Viable Yeast Identification using Bag of Visual Words in Colored images

1st Junior Silva Souza
Instituto Federal de Mato Grosso do Sul (IFMS)
Campo Grande, Brazil
junior.souza@ifms.edu.br

2nd Vanessa Ap. de Moraes Weber
Universidade Catlica Dom Bosco (UCBD)
Universidade Estadual do MS (UEMS)
Campo Grande, Brazil
vamoraes@gmail.com

3rd Ariadne Barbosa Gonçalves
Universidade Federal de MS (UFMS)
Campo Grande, Brazil
ariadne.gon@gmail.com

4th Marco Alvarez
University of Rhode Island
Rhode Island, United States
malvarez@cs.uri.edu

5th Marney Pascoli Cereda
Agro: Laboratories, of Research Processes and Products
Campo Grande, Brazil
mpcereda@gmail.com

6th Wesley Nunes Gonçalves
Universidade Federal de MS (UFMS)
Campo Grande, Brazil
wnunesgoncalves@gmail.com

7th Valguima V. V. Aguiar Odakura
Universidade Federal da Grande Dourados (UFGD)
Dourados, Brazil
valguimaodakura@ufgd.edu.br

8th Hemerson Pistori
Universidade Catlica Dom Bosco (UCBD)
Universidade Federal de MS (UFMS)
Campo Grande, Brazil
pistori@ucdb.br

***Abstract*—In this research it is reported a system to automate the process of identification of viable yeasts whose population control is a crucial task in the ethanol production process. The identification and counting of yeasts made by human vision under a light microscope, is repetitive and susceptible to errors. We used computer vision techniques such as BoVW, Color Coherence Vectors (CCV), Color Moments (CM), Bag-of-Color (BoC) and Opponent Color (OpC) were applied for extracting characteristics that were classified by the Naive Bayes, KNN, SVM and J48 algorithms in 2614 images of yeasts separated into three classes: viable, non-viable and background. The results were analyzed using software R, which in the ANOVA test resulted in a p value equal to $2e^{-16}$ indicating a significant difference between the techniques. The OPC with SVM classifier showed the highest performance using the PCC Percent Correct Classification metric, about 95% compared to other techniques.**

Index Terms*—bag of visual words, color, supervised learning, *saccharomyces cerevisiae

I. Introduction

In the oil crisis in 1973, Brazil has adopted a new source of fuel, the ethanol. Fiscal incentives and funding have been proposed in order to increase the planting of sugar cane and installation of industries for processing. Sugar cane culture was inserted in Mato Grosso do Sul in 1980, after the adoption of the National Alcohol Program (ProAlcool). Due to increase of industries and sugar cane planted area, there is a search for new technologies to improve productivity and quality of ethanol productions [1].

Ethanol production is characterized by the fermentation of the wort (resulting from the dilution of sugarcane juice with water). In this process, the *Saccharomyces cerevisiae* yeasts are added to the wort for ethanol production by fermentation. This process occurs with the sugar consumption from the wort by yeasts, this process enable the ethanol and carbon gas production. To reduce costs in this production process, fermentation "MelleBoinot" is adopted due to its recycling of yeast feature. In this recycling, the yeasts are reused in successive fermentations. The recycling or reuse of yeasts reduce costs in the production process [2].

The quality of ethanol production is related to the yeast viability. Therefore, to ensure the production, it is necessary to check the microbiological control in laboratory, whereas viable yeasts are responsible for fermentation and non-viable yeasts have not action in the fermentation as they should [3].

In activities related to the microbiological control, wort samples are taken from the fermentation tanks and examined in the laboratory by a technical manager. This activity is visually, because requires to identify and counting yeast with the support of a light microscope (LM). Being a repetitive and visual activity, this task is susceptible to human errors, because its process can be tiring and subjective. To facilitate the yeasts identification, samples are mixed in water and methylene blue. The non-viable yeast turns colored in blue [4]. Two types of yeast: viable yeasts are marked by red squares while non-viable yeasts are marked by blue squares shown in Figure 1.

Activities such as the identification and counting of yeasts are repetitive tasks that can be performed automatically by computers programs. The yeast analyzes are done in microscopy images, that is way the proposal of this research is automate this process through computer vision and supervised

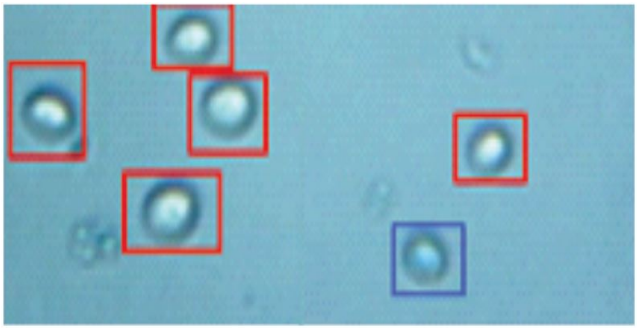

Fig. 1. Example of two types of yeasts: viable (red squares) and non-viable (blue squares). The non-viable yeasts are colored with blue when in contact with methilene blue corant.

learning.

In our research we use the BoVW for features extracting, because it is a widespread technique in many papers that used image classification. However, the color is a very important feature in many images, including yeast dye images [5].

This research aim is to evaluate the performance achieved by BoVW and its color variations on the yeast recognition. In this research, the following color variations were evaluated: color coherence vectors (CCV) [6], and color moments (CM) [7], bag of color (BoC) [8] and opponent color (OpC) [9]. CCV extracts color information by regions or clusters of a single color. The CM extracts color information from the average and variance applied to each image. The BoC is a color histogram, whose goal is to extract the frequency of certain colors. OpC is a variant applying BoVW in each color channel.

In order to evaluate the BoVW variants and classifiers (J48, NB, SVM and KNN), we performed ANOVA hypothesis test. The p-value indicated that the variants differ from each other. According to the experimental results, the OpC with SVM classifier achieved the highest performance. A second experiment was conducted with OpC and SVM, and the result showed that the dictionary with 256 visual words had the best result.

In section II is presented some related works and in section III is explained the materials and methods. The results are reported in section IV, followed by the discussion in section V, conclusion section and future works in the end of this paper.

II. Related Works

The viability of yeasts is commonly used to determine the efficiency of the production process. Usually, the physiological and metabolic changes of yeasts is observed using fluorescence microscopy or flow cytometry [10] [11]. However, these methods are time-consuming and prone to human-error, since they are not automatic or do not quantitatively analyze a large number of cells.

Recently, image-based methods and systems have been proposed to overcome these issues. Including [12] [13] who demonstrated the use of a fluorescence-based image cytometry system Cellometer Vision [10] for the analysis of vitality of *S. cerevisiae*. In order to observe the behavior of *S. cerevisiae*, [14] presented the CellStar, a tool for tracking yeast cells in long-term experiments. They compared CellStar with six other tools and demonstrated its high performance and accuracy. In addition, [15] [16] presented a platform for measuring viability of yeast cells by capturing an in-line hologram of the sample. This hologram sample is classified as live or dead by a Support Vector Machine for measuring viability as well as concentration. Furthermore [17] developed an automated cell counting for estimating the total number and the viability of yeast cells. To avoid reagent-based methods, [18] proposed a novel system to classify yeast viability based on wavelet features, feature selection and Support Vector Machine classifier.

Image-based methods are also proposed to classify yeast cells based on morphological characteristics. In this sense [19] proposing an image processing method designed to classify microscopic images of yeast cells in no budding, small bud, and large bud cells, which has been improved and included in a device [20]. The method is composed of four parts: image preprocessing to remove background noises, segmentation to separate yeast cells from background, extraction of morphological features (compactness, axis ratio and bud size) from each cell and classification using k-nearest neighbors. Texture features has also been used in the analysis of yeast cells. Since [21] proposed an image-based method for determining yeast floc dimension using co-occurrence matrices. They show that the energy of these matrices can be correlated with the mean particle diameter and therefore can be used to quantify changes in yeast floc size during fermentations.

III. Materials and Methods

In the classification stage, we use the following supervised learning techniques: decision tree (J48), naives bayes (NB), support vector machine (SVM) and k-nearest neighbor (KNN). These supervised learning techniques were used with different BoVW variants. Thus, we use computer vision to extract features and supervised learning algorithms to generate classifiers and make the yeast identification. In the feature extraction, we used an image database with 2614 yeast images. The images were manually segmented and defined in three classes: background without yeast, non-viable and viable yeast.

A. Yeast Database

Yeast samples were obtained from the fermentation process the which *Saccharomyces cerevisiae* yeast were added to the wort (water with sugarcane juice) at a concentration of 1% (w / v). The wort was set to 12 Brix and also used the samples when the value of Brix was at 6 and 3.

The yeast images used in this study have been built by INOVISO group (Development and Innovation in Computer Vision). The yeast images from Brix 03 were taken by LM at a 100x magnification. It was obtained 30 yeast images, they were manually segmented and separated in three classes: non-viable with 727 images, 292 images viable and background with 1595 images, totaling 2614 images for the yeast database.

IV. RESULTS

The main hypothesis testing used in this study were the ANOVA and Friedman test with Tukey and Wilcoxon post-tests. The metric used to compare the techniques was the percentage of correct classification (PCC). It is the number of correctly identified images of all class, divided by the total number of images.

The techniques performance were analyzed using ANOVA and Friedman hypothesis test to compare the performance of techniques regarding the percentage of correct classification metric. The dictionary size used by BoVW and variants using the color information has been set to the value 512, because showed better results in relation to values comprehended 64 and 1024.

Combinations between features extractors and classifiers found at Weka software were carried out. The chosen classifiers were: KNN, J48, NB and SVM. The used features extractors were: BoC, BoW, CCV, CM and OpC. Thus, for example, the abbreviation KNNBoC indicates the combination between KNN classifier with BoC features extractor, giving origin to the KNNBoC technique. Similarly, the following abbreviations were defined: KNNBoC, KNNBoW, KNNCCV, KNNCM, KNNOpC, J48BoC, J48BoW, J48CCV, J48CM, J48OpC, NBBoC, NBBoW, NBCCV, NBCM, NBOpC, SVMBoC, SVMBoW, SVMCCV, SVMCM and SVMOpC.

A box-plot diagram obtained through R software shown in Figure 2. In this diagram is shown the performance of each technique. It was done by comparison of the medians, that are the darker strip located on each box. In the diagram we can see that the SVMOpC technique had the highest median performance. This technique results from the combinations between SVM classifier with the Opponent Color feature extractor. We can observe some unusual behavior, such as the combinations of the BoC feature extractor with the classifiers, almost have the best performances, except with the SVM classifier. The combination between KNN classifier with any features extractors (BoW, BoC, CCV and CM), presented almost all the worst results.

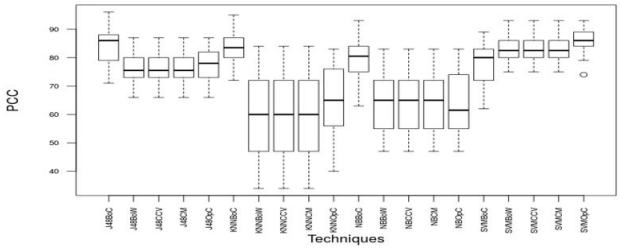

Fig. 2. Results of experiments. The performance is described in y-axis and all techniques are viewed in x-axis.

The variance analysis it was found a p-value $<2e^{-16}$. This result indicates that the null hypothesis can be ruled out, the medians indicate that there is a statistical difference between the techniques. The results of techniques with Turkey post-test showed with SVMOpc is similar with the follows techniques: KNNBoC, J48BoC, SVMBoW, SVMCCV and SVMCM. It happened because these techniques had a good performance in some sets of images than others, however SVMOpC technique maintained a higher. SVMOpC technique had the best performance. The SVMOpC is a technical variation of the BoW algorithm, allows the change of the dictionary size. The dictionary size Opponent Color was adjusted to: 128, 256, 512, 1024 and 2048 dictionary values.

The best result with SVMOpC technique, using the dictionary 256 shown in Figure 3. The p-value obtained by ANOVA was $1.58e^{-8}$, which indicates that the null hypothesis can be discarded, so these variations were very different from each other.

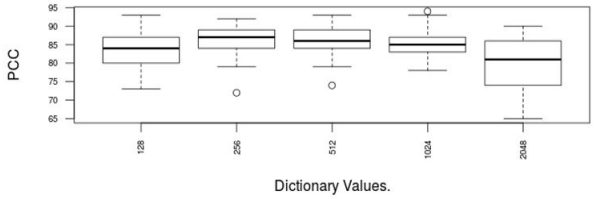

Fig. 3. Performance variations of SMOOpC technique. The y-axis represent performance and x-axis is size of dictionary.

The results with the dictionary size 256 have shown that the SVMOpC technique had the best performance in the yeast identification of colorless images, viable. The confusion matrix was obtained of image wherein the SVMOpC technical presented the best performance, since we used 121 images from sub-sample from database images, that show better result, for show the confusion matrix. In Table 1 is showed the confusion matrix, which there was 82 images identified as background, 16 images identified as viable yeast and 2 images identified as non-viable yeast.

TABLE I
MATRIX OF CONFUSION SHOWING THE YEASTS CLASSIFICATION.

a	***b***	***c***	
82	1	12	a = background
3	2	1	b = non-viable
4	0	16	c = viable

V. DISCUSSION

Two yeasts classified as viable, but it is a wrong identification, because image 6b is a non-viable yeast shown in Figure 4. This is one of the problems encountered, where two images were confused because of both yeasts have the same shape, but the color of the central region of each yeast is the factor to rank them. As we have no control over the regions detected by Color Opponent algorithm, then the interest points can be identified in any region of the image. This means that we do not have the spatial information, and this is one of the main problems found in the histograms.

As shown in Figure 5 both images were classified as viable yeast, but Figure 5b is a background. This is an example where

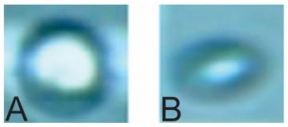

Fig. 4. Both images were classified as viable. A) Viable yeast. B) Non-viable yeast.

the classifier error, since the background color contrasts with the color of the center of viable yeast. This is an example where the shape is a factor that best distinguish viable yeast from image background. As we are working with the Color Opponent algorithm, we look for local changes in every image, leaving the feature concerning the form, which is an important feature in the image identification.

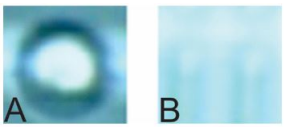

Fig. 5. Both images were classified as viable. A) Viable Yeast. B) Background.

The results showed that the color information added to BoVW algorithm improves the results in the yeasts identification. Although some problems were found in the identification, the SVMOpC technique with dictionary size of equal 256 showed good results when related to [1]. Our technique is invariant to image rotation and with a performance above 85% while a performance of 80% was reported by [1]. In relation to the [22], our technique is better to viable yeast identification.

VI. Conclusion

The counting activities and sorting of viable and non-viable yeast through the blue dye methylene are crucial for the ethanol production guarantee. One way to automate this process is to use the computer vision. In this research we analyzed the BoVW algorithm with some techniques that capture the color information for extracting features that have been used with a combination of classifiers.

The results showed that the opponent color feature extraction with SVM classifier achieved the best results. The metric used was the percentage of correct classification. ANOVA p-value were $2e^{-16}$ with both hypothesis tests. At confusion matrix the SVMOpC technique with 256 size dictionary identified the best viable yeast and the background, even with errors in the non-viable yeasts identification , the technical SVMOpC had the best performance than the others analyzed techniques. This work can be extended through the application in 3D images, thus increasing the amount of information of the image and guaranteed more real images.

Acknowledgment

This work was carried out with the financial support of the Coordination of Improvement of Higher Education Personnel - Brazil (CAPES) - Financing Code 001, National Council for Scientific and Technological Development (CNPq), Foundation for Support for the Development of Education, Science and Technology from the State of Mato Grosso do Sul (FUNDECT), Federal University of Mato Grosso do Sul (UFMS) and Catholic University Don Bosco (UCDB).

References

[1] Quinta, L.N.B., Queiroz, J.H.F.S., Souza, K.P., Pistori, H., Cereda, M.P. 2010. Classificao de Leveduras para o Controle Microbiano em Processos de Produo de Etanol. VI Workshop de Viso Computacional, p. 90-94.

[2] Boinot, M. 1939. Process of alcoholic fermentation with re-use os the yeass. The internatioal Sugar Journal.

[3] Stratford, M., 1996. Yeast flocculation: restructuring the theories in line with recent research. Belgian J. Brew. Biotechnol.

[4] Ceccato-Antonini, S.R., 2011. Microbiologia da fermentao alcolica: a importncia do monitoramento microbiolgico em destilarias. So Carlos: Universidade Federal de So Carlos, p. 105.

[5] Van Weijer, J., Khan, F. S. 2013. Fusing color and shape for bag-of-words based object recognition. In Computational Color Imaging, Springer, p. 2534.

[6] Pass, G., Zabih, R., Miller, J., 1997. Comparing images using color coherence vectors, in: Proceedings of the Fourth ACM International Conference on Multimedia. p. 6573.

[7] Bahri, A., Zouaki, H., 2013. A SURF-color moments for images retrieval based on bag-of features. Eur. J. Comput. Sci. Inf. Technol. 1, 1122.

[8] Wengert, C., Douze, M., Jgou, H., 2011. Bag-of-colors for improved image search, in: Proceedings of the 19th ACM International Conference on Multimedia. p. 14371440.

[9] van de Sande, K.E.A., Gevers, T., Snoek, C.G.M., 2008. Color descriptors for object category recognition, in: Conference on Colour in Graphics, Imaging, and Vision. p. 378381.

[10] Chan, L.L., Lyettefi, E.J., Pirani, A., Smith, T., Qiu, J., Lin, B., 2011. Direct concentration and viability measurement of yeast in corn mash using a novel imaging cytometry method. J. Ind. Microbiol. Biotechnol. 38, p. 11091115. doi: 10.1007/s10295-010-0890-7.

[11] Zhang, T., Fang, H.H.P., 2004. Quantification of Saccharomyces cerevisiae viability using BacLight. Biotechnol. Lett. 26, 989992.

[12] Chan, L.L., Kury, A., Wilkinson, A., Berkes, C., Pirani, A., 2012. Novel image cytometric method for detection of physiological and metabolic changes in Saccharomyces cerevisiae. J. Ind. Microbiol. Biotechnol. 39, p. 16151623. doi: 10.1007/s10295-012-1177-y.

[13] Saldi, S., Driscoll, D., Kuksin, D., Chan, L.L.-Y., 2014. Image-based cytometric analysis of fluorescent viability and vitality staining methods for ale and lager fermentation yeast. J. Am. Soc. Brew. Chem. 72, 253260. doi: https://doi.org/10.1094/ASBCJ-2014-1015-01.

[14] Versari, C., Stoma, S., Batmanov, K., Llamosi, A., Mroz, F., Kaczmarek, A., Deyell, M., Lhoussaine, C., Hersen, P., Batt, G., 2017. Long-term tracking of budding yeast cells in brightfield microscopy: CellStar and the Evaluation Platform. J. R. Soc. Interface 14. doi: 10.1098/rsif.2016.0705.

[15] Feizi, A., Zhang, Y., Greenbaum, A., Guziak, A., Luong, M., Chan, R., Berg, B., Ozkan, H., Luo, W., Wu, M., others, 2017a. Lensfree on-chip microscopy achieves accurate measurement of yeast cell viability and concentration using machine learning, in: CLEO: Applications and Technology. p. ATh4B–4.

[16] Feizi, A., Zhang, Y., Greenbaum, A., Guziak, A., Luong, M., Chan, R.Y.L., Berg, B., Ozkan, H., Luo, W., Wu, M., others, 2017b. Yeast viability and concentration analysis using lens-free computational microscopy and machine learning, in: Optics and Biophotonics in Low-Resource Settings III. p. 1005508.

[17] Hong, D., Lee, G., Jung, N.C., Jeon, M., 2013. Fast automated yeast cell counting algorithm using bright-field and fluorescence microscopic images. Biol. Proced. Online 15, 13. doi: 10.1186/1480-9222-15-13.

[18] Wei, N., Flaschel, E., Friehs, K., Nattkemper, T.W., 2008. A machine vision system for automated non-invasive assessment of cell viability via dark field microscopy, wavelet feature selection and classification. BMC Bioinformatics 9, 449. doi: 10.1186/1471-2105-9-449.

[19] Yu, B.Y., Elbuken, C., Ren, C.L., Huissoon, J.P., 2011. Image processing and classification algorithm for yeast cell morphology in a microfluidic chip. J. Biomed. Opt. 16, 66008. doi: 10.1117/1.3589100.

[20] Yu, B.Y., Elbuken, C., Shen, C., Huissoon, J.P., Ren, C.L., 2018. An integrated microfluidic device for the sorting of yeast cells using image processing. Sci. Rep. 8, 3550. doi:10.1038/s41598-018-21833-9.

[21] Mas, S., Ghommidh, C., 2001. On-line size measurement of yeast aggregates using image analysis. Biotechnol. Bioeng. 76, p. 9198. doi: 10.1002/bit.1148.

[22] Mongelo, A.I., Da Silva, D.S., Quinta, L.I.A.N.B., Pistoti, H., Cereda, M.P., 2011. Validao de mtodo baseado em viso computacional para automao de contagem de viabilidade de leveduras em indstrias alcooleiras, in: VIII Congresso Brasileiro de Agroinformtica SBIAGRO.