REDUCING THE FALSE POSITIVE RATE OF COMPUTER AIDED DETECTION FOR CT COLONOGRAPHY USING CONTENT BASED IMAGE RETRIEVAL

Javed Aman, Jianhua Yao, Ronald M. Summers

Radiology and Image Sciences Department, Clinical Center
National Institutes of Health, Bethesda, MD 20892, USA

ABSTRACT

We present a method to reduce the false positive rate of our computed tomographic colonography computer aided detection (CTC-CAD) system based on the paradigm of Content Based Image Retrieval (CBIR). Using the feature vectors generated from our CTC-CAD system in conjunction with the distance metric of our CBIR system we rank detections based on their likelihood of being true positive. By eliminating detections with low rank, we prune approximately half of the false-positive detections from our CAD system.

Index Terms— Content Based Image Retrieval, CT Colonography, Computer Aided Detection

1. INTRODUCTION

Computer aided detection (CAD) systems have proven their usefulness in clinical environments by increasing the sensitivity of interpretations. In particular, the field of computed tomographic colonography (CTC) has benefited from CAD systems [1].

An interesting area of research involves reducing the false positive (FP) rate of the CAD systems. False positive detections increase overall observer time and can lead to inappropriate referrals and more invasive follow-up procedures.

Methods based on medical preparation and elimination of detections on ileocecal valves or rectal tubes have been shown to reduce the FP rate [2]. However, many FPs cannot be attributed to any of the aforementioned causes [3]. In this study we present a more general approach to remove false positives based solely on the features attributed to our CTC-CAD detections. Detections are structures in the colon which the CTC-CAD system believes are polyps.

Our approach builds on the content based image retrieval (CBIR) technique. CBIR has been adapted for a variety of applications including general image searching (i.e. GiFT, ImgSeek) [4], artwork retrieval [5], and medical applications [6,7]. Most notably, CBIR was effective in evaluation of mammogram images [8].

Unlike traditional text-based search engines and image retrieval systems based on them, CBIR relies on the content of the image itself. Considering the vast amounts of imaging data available today, it would seem improbable to have it all labeled based on features of the image. Instead, CBIR utilizes features or numerical values given to different aspects of an image. The features used in this study are generated by our CAD system and are used to describe detections. A feature vector is the list of features associated with each detection.

The goal of this study is to demonstrate that detections can be classified based on their features (i.e. FPs have similar feature vectors) using CBIR and with this information can remove many of the FP detections.

Figure 1. System flow chart

2. METHODS

The steps of the system are described in Figure 1. There are two phases: database construction and false positive reduction.

2.1. CTC-CAD

The detail of our CAD system is described in [9]. Given a CT colonography data, the CAD software segments the colon and produces a list of detections based on shape, curvature, intensity and texture features. For each detection, we extract a 64*64*64 image block and form a corresponding feature vector.

2.2. Search Database Construction

After the CAD is run on a test set of CT colonography data, each detection is semi-automatically labeled as either false positive (FP) or true positive (TP). We manually trace the boundaries of polyps on CTC, and all detections that fall inside the manual segmentation are considered TPs. A MySQL database stores the paths to images and their corresponding feature vectors.

Generally, FP detections outnumber TP detections creating a bias in the database. The search database contains the same number of TPs and FPs. We select the FPs based on their distances to the TPs in the database. The ones with greater distance are selected since they are unlikely to be in the result sets of TP queries. The distance metric will be described in Section 2.3.1. Figure 2 shows examples of TPs and FPs in the database.

Figure 2: Examples of two CBIR queries and their results. First row: a TP query; Second row: first five retrieved detections (all TPs) for the TP query; Third row: a FP query; Fourth row: first five retrieved detections (all FPs) for the FP query

2.3. Database Retrieval

Given a new detection, the CBIR retrieves similar ones from the database. This is accomplished by first computing the

distance of the query to all detections in the database using a distance metric. Detections in the database are ranked in ascending order with respect to distance to the query image; the most similar images are at the top, while the least similar ones are at the bottom. The first few images are returned in the result set.

2.3.1. Distance Metric

The *distance metric* is a function which defines the distance between two images. Examples of this function include Euclidean distance [10], Hausdorff distance [10], and others. In general a smaller distance implies a better match. Furthermore, a thresholding method can be implemented to further separate the results. Determining the threshold value requires statistical analysis of the database and the distance metric.

Our distance metric is the weighted absolute distance between two detections: $dist = \sum_{i=1}^n w_i |f_q^i - f_r^i|$, where f_q^i and f_r^i represent the values of feature i for the query and the database detections respectively. The weights, w_i , are set to be the reciprocals of the standard deviations of the feature spanning the entire database. Doing this removed the bias towards features with larger ranges and orders of magnitude.

2.3.2. Search Depth

Search depth is synonymous with the result set size, the maximum number of results the CBIR will return for a given query. By maximizing the percentage of TPs for TP queries and minimizing the percentage of TPs for FP queries, one can determine the optimal search depth.

2.4. False Positive Reduction

For the purpose of false positive reduction, given a new detection generated by the CAD system, a CBIR search is conducted against the search database.

The CBIR generates a result set of database detections for all test detections or query detections. The results are ranked (in ascending order) based on the distances of their feature vectors from the query detection's feature vector. Distances are computed using a specified distance metric. Test detections are considered strong false positives if a majority of their result set are false positives. These false positives can then be trimmed from the set.

Two queries and result sets are shown in Figure 2. The first represents a typical result set of a TP query and the second that of a FP query. Notice in general a TP query produces a majority of TP detections in its result set, while an FP query does the opposite.

2.5. Evaluation Methods

To evaluate the system we conducted two experiments. The first one, the *leave-one-out* test, is that the detections in the search database are used for retrieval test against the rest of

the database. The second one, the *independent* test, is that a far greater number of detections not in the database and better represents a typical CTC-CAD output from a new data set are used for retrieval test.

Two statistics are collected: the true positive ratio (TPr), i.e., number of retrieved TPs/search depth, and the index of the first incident of a TP (TPi). A detection with a high TPr and low TPi is predicted to be a TP.

3. RESULTS

3.1. Search Database Initialization

Our data set contained approximately 9100 detections from 162 CTC studies by our CTC CAD. Of them, 160 were defined to be true positives (larger than 6mm and in all histology types), while the rest were false positives. By simple probability, this discrepancy skewed our initial testing results towards false positive predictions. To overcome this issue, we trimmed the search database down to 160 TPs and 160 FPs. For all FPs, we computed the distance to its closest TP in the data set and picked the 160 FPs with largest distance. Of these 160 TPs and 160 FPs, 80 from each are randomly chosen to create the search database.

3.2. Search Depth Analysis

The optimal search depth can be found by comparing the average TPr for differing search depths. As the search depth increases the average TPr for TP queries decreases. In addition, the difference between TP and FP queries decreases. A plot is shown in Figure 3. Base on this result, optimal search depth for retrieval tests is ten.

Figure 3: TPr (%) versus Search Depth

3.3. Leave-one-out and Independent Test

The leave-one-out (LOO) test is applied to the search database and has 80 TPs and 80 FPs. Using a search depth of ten and the weighted absolute distance metric the LOO test was run through the CBIR. The statistics of the test are shown in Table 1. A histogram showing the distribution of TP and FP queries is shown in Figure 4.

The independent set is the collection of CTC-CAD detections which are not in the search database. There were 80 TPs and 8860 FPs in the test set. The other 80 TPs and 80 FPs made up the search database. The statistics of the independent test is in Table 2. A histogram of TP and FP queries is shown in Figure 5.

Table 1. Statistics of CBIR with LOO test

TP/FP	Count	TPr	TPi
TP	80	0.81 ± 0.24	1.3 ± 6.9
FP	80	0.28 ± 0.23	3.5 ± 3.1

Table 2. Statistics of CBIR with Independent test

TP/FP	Count	TPr	TPi
TP	80	0.80 ± 0.23	1.3 ± 6.9
FP	8860	0.27 ± 0.32	3.2 ± 2.9

Figure 4: Histogram of the LOO test. There are 80 FPs and 80 TPs in the test.

Figure 5: Histogram of the independent test. There are 8860 FPs and 80 TPs in the test.

The likelihood a detection is TP depends on the TPr of the result set. FP detections averaged TPr of .28 and .27 of the LOO and independent tests respectively. This result may seem surprising, but can be explained as the number of FPs in the independent test is significantly larger than that in the LOO test. Although the LOO test contained a subsample of the “best” FPs, many of the other FPs matched with them are not included. Furthermore, the FPs in the LOO test do not necessarily match with the other FPs in the database. The

independent test includes many of the missing FPs matched with the FPs in the LOO test.

TP queries averaged 0.81 TPr for the LOO test and 0.80 for the independent test. In both situations, TP queries had significantly more TPs in their result sets than FP queries.

Examination of the TPi statistics reveals TP results appear closer and earlier with TP queries than with FP queries. Nevertheless, the difference between the TP and FP queries considering these fields is minimal. However, the results are somewhat misleading; they do not include cases where no TP appeared in the result set. In those situations, the TPi is not defined. This is especially important since there is a large fraction of FP queries with no TP in their result set.

Figure 6: ROC curves of both LOO and Independent Sets

3.4 False Positive Reduction and ROC Analysis

Receiver operating characteristic (ROC) curves are generated using the histogram information given from both the LOO and independent tests. The curves are shown in Figure 6. While maintaining 100% sensitivity, nearly 15% of the FPs can be removed. This is achieved by removing the queries with zero TPs in their result set. In the case of the training set, up to 57.5% of the FPs can be removed while maintaining 96.3% sensitivity. The pruned detections correspond to queries which have only zero to three TPs in their result sets. Removing these from the test set results in a sensitivity of 97.5% and 51.1% FPs removed.

4. CONCLUSION

The false positive rate of our CTC-CAD can be reduced using a CBIR paradigm. Our study shows over a 50% reduction in false positives while maintaining a high sensitivity. This is achieved by eliminating query detections which match poorly with the true positives in the search database.

The CBIR approach is fast, simple, and can easily be generalized. However, it is dependent on the strength of the feature selection, distance metric, and database. Once the

search database is established the system is fully automatic. In combination with other FP removal techniques, a CAD system can be highly effective. Furthermore, with proper labeling, this system can be extended to differentiate types of polyps (i.e. hyperplastic/adenoma). These features will be implemented in the future.

5. ACKNOWLEDGEMENT

The authors thank Perry J. Pickhardt, William R. Schindler and Richard Choi for providing computed tomographic colonography and supporting data. This research was supported by the Intramural Research Program of the National Institutes of Health, Clinical Center.

6. REFERENCES

- 1) Summers RM, et al. Automated polyp detection at CT colonography: feasibility assessment in a human population. *Radiology*; 219:51-59, (2001).
- 2) Summers RM, Yao J, Johnson CD. CT Colonography with Computer-Aided Detection: Automated Recognition of Ileocecal Valve to Reduce Number of False-Positive Detections. *Radiology* 233:266-272, (2004).
- 3) Yoshida H., Dachman A.H. "CAD Techniques, Challenges, and Controversies in Computed Tomographic Colonography" *Abdominal Imaging* 30:26-41, (2005).
- 4) D.M. Squire, W. Müller, H. Müller, T. Pun, Content-based query of image databases: inspirations from text retrieval, *Pattern Recognition Letters*, vol. 21, pp. 1193-1198, (2000).
- 5) Cuicca G, Schettini R. "A relevance feedback mechanism for content-based image retrieval" *Info. Proc. and Manag.* 35:605-632 (1999)
- 6) Muller H., et al., Review of content-based image retrieval systems in medical applications—clinical benefits and future directions, *IJMI* 73 1-23, (2004).
- 7) Lehman T., Automatic Categorization of Medical Images for Content-Based Retrieval and Data Mining, *Computerized Medical Imaging and Graphics* 29, 143-155, (2005)
- 8) Azevedo-Marques P. M., Reducing the semantic gap in content-based image retrieval in mammography with relevance feedback and inclusion of expert knowledge, *Int J CARS*: 3 123-130, (2008)
- 9) Summers R., et al.. "Computed Tomographic Tomographic Virtual Colonoscopy Computer-Aided Polyp Detection in a Screening Population" *Gastroenterology* 129:1832-1844, (2005)
- 10) Datta R., Joshi D., Li J., Wang J., Image Retrieval: Ideas, Influences, and Trends of the New Age, *ACM Computing Surveys* 40 5:1-60, (2008).