Estimating Marbling Score in Live Cattle from Ultrasound Images Using Pattern Recognition and Neural Network Procedures

John R. Brethour
Kansas State University, Fort Hays Branch Experiment Station, Hays 67601

ABSTRACT: Neural network processing of texture statistics (which parameterized longissimus muscle echograms of live cattle) resulted in marbling estimates that differed from corresponding USDA carcass marbling scores by an average of .42 marbling score units. This was more accurate ($P < .001$) than using the same features in a multiple regression model. Images were used from 53 cattle in the training set and from 108 cattle in the validation set. Over 500 texture statistics (including variations in direction, resolution, and step size) were screened to identify three candidates (Markovian homogeneity – step size = one; third quadrant emphasis from the bit-4, normalized run length/gray level matrix; and 12-pixel local standard deviation) for intensive analysis. The differences between the live animal estimates and carcass marbling were not much greater than the human error in assigning carcass marbling scores. When the results were subjected to receiver operating characteristic analysis, accuracies in grade classification were comparable to clinical, diagnostic imaging evaluations. It is feasible to incorporate this procedure into a computer interfaced with an ultrasound system to provide unsupervised instrument evaluation of live cattle in “near real time” (2 or 3 s).

Key Words: Ultrasound, Marbling, Beef Cattle, Image Analysis, Neural Network, ROC Analysis

Introduction

The amount of marbling (intramuscular fat in the longissimus muscle) is the most important element in determining quality grade in the U.S. system and putatively is associated with both beef palatability and product consistency. It is an economically important trait, as evidenced by the monetary premiums that are frequently paid for USDA Choice and Prime.

However, visual assessment of marbling in the live animal is virtually impossible, so cattle often are overfed to assure acceptable marbling. This results in excessive fat, which is the primary problem of product quality in the beef industry. An earlier paper (Brethour, 1990) indicated that ultrasonic speckle, backscattered from the small and irregularly shaped marbling deposits, was related to the degree of marbling. A skilled sonographer can visually interpret an echogram and estimate marbling in a live animal with fair accuracy (Figure 1). Pictures that depict ultrasound images corresponding to different degrees of marbling have been published (Brethour, 1992). However, visual image interpretation is subjective and not easily taught, scaling is imperfect, and absolute values are often inaccurate. Furthermore, maintaining constancy across insonation sessions is difficult, so the method is not readily adaptable to serial measures that monitor marbling development over time.

A pilot study (Figure 2) indicated that subjecting the ultrasound image to pattern recognition analysis (computer vision) combined with artificial neural network procedures might enable a computer program to make unsupervised estimates of marbling in the live animal. Results were sufficiently promising that a larger study was conducted with more subjects and greater diversity in marbling score to determine the accuracy of the interfaced technologies in estimating marbling.

Materials and Methods

This section describes in detail how ultrasound images were captured and processed. In pattern recognition (computer vision) procedures, images are parameterized to enable identification of desired

\[1\] Contribution no. 94-135-J of the Kansas Agric. Exp. Sta. The author acknowledges the efforts of Jerry D. Gresham, Univ. of Tennessee at Martin, for assistance in data collection; John B. O'Loughlin, Director CAD/CAM Laboratory, Wichita State Univ., for writing software to convert TIFF files to ASCII format; and to Charles E. Metz, Dept. of Radiology, Univ. of Chicago, for providing software to perform ROC analysis.

Received September 29, 1993.
Accepted March 2, 1994.
attributes. In this research, a special application of pattern recognition, texture analysis, was used for this parametrization. Statistical features that described differences in image texture were subjected to neural network analysis in attempts to refine the ability to associate the texture statistics with marbling score.

Different sets of cattle and images were used for developing prediction models and training neural networks than were used in validating and testing procedures; that was more rigorous than splitting a data set. The calibration images were taken from two sets of cattle, one at Hays, KS, and one at Martin, TN. The test set of images was taken from 108 market-ready steers that had been on full feed for several months. Animals represented a cross-section of *Bos taurus* breeds and were between 15 and 20 mo old. Cattle were slaughtered in a commercial packing plant within 1 wk of insonation, and marbling scores were called to tenths of a score by experienced USDA graders after a 24-h chill and approximately a 15-min bloom following ribbing. Marbling was quantified by assigning a score of 4.0 to Slight 00, 5.0 to Small 00, and 6.0 to Modest 00.

Images were acquired with an Aloka 210 ultrasound system (Corometrics Medical Systems, Inc., Wallingford, CT) equipped with a 3.5-MHz general purpose transducer array (UST 5021 - 125 mm window). Mineral oil was the couplant, hair was not clipped, and no standoff was used. A tomogram of the longissimus muscle lateral to the backbone was obtained by placing the probe both parallel and orthogonal to the 12th rib. Approximately 30 s of real-time imaging from each animal was captured on conventional videocassette tape and transported to an image analysis system.

A Targa M8 (Truevision, Inc., Indianapolis, IN) image capture board was used in conjunction with Java (Jandel Scientific, Corte Madera, CA) video analysis software to digitize a region of interest (ROI) from the tape. The ROI was a rectangle about 40 x 80 mm (20,000 pixels) from a uniformly echoic area over the longissimus muscle (Figure 3). This was chosen after preliminary evaluation of several other geometries and locations. The TIFF format of the digitized file was converted to ASCII format, and image texture analysis was performed in Microsoft FORTRAN. Regions of interest were captured from three separate frames from each animal, and texture statistics from those three pictures were averaged for statistical and neural network analyses.

A large number of statistics to parameterize the image were screened from the training set. Those included first-order pixel value parameters such as mean, standard deviation, skewness, and kurtosis. Markovian texture parameters, which detect nonrandomness in values of pixel pairs, are frequently used.
in pattern recognition to quantify optical texture (Haralick, 1979) and do not rely as heavily on periodicity as do Fourier transforms (Pressman, 1976). Correlation and homogeneity from the co-occurrence matrix were the only Markovian features evaluated, because they had been identified in an earlier study (Brethour et al., 1992) as especially efficient. Autocorrelations from Box-Jenkins time series methodology also identify regularities in wave structure (Pankratz, 1983). However, the second lag partial autocorrelation function, which behaved as a gradient operator, emerged as a better predictor than individual autocorrelations. Other features included fractal dimension, mean run length, and the gray level distributions of run lengths proposed by Chu et al. (1990). A number of novel measures were devised, and two were especially effective predictors: the standard deviation within 12-pixel primitives and either third or fourth quadrant emphasis in the run length/gray level matrix, which was inspired by the work of Chu et al. (1990). The latter statistic measured bias in the distribution of run lengths by gray levels by accumulating the squared values of the run lengths (normalized to the gray level distribution) divided by the squared values of the gray levels. All texture statistics were normalized so they would be invariant to ROI size and shape.

The scan converter of the ultrasound system stored images in 4-bit (16 gray levels) format, but they were digitized into 8-bit format by the image capture board. The examination of statistical features included computed resolutions of 7, 6, and 4 bits; the last was most appropriate for run length features. Vectors that were 45, 90, and 135 degrees to the axial direction also were checked, but they produced inferior measures. Co-occurrence matrices were built to canvass Markovian features in steps from 1 to 20 from which 11th or 12th order appeared most preferred. The time gain compensation of the ultrasound system always was set to eliminate beam attenuation, was included as a texture measure (beam attenuation, was included as a texture feature).

Some image preprocessing was applied. Synchronizing the digitizing system with the playback videocassette recorder was difficult, and resulted in raster interlacing, so each axial vector (A-line) was normalized by the FORTRAN program. The time gain compensation of the ultrasound system always was set high to acquire a level image, but additional trueing was done on the ASCII file with mathematical regression (the regression coefficient, which ought to measure beam attenuation, was included as a texture feature).

Texture measures were screened with multiple regression analysis (using carcass marbling as the dependent variable) to identify the best combinations of two or three high-quality variables. The search objective was a parsimonious model with as few parameters as possible to enhance stability during validation. Criteria for selection of texture statistics included repeatability within images from the same animal, constancy of predictive efficiency across sets of cattle, and absence of multicollinearity.

A model with three texture statistics (Markovian homogeneity - step size = one; third quadrant emphasis from the bit-4, normalized run length/gray level matrix; and 12-pixel local standard deviation) was chosen for development and testing. In addition to validating a conventional multiple regression equation, the same feature vectors were subjected to neural network analysis using Neuralworks Explorer software (NeuralWare Inc., Pittsburgh, PA).

Artificial neural networks (ANN) mimic brain learning processes. A set of input variables (in this application, the numerical texture statistics) and desired responses are presented repeatedly to the ANN program until it is able to relate entry patterns to correct outcomes. This is done through a back propagation procedure that adjusts weights of connections between nodes until algorithms that minimize output errors converge. The NeuralWare documentation provides an excellent introduction to this technology, and an especially lucid and thorough explanation of a neural network application can be found in the paper by Jando et al. (1993).

Simple architectures that generalized the data predicted more accurately than complex ones, which tended to memorize. The most useful topology was a three-layer structure with three input nodes, three hidden nodes, and one output node, which contained the numerical marbling score values. A bias node was connected to the processing elements in the hidden and output layers. Processing elements were connected only to those in the next higher layer, and sine transfer functions were used. The best performance was obtained when only about 2,200 presentations (40 epochs) were made during training. In building the multiple regression model and training the ANN, the data set was reduced from 82 to 53 animals by discarding outliers (individuals that did not fit the original multiple regression model) and also randomly casting out some of the near-average values to flatten marbling score distributions. Training with well-behaved examples should be more effective than with promiscuous data.

Often, categorizing carcasses into discrete grades is of as much interest as estimating marbling scores per se. A decision matrix was created by assuming that all carcasses with Small\(^{0}\) marbling or better would grade Choice. The texture data also were subjected to discriminant analysis (SAS, 1985) in which the PRIORS statement was set proportional to the distribution in the training set (32% Select and 68% Choice). This approximated the distribution in the validation set (22% Select and 78% Choice).

However, conventional sensitivity/specificity analysis has pitfalls related to variation in prior probabilities and decision criteria. A preferred procedure is
receiver operating characteristic analysis (ROC), which is explained by Metz (1978) and also Swets (1988). Results were subjected to ROC analysis using the LABROC1 software provided by Charles E. Metz.

Results and Discussion

Figure 4 shows pixel values along axial vectors of ultrasound echograms. Greater periodicity occurs in the vector from the Prime animal; in fact, Wilson et al. (1993) have successfully exploited spectral analysis of ultrasound images to estimate marbling score. Many of the texture features quantified the pattern differences that are visibly apparent between the two projections in Figure 4. For example, the strongest single correlation of a feature with marbling was the variance in wave amplitude. Images from poor-marbling animals contained blotches of low echogenicity, which corresponded to long runs at low gray levels. This was quantified with fourth quadrant emphasis in the run length/gray level matrix. The regular waves in the Prime animal probably were related to coherent speckle caused by backscatter from marbling deposits, whereas the peaks in the Select animal resulted from specular echoes associated with the vascular system.

Table 1 shows a sample of the simple linear correlation coefficients between texture statistics and marbling score, including those features that consistently emerged as the best single predictors. The coefficients in this table are from three sets of cattle (n = 202) with a mean marbling score of Small 48 and a SD of .88 marbling score units. Mean gray level, putatively associated with marbling, was only about 25% as efficient as the most powerful predictors (relative size of squared coefficients). Attenuation associated with high degrees of marbling often reduced the image gray level. Liu et al. (1993) also obtained little predictive efficiency from mean image intensity; conversely, Gresham (J. D. Gresham, University of Tennessee at Martin, personal communication) found a good relationship between mean gray level and marbling score. Second-order features should be more invariant than first-order measures to differences in echogenicity related to transducer contact, hair coat, and attenuation. Considerable mimesis seemed to exist among texture features because little increase in prediction power occurred when more than three elements were added to the multiple regression model (Figure 5). Other sets of two or three texture statistics might have performed as well as the model chosen for evaluation.

Figure 6 shows results from validating the multiple regression model that was composed of three texture features. The figure includes both a regression line and the isopleth. Although the former has been the standard in most reports, the deviation from the isopleth is a better measure of accuracy because it represents a plot with the origin at zero and unity slope. The coefficients of determination state the percentage of total variance associated with each function. The average difference between the predicted and actual marbling scores with the multiple regression procedure was .59 marbling score units.

The ANN procedure was more accurate than the multiple regression technique and analysis of the reduction in variance showed that it reduced
ESTIMATING MARBLING FROM ULTRASOUND IMAGES 1429

Figure 5. Size of the coefficient of determination for estimating marbling score as a function of the number of texture features in the model.

(P < .001) the average difference between estimated and carcass marbling scores to .49 units (Figure 7). Two attributes of the ANN probably accounted for its superiority. It was not limited to a single prediction equation, but contained a model for each pathway from the input nodes to the output node; the minimally connected topology used here contained three simultaneous models. Also, the sine transfer function in the ANN allowed nonlinearity in weights from layer to layer. After training, weights could be retrieved from the ANN platform and used in direct computations. However, both the values of the texture statistics and the ANN weights will vary among different ultrasound systems, so recalibration will be necessary when an image acquisition system is changed.

The ANN may be even more accurate than Figure 7 indicates. The training set had no Prime cattle, so it may be naive to expect the ANN to classify that grade correctly. If the four Prime animals had been omitted from the test set, the ANN would have performed with an average error of .42 marbling score units compared to an error of .56 with the multiple regression model.

However, the ANN was very sensitive to small changes, such as the number of presentations, in the operating protocol. Even though a separate testing set was used, the results were obtained retrospectively and the ANN model may have been spurious because convergence had not been attained during training. When the ANN was trained with a sigmoid transfer function, with which tighter convergence could be obtained than with the sine function (after about 100,000 presentations), the error between predicted and carcass marbling score was only 3% smaller than that achieved with the multiple regression procedure. Possibly, the ANN merely learns the multiple regression model when relationships are highly linear, which was the situation in these data because adding quadratic components to the multiple regression model did not improve its fit. A need remains to test ANN capabilities in an unbiased, a priori environment.

Figure 8 portrays the decision matrix for classifying the ANN results into Select or Choice (positive). The sensitivity (ratio of predicted Choice to total Choice), specificity (ratio of predicted Select to total Select), and accuracy (percentage of correct decisions) were 88%, 71%, and 84%, respectively. Classification from discriminant analysis (using the same input variables and building a discriminant function from the training set) resulted in values for the respective categories of 95%, 50%, and 85%. The reduction in specificity with the discriminant analysis procedure seemed to more than offset the gain in sensitivity. Also, discriminant analysis did not provide the direct estimates of marbling score obtained with the other procedures.

Receiver operating characteristic analysis is superior to the conventional decision matrix for evaluating diagnostic accuracy, because results are not contingent on event distribution and the procedure
allows a user to compare classification sensitivity with the false positive ratio at different levels of decision criteria. The ANN procedure for discriminating between Select and Choice had an $A_z$ value of 86.4 (Figure 9). The $A_z$ value (area under the curve) is the probability of a correct decision in a two-alternative choice and often is cited as a quantitative measure of the accuracy of a diagnostic system. $A_z$ values from multiple regression and discriminant analysis (using the arcsine of the probability of Choice classification) were equal (85.7, SD = 3.9). The high $A_z$ values confirmed that any of the three methods might be used for transforming image texture statistics into carcass grade classifications.

Receiver operating analysis of the ANN results also was performed to determine accuracy in identifying cattle with Modest or more marbling (Figure 10). This is a requirement in the Certified Angus Beef program and also in specialty products from some packing plants. The resulting $A_z$ value was 94.4, indicating exceptional discriminating ability. From a clinical perspective, the diagnostic accuracy of 86.4% for distinguishing between Select and Choice is comparable to the ability to correctly classify mammograms and the $A_z$ value of 94.4% is equivalent to the accuracy of radiologists in detecting lesions in lung x-ray films (Swets, 1988).

The derivative of the ROC curve (Figure 11) provides an estimate of classification accuracy at specific points on the criteria axis. This can be especially valuable in upstream evaluations that predict future carcass merit and where conditional probabilities of quality grade premiums must be matched to risks of yield grade penalties. Metz (1978) explains the application of ROC curves to cost-benefit analysis.

Pattern recognition involving texture analysis has been used extensively in clinical research on ultrasonic tissue characterization. Raeth et al. (1985) reported that computerized image analysis was superior to subjective visual evaluation of liver echograms in characterizing liver pathology. Another excellent paper on texture analysis of ultrasound liver images was published by Nicholas et al. (1986). Both papers provide formulas for calculating texture statistics. Finette et al. (1983) discussed similar techniques for characterizing breast disease in humans. A clinical application combining image analysis and neural network procedures similar to the protocol in this report can be found in Goldberg et al. (1992).

Image analysis of echograms has not been exploited widely for estimating marbling in cattle, even though researchers observed as early as 1980 (Anselmo et al.,...
1980) that differences related to marbling occurred in the ultrasonic A-scan signature. Green et al. (1991) used a discriminant function built from histograms of pixel gray levels to predict marbling score in live cattle. Whittaker et al. (1992) reviewed the extensive efforts in their laboratory to evaluate both echographic techniques and frequency analysis for estimating marbling score with ultrasound in both live and slaughtered animals. Autocorrelation procedures were used to parameterize speckle in B-mode images and estimate marbling by Liu et al. (1993). Differences among research protocols make it difficult to compare those methods with the system presented in this paper. Wilson et al. (1993) have exploited spectral analysis of image pixel values to estimate fat content in the longissimus muscle of live animals. Equipment to estimate marbling from image analysis has been developed in Canada (David R. C. Bailey, Agriculture Canada, Lethbridge, personal communication). Also, a project is under way at the University of Illinois to measure beef carcass characteristics with ultrasound (Hein et al., 1992).

The author's work was performed with relatively low-grade equipment, so it was encouraging that the procedures estimated carcass marbling score in the live animal with about the same accuracy that might be expected from two individuals grading the same carcasses. The Aloka 210 has only 16 gray levels, and the ultrasound beam is focused poorly. Considerable smearing occurs in the format conversions from digital to analog and back to digital during image transport on videocassette tape. Better control of focusing and signal preprocessing along with higher gray level resolution and the ability to transfer digital data directly from the ultrasound scan converter into computer processing should enhance the accuracy of this method. Visual marbling score is a subjective target; percentage of fat in the longissimus muscle would be more objective.

Furthermore, off-the-shelf ultrasound systems designed to portray an image probably discard as much as 95% of the information conveyed in the original signal (Powis and Powis, 1984). Ultrasound interacts at cellular and molecular levels, as well as echoing from specular boundaries, but analysis of the unprocessed signal will be needed to extract this additional information. The method reported by Perry et al. (1990) was based on measuring attenuation, which is correlated with marbling, from frequency analysis. Waveform analysis (Park and Whittaker, 1990) may complement the echographic procedures described in this paper and enable direct assessment of beef palatability factors such as juiciness and tenderness in live animals.

Automatically estimating marbling upstream in the production process may be more important than evaluating market-ready cattle. Marbling estimates at weaning from visual interpretation of the echogram were correlated ($R^2 = .35$) with carcass marbling collected 9 mo later (Brethour, 1992). Models that predict marbling as a function of time in the feedlot probably can be developed. They might be similar to existing equations for predicting future backfat thickness (Brethour, 1988). Combining marbling prediction with backfat estimates, which are highly related to yield grade, would enable producers to target narrow carcass specifications with appropriate clustering strategies.

In conclusion, this study indicates that it is feasible to exploit available technologies, including ultrasound, image analysis, and neural network procedures, to automate marbling estimation in live cattle. A computer equipped with programs containing parameters from research such as this can be interfaced with an ultrasound system to provide instrument grading of live animals in "near real time" (2 or 3 s). Such monitoring of this economically important trait would create opportunities to manage cattle to attain carcass marbling targets and to focus on marbling selection in seed stock herds.

**Implications**

The research summarized in this paper divulges methods that might be incorporated into an instrument to estimate marbling in the live animal. By exploiting and interfacing ultrasound, computer vision, and neural network analysis, a system can be constructed that provides unsupervised image interpretation and reduces the need for trained sonographers. A user-friendly system that accurately and rapidly measures marbling in live cattle would be a welcome tool for producers who strive to improve carcass merit and scientists who desire to monitor marbling development.
Literature Cited


