation of dense two-frame stereo algorithms. International Journal of Computer Vision 47 (1-3) (2002) 7-42.

[19] Sizintsev, M., Kuthirummal, S., Samarasekera, S., Kumar, R., Sawhney, H., Chaudhry, A.: GPU accelerated realtime stereo for augmented reality. In Proceedings of the 3D Data Processing, Visualization and Transmission Conference (2010).

[20] Sizintsev, M., Wildes, R.P.: Coarse-to-fine stereo vision with accurate 3D boundaries. Image and Vision Computing 28 (3) (2010) 352-366

[21] Sizintsev, M., Wildes, R.P.: Spatiotemporal Oriented Energies for Spacetime Stereo. In Proceedings of the IEEE International Conference on Computer Vision (2011).

[22] Sizintsev, M., Wildes, R.P.: Stereoscopic Datasets and Algorithm Evaluation for Driving Scenarios. York University Technical Report CSE-2013-06 (2013).

[23] Spinczyk, D., Karwan, A., Rudnicki, J., Wroblewski, T.: Stereoscopic liver surface reconstruction. Videosurgery Miniinv 7 (2012) 181-187.

[24] Stoyanov, D., Darzi, A., Yang, G.: A practical approach towards accurate dense 3D depth recovery for robotic laparoscopic surgery. Computer Aided Surgery 10 (4) (2005) 199-208.

[25] Stoyanov, D., Visentini-Scarzanella, M., Pratt, P., Yang, G.Z.: Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (2010).

[26] Tappen, M.F., Freeman, W.T.: Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters. In Proceedings of the IEEE International Conference on Computer Vision (2003).

[27] Tombari, F., Mattoccia, S., Di Stefano, L.: Stereo for robots: Quantitative evaluation of efficient and low-memory dense stereo algorithms. In Proceedings of the International Conference on Control Automation Robotics & Vision (2010).

[28] Vagvolgyi, B., Su, L., Taylor, R., Hager, G.: Video to CT registration for image overlay on solid organs. In Proceedings of the Augmented Reality in Medical Imaging and Augmented Reality in Computer-Aided Surgery Conference (2008) 78-86.

[29] Vaudrey, T., Rabe, C., Klette, R., Milburn, J.: Differences between stereo and motion behaviour on synthetic and real-world stereo sequences. In Proceedings of International Conference on Image and Vision Computing New Zealand (2008).

[30] Wu, C., Sun, Y. and Chang, C.: Three-dimensional modeling from endoscopic video using geometric constraints via feature positioning. IEEE Transactions Biomedical Engineering 54 (7) (2007) 1199-1211.

# Appendices

## A Datasets and Stereo Algorithm Results

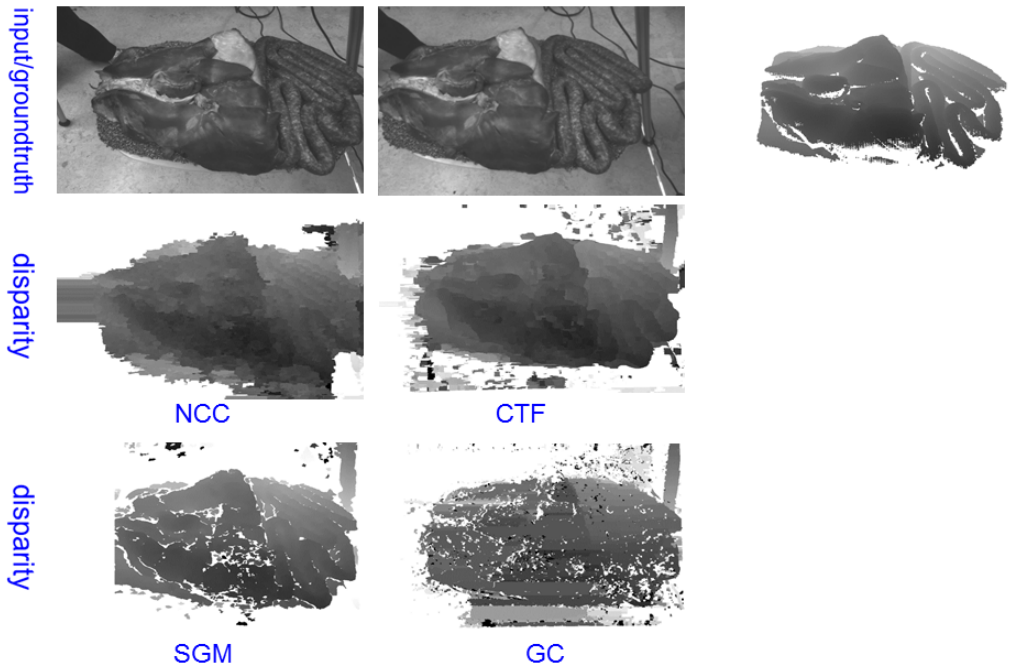


Figure 10: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Full Anatomical 33 cm Data Set.

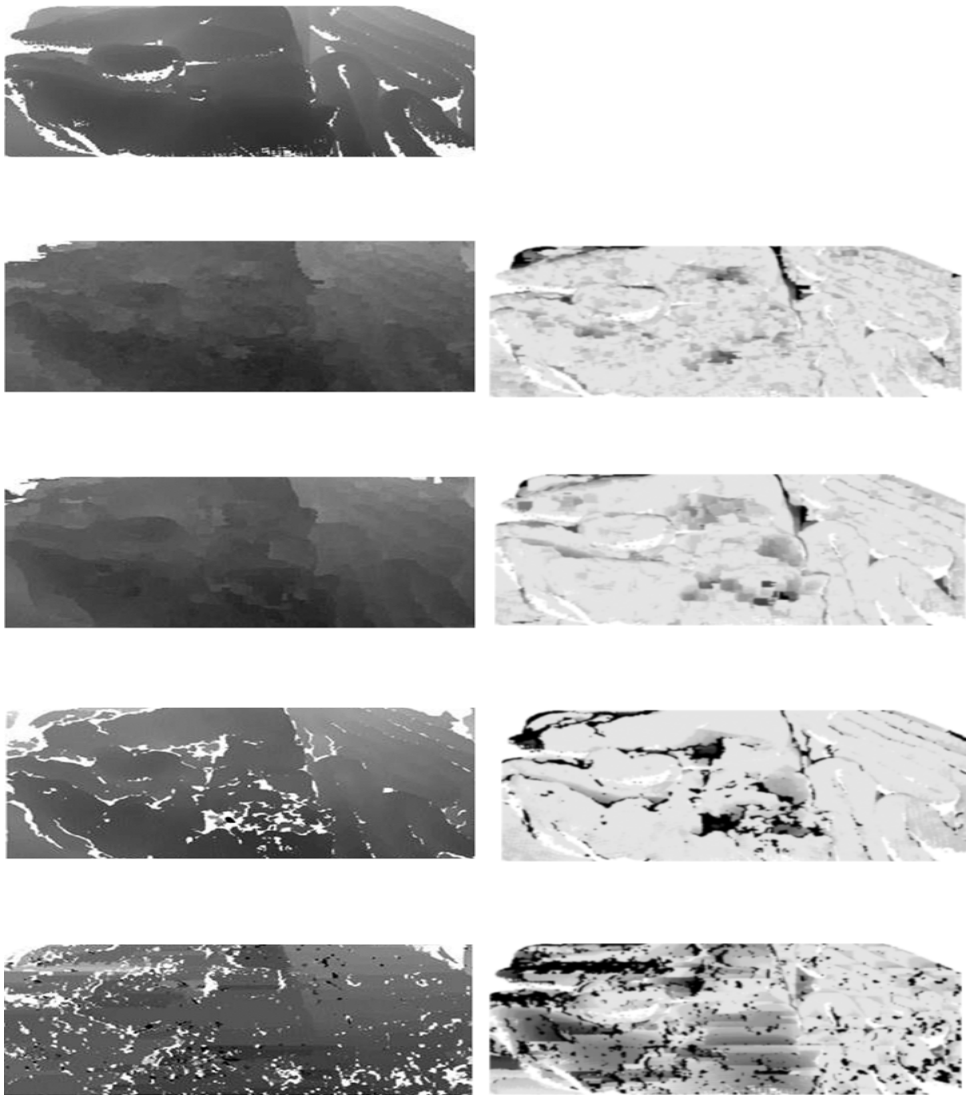

Figure 11: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Anatomical 33 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

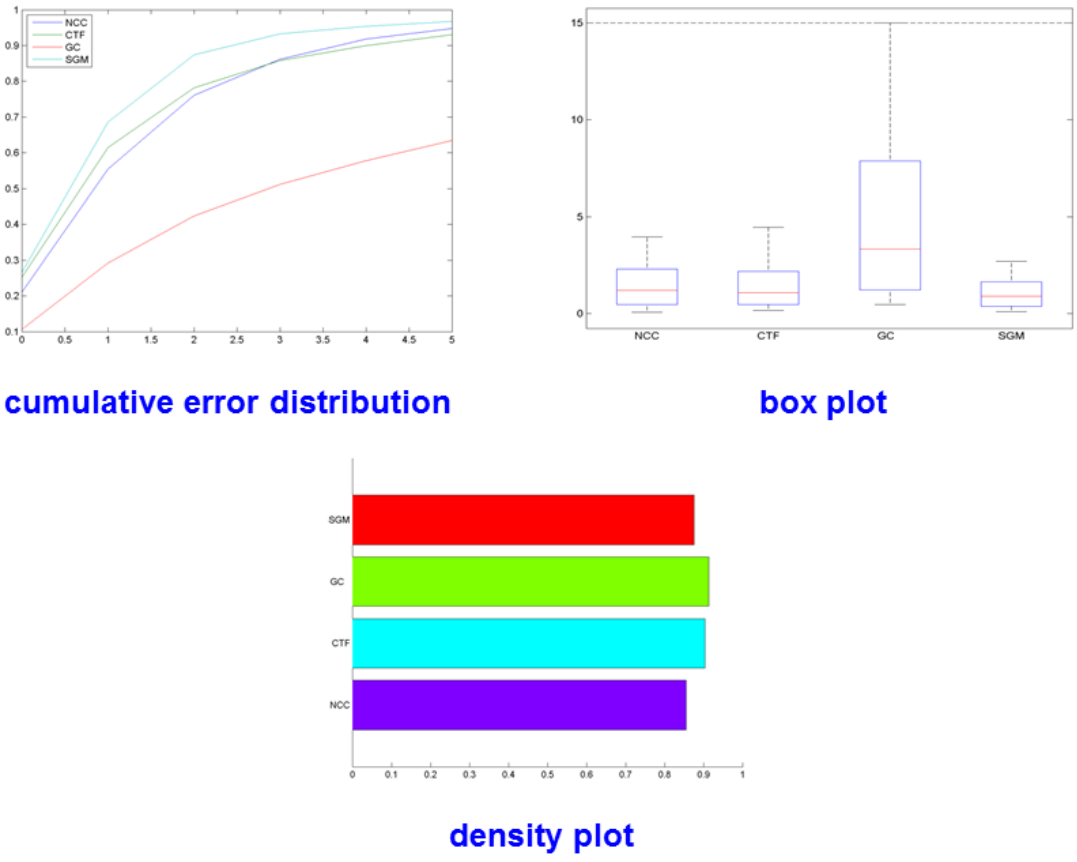


Figure 12: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Full Anatomical 33 cm Data Set.

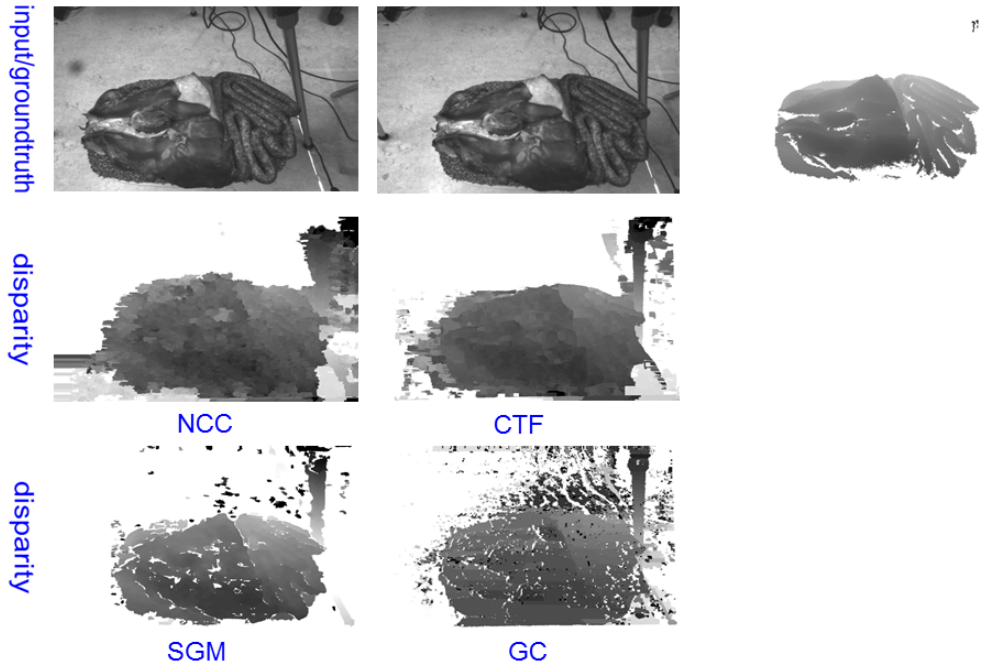


Figure 13: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Full Anatomical 41 cm Data Set.

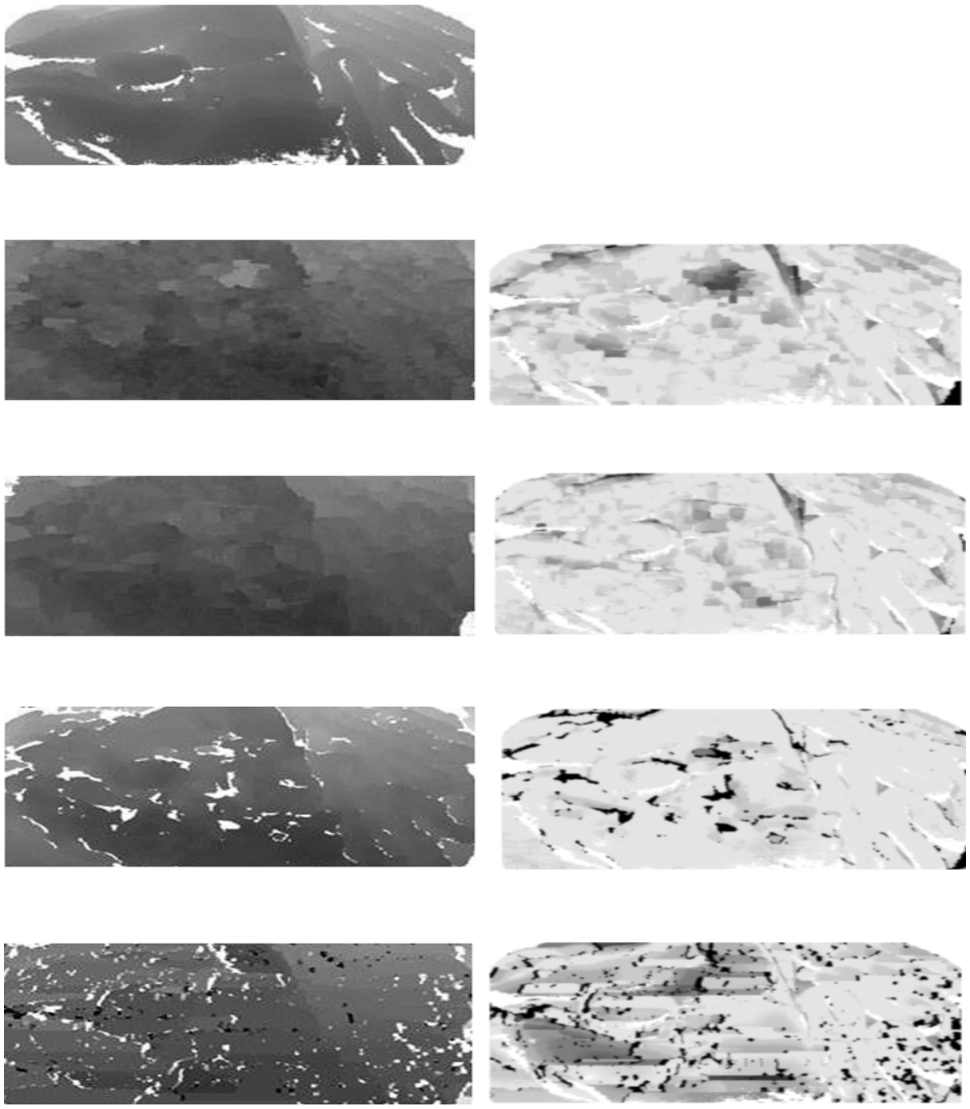

Figure 14: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Anatomical 41 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

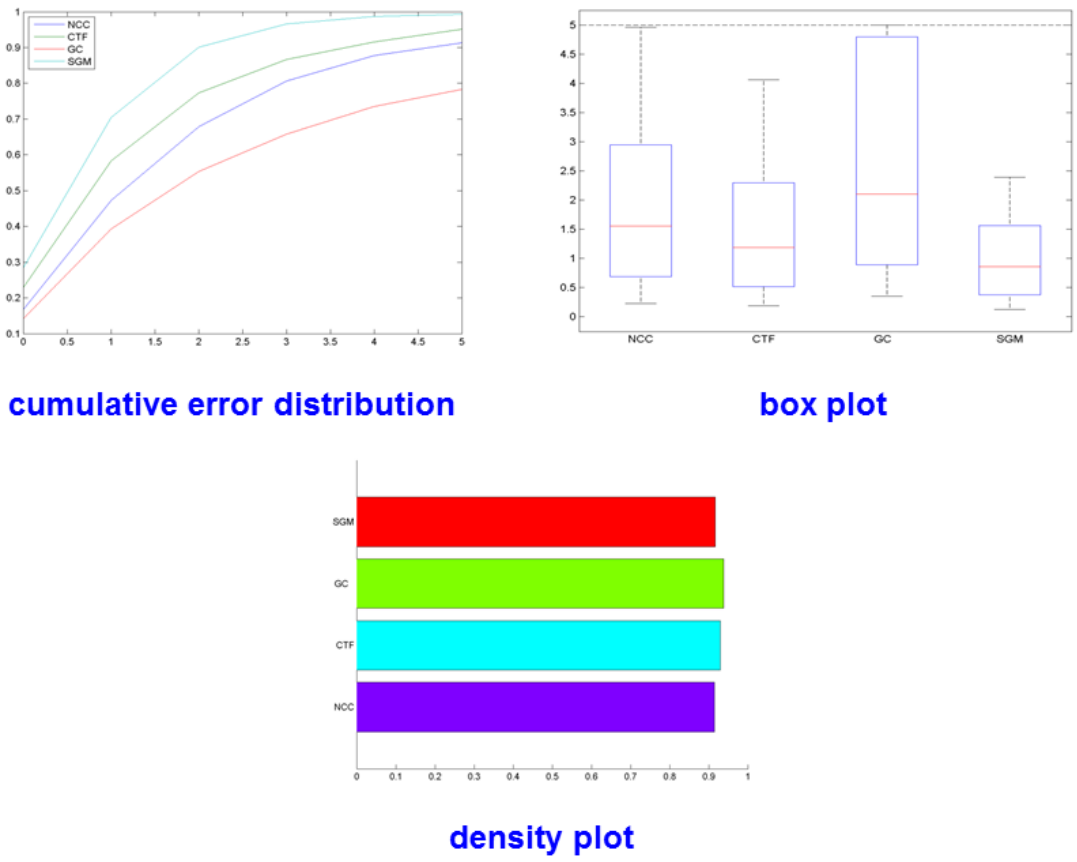


Figure 15: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Full Anatomical 41 cm Data Set.

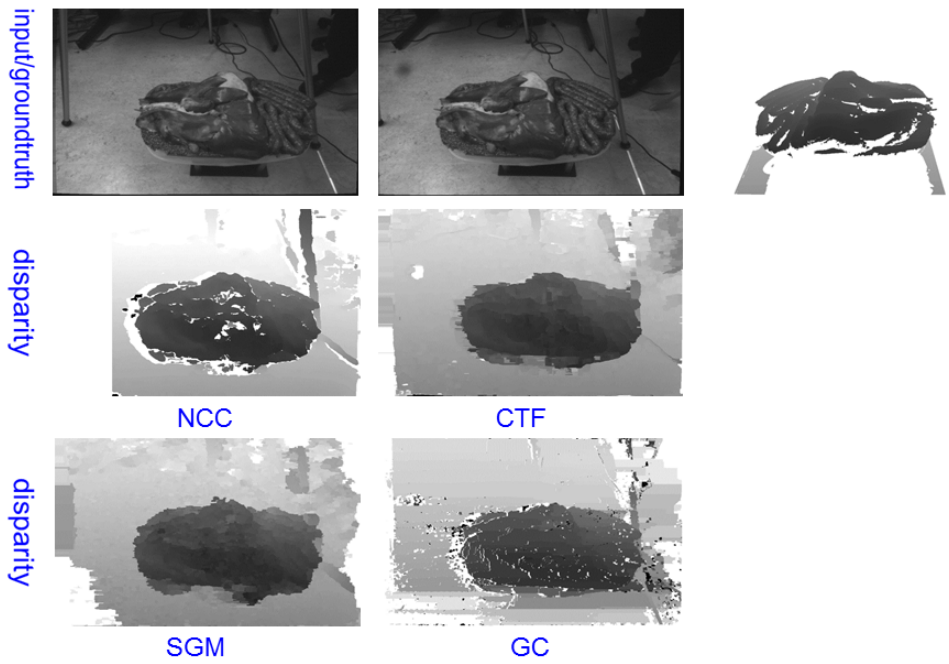


Figure 16: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Full Anatomical 48 cm Data Set.

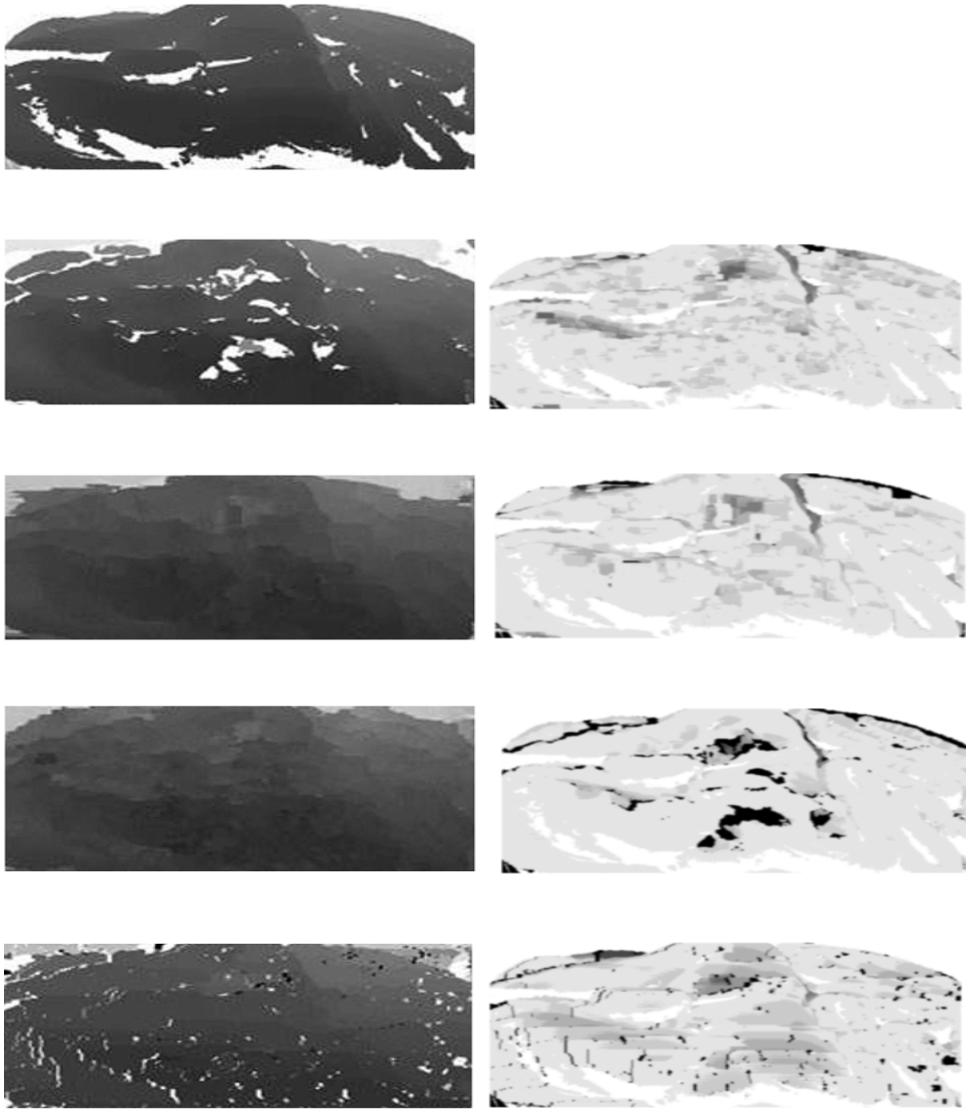

Figure 17: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Anatomical 48 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

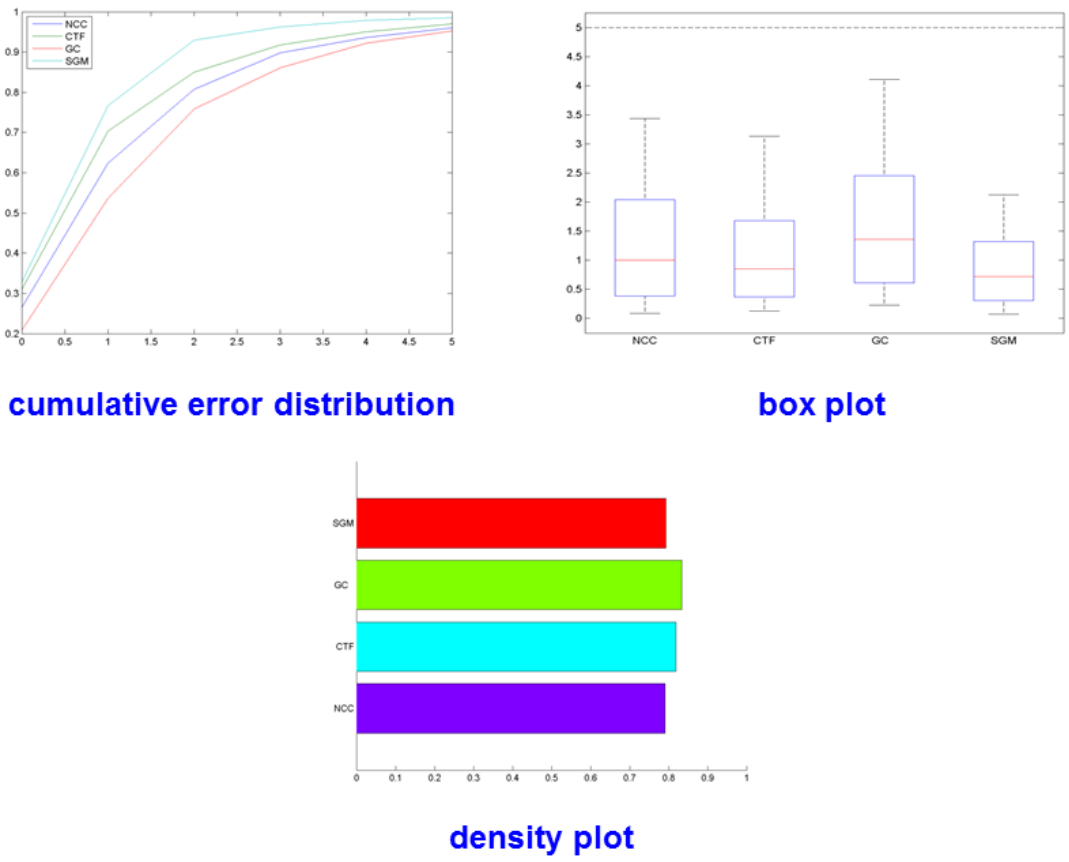


Figure 18: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Full Anatomical 48 cm Data Set.

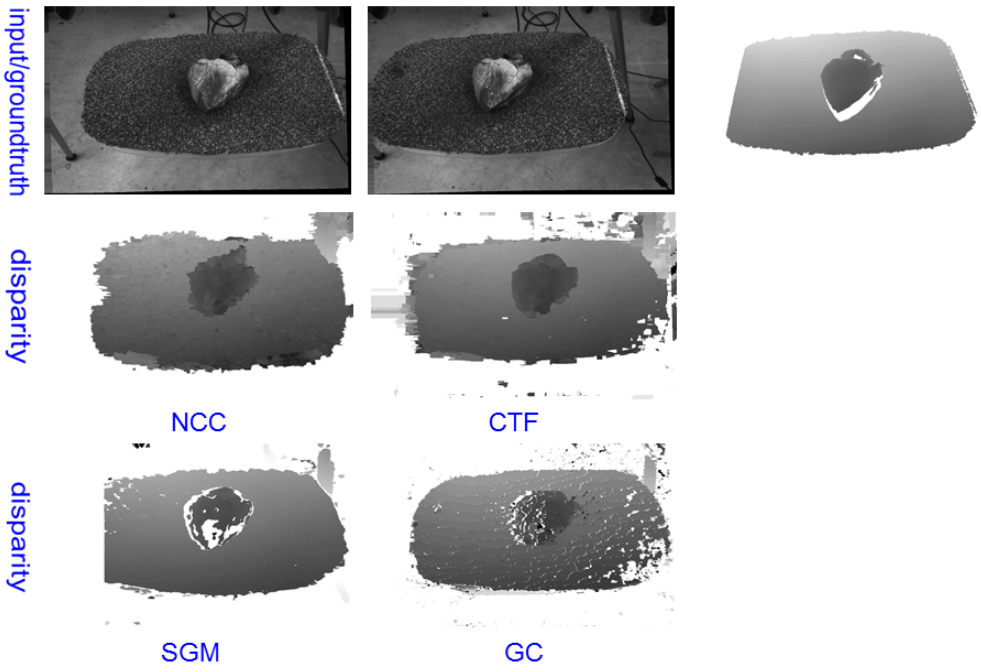


Figure 19: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Heart Closed 33 cm Data Set.

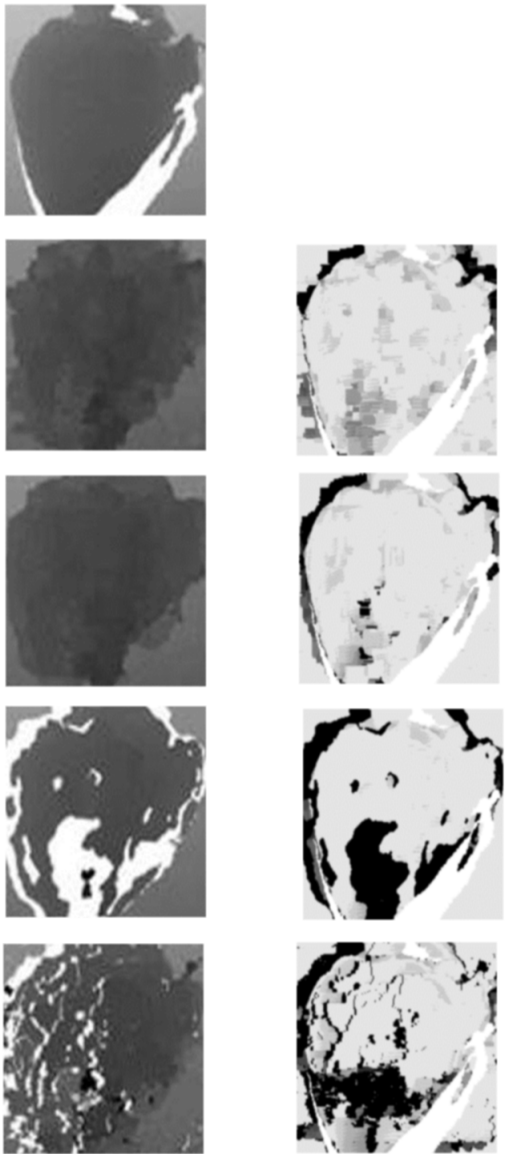

Figure 20: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Heart Closed 33 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

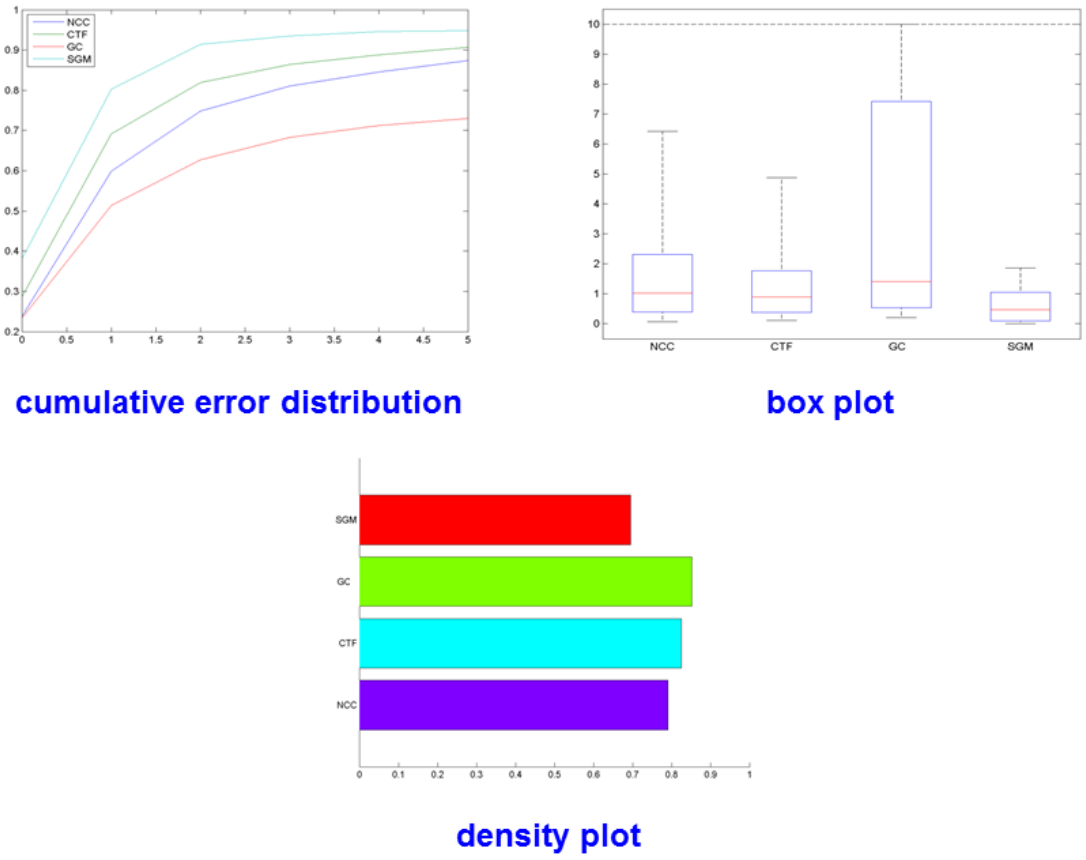


Figure 21: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Heart Closed 33 cm Data Set.

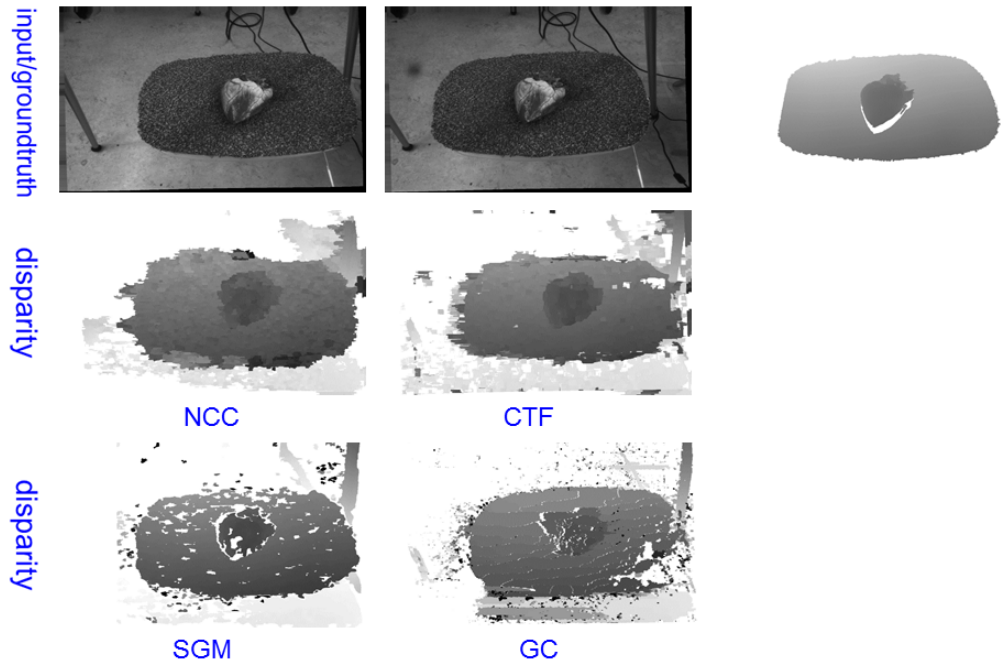


Figure 22: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Heart Closed 41 cm Data Set.

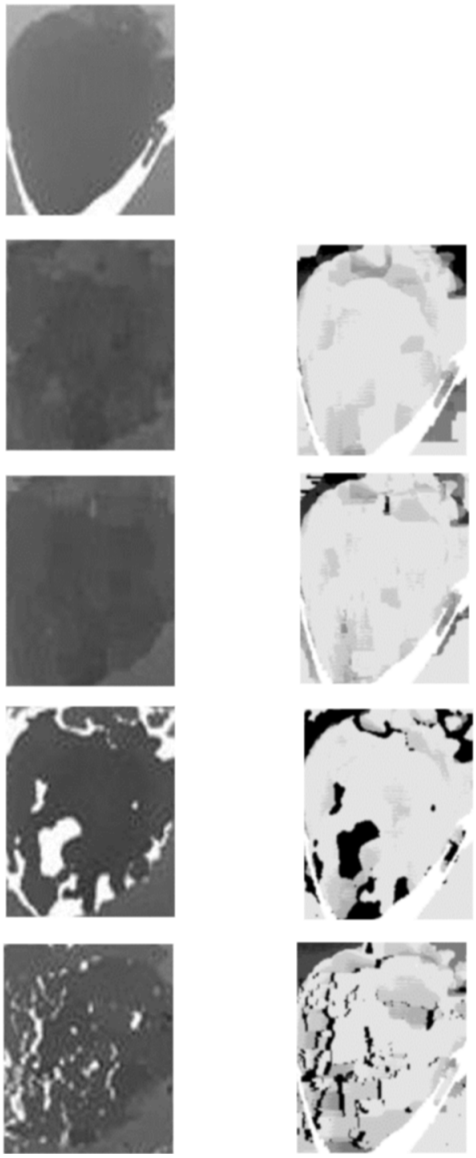

Figure 23: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Heart Closed 41 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

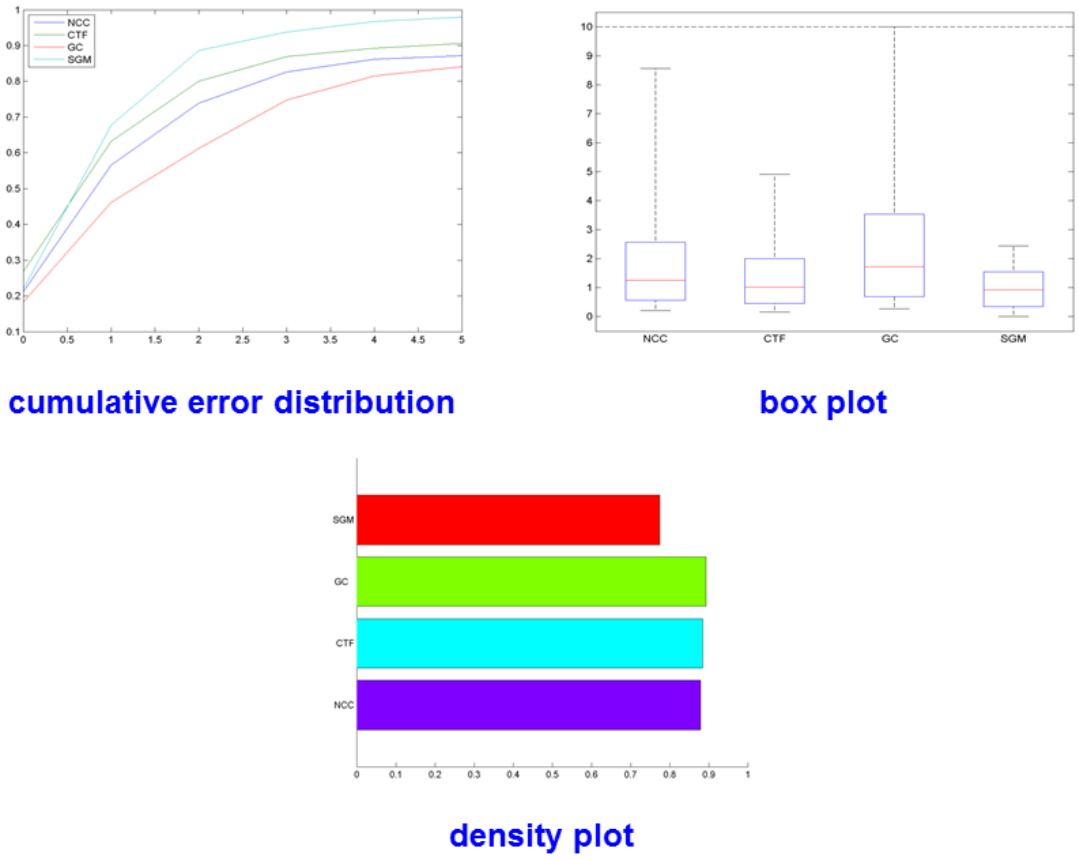


Figure 24: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Heart Closed 41 cm Data Set.

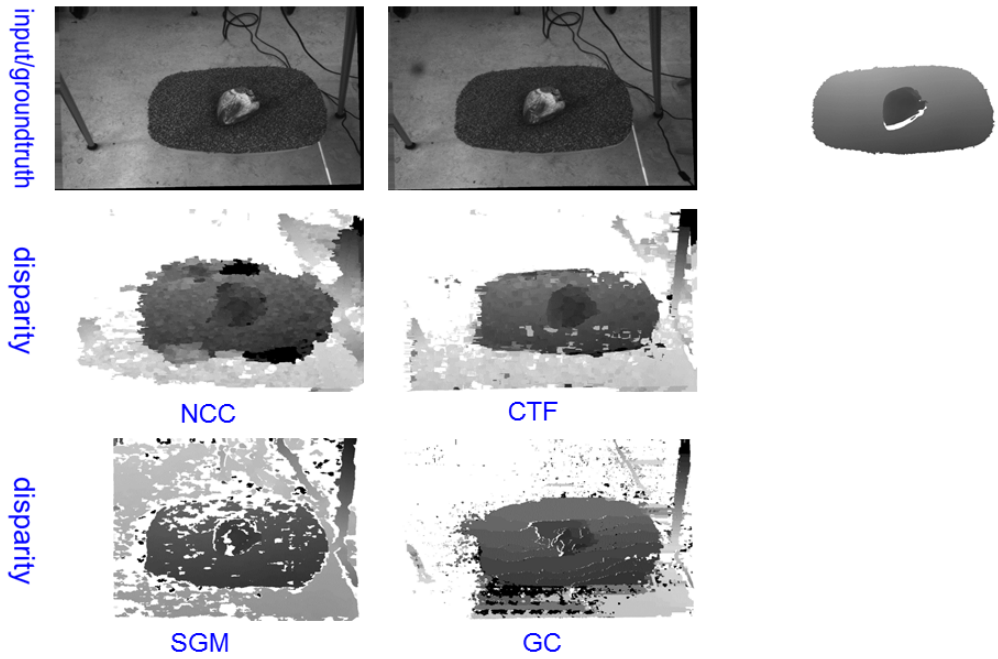


Figure 25: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Heart Closed 48 cm Data Set.

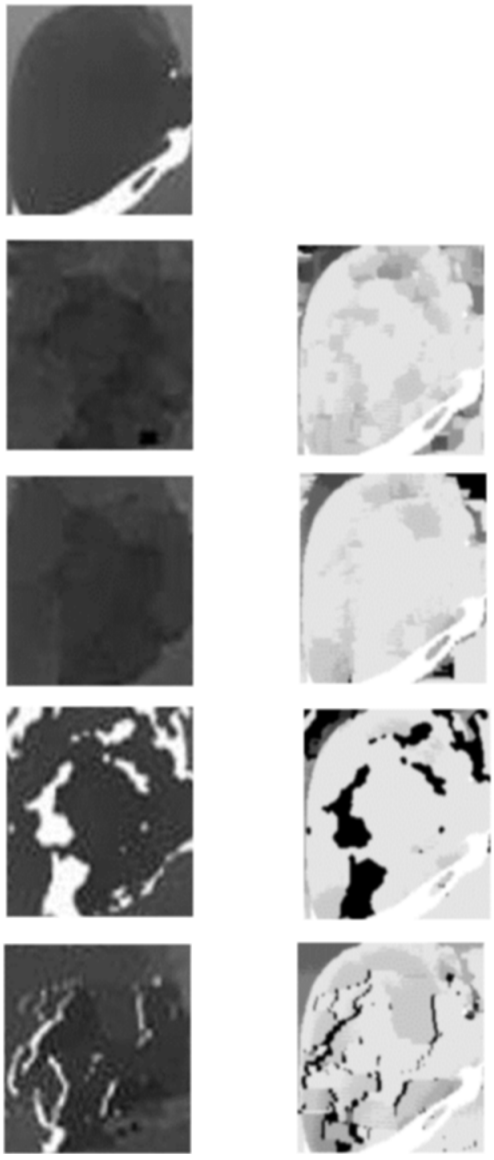

Figure 26: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Heart Closed 48 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

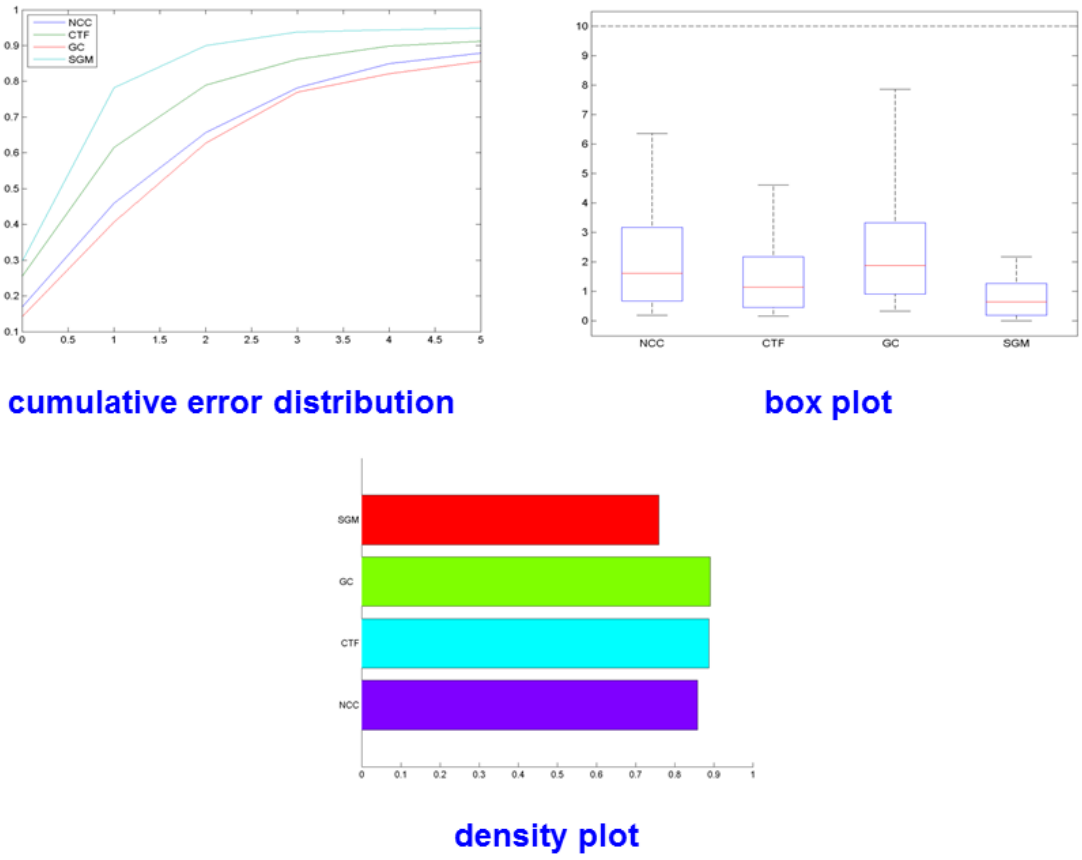


Figure 27: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Heart Closed 48 cm Data Set.

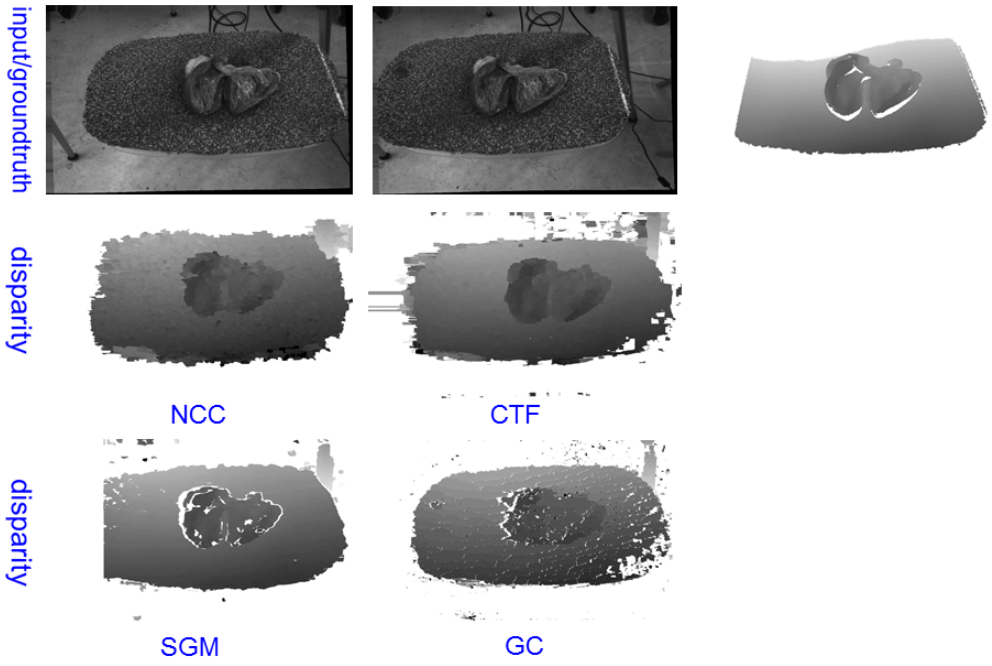


Figure 28: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Heart Open 33 cm Data Set.

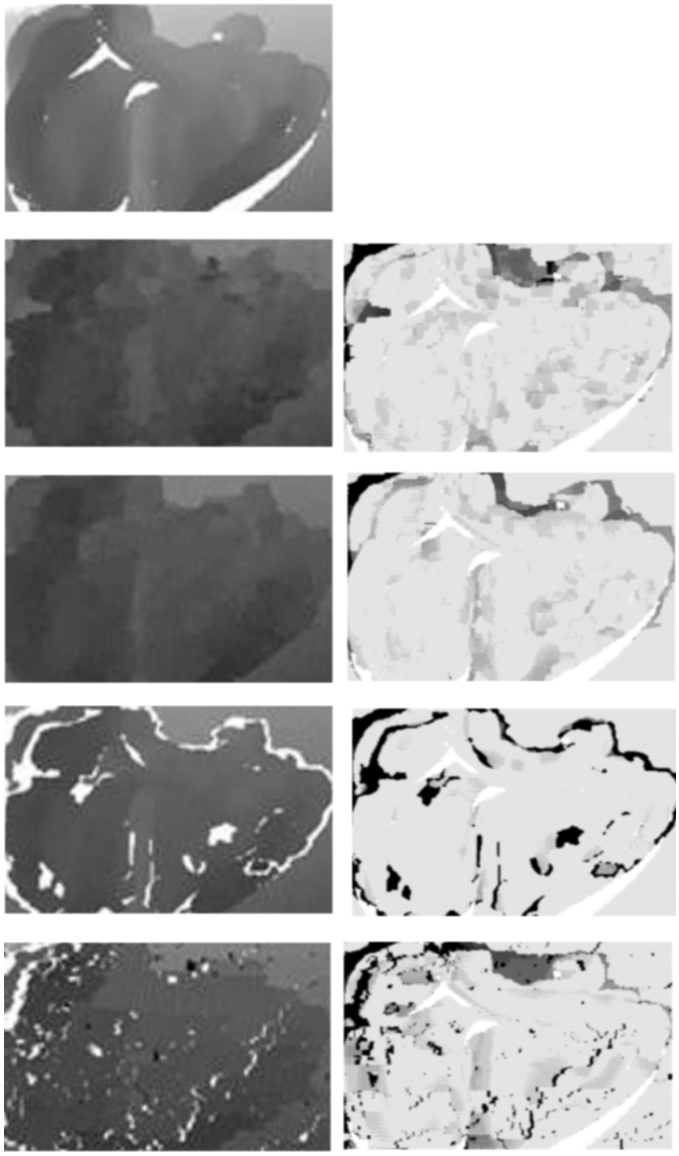

Figure 29: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Heart Open 33 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

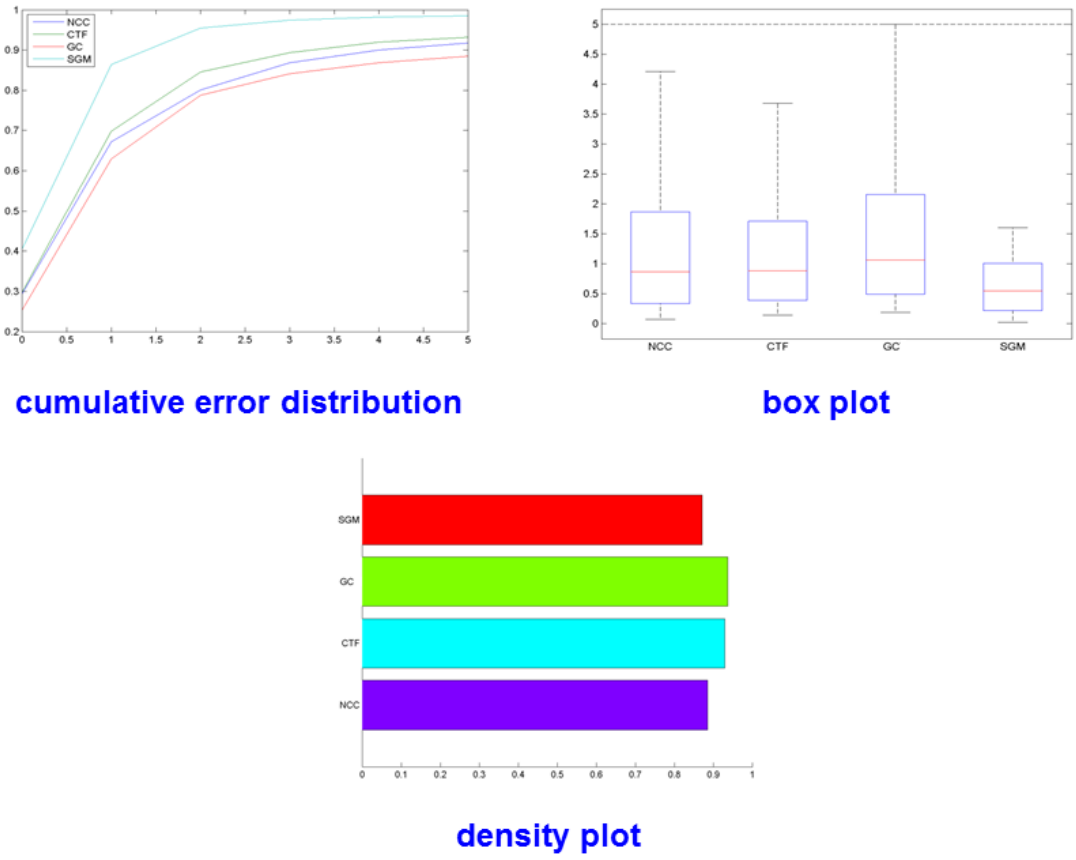


Figure 30: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Heart Open 33 cm Data Set.

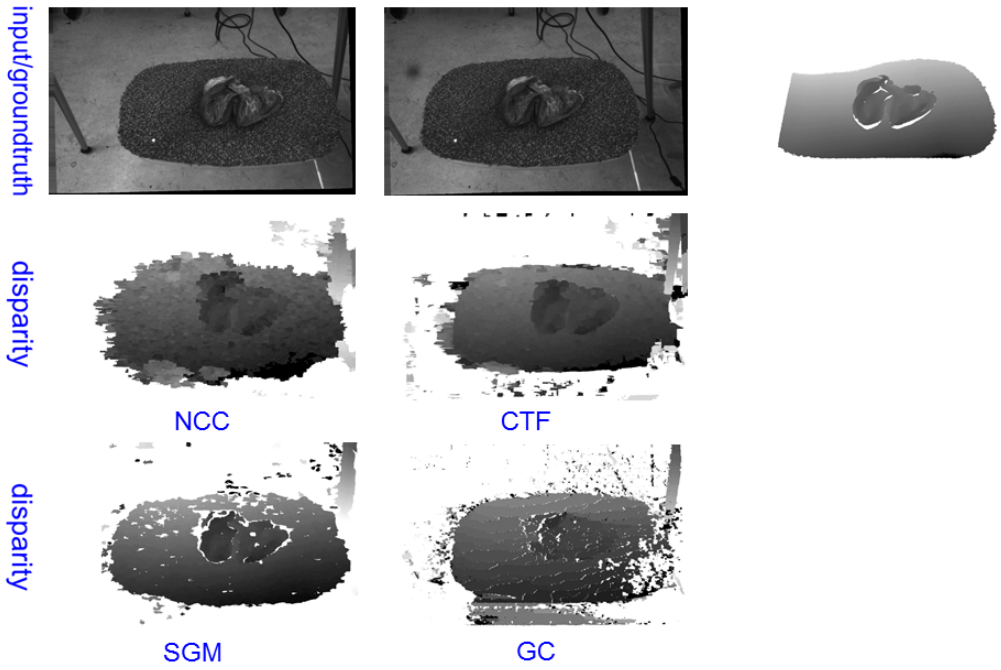


Figure 31: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Heart Open 41 cm Data Set.

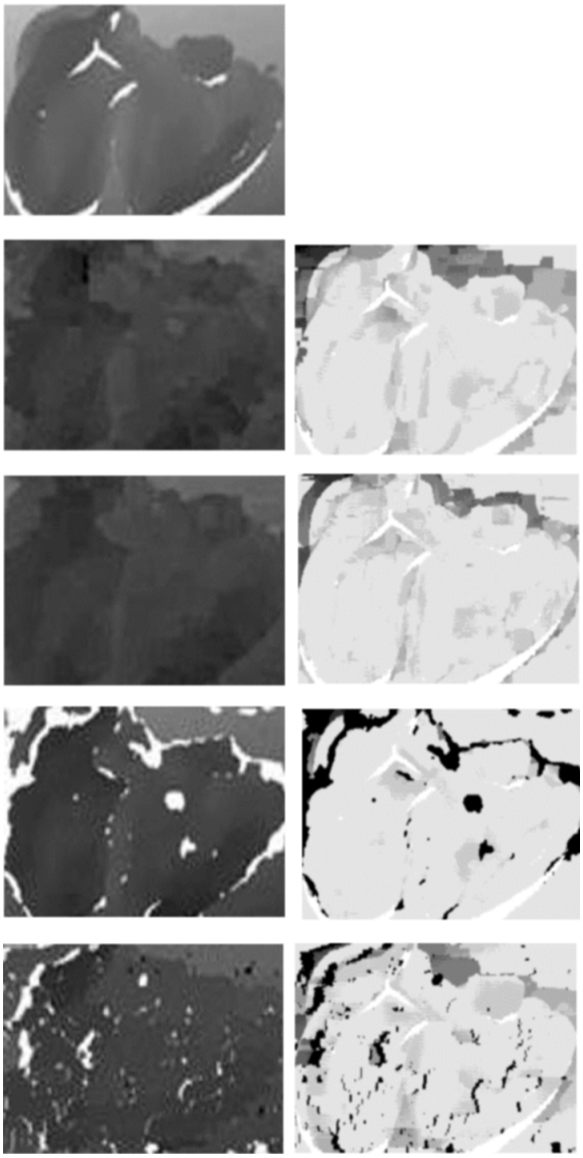

Figure 32: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Heart Open 41 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

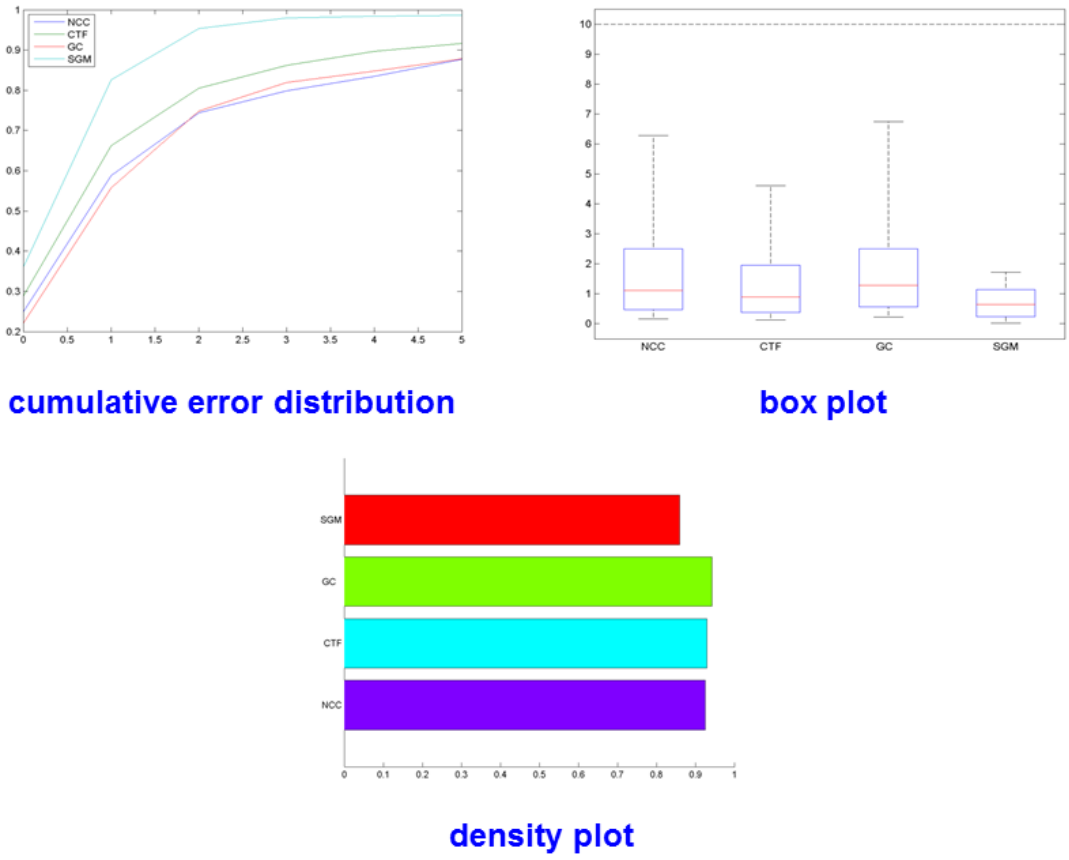


Figure 33: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Heart Open 41 cm Data Set.

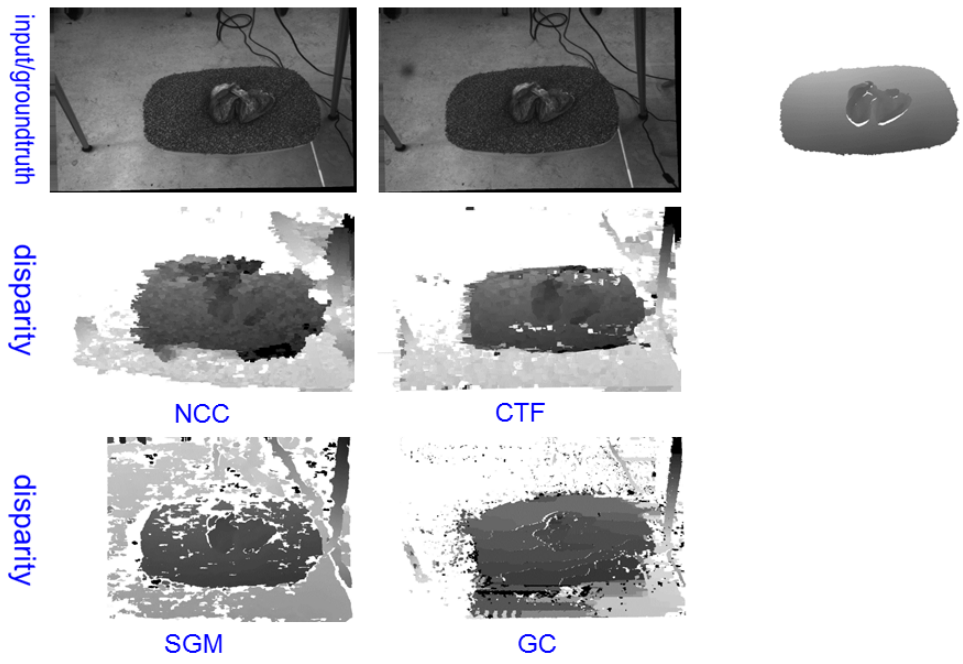


Figure 34: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Heart Open 48 cm Data Set.

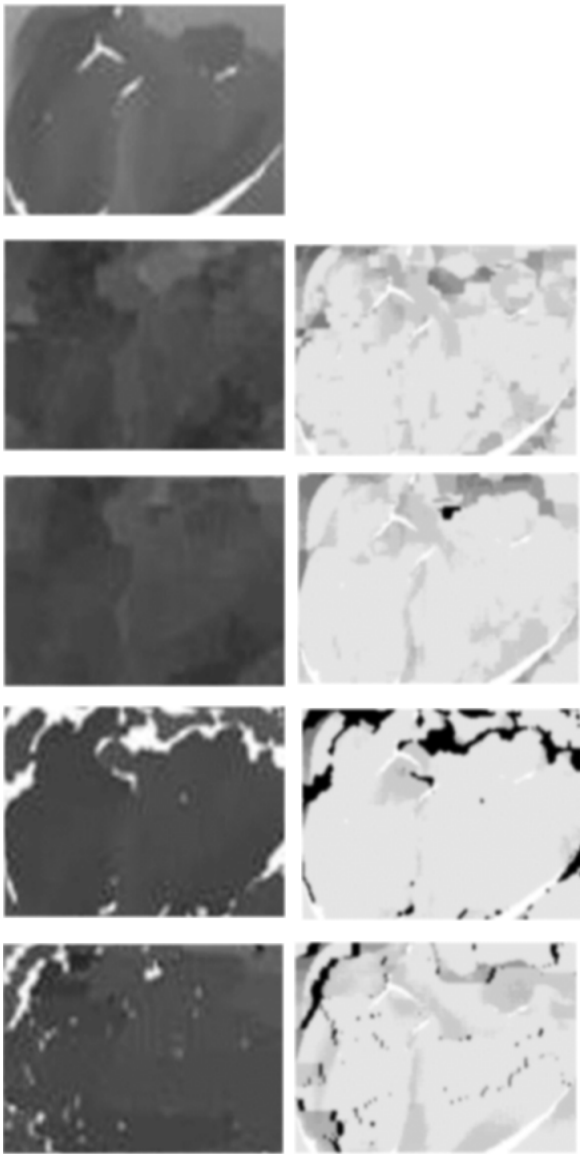

Figure 35: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Heart Open 48 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

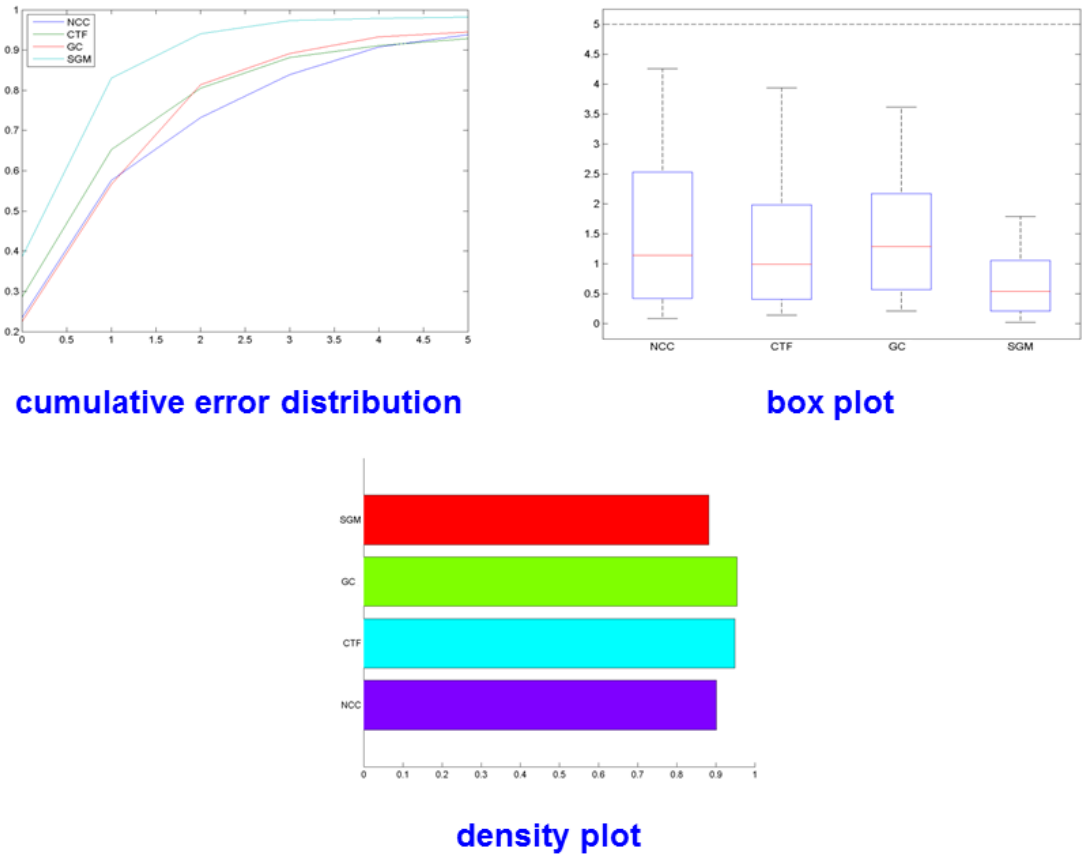


Figure 36: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Heart Open 48 cm Data Set.

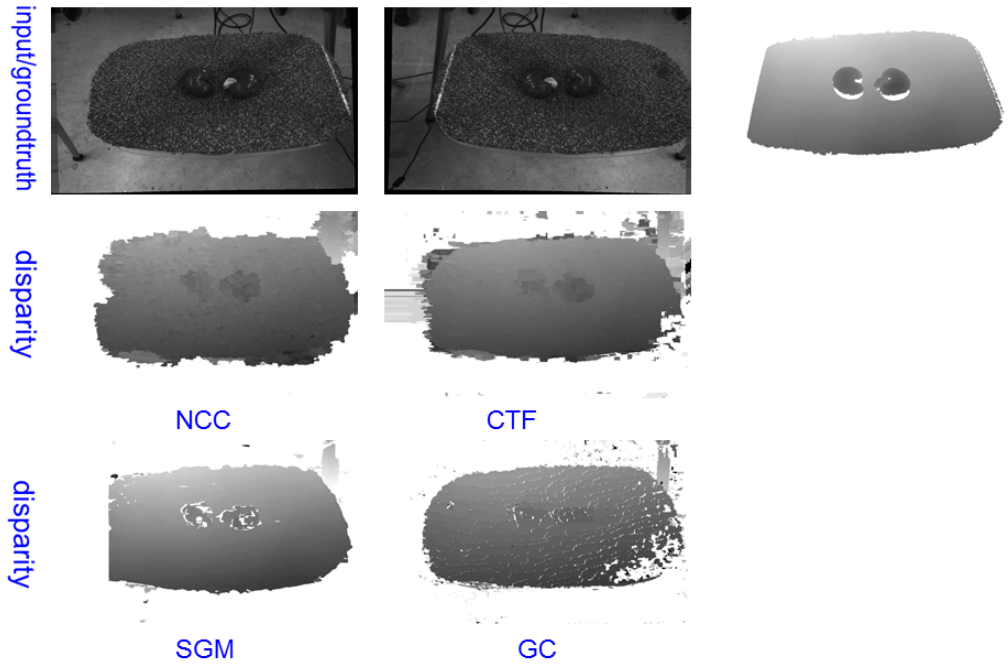


Figure 37: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Kidney 33 cm Data Set.

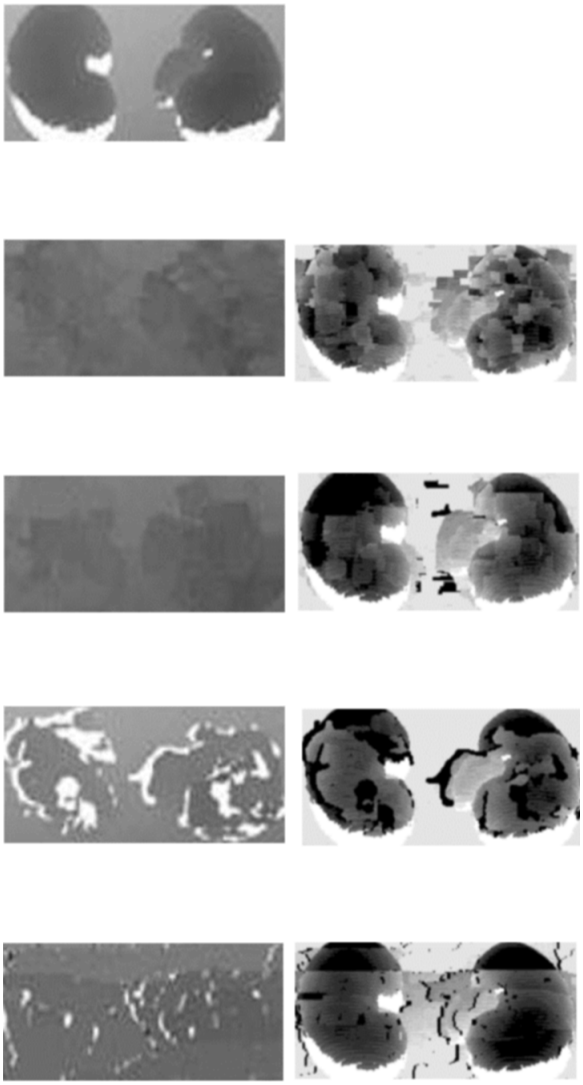

Figure 38: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Kidney 33 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

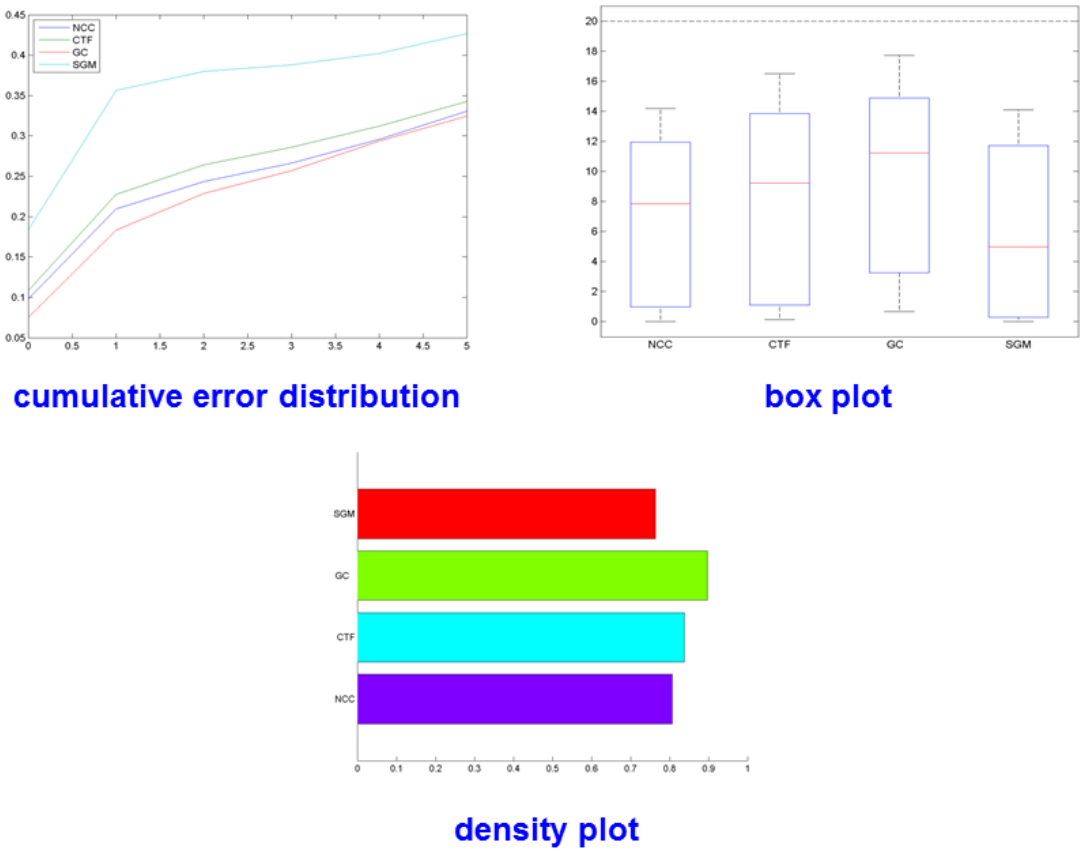


Figure 39: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Kidney 33 cm Data Set.

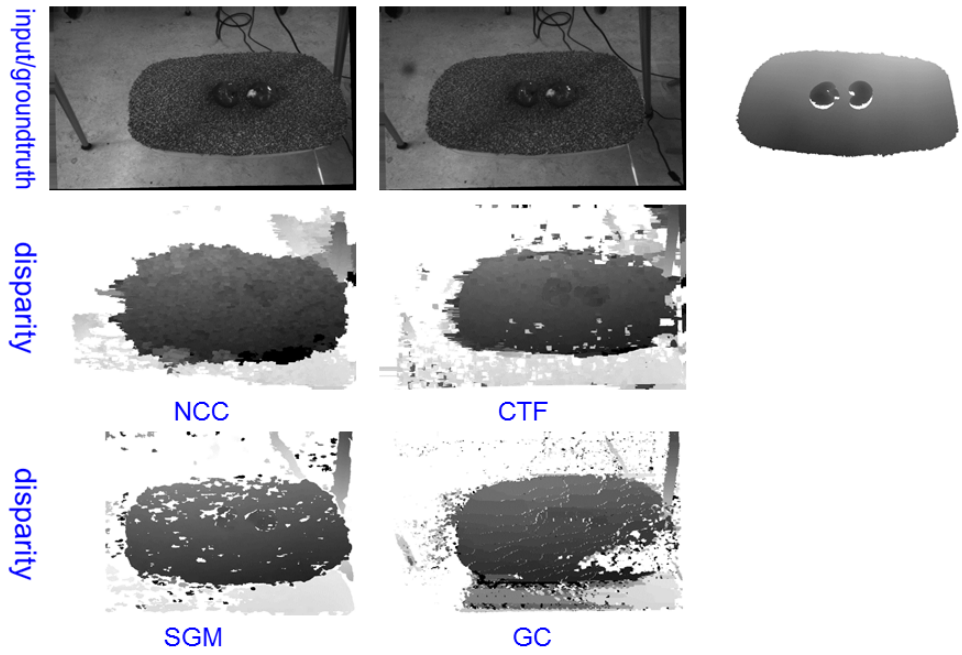


Figure 40: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Kidney 41 cm Data Set.

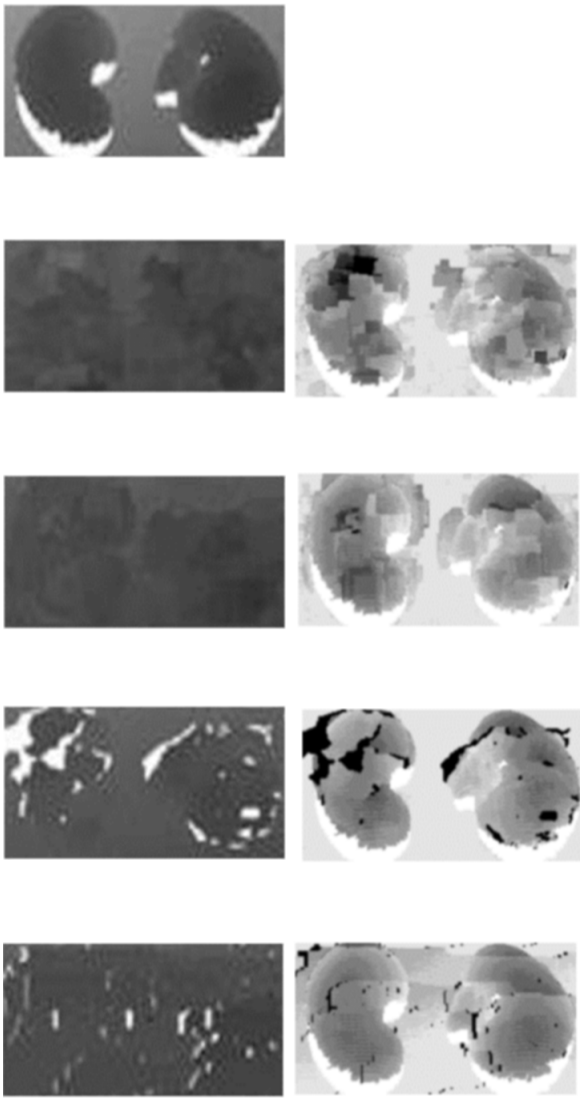

Figure 41: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Kidney 41 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.


Figure 42: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Kidney 41 cm Data Set.

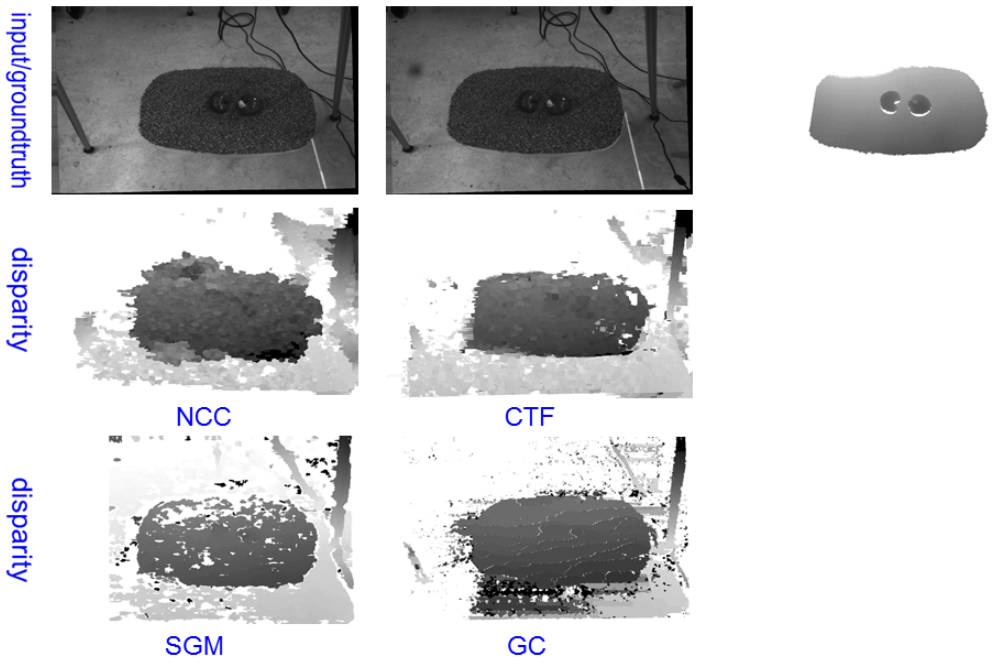


Figure 43: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Kidney 48 cm Data Set.

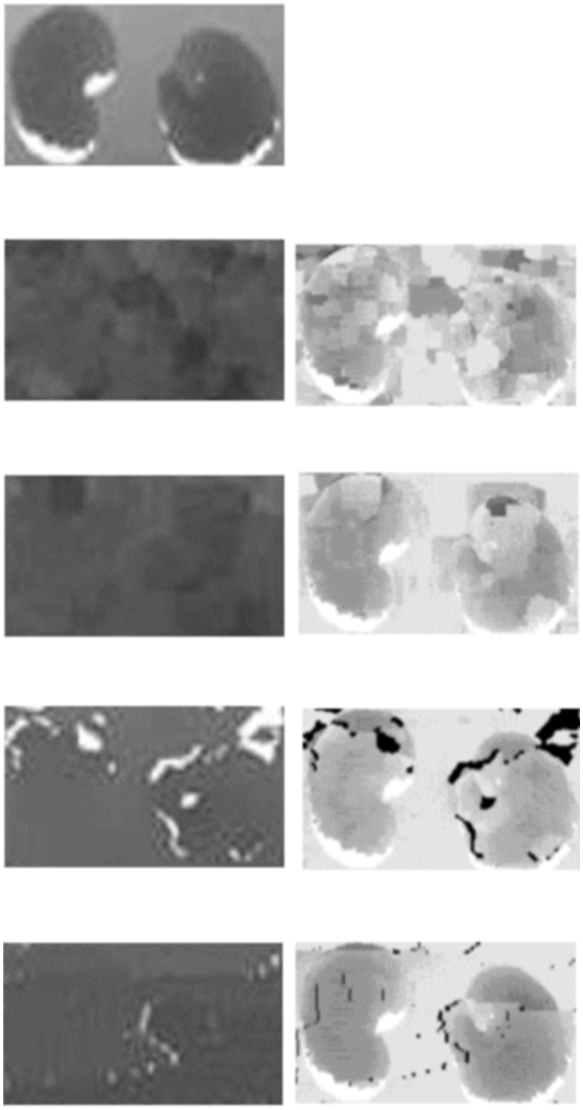

Figure 44: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Kidney 48 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

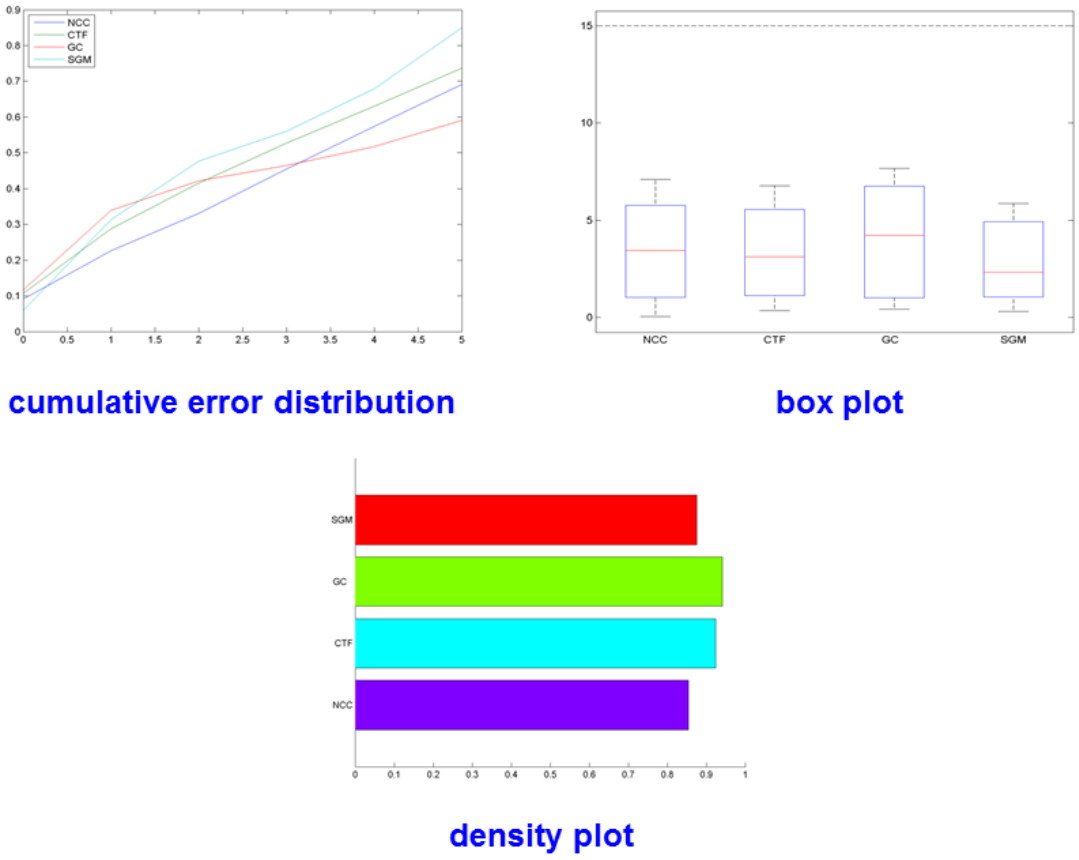


Figure 45: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Kidney 48 cm Data Set.

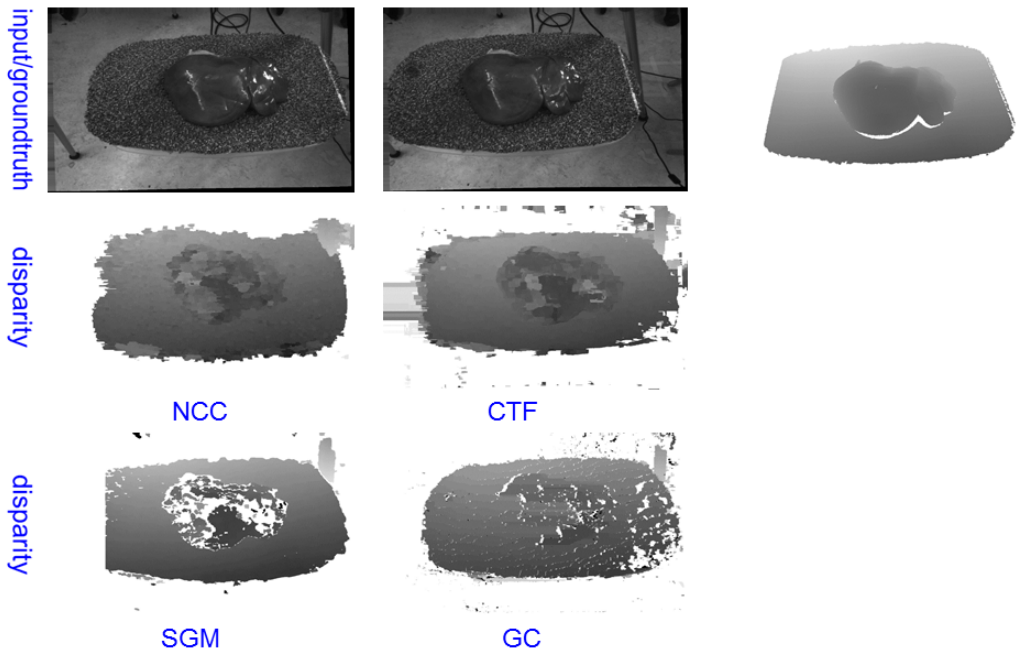


Figure 46: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Liver 33 cm Data Set.

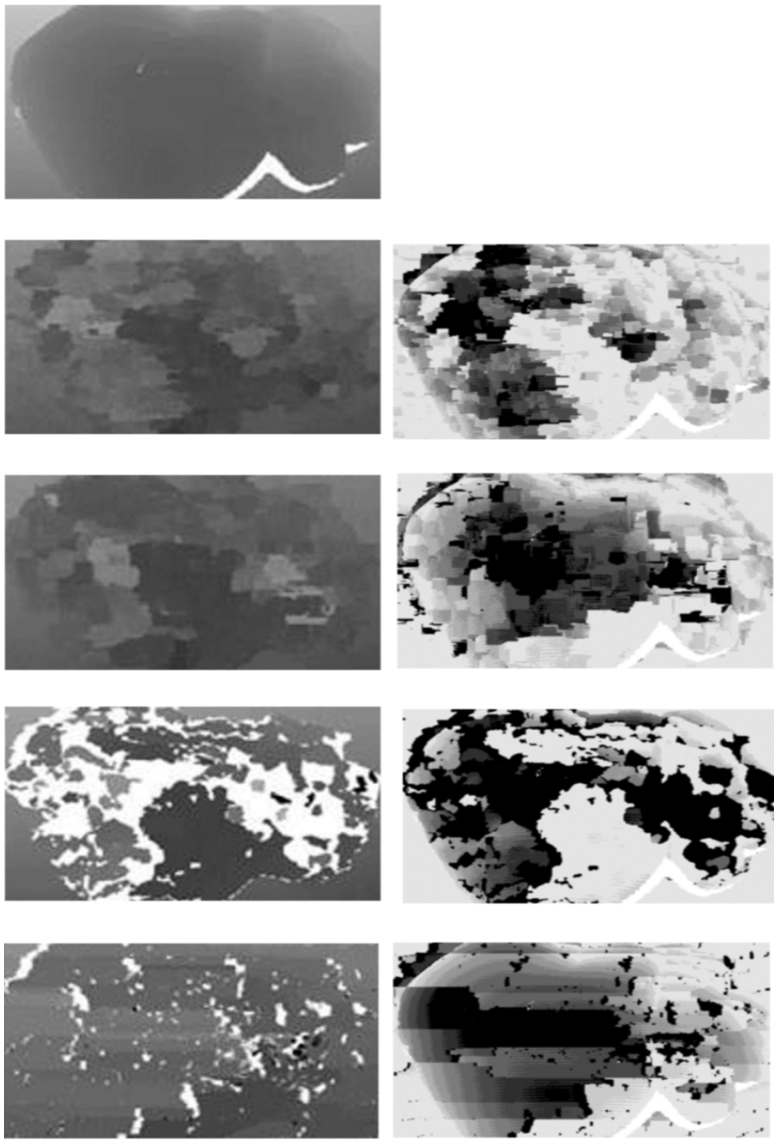

Figure 47: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Liver 33 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.


Figure 48: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Liver 33 cm Data Set.

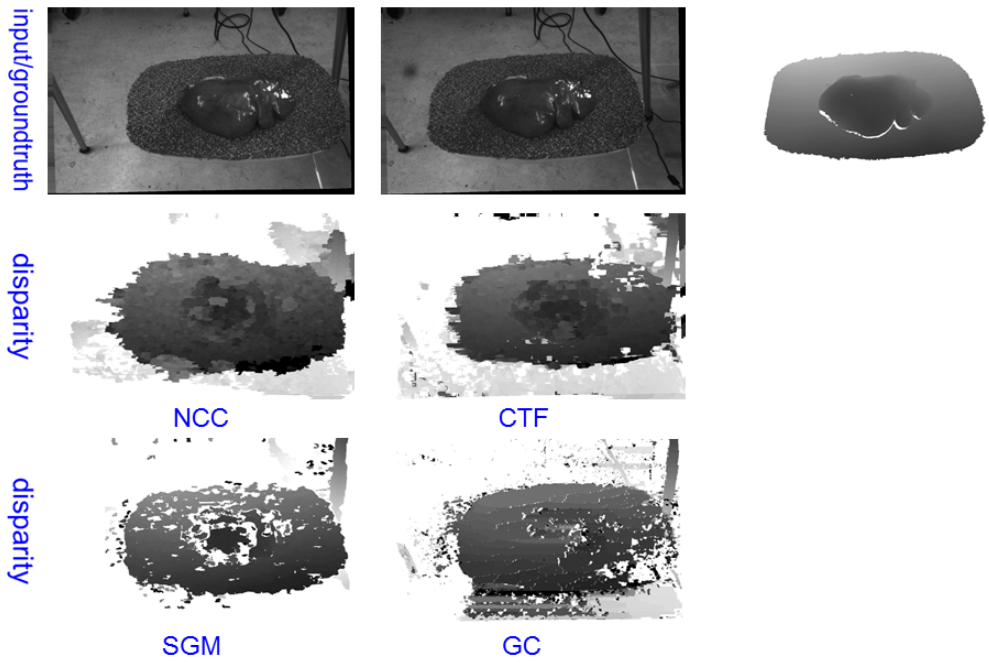


Figure 49: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Liver 41 cm Data Set.

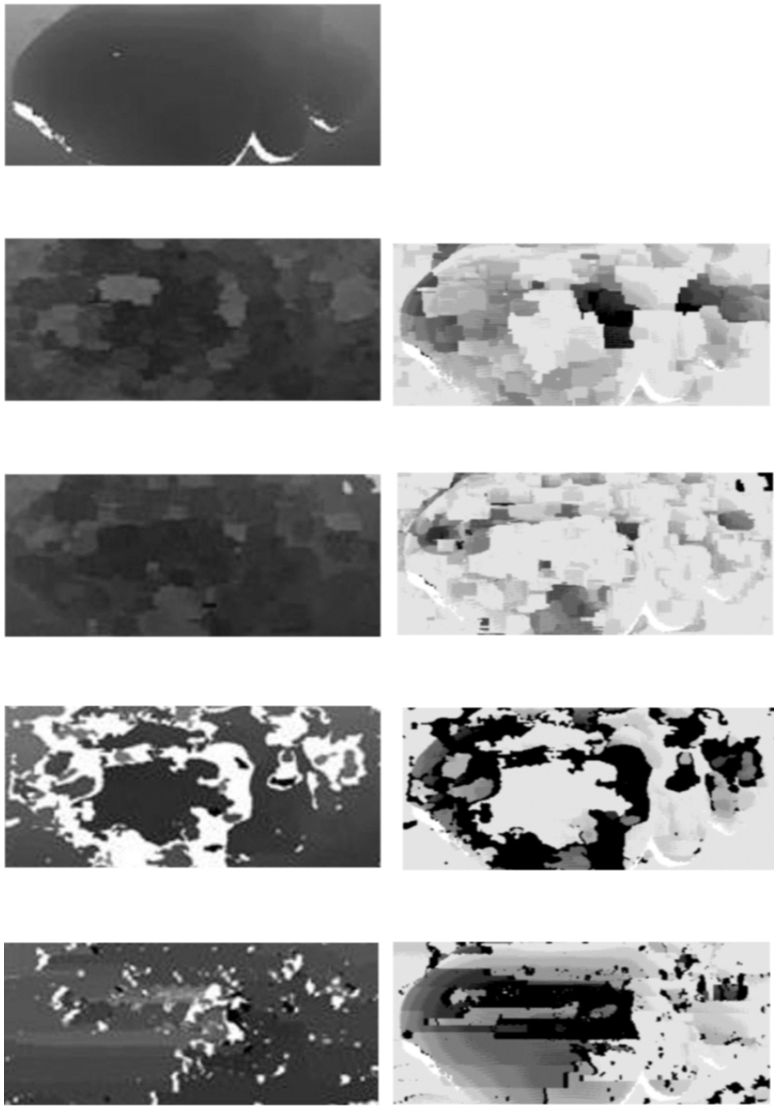

Figure 50: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Liver 41 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

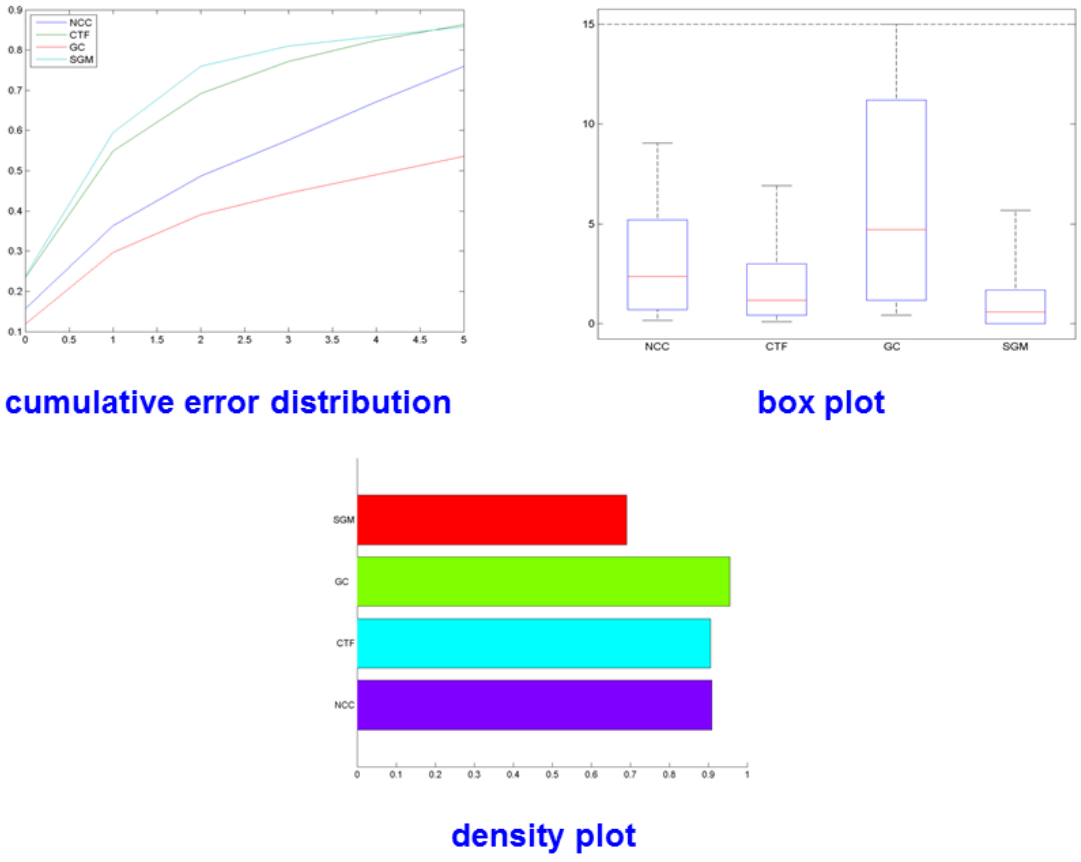


Figure 51: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Liver 41 cm Data Set.

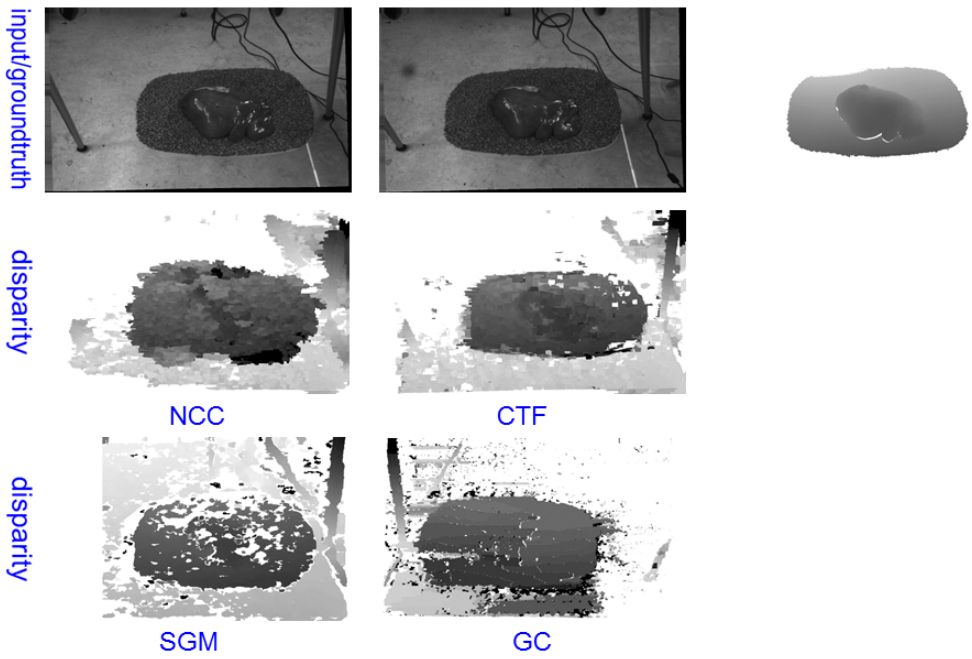


Figure 52: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Liver 48 cm Data Set.

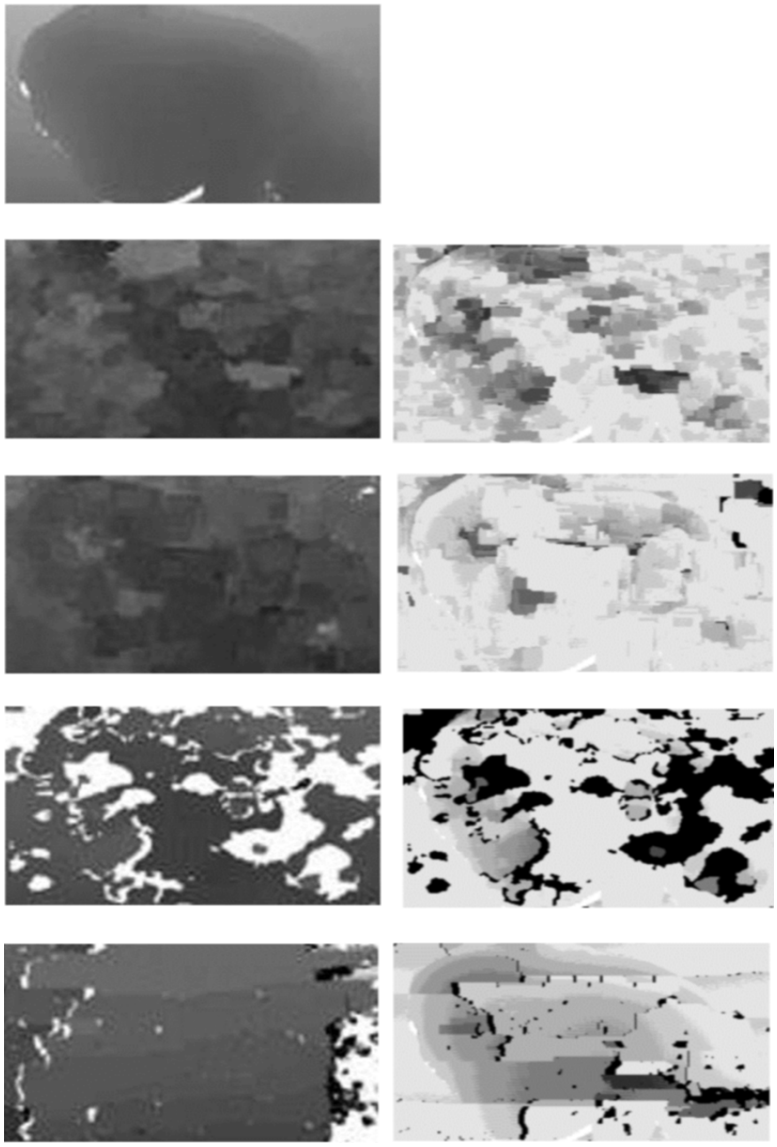

Figure 53: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Liver 48 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

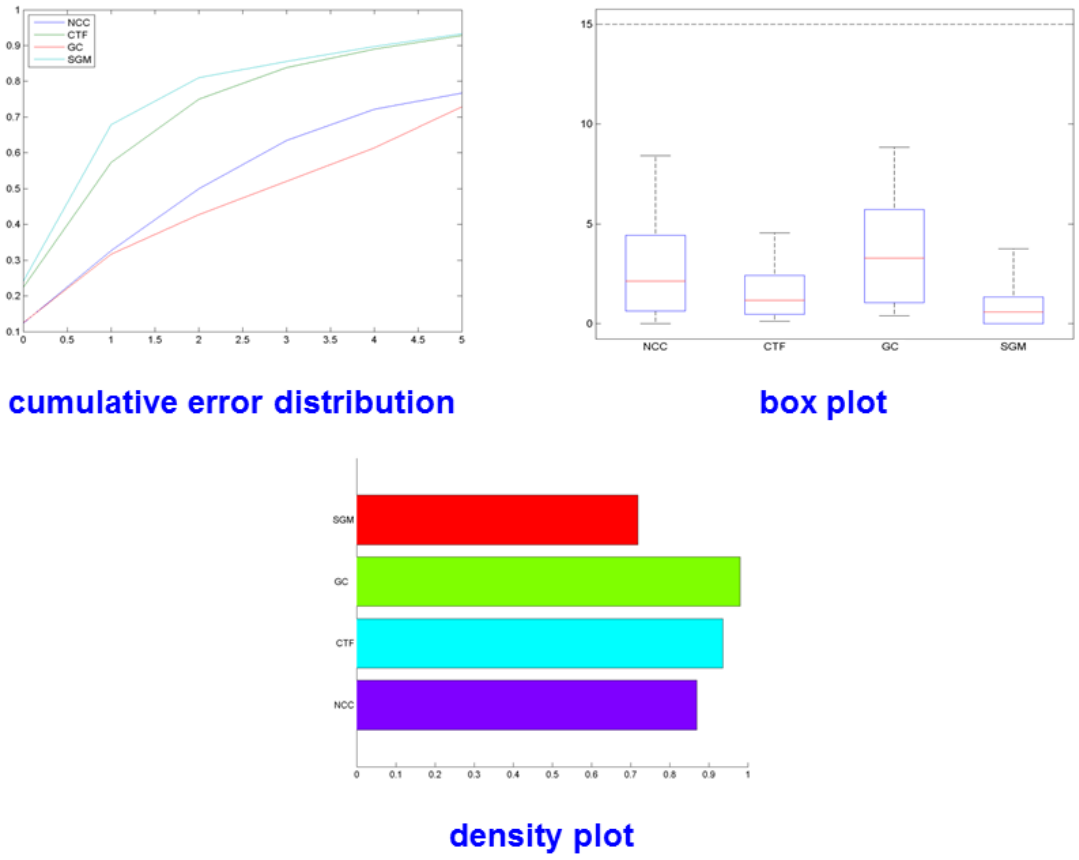


Figure 54: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Liver 48 cm Data Set.

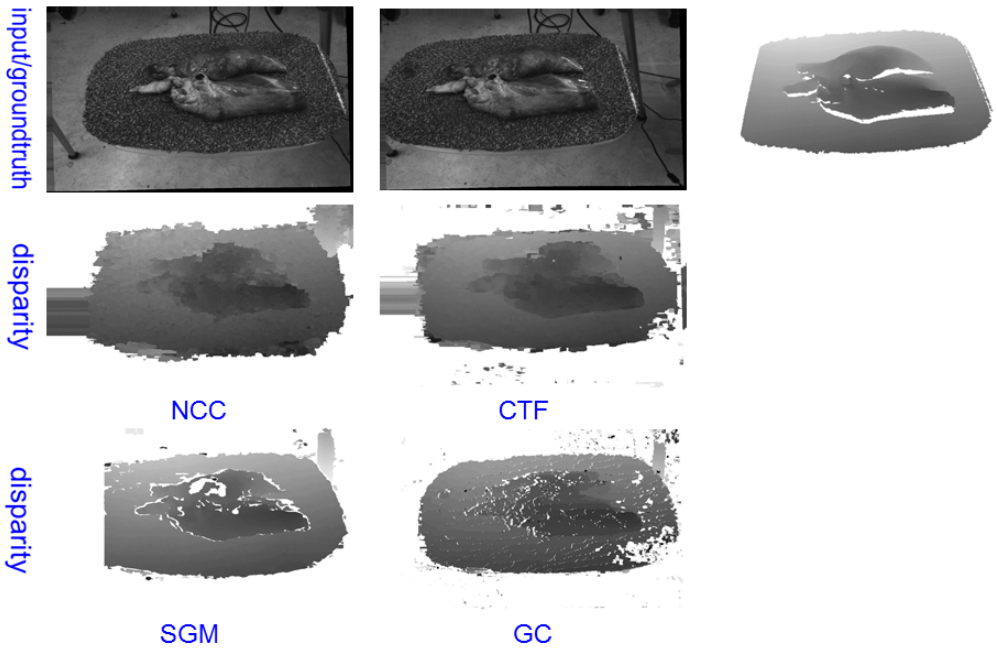


Figure 55: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Lungs 33 cm Data Set.

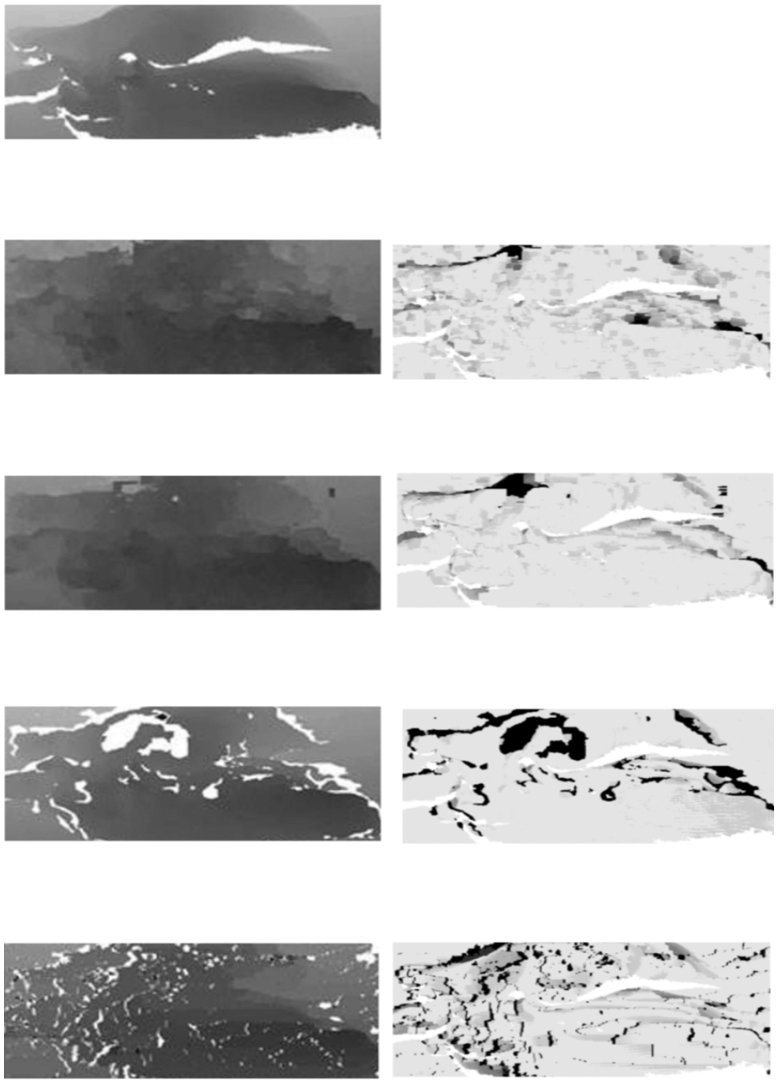

Figure 56: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Lungs 33 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

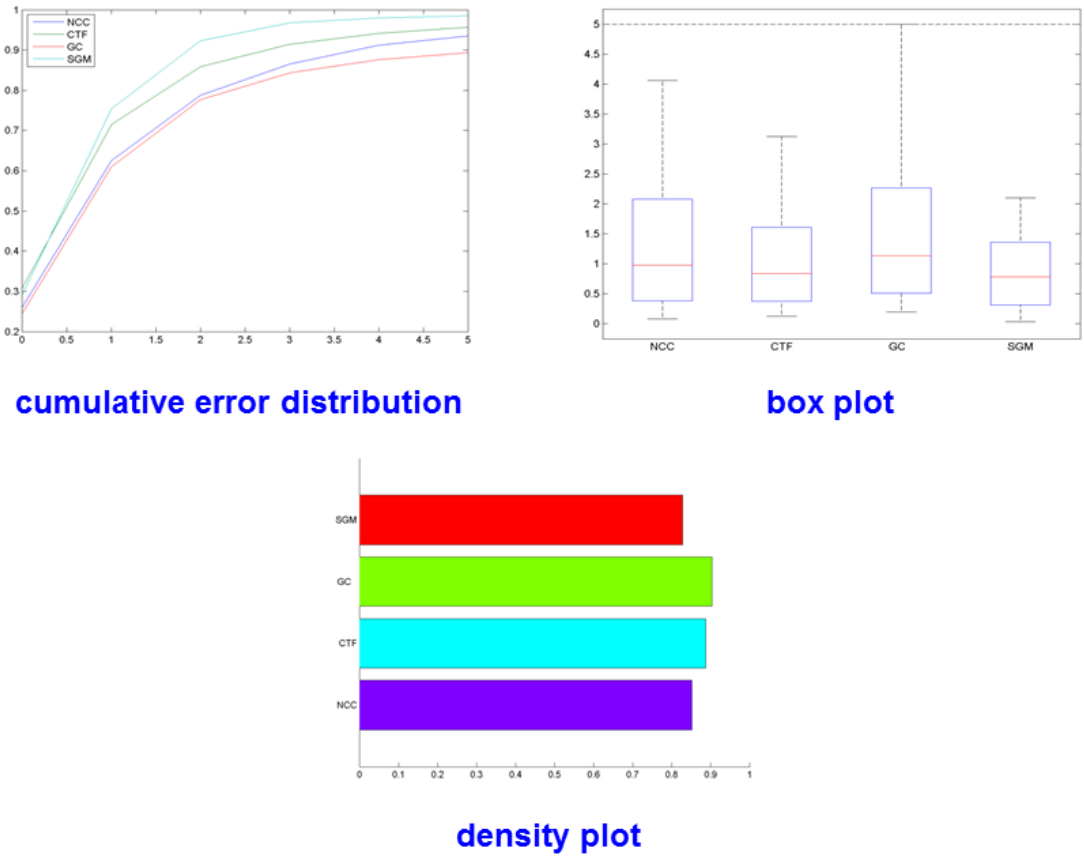


Figure 57: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Lungs 33 cm Data Set.

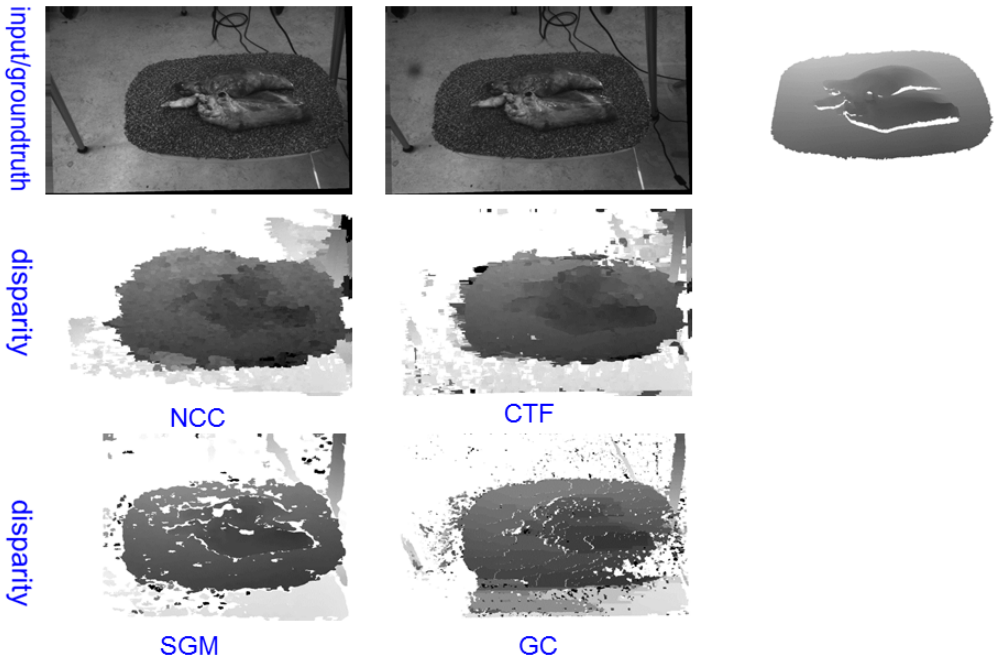


Figure 58: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Lungs 41 cm Data Set.

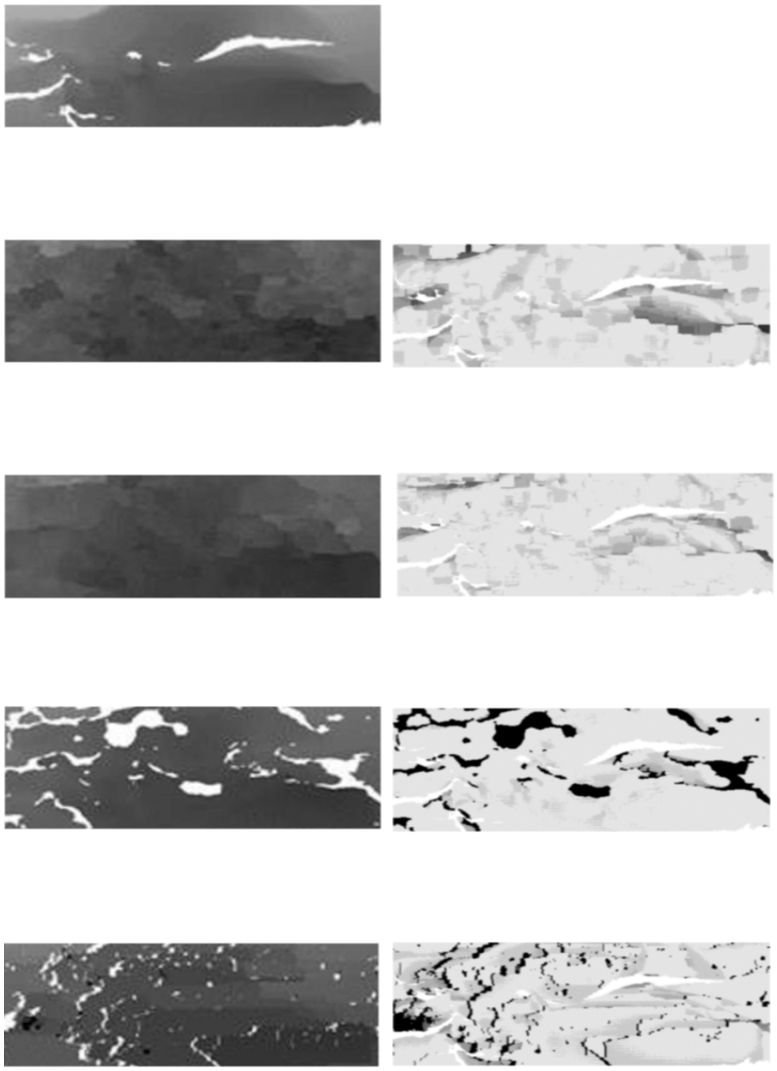

Figure 59: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Lungs 41 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

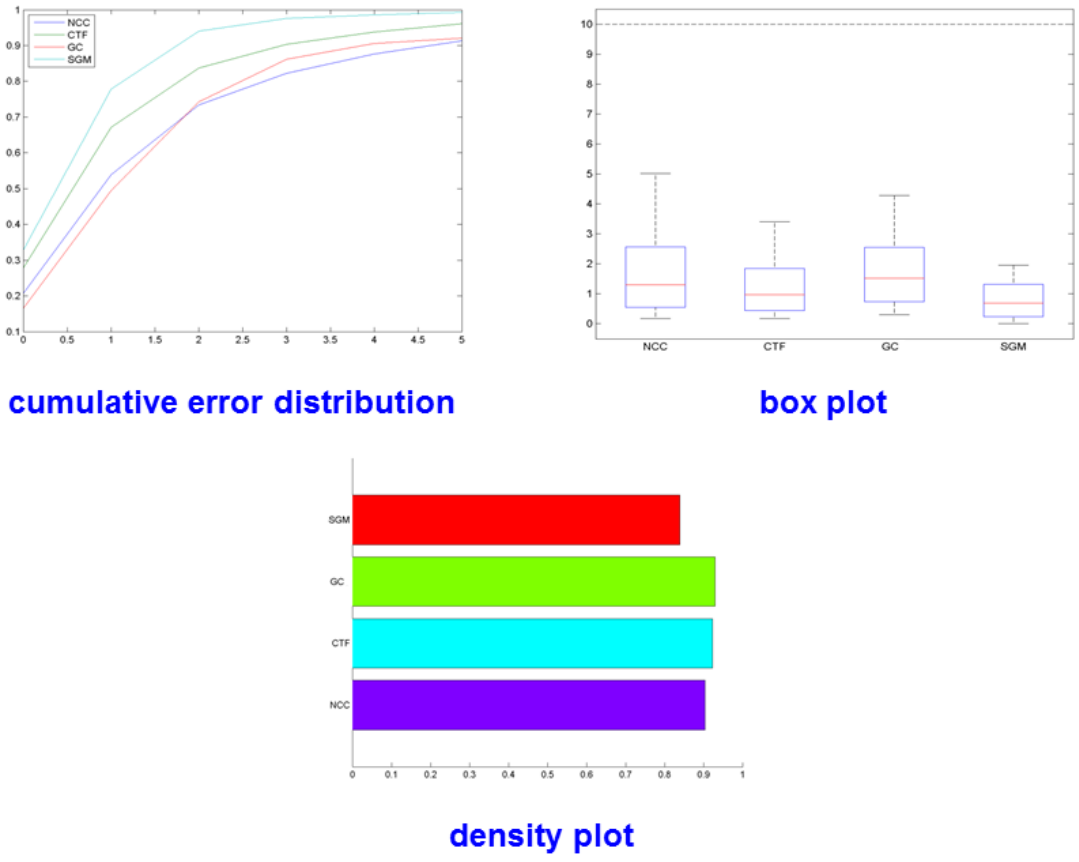


Figure 60: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Lungs 41 cm Data Set.

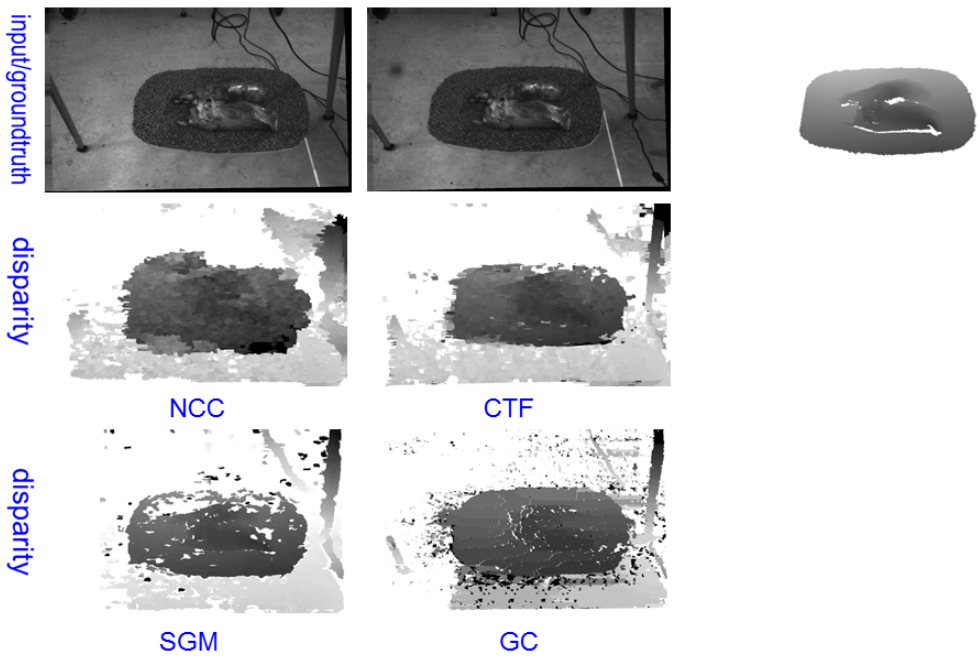


Figure 61: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Lungs 48 cm Data Set.

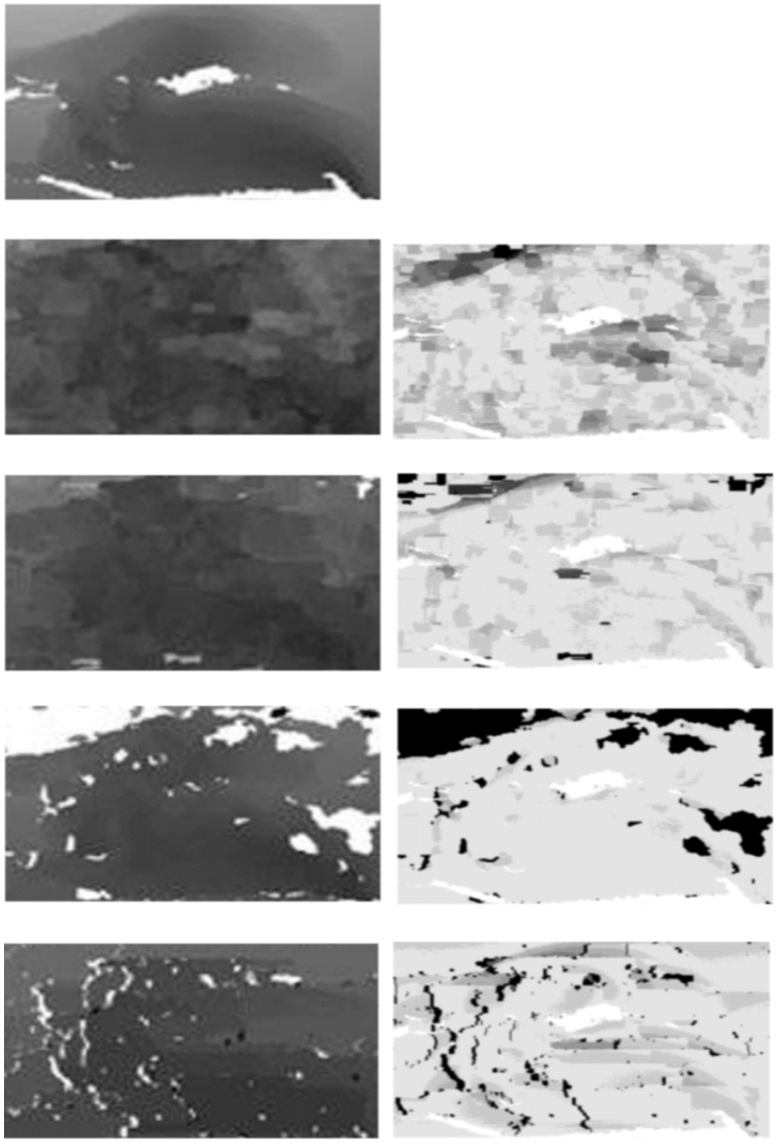

Figure 62: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Lungs 48 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

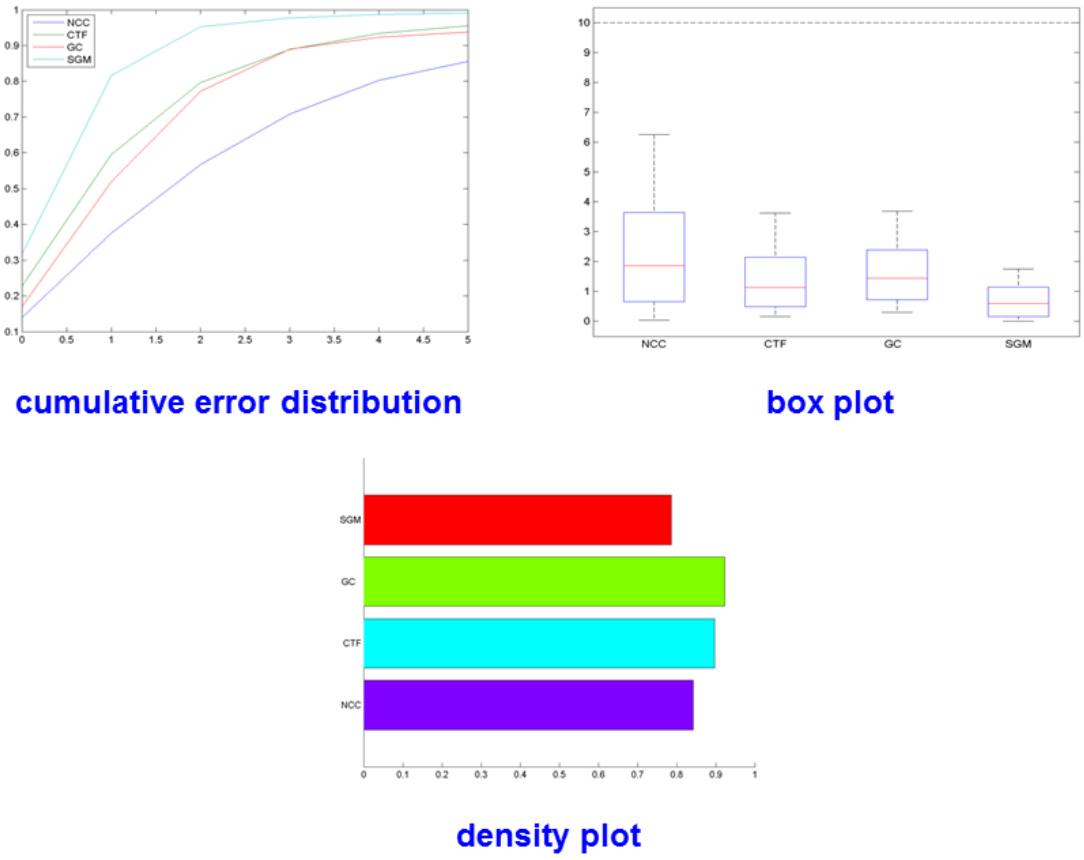


Figure 63: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Lungs 48 cm Data Set.

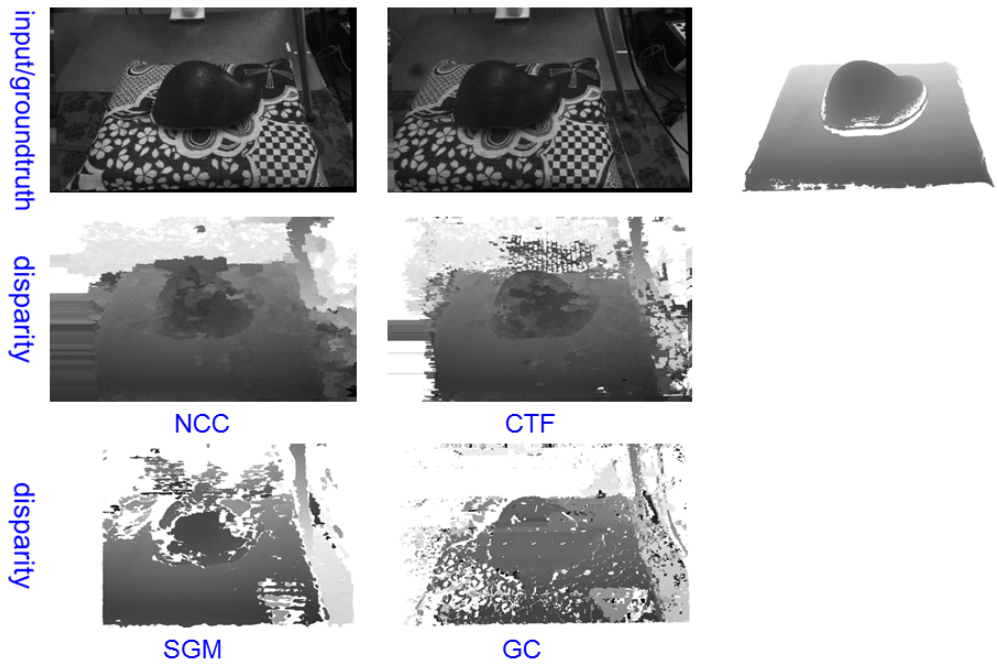


Figure 64: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Phantom Liver 33 cm Data Set.

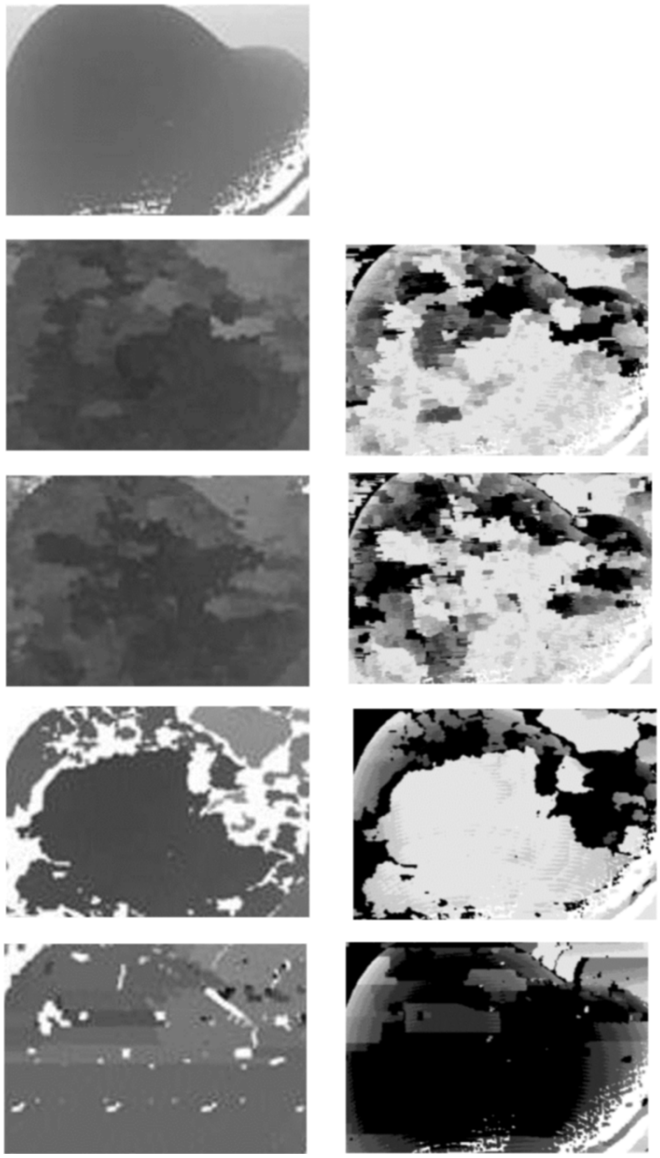

Figure 65: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Phantom Liver 33 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

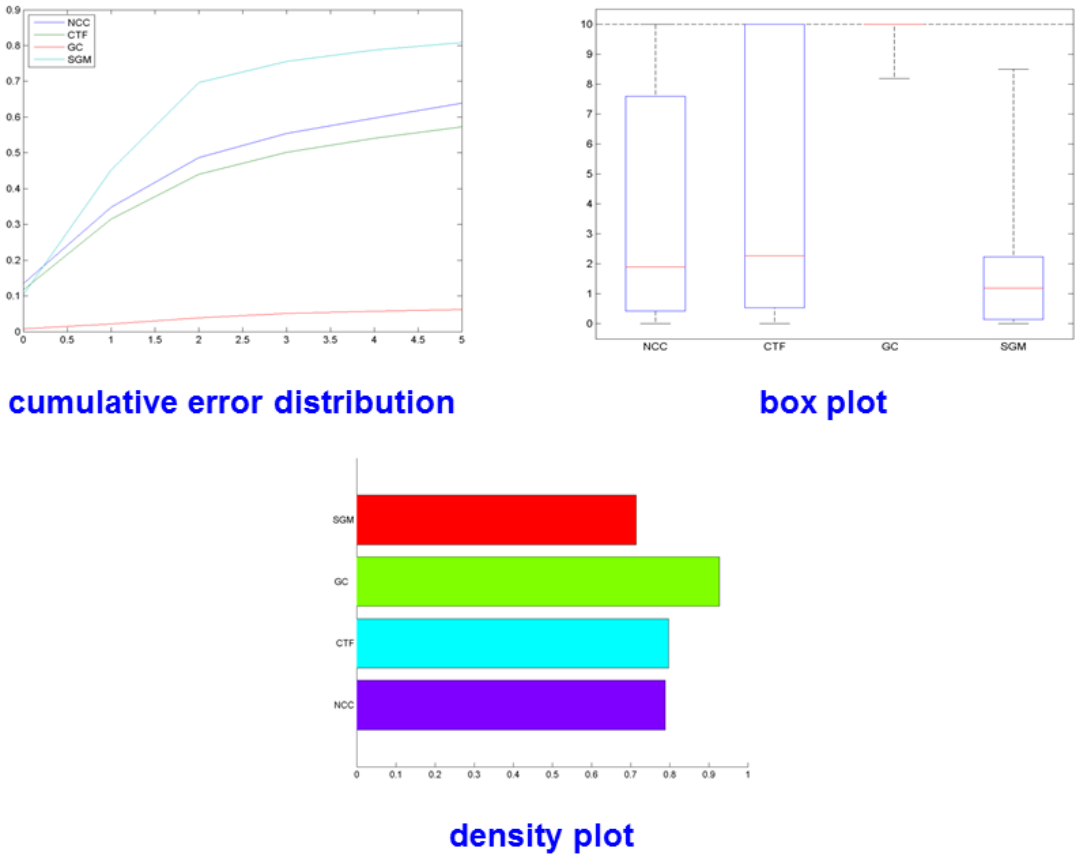


Figure 66: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Phantom Liver 33 cm Data Set.

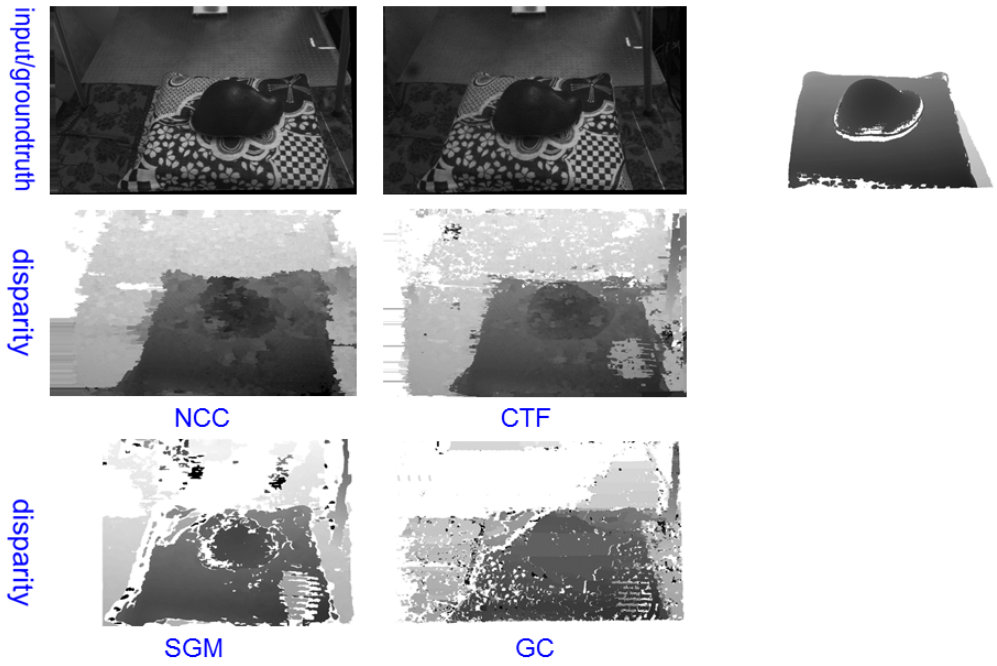


Figure 67: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Phantom Liver 41 cm Data Set.

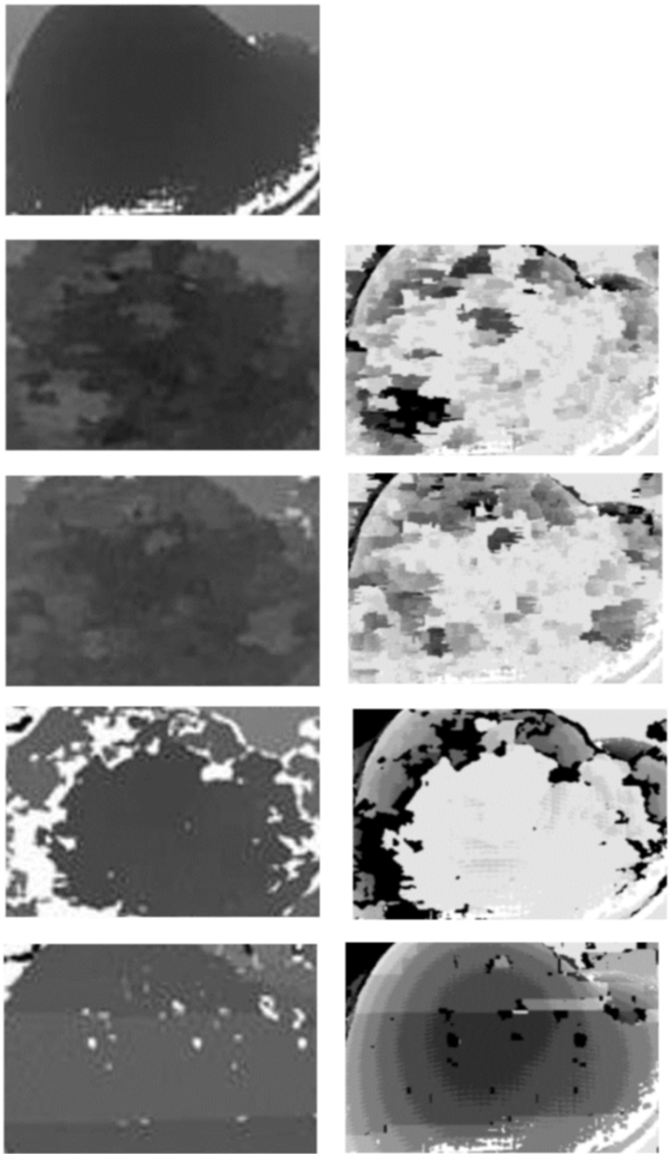

Figure 68: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Phantom Liver 41 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

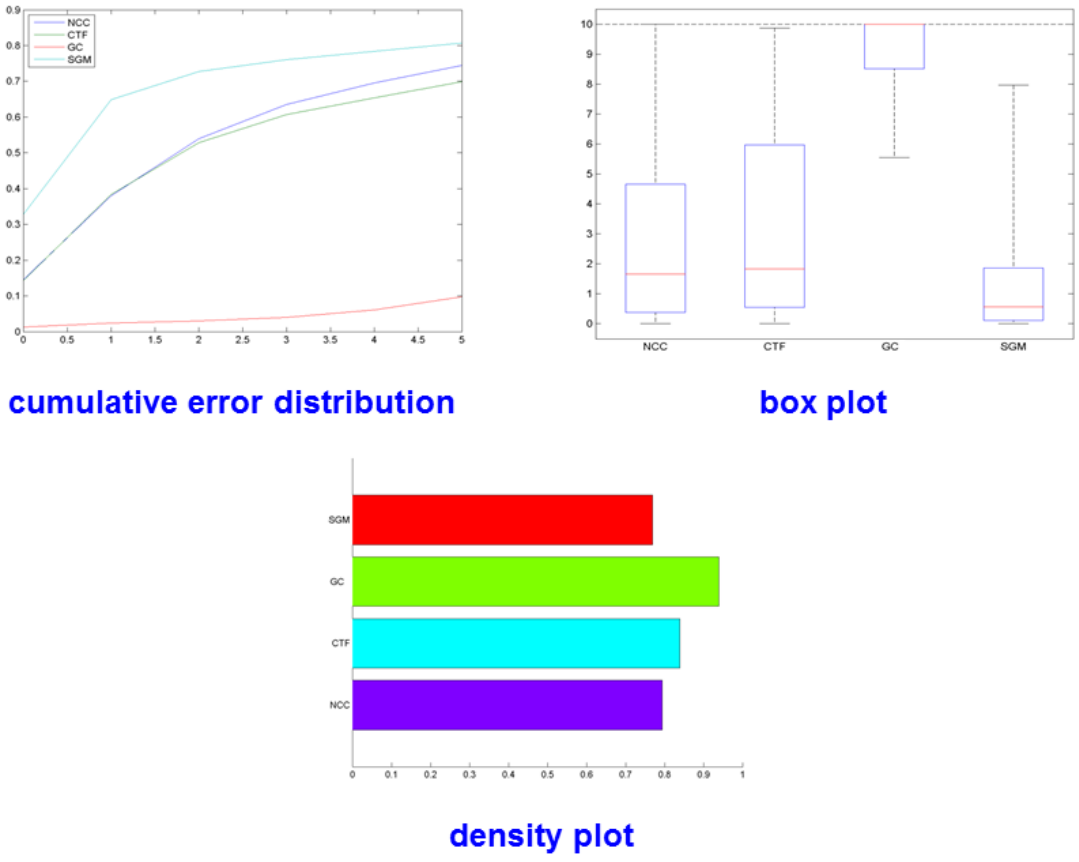


Figure 69: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Phantom Liver 41 cm Data Set.

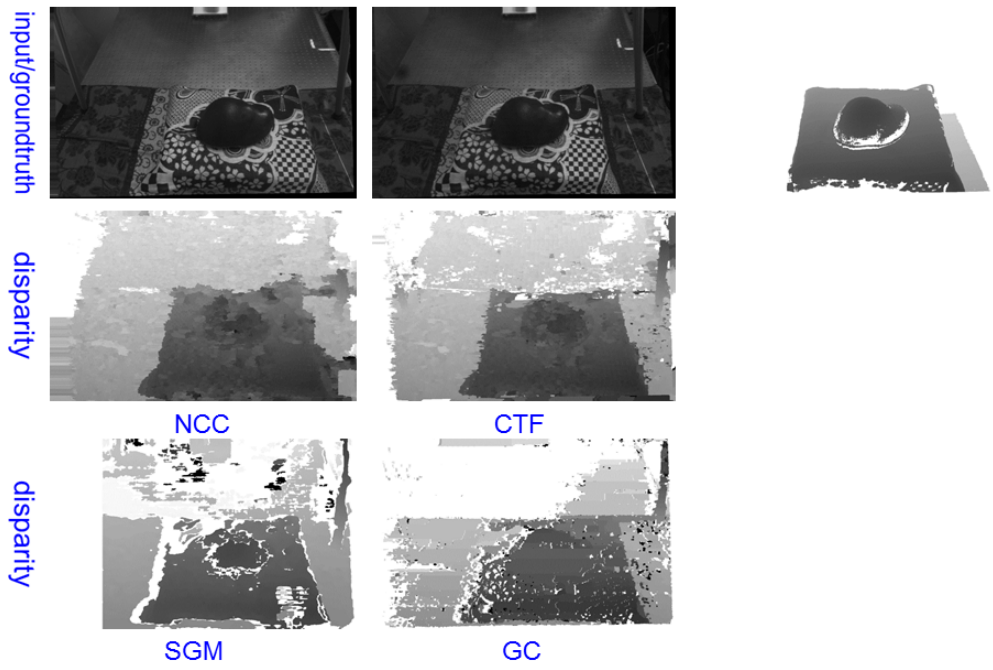


Figure 70: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Lab Phantom Liver 48 cm Data Set.

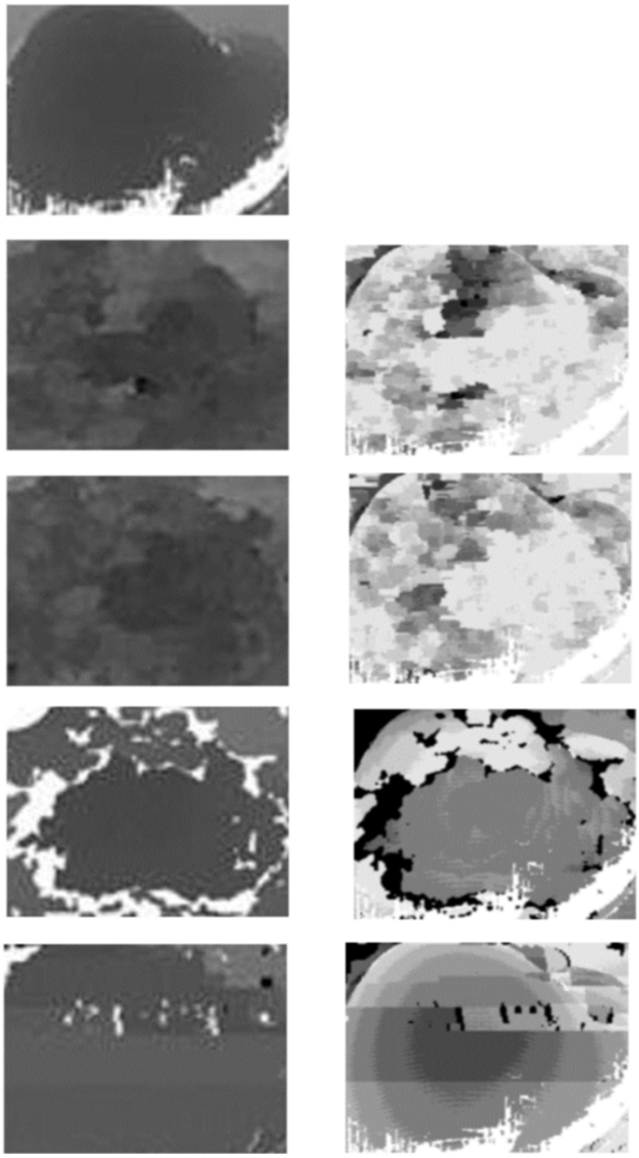

Figure 71: Absolute Difference Map Between the Groundtruth and Disparity Recovered by each of the Four Algorithms for the Lab Phantom Liver 48 cm. Data Set, as Restricted to the Region of Interest. The top image is the ground truth disparity. The second row is the NCC disparity and difference map. The third row is the CTF disparity and difference map. The fourth row is the SGM disparity and difference map. The fifth row is the GC disparity and difference map. In difference maps, brighter intensity corresponds to smaller error.

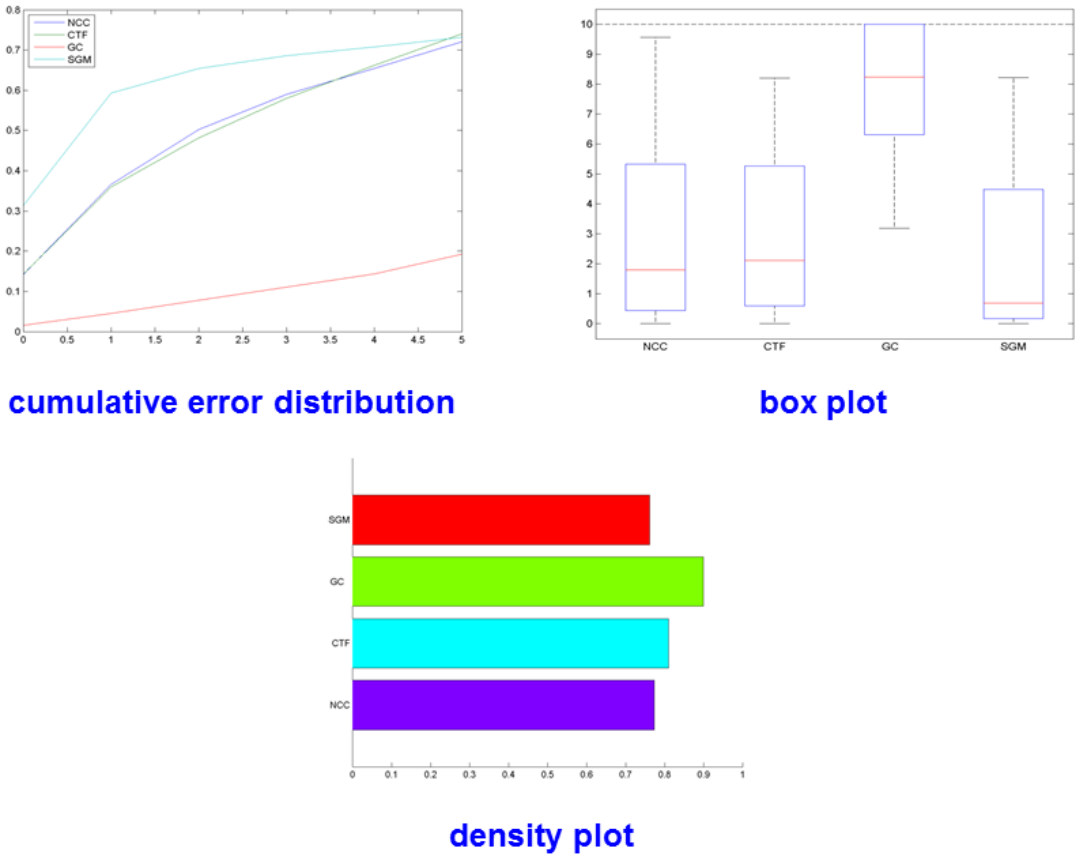


Figure 72: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Phantom Liver 48 cm Data Set.

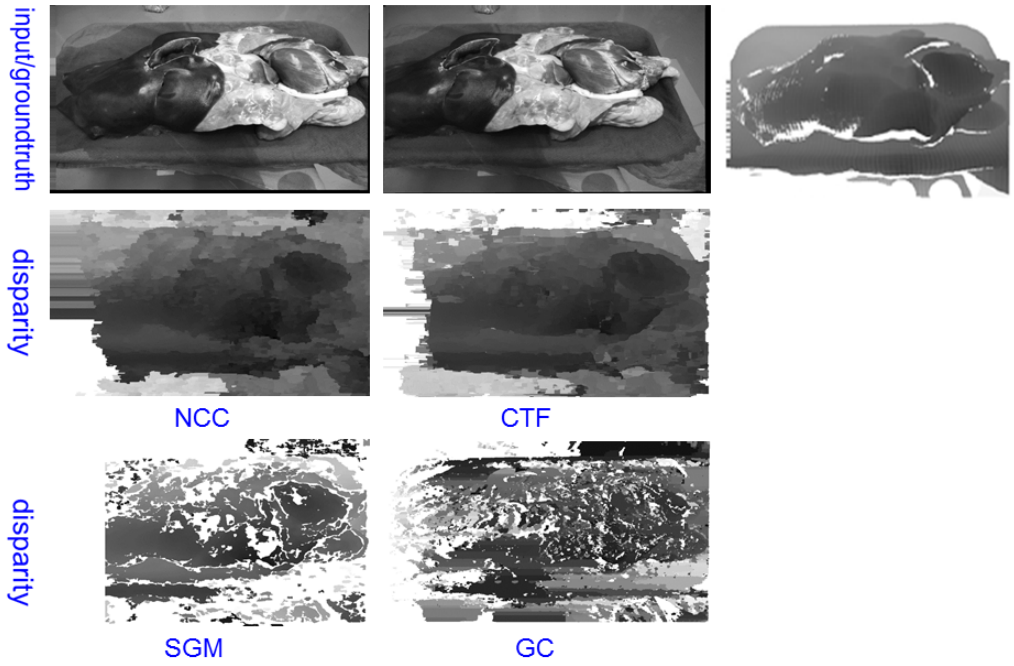


Figure 73: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Hospital Anatomical Front View 41 cm Data Set.

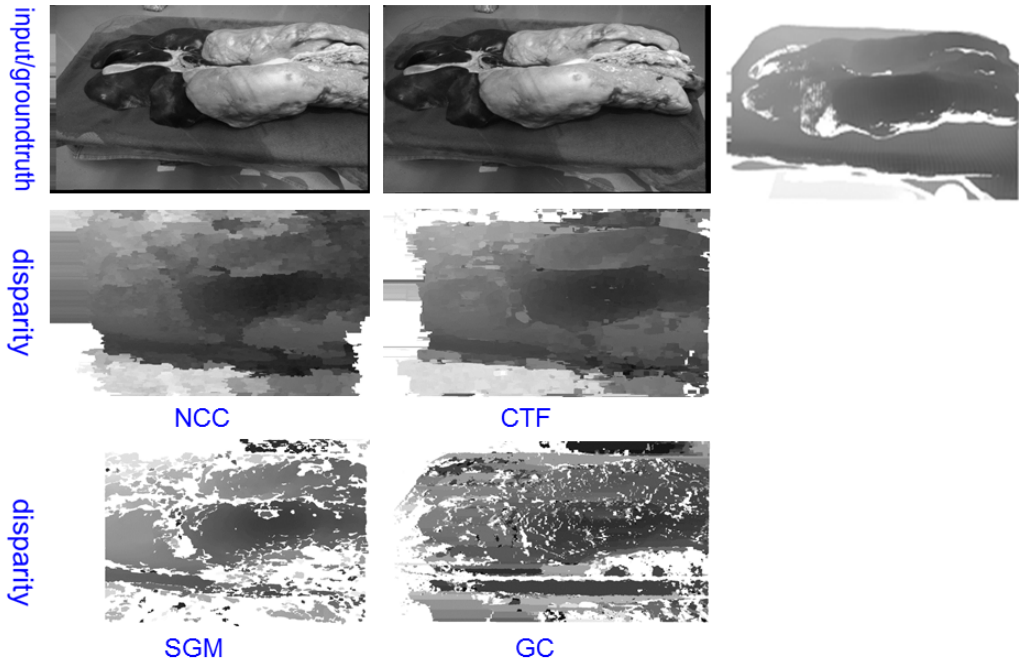


Figure 74: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Hospital Anatomical Back View 41 cm Data Set.

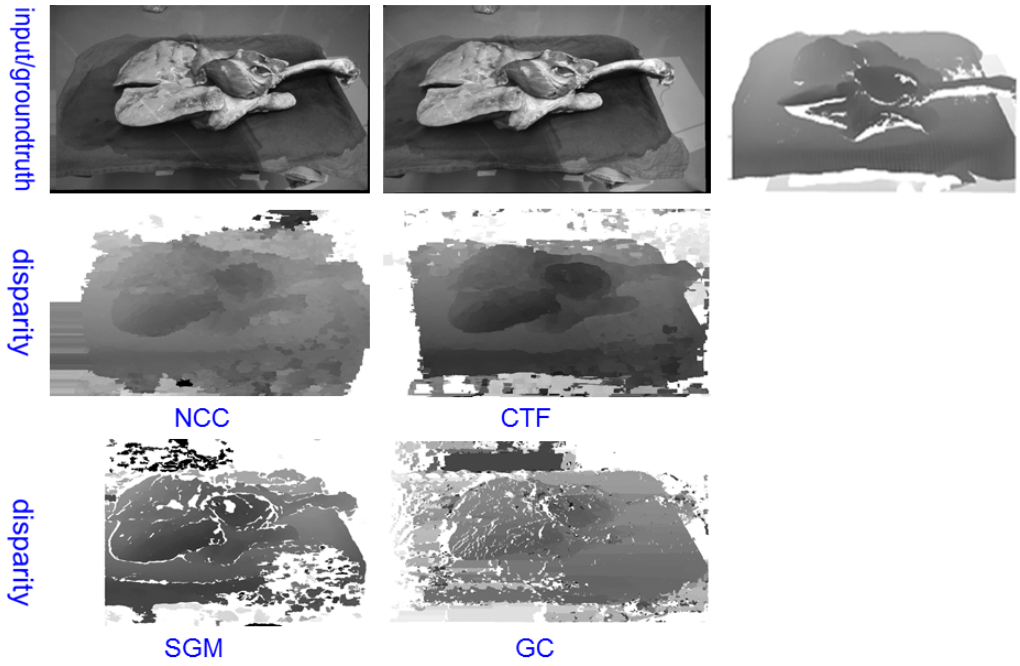


Figure 75: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Hospital Heart/Lungs Front View 41 cm Data Set.

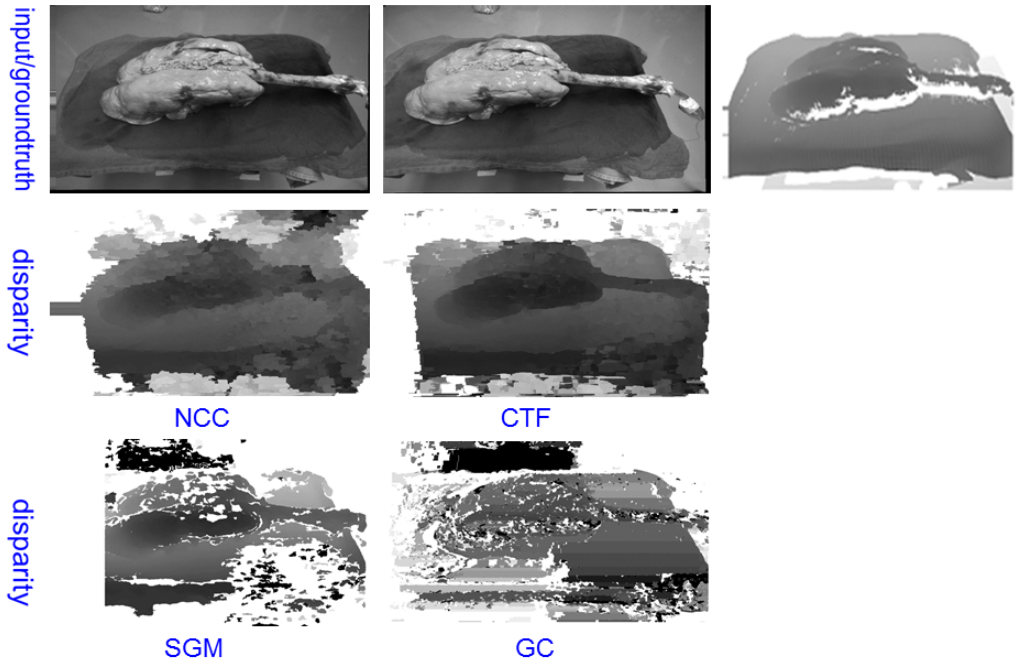


Figure 76: Left-Right Image Pair, the Left Based Groundtruth and the Disparity Results of 4 Algorithms on the Hospital Heart/Lungs Back View 41 cm Data Set.

# B Stereo Algorithm Results Empirical Evaluation - Averages

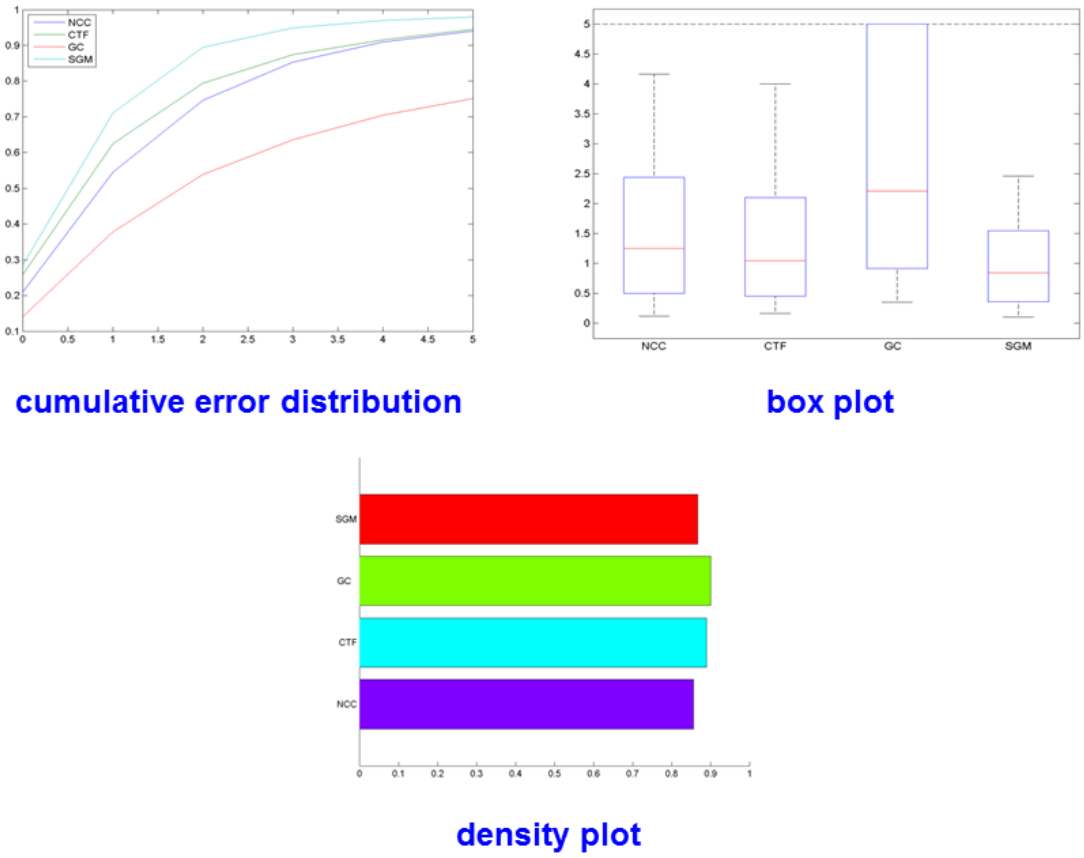


Figure 77: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Full Anatomical - Collapsed Across Distance.

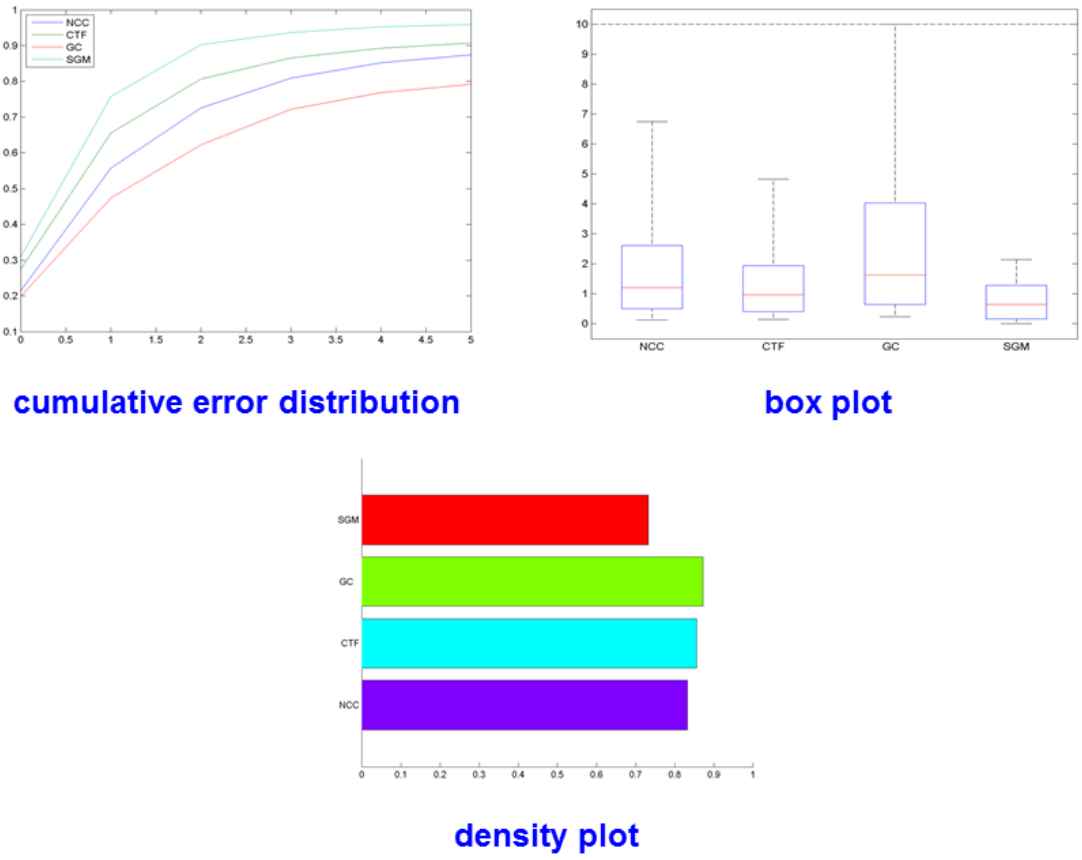


Figure 78: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Heart Closed - Collapsed Across Distance.

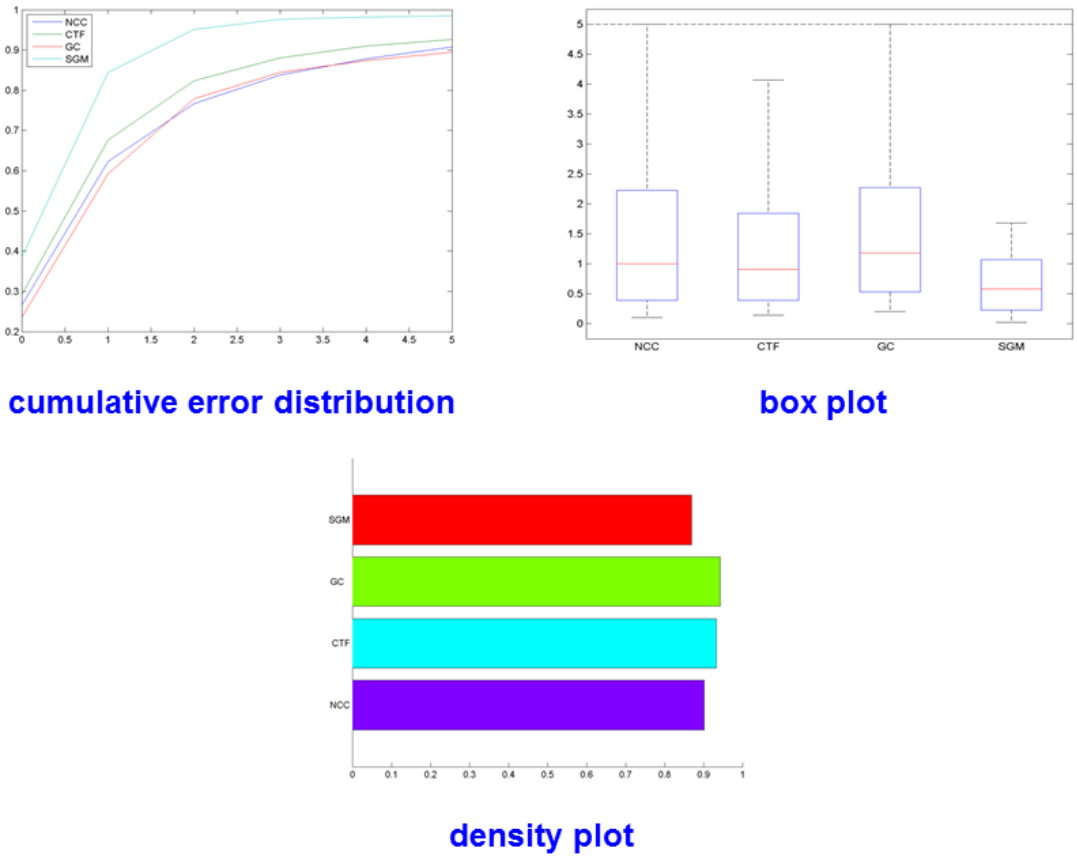


Figure 79: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Heart Open - Collapsed Across Distance.

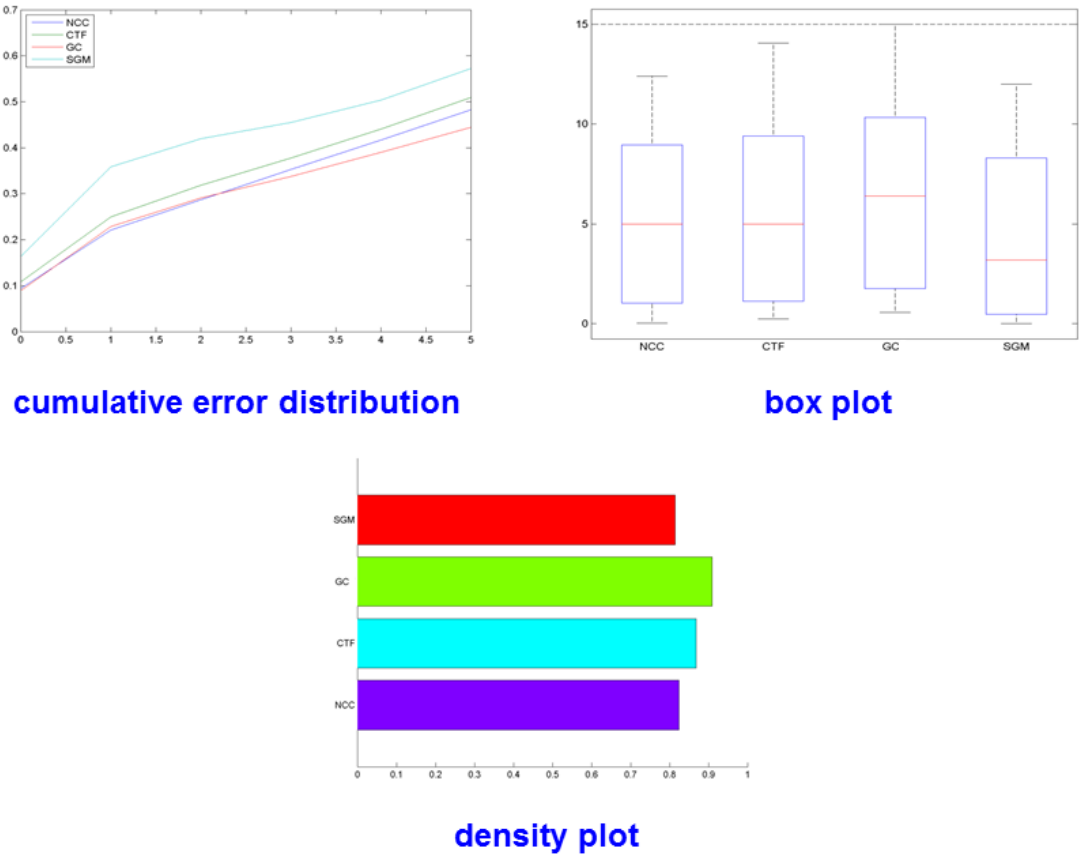


Figure 80: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Kidney - Collapsed Across Distance.

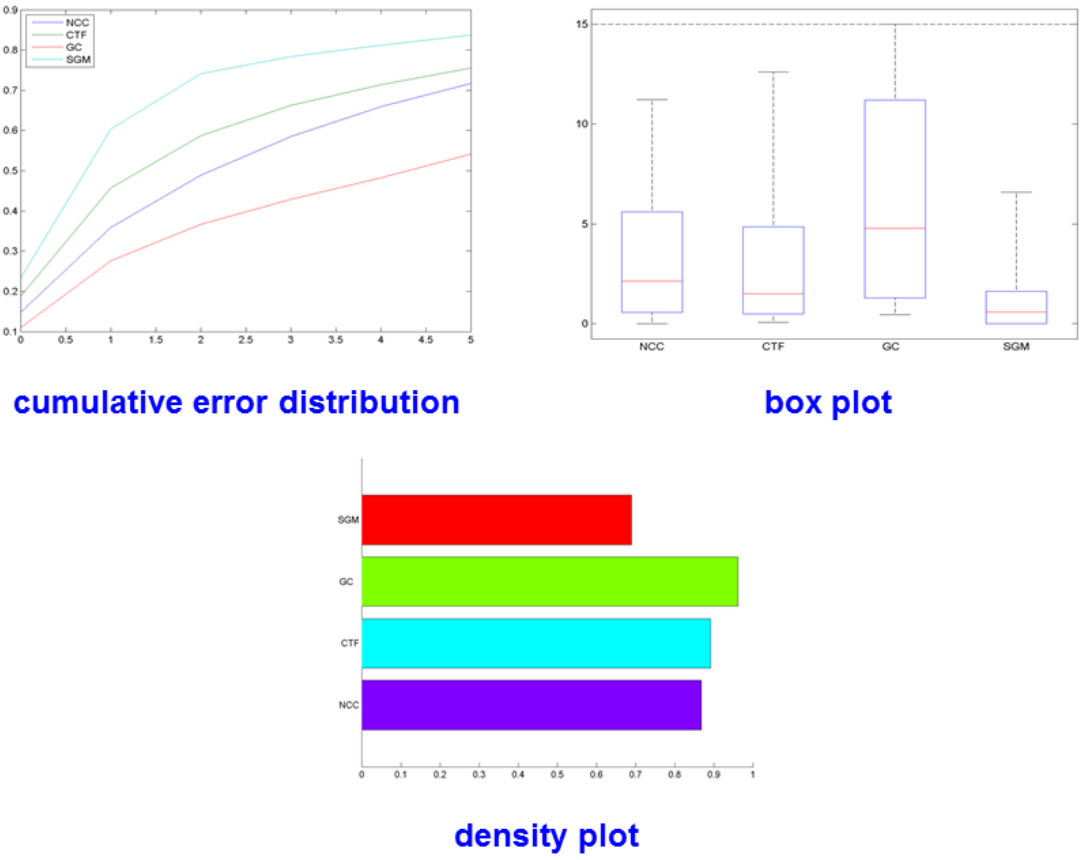


Figure 81: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Liver - Collapsed Across Distance.

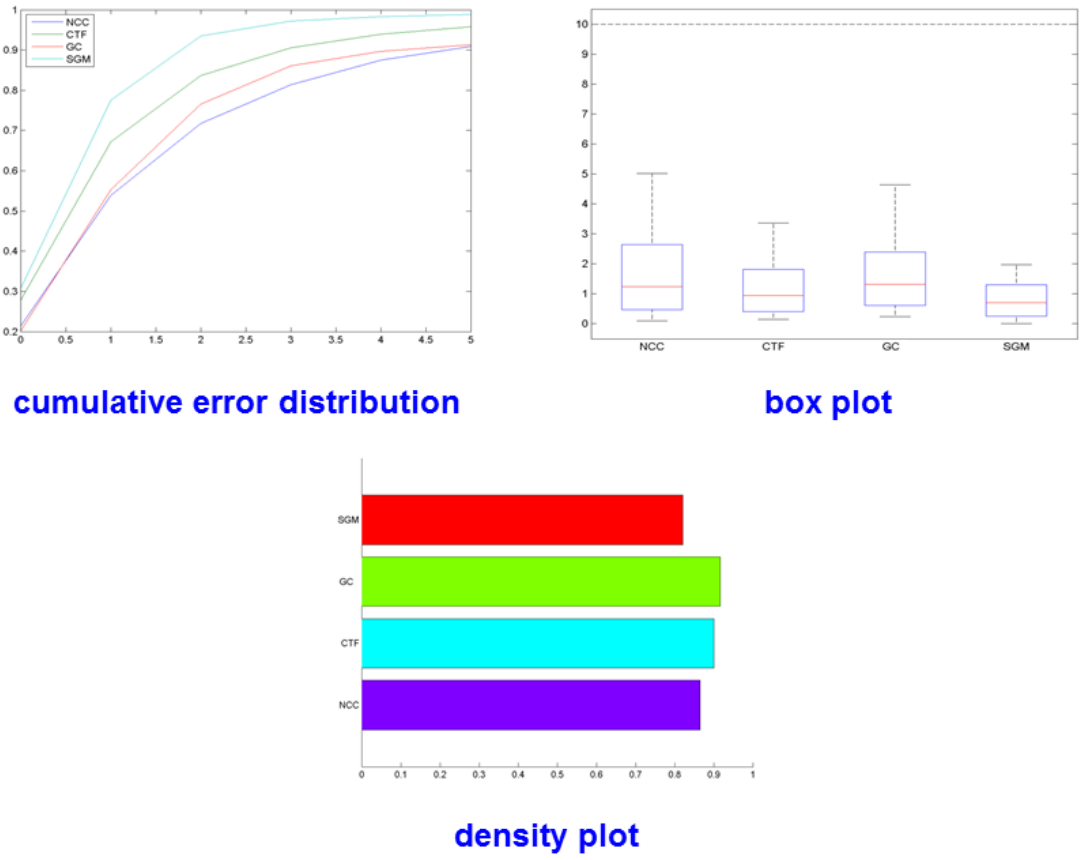


Figure 82: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Lungs - Collapsed Across Distance.

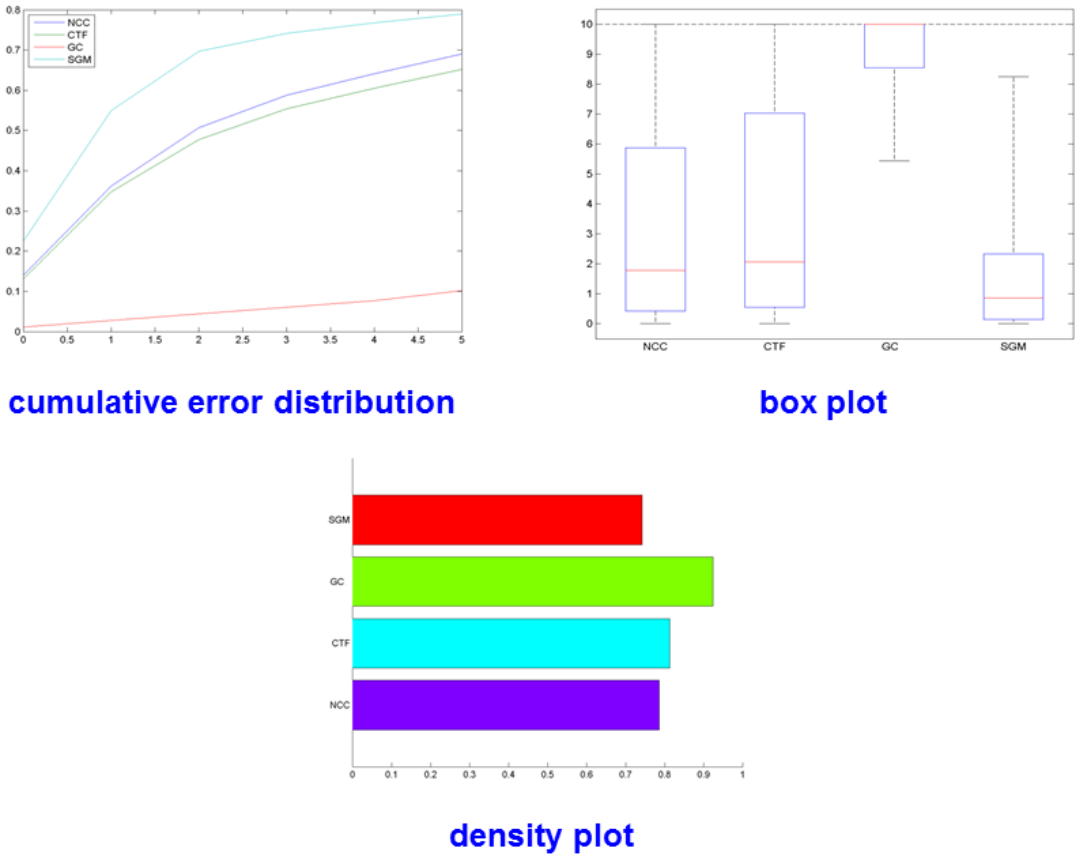


Figure 83: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab Phantom Liver - Collapsed Across Distance.

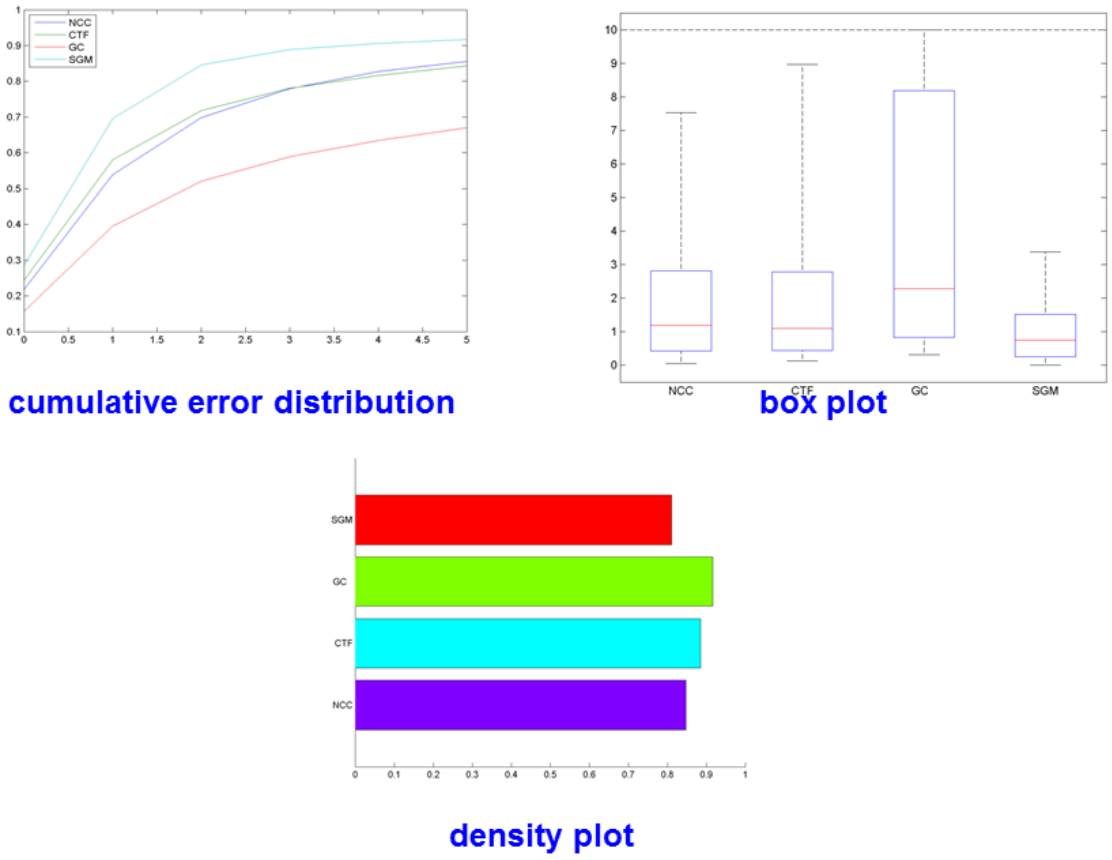


Figure 84: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab 33 cm Distance - Collapsed Across Organs (Excluding the Phantom).

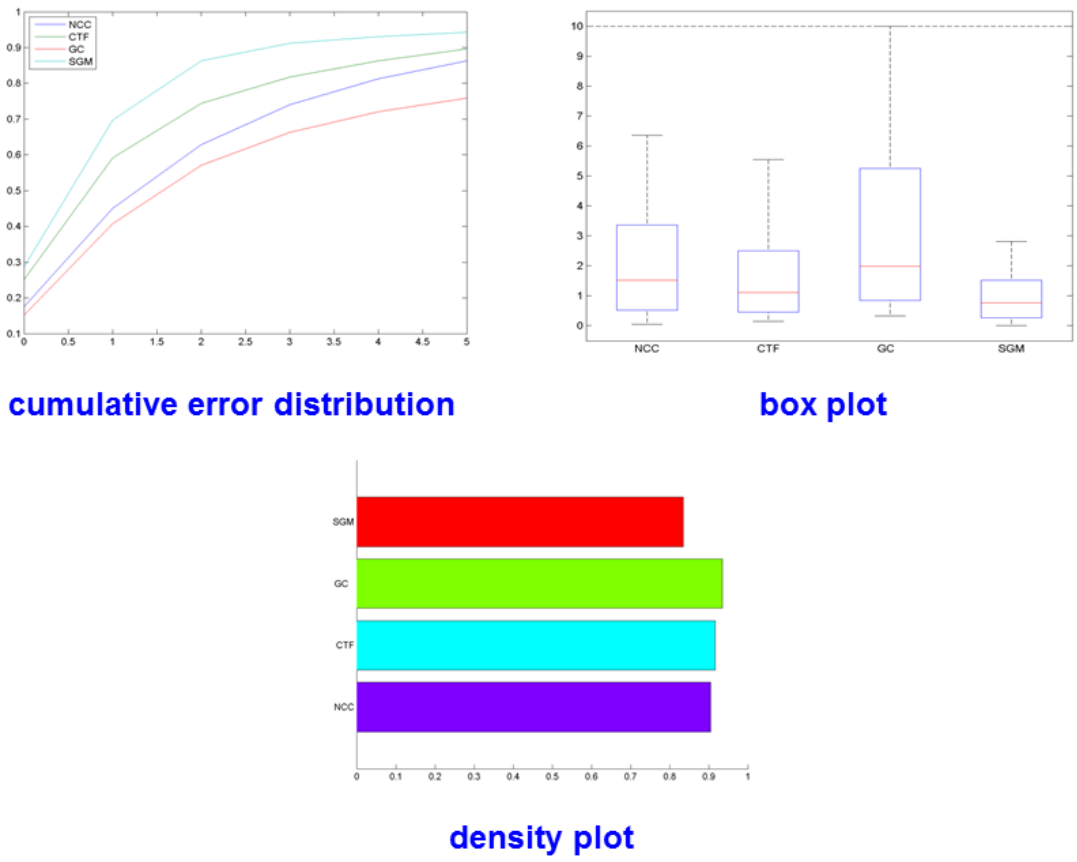


Figure 85: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab 41 cm Distance - Collapsed Across Organs (Excluding the Phantom).

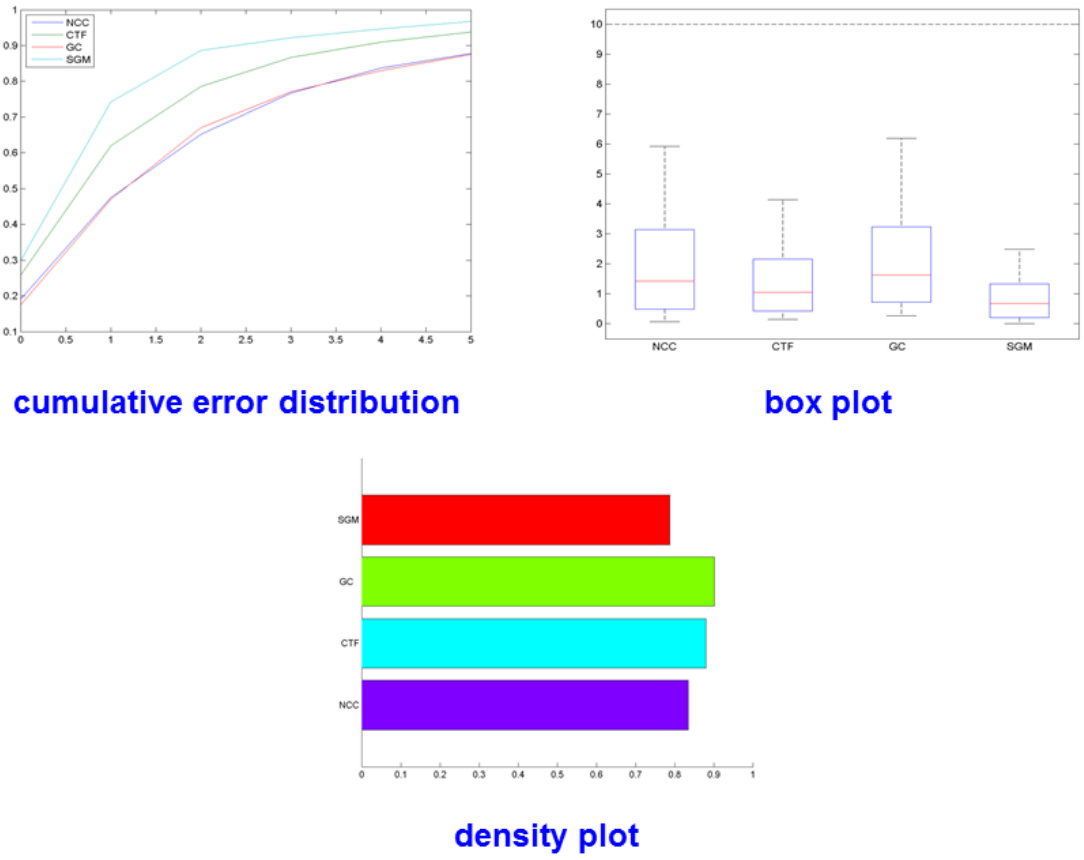


Figure 86: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on the Lab 48 cm Distance - Collapsed Across Organs (Excluding the Phantom).

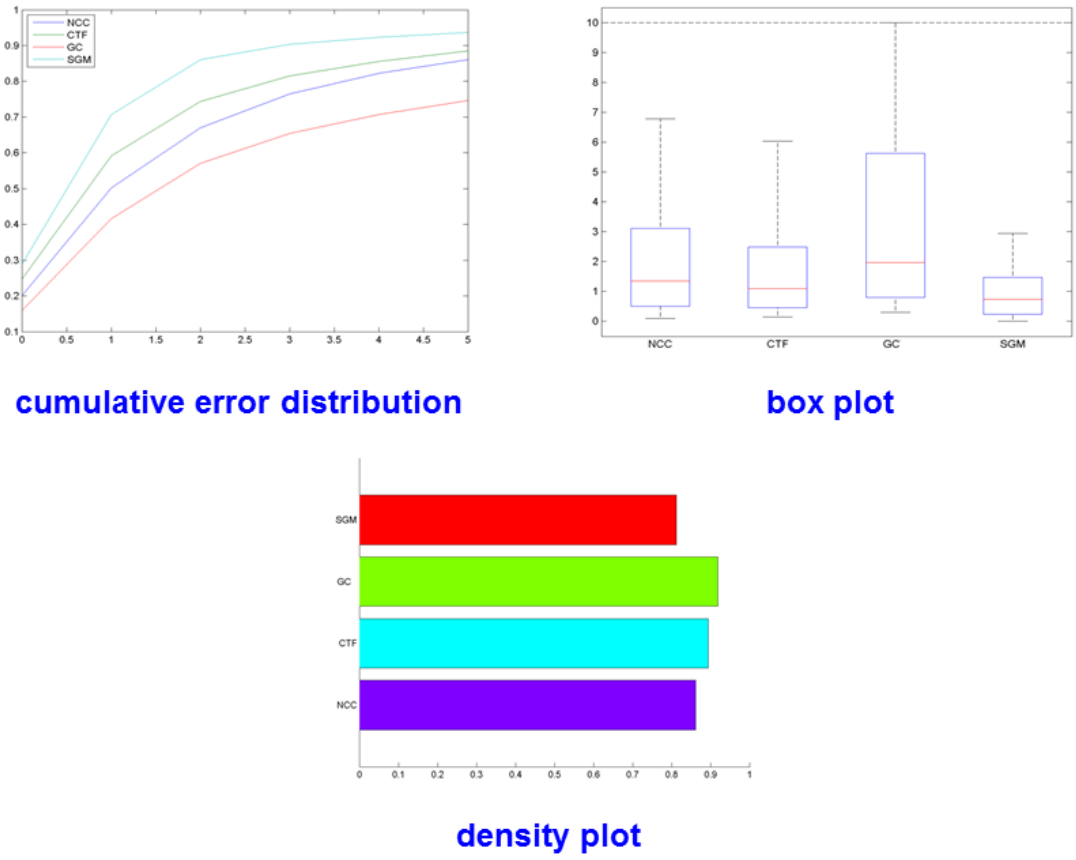


Figure 87: Cumulative Error Graph, Box Plot and Density Plot of 4 Algorithms on All the Datasets Acquired in the Lab (Excluding the Phantom).