International Journal of Electronic Commerce Studies
Vol.6, No.1, pp.137-144, 2015
doi: 10.7903/ijecs.1400

MARKET SEGMENTATION USING COLOR INFORMATION OF IMAGES

Ines Daniel
Brandenburg University of Technology
Erich-Weinert-Straße 1, 03046 Cottbus, Germany
ines.daniel@tu-cottbus.de

Sarah Frost
Brandenburg University of Technology
Erich-Weinert-Straße 1, 03046 Cottbus, Germany
sarah.frost@tu-cottbus.de

Daniel Baier
Brandenburg University of Technology
Erich-Weinert-Straße 1, 03046 Cottbus, Germany
daniel.baier@tu-cottbus.de

ABSTRACT

Market segmentation is an important area of marketing. In this field, researchers use clustering algorithms to divide customers into homogeneous groups. Traditionally, these groups are formed on the basis of survey data. In these surveys, the test persons often have to answer a variety of questions. With the increasing amount of digitalization and improved technical capabilities, new databases are now available for this purpose. For example, potential customers might provide photos that describe their activities, interests, or opinions. In the area of content-based image retrieval (CBIR) there are various methods that currently exist to analyze the similarity of such photos, e.g., using distributional descriptors of colors, textures, or shapes. In this paper we discuss which dissimilarity measures could be used to segment photos by hierarchical clustering on the basis of color. For this purpose we analyzed 2,100 images concerning three color spaces RGB, HSV and CIE L*a*b* using different distance measures as the basis for hierarchical clustering.

Keywords: Color Space, Image Clustering, Market Segmentation

1. INTRODUCTION

Clustering algorithms are standard tools to divide customers in homogeneous groups for market segmentation. Currently, databases that contain demographic, psychographic or behavioral attributes are used for classification. With increasing digitalization and the spread of digital photography, there are new databases available for segmentation. In the area of content-based image retrieval (CBIR), there are currently various methods to analyze images. In this paper we combined the techniques of market segmentation with the techniques of CBIR.

This paper is organized as follows: Section 2 gives a short introduction to market segmentation and describes the idea behind using images for market segmentation. In section 3 we give an overview of comparing images by color. For this purpose we describe different color spaces and image equations by color histograms. In section 4 we present the different distance measures that we used in our experiments. The investigation and results are shown in section 5. Finally we give a short conclusion and outlook.

2. MARKET SEGMENTATION

The roots of market segmentation can be traced back to the 1930s. But the final breakthrough in the concept of market segmentation was achieved with the article "*Introduction to Special Section on Market Segmentation Research*" by Wendell R. Smith in 1956[1]. In market segmentation, there are some typical steps involved to accomplish segmented markets[2]:

1. Establishing criteria for the division of the overall market into homogeneous sub-markets
2. Extracting data based on the established criteria
3. Dividing the total market with the help of the data obtained
4. Selecting of sub-markets for the segment-related processing
5. Performing segment-oriented marketing activities in selected markets.

To divide the total market with the help of obtained data, clustering algorithms are one of the most popular methods for post-hoc descriptive methods[3]. With the increasing digitalization it is possible to use new segmentation bases like private photos for market research. A study of BITKOM shows that more than 20 million Germans uploaded personal pictures to the Internet. This is also especially the case among young people. More than four out of five people between 10 and 17 years old have uploaded photos to the Internet[4]. On Facebook, more than 250 million photos are uploaded every day[5]. These facts illustrate that a multitude of

images are taken. As such, it is interesting to find that in market research, we only found one example where images to segment customers were used: Sinus Sociovision has made a survey about the living styles of Germans. Therefore they took pictures of the apartments of different people. These pictures helped to distinguish socio-demographic twins. Our idea is to combine the images with other databases for market segmentation. First, we analyzed the reduction of a questionnaire based lifestyle analysis with the help of images[6]. The authors show that it is possible to reduce item batteries thru the use of images. To analyze these images it is possible to use feature extraction and similarity indices. Some options will be shown in the next chapters.

3. COMPARING IMAGES BY COLOR

Many of the general features we use in our experiments are based on the distribution of color intensities across a digital image. Each single raster point of a digital image (each pixel) is characterized by intensities with respect to three colors of a selected color model. The intensities have a range, e.g., from 0 to 255 in an 8-bit image representation. In our experiments we used distances between color histograms to compare images, since the similarity of colors has proven to be very important for humans[7]. A color space can be seen as a vector space, in which every point represents a color. In our test application we used – as proposed in the literature[8] – the three-dimensional RGB, HSV and CIE-L*a*b* color spaces as candidates for powerful color models. Further information on them can be found in Batchelor[9] or Wyszecki and Stiles[10].

In the following we assume an image A that consists of T_A pixels. The color of all pixels can be stored in a color measurement set $C_A = \{c_1^A, \dots, c_{T_A}^A\}$. In a three dimensional color space the color measurement $c_t^A (t = 1, \dots, T_A)$ takes values from $X = \{x_1, \dots, x_M\} \subset \mathbb{R}^3$ which reflects the possible color intensities. For example in the 8-bit-coded RGB (Red-Green-Blue) color space $\mathrm{X} = \{(0,0,0), (0,0,1), \dots, (255,255,255)\}$ is possible, i.e. $\mathrm{M} = |\mathrm{X}| = 256^3 = 16{,}777{,}216$ different colors. We use color histograms for characterizing the distribution of these measurements across the image. A histogram is a fixed-size discrete distributional function with an a priori declared number N of disjoint color ranges $X_i \subset X\ (i = 1, \dots, N)$, the so-called bins. So, e.g., if each possible color of the 8-bit-coded RGB space would be declared as a bin, we would have $N = M = 16{,}777{,}216$ bins. Alternatively, if 8 subsequent intensities in each of the three dimensions would be summarized to one bin, one would receive only $\frac{256}{8} = 32$ bins per dimension. Thus we get $N = 32^3 = 32{,}768$ bins in the whole color space. For each bin i of histogram

$\mathbf{h}^A = h_1^A, \ldots, h_N^A$ we can calculate its number of colors in its range N_i and the corresponding number of pixels h_i^A in image A according to

$$N_i = \sum_{m=1}^{M} 1_{\{\mathbf{x}_m \in X_i\}}, \; h_i^A = \sum_{t=1}^{T^A} 1_{\{\mathbf{c}_t^A \in X_i\}}. \quad (1)$$

4. DISTANCE MEASURES

There are many different distance measures presented in the literature. We distinguish between bin-by-bin and cross-bin distances. The main characteristic of bin-by-bin distances is that only corresponding bins h_i^A and h_i^B of images A and B are compared. In our investigation we used the three established bin-by-bin distance measures *Euclidean distance* (or L_2 distance), the *Manhattan distance* (also called city block distance or L_1 distance) and the *Jeffrey-divergence* (JD). We also compared cross-bin distances. In contrast to bin-by-bin distances, they do not only compare corresponding bins. For example the *quadratic-form distance* (QF) uses a similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ to weight the distance between the bins. Another cross-bin distance, that is very useful because of the linear computation complexity, is the *match distance*[11]. We also tested the *Kolmogorov-Smirnov distance* (KS)[12]. and the *Earth Mover's Distance* (EMD)[13]. But in our experiments we used large image databases and a high number of bins. That is why the EMD computation complexity prevents its usage for histograms with more than three bins per dimension. The computation complexity is caused by the optimization problem that has to be solved. As such, it was not possible to use the EMD in every case during our investigation.

Table 1 lists the traditional distance measures we used in our evaluation. All distances between image A and B make use of the corresponding normalized histograms $\mathbf{h}^A$, $\mathbf{h}^B$. The quadratic-form distance additionally uses the symmetric similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ with $s_{ij} = 1 - \frac{g_{ij}}{\max g}$[13], where g_{ij} is the distance (we used Euclidean dist.) between the bins i and j. The main idea of the EMD is to treat the bins as mass of earth. One histogram is defined as piles of earth and the other one as holes. The goal is to fill the holes with earth from the piles with the minimal amount of work, where work is defined as amount of mass that has to be transported multiplied by the distance that has to be covered. The *flow* f_{ij} is the amount of mass which is transported from bin i of histogram $\mathbf{h}^A$ to bin j of histogram $\mathbf{h}^B$. The *ground distance* g_{ij} is the corresponding distance between these bins i and j, i.e. it is the distance that has to be covered when mass is transported from bin i to j.

Table 1. Overview of the distance measures used in our experiments. The distance is calculated on the basis of the normalized histograms h^A and h^B with $m_i = (h_i^A + h_i^B)/2$ and $\hat{h}_i = \sum_{j=1}^{i} \hat{h}_i$.

Bin-by-bin distances		Cross-bin distances	
L_2	$d_{L_2}(\boldsymbol{h}^A, \boldsymbol{h}^B) = \sqrt{\sum_i (h_i^A - h_i^B)^2}$	QF	$d_{QF}(\boldsymbol{h}^A, \boldsymbol{h}^B) = \sqrt{(\boldsymbol{h}^A - \boldsymbol{h}^B)^T S (\boldsymbol{h}^A - \boldsymbol{h}^B)}$
L_1	$d_{L_1}(\boldsymbol{h}^A, \boldsymbol{h}^B) = \sum_i \lvert h_i^A - h_i^B \rvert$	Match	$d_M(\boldsymbol{h}^A, \boldsymbol{h}^B) = \sum_i \lvert \hat{h}_i^A - \hat{h}_i^B \rvert$
JD	$d_J(\boldsymbol{h}^A, \boldsymbol{h}^B) = \sum_i \left(h_i^A \log\left(\frac{h_i^A}{m_i}\right) + h_i^B \log\left(\frac{h_i^B}{m_i}\right) \right)$	KS	$d_{KS}(\boldsymbol{h}^A, \boldsymbol{h}^B) = \max_i \lvert \hat{h}_i^A - \hat{h}_i^B \rvert$
		EMD	$d_{EMD}(P, Q) = \min \sum_i \sum_j d_{ij} f_{ij}$

5. INVESTIGATION

In our investigation we generated a data set of 2,100 typical holiday images. The images were classified into three categories: mountains, sunset and city lights. The pictures have different sizes and resolutions. To analyze the pictures we calculated several histograms of the images:

- RGB color bands (3x256 bins) and Compression of color bands (3x3 bins)
- RGB-, HSV-, L*a*b* color cubes (3x3x3-, 4x4x4-, 8x8x8-, 16x16x16 bins)
- Statistical moments of each color component (mean, variance and skewness)

After calculating the histograms we applied the hierarchical clustering using Ward's algorithm. For clustering we used the distance measures shown in Table 2. To make the clustering result comparable we used the adjusted Rand index[14]. The adjusted Rand index compares the calculated clusters with the reference clusters. A result of 1.0 means a perfect clustering. Table 2 shows the clustering results using the different distances. In the case of statistical moments we only calculated bin-by-bin distance because there is no cross-bin relationship. From our data a general statement about the "best" distance or the "best" feature could not be made. In many cases the bin-by-bin distances L_1 and JD work very well. Bin-by-bin distances have the advantage that they have shorter computation times than cross-bin distances. However they reach good results. In some cases the cross-bin

distance Match distance reaches very good results. Overall the three distances JD, QF distance and EMD reach the best results. They obtained a mean value over 0.57 across all features. The EMD reaches the best mean value at 0.596. A view to the mean values of the features over all distances show, the L*a*b* color cube (4x4x4 bins) reaches the best mean value (0.561). The best combination to cluster our images is the RGB color cube with 27 bins in combination with JD. This combination reaches an adjusted Rand index of 0.786. That means that 1,940 out of 2,100 images were classified into the right cluster. The worst result was found in the combination of the HSV color cube with 4,096 bins and L_2 distance. This combination reached an adjusted Rand index of 0.010. In this case 833 out of 2,100 images were classified into the right cluster. This shows that the adjusted Rand index decreases very quickly.

Table 2. Adjusted Rand indices of the clustered images (**** = no values because of very high computing duration; xxxx = cross-bin distances may not be appropriate)

	L2	L1	JD	QF	Match	KS	EMD	Ø
RGB color bands	0.331	0.654	0.337	0.523	0.610	0.359	****	0.469
Comp. RGB bands	0.458	0.434	0.635	0.456	0.640	0.549	0.640	0.545
RGB Cube -3	0.595	0.634	0.673	0.626	0.325	0.340	0.641	0.548
RGB Cube -4	0.602	0.746	0.691	0.671	0.318	0.295	0.586	0.559
RGB Cube -8	0.130	0.508	0.786	0.600	0.351	0.287	0.649	0.473
RGB Cube -16	0.115	0.487	0.457	****	0.252	0.257	****	0.314
HSV Cube -3	0.223	0.465	0.510	0.557	0.246	0.193	0.477	0.382
HSV Cube -4	0.173	0.477	0.538	0.657	0.246	0.272	0.680	0.435
HSV Cube -8	0.288	0.680	0.531	0.595	0.256	0.239	0.524	0.445
HSV Cube -16	0.010	0.491	0.599	****	0.271	0.288	****	0.332
L*a*b* Cube -3	0.467	0.456	0.689	0.536	0.408	0.349	0.537	0.492
L*a*b* Cube -4	0.633	0.644	0.702	0.563	0.406	0.372	0.609	0.561
L*a*b* Cube -8	0.480	0.555	0.558	0.630	0.411	0.332	0.614	0.511
L*a*b* Cube -16	0.440	0.603	0.448	****	0.407	0.330	****	0.446
Moments	0.418	0.449	0.426	xxxx	xxxx	xxxx	xxxx	0.431
Ø	0.358	0.552	0.572	0.583	0.368	0.319	0.596	

Furthermore Table 2 shows that a high number of bins in the image histograms do not guarantee a good classification. Mostly color cubes with 4,096 bins reach worse results than color histograms with 512 or 64 bins. However it is not possible to say which number of bins is the best. But with increasing number of bins the calculation time increase enormously. For example to calculate the distances of the cubes with 4,096 bins per

histogram (that means cube -16) with EMD, the calculation takes 27,789 days (using Intel(R) Core(TM) i7-2600K, 3.40GHz, 12GB Ram).

6. CONCLUSION AND OUTLOOK

This paper gives an overview of the development of databases for market segmentation. With increasing digitalization, there are many new databases. For example, one can utilize a database of images from possible consumers. This paper provides some color features and distances to automatically analyze these images. In our approach we used 2,100 images from potential consumers and classified these images thru hierarchical clustering. In further works we will test the influence of cutting bins with small entries. We also want to analyze the influence of using single linkage clustering before using Ward's algorithm. Additionally, we want to analyze other features like edge detection or shape detection. In a marketing context we would like to develop the ability to compare between clustering of digital images and questionnaire-based grouping of consumers. Another approach is to develop the usage of uploading images during online interviews.

7. ACKNOWLEDGEMENTS

This research was funded by the Federal Ministry for Education and Research under grants 03FO3072. The authors are responsible for the content of this paper.

8. REFERENCES

[1] W.R. Smith, Product differentiation and market segmentation as alternative marketing strategies. *Journal of Marketing*, 21(1), p3-8, 1956. http://dx.doi.org/10.2307/1247695.

[2] D. Baier, and M. Brusch, Marktsegmentierung. In A. Hermann et al. (Eds.), *Handbuch Marktforschung* (p769-790). Gabler: Wiesbaden, 2008.

[3] M. Wedel, and W. Kamakura, *Market segmentation: Conceptual and methodological foundations (2nd ed.)*. Boston: Kluwer Academic Publishers, 2000.

[4] C. Hallerberg, and F. Koch, *20 Millionen Deutsche stellen eigene Fotos ins Internet*. Retrieved on January 16, 2012, from http://www.bitkom.org/files/documents/bitkom-presseinfo_einstellen_von_videos_und_bildern_04_09_2009.pdf.

[5] Facebook, *Statistik*. Retrieved on January 16, 2012, from http://www.facebook.com/press/info.php?statistics.

[6] I. Daniel, and D. Baier, Lifestyle segmentation based on contents of

uploaded images versus ratings of items. In B. Lausen, D. Van den Poel, and A. Ultsch (Eds.), *Algorithms from and for Nature and Life Studies in Classification, Data Analysis, and Knowledge Organization* (p439-447). Heidelberg: Springer, 2013. http://dx.doi.org/10.1007/978-3-319-00035-0_44.

[7] M. Swain, and D. Ballard, Color indexing. *International Journal of Computer Vision*, 7(1), p11-32, 1991. http://dx.doi.org/10.1007/BF00130487.

[8] J. Schanda, CIE colorimetry. In J. Schanda (Ed.), *Colorimetry* (p25-78). Hoboken: John Wiley & Sons, 2007. http://dx.doi.org/10.1002/9780470175637.ch3.

[9] B.G. Batchelor, Machine vision for industrial applications. In B.G. Batchelor (Ed.), Machine Vision Handbook (p3-59). London: Springer, 2012. http://dx.doi.org/10.1007/978-1-84996-169-1_1.

[10] G. Wyszecki, and W. Stiles, *Color science. Concepts and methods, quantitative data and formulae* (2nd ed.). New York: Wiley, 2000.

[11] M. Werman, S. Peleg, and A. Rosenfeld, A distance metric for multidimensional histograms. *Computer Vision, Graphics, and Image Processing*, 32(3), p328-336, 1985. http://dx.doi.org/10.1016/0734-189X(85)90055-6.

[12] D. Geman, S. Geman, C. Graffigne, and P. Dong, Boundary detection by constrained optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7), p609-628, 1990. http://dx.doi.org/10.1109/34.56204.

[13] Y. Rubner, C. Tomasi, and L.J. Guibas, The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2), p99-121, 2000. http://dx.doi.org/10.1023/A:1026543900054.

[14] N.X. Vinh, J. Epps, and J. Bailey, Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In L. Bottou and M. Littman (Eds.), *Proceedings of the 26th International Conference on Machine Learning* (p1073-1080). Canada: Montreal, 2009. http://dx.doi.org/10.1145/1553374.1553511.