Cartridge Clustering for Improving Tone Prediction Accuracy in Calibration for Color Electrophotography*

Chao-Lung Yang[a], Yan-Fu Kuo[b], Yuehwern Yih[c] , George T.-C. Chiu[d], , and Jan P. Allebach[e]; a: Department of Industrial Management, National Taiwan University of Science and Technology, Taipei, Taiwan ROC, b: Department of Bio-industrial Mechatronics Engineering, National Taiwan University, Taipei, Taiwan ROC, c: School of Industrial Engineering, d: School of Mechanical Engineering, e: School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana USA

Abstract

The variation of supplementary components, such as cartridges, has been observed to have an impact on the color electrophotographic (EP) process. The preliminary study shows that the cartridge variation affects the calibration sensor mapping which predicts colorimetric tone reproduction (CTR) values on printing media as a function of internal sensor readings from halftone patterns printed on the printer transfer belt. This paper presents the result of the data-driven study to identify the cartridge patterns and further improve the tone prediction accuracy in calibration of EP platform. First, the constrained hierarchical clustering method was applied to generate a variety of cartridge groups according to the characteristics on the sensor mapping. A new sensor mapping model including a cartridge clustering module was proposed not only to compensate for environmental and consumable conditions, but also to consider cartridge variation. The experimental results show that overall accuracy of the sensor mapping can be improved by ~ 40% on average after considering the cartridge factor.

Introduction

A color laser printer, i.e., a color electrophotographic (EP) printing system, is physically a binary process which produces all output colors by the combinations of certain primary colors such as cyan, magenta, yellow, and black (CMYK). To reproduce a primary color with a desired tone ranging from 0 to 255, the half-toning process is proceeded to translate the desired continuous tone image into a half-toned image labeled with a half-toned density. Then, the appropriate amount of toner is laid down on the media by a particular pattern with the given half-tone density [1]. The colorimetric tone reproduction (CTR) measurement such as CIE L*a*b* is applied to represent colorimetric characterization of the EP printing output on media.

Based on previous research [2-4], temperature, relative humidity (RH), cartridge toner consumption (CTC), organic photoconductive drum age, and developer bias voltages have been recognized as factors influencing color reproduction quality in terms of color consistency. Performing a calibration process periodically is a prevailing approach to maintain color consistency by restoring the printer control parameters to a desired state. To avoid the user involvement, the "off-media" calibration is performed. Basically, during a calibration, a number of color patches are printed on the transfer belt rather than output media and measured by on-board sensors to obtain "indirect" measurements, densitometer readings S_l (l denotes the tone level). Based on these readings S_l, a sensor mapping is constructed to determine how measured density S_l on calibration media relates to the actual printing density CTR_l on paper media. The sensor mapping result is then used in calibration algorithms to generate appropriate adjustments to EP control parameters

Fig. 1 shows an example of an off-line-generated calibration sensor mapping which predicts CTR values on printing media by densitometer readings S measured on substitute media. Note that the CTR values are measured during off-line development from printed uncalibrated test pages and are not available when the calibration is performed online. The prediction model can be constructed by linear regression to predict CTR from S. The variation caused by RH and CTC disturbances can be considered in the prediction model to improve the tone prediction accuracy [2].

However, as shown in Fig. 1, the distinguishable data clusters representing different cartridges show that the cartridge variation affects the sensor mapping accuracy. We can observe that the cartridge #2 cluster (square indicator) tends to shift to lower-left bound; while cartridge #1 (solid cicle) stays at the opposite location (top-right). The discrepancy among cartridge clusters in fact spreads out the sensor mapping variation and leads to high prediction error.

This research attacked this cartridge issue and proposed a clustering analysis framework to study cartridge impact by analyzing the sensor and measurement data collected from off-the-shelf color EP printers. Considering the variation caused by cartridge, a new sensor mapping model is developed for each cartridge cluster for improving the prediction accuracy during calibration.

* Research is supported by the Hewlett-Packard Company. Research partially conducted while Chao-Lung Yang was a graduate student at Purdue University.

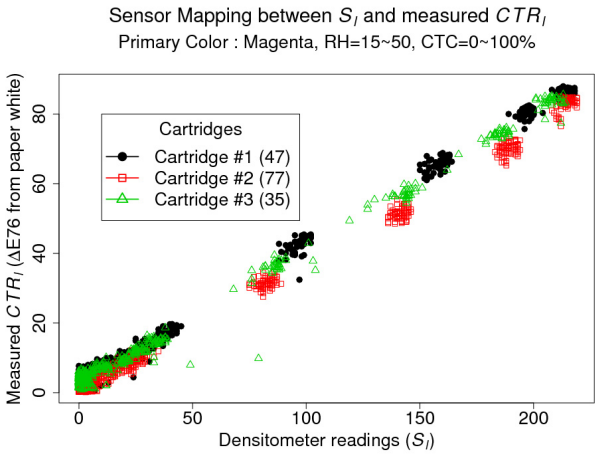

Figure 1. *An example of sensor mapping between CTR values (ΔE76 values from paper white) measured from the printed halftone pattern on substitute media and the internal densitometer sensor readings S (magenta cartridges in this case) from halftone patterns printed on the printer transfer belt. Indicators denote different cartridges and parentheses show the number of data points. The variation of the data mapping among different cartridges can be clearly seen.*

Cartridge Clustering

Clustering is one of unsupervised data mining tools. The goal of clustering is to partition the data points into groups (clusters) such that the instances in the same cluster have smaller pairwise dissimilarities than those in different clusters [5]. In general, there are two kinds of clustering algorithms: partitional and hierarchical algorithms. The partitional algorithm (such as K-means and its derived methods) simultaneously sections all data points to desired number of portions (clusters) based on similarity measures and the pre-determined number of clusters. On the other hand, the hierarchical algorithm recursively searches nested clusters according to the distance (or similarity) matrix which specifies the distance (or proximity) among data instances. The hierarchical algorithm constructs a tree structure, called dendrogram, in either agglomerative or divisive mode. The agglomerative mode starts from an individual data point in its own cluster and merges the most similar pair of clusters to form an upper-level cluster; the divisive mode starts from all data points in one cluster and divides clusters into smaller clusters (lower-level) based on similarity measurement.

In this research, the agglomerative hierarchical clustering algorithm was chosen to perform cartridge clustering because it does not use random initialization which may lead to different clustering result (note that K-means method assigns initial clustering centroid randomly). In addition, the dendrogram generated by hierarchical clustering can visually provide the proximity relationship among cartridges. Fig. 2 (a) shows an example of a dendrogram which indicates instance #3 are #5 are grouped first based on the shortest distance (2 in this case) between them, then the tree continues to merge instance #1 (distance is 3) into the nested cluster, and so on. The horizontal dash line denotes three clusters can be determined if the distance level = 4 is chosen as a cutoff point. This proximity relationship can help on data exploration of cartridge analysis.

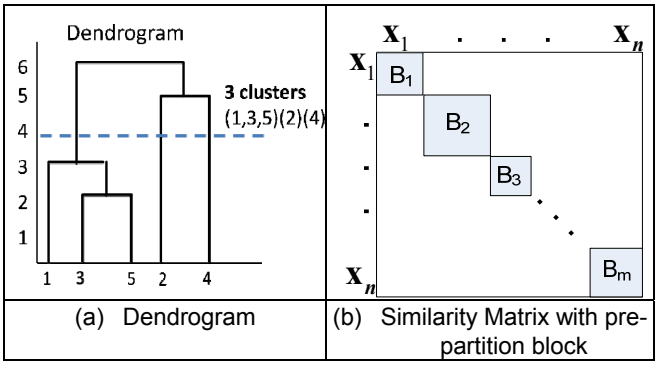

Figure 2. *(a) An example of hierarchical clustering dendrogram; (b) an example of distance matrix with pre-partition block*

In this work, given a set of data $\boldsymbol{D}$ with n data instances denoted as $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, Euclidean distance is used to calculate the distance between instances under a particular tone level l.

$$d(\mathbf{x}_{i,l}, \mathbf{x}_{j,l}) = \sqrt{(CTR_{i,l} - CTR_{j,l})^2 + (S_{i,l} - S_{j,l})^2} \quad (1)$$

Where i and j denotes different data points.

In order to guarantee the data points from the same cartridge are clustered together, the constrained version of hierarchical clustering method is used [6]. Essentially, the instance-level must-link (ML) constraints are constructed to specify that two instances must be placed in the same cluster if they are from the same cartridge. All ML constraints constructed by cartridge partitions in fact create pre-partition blocks with the similarity (or distance) matrix; see Fig. 2 (b) as an example. Each sub-square-matrix in Fig. 3(b) represents a cartridge partition. Due to this partition, the constrained agglomerative hierarchical clustering algorithm, in fact, uses centroids of cartridge blocks to produce a set of nested clusters.

Number of Clusters

In order to determine the number of clusters, the principal component analysis (PCA) is applied to investigate the sensor mapping variation. PCA, essentially, is a linear orthogonal transformation which converts a set of correlated $CTR_l - S_l$ variables into a set of values of linearly uncorrelated variables called principal components [7]. Fig. 3 shows an example of converting 5-cartridge dataset to two orthogonal principal components (PCs). The different hollow symbols indicate data points of 5 cartridges. The solid symbols denote the centroids of cartridge clusters. As seen in Fig. 3, the data points of cartridge #4 seem to be relatively far from other cartridges because all data points of cartridge #4 have negative projecting scores on the 2nd PC axis, while most of data points of other cartridges (#1, #2, #3, and #5) are on positive coordinate of the 2nd PC axis.

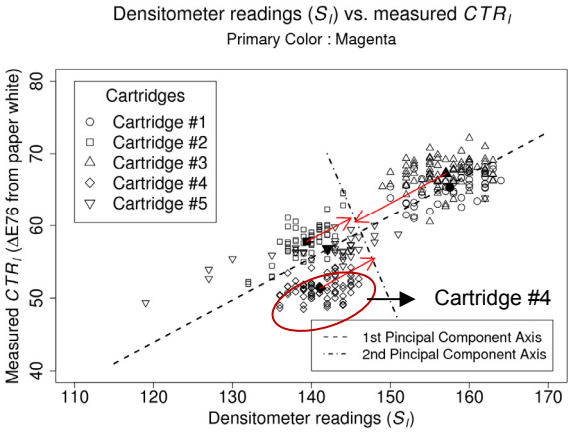

Figure 3. *An example of 5-cartridge sensor mapping on two orthogonal principal component axes.*

By using PCA, we can evaluate the projecting scores on the 2nd PC axis to determine the number of clusters which can contribute less overlapping of cartridge clusters on the 2nd PC axis. For each cluster, the 95% confidence interval (CI) of projecting scores on the 2nd PC axis can be constructed. In Fig. 4, each plot shows the mean (solid cycle) and the associated 95% CI of each cartridge group under different k. We can see when the number of clusters (k) increases from 2 to 5, the 95% CI among cartridge groups starts to overlap (when k = 5). In this case, k = 4 is the largest number of clusters before overlapping occurs.

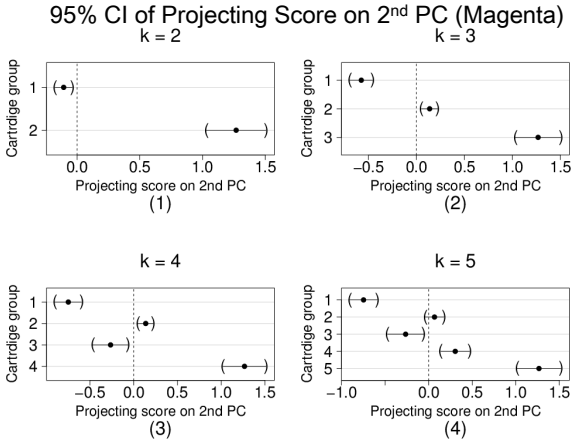

Figure 4. *95% CI of projecting scores of each cartridge cluster on the 2nd PC for different number of clusters (k).*

Experiment

The dataset used in this study was collected from three midrange off-the-shelf EP printers, which were all the same model, between June 2005 and October 2006. These three printers were operated and managed by the same entity within Purdue University. The data collecting procedure consists of three processes: (1) performing calibration, (2) retrieving sensor information including S_l from the printer, and (3) printing the test page. The test page contains thirteen color patches with various tone levels ranging from 0 to 160 according to an 8 bits/pixel tone scale (256 colors) for each primary color. 2083 test pages were printed and graded to obtain CTR_l values. The corresponding sensor information such as S_l, RH, CTC, and etc. were collected accordingly. Totally, the data with 18, 18, 16, and 18 cartridges were collected from three printers for CMYK, respectively.

For each primary color, the agglomerative hierarchical clustering algorithm was performed on mid-range tone level to determine the cartridge group. For CMYK, 3, 4, 3, and 3 cartridge groups were designated, respectively. For each found cartridge group, a sensor mapping model is developed to predict *CTR* values. The existing method, the proposed model in [2] which considers RH, CTC, and tone-level factors, and the new cluster-based model are compared by their prediction accuracy.

The root mean square error on cross validation (CVRMSE) defined as Eq. (2) is a conventional performance measure to compare the prediction models [8]. Essentially, the whole data set is divided by *k* folds. Each fold of data is selected randomly from the entire data pool without replacement. Then, the model training and validation are performed iteratively *k* times with different data sets. In each of the *k* iterations, one fold of the data is used as the validation set; and the rest of the data is the training set for developing the model. The RMSE of all predictions on each testing set is computed. In this article, 10-fold cross validation is used;

$$\mathrm{CVRMSE}_l = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(CTR_{l,i} - CTR'^{\,k(x)}_{l,i})^2}\,, \qquad (2)$$

where i=1,…,n; and n is the total number of data points. $CTR_{l,i}$ is the measured *CTR* value on printing media under tone level *l*; $CTR'^{\,k(x)}_{l,i}$ is the predicted *CTR* value by k-fold cross validation, and *k(i)* indicates the sensor mapping function developed by the *k*th training set.

Result

Because the results of CMYK behave similarly, we used magenta cartridges as an example to present the results. All 18-magenta-cartridge dataset are clustered into 4 groups based on the cartridge clustering result. The distinct sensor mapping model was developed for each cartridge group. Totally, 4 sensor mapping models for magenta were developed. Fig. 5 compares the CVRMSE of three models mentioned above. In order to compute the 95% confidence interval of CVRMSE, the bootstrapping technique was applied to generate 100 samples from the original dataset. The model training and testing were repeated 100 times on generated 100 samples.

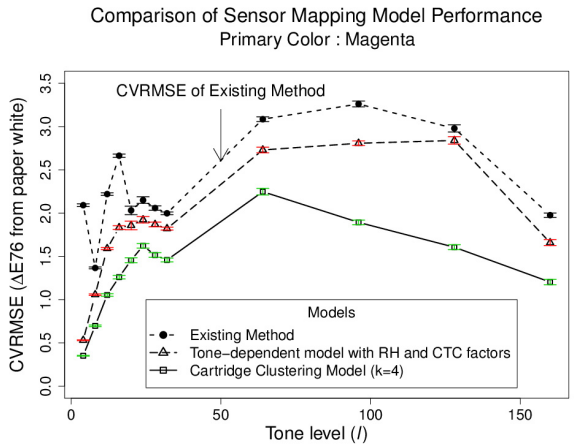

***Figure 5**. The sensor mapping performance computed by CVRMSE (ΔE76) on 18-magenta-cartrdige dataset. Each indicator shows the mean of CVRMSE for different models on a particular tone level. The associated 95% confidence intervals of mean CVRMSE are shown as horizontal lines.*

In Fig. 5, the black solid points show the CVRMSE of the existing method. They represent the worse case which the proposed method competes with. The triangle indicators represent the CVRMSE of the tone-dependent model with RH and CTC factors suggested in [2]. Furthermore, the square symbols indicate the CVRMSE of the proposed cartridge clustering model. Note that the proposed cartridge clustering model has considered tone-dependent characteristics and RH and CTC factors because the distinct model for each cartridge group was generated by the method proposed in [2]. The cartridge clustering model outperforms other methods. It implies that the prediction accuracy of sensor mapping can be further improved by considering cartridge factor.

Table I: Average root mean square error on cross validation (CVRMSE) comparison of sensor mapping models

Model	CVRMSE (ΔE76)				
	Cyan	Magenta	Yellow	Black	mean
# of Cartridge Group	3	4	3	3	
Existing Method	1.98	2.31	2.82	3.20	2.58
Tone-level-dependent Model with RH and CTC factors	1.02 (48.30%)	1.86 (19.41%)	2.22 (21.19%)	2.17 (32.20%)	1.82 (29.46%)
Cartridge Clustering Model	0.92 (53.34%)	1.36 (41.18%)	1.92 (31.89%)	1.87 (41.34%)	1.52 (41.10%)

Table I shows the CVRMSE results of three different models for CMYK. The CVRMSE shown in the table is the average across all tone levels. The parentheses denote the percentage improvement of the proposed models against the existing method. It is noted the proposed cartridge clustering model can improve the accuracy by ~10% against the tone-level-dependent model with RH and CTC factors proposed in [2]. There is a reduction of mean ΔE76 from 1.82 to 1.52. The overall accuracy against the existing method can be improved by ~ 40% on average. This is a reduction of mean ΔE76 from 2.58 to 1.52.

Conclusion

In this research, the cartridge impact on prediction accuracy of sensor mapping in EP printer calibration was observed. An agglomerative hierarchical clustering algorithm was applied to perform a cartridge clustering. For each cartridge cluster, the distinct sensor mapping model was developed. The experimental results show that the proposed clustering model is able to significantly further improve the prediction accuracy of sensor mapping.

Acknowledgements

We gratefully acknowledge the support from the Hewlett-Packard Company. We would like to especially thank Dennis Abramsohn for his valuable guidance in this research.

References

[1] R. Ulichney, Digital Halftoning (MIT Press, Cambridge, MA, 1987).
[2] C.-L. Yang, Y.-F. Kuo, Y. Yih, G. T.-C. Chiu, D. A. Abramsohn, G. R. Ashton, and J. P. Allebach, "Improving tone prediction in calibration of Electrophotographic printers by linear regression: environmental, consumables, and tone-level factors," J. Imaging Sci. Tech., 54(5): 050301 (2010).
[3] Y.-F. Kuo, C.-L. Yang, G. T.-C. Chiu, Y. Yih, J. Allebach, and D. Abramsohn, "Model-based calibration approach to improve tone consistency for color Electrophotography," J. Imaging Sci. Tech., 55(6): 060505 (2011).
[4] Y.-F. Kuo, C.-L. Yang, G. T.-C. Chiu, Y. Yih, J. Allebach, "Improving tone prediction in calibration of Electrophotographic printers by linear regression: using principal components to account for collinearity of sensor measurements," J. Imaging Sci. Tech., 54(5): 050302 (2010).
[5] A. K. Jain, "Data Clustering: 50 Years Beyond K-Means," Pattern Recognition Letters, 31(8), pg. 651-666 (2010).
[6] I. Davidson, and S. S. Ravi, "Agglomerative Hierarchical Clustering with Constraints: Theory and Empirical Resutls. In A. Jorge, L. Torgo, P. Brazdil, R. Camacho, & J. Gama, 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 2005 (pp. 59-70). Porto, Portugal: Birkhäuser (2005).
[7] Johnson, R. A., and Wichern, D. W. (2001). Applied Multivariate Statistical Analysis. (Prentice Hall, NJ, 2001).
[8] T. Hastie, R. Tibshirani, and J. Freidman, The Elements of Statistical Learning – Data Mining, Inference, and Prediction (Springer-verlag, New York, 2001) pp. 253-268.

Author Biography

Chao-Lung Yang is an Assistant Professor in Department of Industrial Management at National Taiwan University of Science and Technology, Taipei, Taiwan. His research interests include data mining on product design, machine learning, statistical analysis, and quality improvement. He is a current IS&T member.