Segmentation Techniques based on Background Subtraction and Supervised Learning: A Comparative Study for Images of Mice and Human Skins

Bruno Brandoli Machado, Wesley Nunes Gonçalves and Jonathan de Andrade Silva
Institute of Mathematical Sciences and Computing (ICMC)
Physics Institute of São Carlos (IFSC)
University of São Paulo (USP) – São Carlos, SP, Brazil
{brandoli,jandrade}@icmc.usp.br
{wnunes}@ursa.ifsc.usp.br

Vinícius Saueia, Kleber Padovani de Souza, Bruno Toledo and Hemerson Pistori
INOVISAO and Biotechnology Department
Dom Bosco Catholic University (UCDB) – Campo Grande, MS, Brazil
{vsaueia,padovani,bc.toledo,pistori}@ucdb.br

Abstract

This paper presents a comparison between two image segmentation approaches based on background subtraction and supervised learning. Real images from two important issues, which have been studied by several computer vision research groups, were used in our experiments: namely, sign language interpretation and mouse behavior classification. According to performance measures, such as accurate rate, Jaccard coefficient, Yule coefficient, relative area error, and misclassification error, best results were obtained by background subtraction segmentators using images with complex background, otherwise, segmentators based on support vector machines outperformed when simple background were used.

keywords: *Mouse image segmentation, human skin segmentation, background subtraction, supervised learning*

1. Introduction

Segmentation is an important step in several computer vision systems. The essential idea is to split the image content into interesting and irrelevant objects. The classification of objects as interesting or irrelevant is highly dependent on the application domain. Thus, evaluating segmentation techniques in different domains is an important issue. Gesture recognition is an important element for many vision based human-to-computer interaction or computer mediated human-to-human communication [9]. Segmentation of human skin is an essential stage for vision-based gesture recognition, mainly when signers do not use additional resources, including colored gloves or markers. Another application addressed in this paper is related to behavior identification of mice under drug effect during experiments for medicine development, such as open-field experiment [10].

Different applications of image segmentation were reported in the literature. A color based segmentation method was proposed in [4], employed for a real-time system using a frame rate of 10 fps. The method was evaluated on synthetic and real images by means of outcome and manually segmented images. The metrics used were processing time, false positive, false negative, medium error of area and center of object. Terrillon and Fukamachi [15] compared nine color spaces for image segmentation. They presented two colors-based segmentation models: Gaussian model based on Mahalanobis distance and Gaussian mixture model. The aim was to segment human faces on colored images.

A segmentation algorithm for moving objects was presented in [3]. This algorithm combines the adaptive background subtraction with frames differentiation algorithm. Differentiation technique was used to determine which regions are moving and adaptive background subtraction assists on the identification of the whole moving region. In [8], a comparison of many subtraction background techniques described in the literature is showed. One of the eight compared techniques was Adaptive Gaussian mixture. For this technique, a K Gaussian mixture is built for each pixel, representing the background model. After object detection, the Gaussian parameters are updated based on an update constant. The techniques were evaluated using three parameters: distinction between foreground pixels and background

pixels, background storage and updating over time and post-processing of the objects to eliminate false positive of the resultant image.

This paper presents a comparative study of segmentation techniques applied to sign language recognition and mouse behavior analysis. The experiments were performed on a set of 40 images of mice and on another set of 240 images of Brazilian Sign Language gestures. Ground-truth segmented images were produced for all samples with the help of field specialists. Here we use five well-know measures to help us to assess the quality of the output images, thus reducing the subjectiveness of evaluations based purely on the visual analysis. The five performance measures are: accurate rate, Jaccard coefficient, Yule coefficient, relative area error and misclassification error [14]. In this sense, the main contributions of our study is to compare both groups of techniques and give directions to the systems in issue.

The remaining of this paper is organized as follows. Section 2 presents a background with regard to segmentation techniques. In Section 3 the performance measures are described. Experiments and results are discussed in Section 4. Finally, conclusions and future works are presented in Section 5.

2. Segmentation Techniques

This section presents an overview of the segmentation techniques used in this work: background subtraction, adaptive background subtraction, Gaussian models, decision trees, artificial neural networks and support vector machines. For more details of the techniques, we refer the reader to the reference papers.

2.1. Background and Adaptive Background Subtraction

Background subtraction is among the most used techniques in computer vision because it is easy to implement and demands low processing time [2, 7]. The underlying idea is to subtract an image from a reference image (a fixed background image) which does not contain interesting objects. Each new image is segmented using a pre-defined threshold. Given an image $\mathbf{I}_t$, at the instant t (frame), and $\mathbf{B}_t$ is the background image, the output image is achieved by:

$$|\mathbf{I}_t(x) - \mathbf{B}_t(x)| > \tau \tag{1}$$

where (x) and τ are the spatial position of each pixel and a predefined threshold, respectively.

However, background subtraction does not update the background information and therefore may not suit image changes on scenes, such as different illumination conditions or irrelevant objects that suddenly appear in images and no longer move. To overcome this drawback, an adaptive variant that updates the reference image over time was proposed in [6], turning this method more robust than the first one. In this variant the reference image is iteratively adjusted as follows:

$$\mathbf{B}_{t+1} = \alpha\mathbf{I}_t + (1-\alpha)\mathbf{B}_t \tag{2}$$

where $\alpha \in [0, 1]$ is an updating constant.

2.2. Segmentation based on Supervised Learning

Machine learning techniques [1] can be used for image segmentation purpose by learning a classifier from pixels of an object of interest $\mathbf{O}$. Pixels $\mathbf{p} = [r, g, b] \mid \mathbf{p} \in \mathbf{O}$ are extracted to compose the training set $\mathbf{S} = \{\mathbf{p}\}$, where r, g, b are red, green and blue components, respectively. In order to learn the classifier from $\mathbf{S}$, four supervised learning techniques are briefly described in this study:

Gaussian statistical learning: In this strategy [17], the parameters of a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ are estimated from the object pixels $\mathbf{S}$, defined by:

$$\mu = \frac{1}{|\mathbf{S}|}\sum_{\mathbf{p}\in\mathbf{S}}\mathbf{p}, \quad \Sigma = \frac{1}{|\mathbf{S}|}\sum_{\mathbf{p}\in\mathbf{S}}[\mathbf{p}-\mu][\mathbf{p}-\mu]^t \tag{3}$$

where μ is the mean, Σ corresponds to the covariance matrix, $\mathbf{p}$ is the training set vector. After the training step, a new pixel $\mathbf{p}'$ is assigned to the object of interest if its Mahalanobis distance with respect to $\mathcal{N}(\mu, \Sigma)$ is higher than a pre-defined threshold. The Mahalanobis distance is calculated by:

$$M_d = [\mathbf{p}-\mu]^t\Sigma^{-1}[\mathbf{p}-\mu] \tag{4}$$

Decision trees: Induction of decision trees [12] is a divide-and-conquer strategy for classification tasks. The algorithm splits the instance space into decision regions by generating internal or test nodes. The classification of an input instance is based on the best features that separate the data. The procedure starts by generating the root node, taking into account the whole training set. Interior nodes divide the set of instances by testing a specific feature and each child of that node will correspond to a distinct value represented by that feature. This procedure is repeated until leaf nodes are obtained. Leaf nodes will predict the class of the instances according to the path down of the final tree. Here, we use the C4.5 algorithm that splits each node on the feature with the highest information gain. The gain on a feature F with value v is defined as:

$$Gain(S, F) = Entropy(S) - \sum_{v\in F}\frac{|S_v|}{|S|}Entropy(S_v) \tag{5}$$

where S and S_v denote a set and a subset of instances, respectively. The gain is calculated by using impurity measures for quantifying the quality of the split generated by the trained model. In this study, entropy has been used.

Artificial neural networks: MLP [13] is a particular architecture of feedforward neural network which addresses data that are not linearly separable. A MLP net consists of one or more hidden layers between an input and output layer of neurons. Each neuron is fully connected from one layer to the next and can be described as a processing element which is activated by a nonlinear activation function. The neural network activation is the inner product of the input vector ($\mathbf{p}$) with the weights (connections) at hidden layers. For training the MLP, the backpropagation algorithm has been adopted (we refer to [5] for more details).

Support vector machines: SVM [16] is a widely used technique for data classification. Given a training set of two instance-label (x_i, y_i), for $i = 1, \ldots, l$ where $x_i \in R^n$ and such that $y_i = \pm 1$, a SVM finds a hyperplane that separates a couple of classes with the maximal margin in the higher dimensional space ϕ. Training a SVM requires the solution of an optimization problem with a very large quadratic programming. Alternatively, the sequential minimal optimization algorithm (SMO) [11] is a fast optimization that works by breaking the quadratic algorithm into a subset of smallest problems. For implementing the support vector machines strategy, the SMO has been used, along with polynomial basis function transformations.

3. Evaluation Methods

Visual evaluation of image segmentation algorithms is not a reliable criterion to analyze final results. Hence, there are many different ways for evaluating them. The performance can be measured through correcteness at pixel level or image regions. In this paper, we compute a total of five measures. Next we describe three of them based on pixels and two based on region properties.

3.1. Pixel-based Evaluation

In this approach, the results are based on the comparison between the ground-truth and output images. Each analyzed pixel can be classified into four possible labels, namely: (1) true positive, (2) false positive, (3) true negative, and (4) false negative. A true positive (TP) occurs when the outcome from a prediction corresponds to the ground-truth, otherwise, a false positive (FP) is found. Conversely, a true negative (TN) occurs when the outcome from a prediction is different from ground-truth image, or else, a false negative (FN) is considered.

Following these labels, it is possible to compute three quantitative measures regarding two binary images: the percentage correct classification (PCC), or as known as accurate rate, the Jaccard coefficient (JC) and the Yule coefficient (YC). Such measures are formalized in Equations 6, 7 and 8, respectively.

$$PCC = \frac{TP + FP}{TP + FP + TN + FN} \tag{6}$$

$$JC = \frac{TP}{TP + FP + FN} \tag{7}$$

$$YC = \mid \frac{TP}{TP+FP} + \frac{TN}{TN+FN} - 1 \mid \tag{8}$$

3.2. Region-based Evaluation

This approach judges the quality of the segmentation using blobs. Moreover, the performance measures are scored ranging from 0 (zero), for a totally correct segmentation, to 1 (one), for an incorrect case. The relative foreground area error (RAE) is obtained through shapes and areas from the segmented image regarding ground-truth image [14]. RAE value is 0 wheter a complete matching between overcome and ground-truth images are achieved, while the minimum matching is 1. Equation 9 defines RAE measure, where A_0 is the area of ground truth image, and A_t is the area of segmented image.

$$RAE = \begin{cases} \frac{A_0 - A_t}{A_0}, if A_t < A_0 \\ \frac{A_t - A_0}{A_t}, if A_t \geq A_0 \end{cases} \tag{9}$$

Another considered metric is defined as misclassification error (ME). It is obtained by computing the percentage of background pixels erroneously set as foreground, and foreground set as background. The ME is formalized in Equation 10, in which B_0 and F_0 refer to background and foreground of ground truth image; B_t and F_t indicate the background and foreground of the segmented image; and $|.|$ represents the cardinality for the set.

$$ME = 1 - \frac{|B_0 \cap B_t| + |F_0 \cap F_t|}{|B_0| + |F_0|} \tag{10}$$

4. Experimental Results

In order to evaluate the algorithms described in this paper, experiments are performed on two different colored image datasets. First, the datasets used for evaluation and experimental details are described. Then, the results are discussed.

4.1. Datasets

The Mouse Behavior database consists of two behaviors: vertical exploration and spatial locomotion [10]. 640 $\times$ 480 images are taken of the species Swiss and C57 within

a circular arena, resulting in a total of 40 images. Swiss are white haired animals and C57 black haired animals. We analyze the ability of segmentation algorithms in tasks where foreground objects are strongly correlated with background. In this particular case, the animal is the same color as the background. Both behaviors and two backgrounds are presented in Figure 1.

The Brazilian Sign Language database contains 240 images of postures that explore different gestures separated in 10 classes. The size of images is 800 $\times$ 600 pixels. Images are taken of 6 different signers with distinct skin tones. Each pose is captured with a static and complex background, in this case within a laboratory with artificial illumination. Further, the remaining 120 images are taken in outdoor environments with natural illumination and complex background. Some samples with all sort of actions with complex and static background are shown in Figure 2.

4.2. Experimental Setup and Sampling

Experiments are carried out with two groups of image segmentation techniques, as described in Section 2. The comparison is performed by attaining performance measures. First, we create by hand the ground truth images of both databases, a time consuming and laborious process. Additionally, for background subtraction techniques a reference image is obtained.

Samples with size of 40 $\times$ 40 pixels are collected in order to train the supervised learning models. For gesture images, samples of different parts of the body are extracted, such as parts of face, right arm, left arm and neck. For mouse behavior images, five samples of the animal and five samples of the arena are selected. Parameters for both approaches are chosen by means of empirical range for each algorithm, which result over 70,000 segmented images. Then, performance measures are calculated over segmented images. For our datasets results, the average precision is taken as the performance metric for determining the accuracy of the segmentation algorithms.

4.3. Results

Experiment 1: First, a comparison with different combinations of backgrounds and coat color of the Mouse Behavior database is shown in Figure 3. The combinations are indicated in the y axis of the graph, while the average precisions are indicated in the x axis. The highest values are obtained by the contrast between background and coat color, for instance: the black background and the Swiss mouse. For the supervised learning techniques, only the combination white background obtained good results and the remaining degrade in performance.

Experiment 2: Figure 4 shows the results achieved using the combination between backgrounds and skin tones under natural and artificial illumination. The combinations are indicated in the y axis of these graphs, while the average precisions are indicated in the x axis. We can observe that the background subtraction techniques in presence of artificial illumination do not perform well for these four combinations. There is a clear distinction in performance for simple backgrounds between these techniques and the supervised learning ones. Within the group of supervised learning, decision tree, artificial neural networks and support vector machines techniques perform much better than background subtraction and adaptive background subtraction. For natural illumination, it is also observed that the combination for both simple backgrounds perform better compared with artificial illumination. This is consistent with the shadow done for the wall. Conversely, for this case we can see that the supervised learning approach has the same behavior independent of the illumination condition, which they obtained the higher values.

5. Conclusions and Future Works

This paper has evaluated two different groups of segmentation techniques applied on two important real-world applications: gesture recognition for Brazilian Sign Language and mouse behavior analysis. We examined the challenges including illumination change and strong correlation between foreground objects and background. The evaluation was based on well known performance metrics which compare the segmented images with ground truth ones. The results presented here support that each algorithm performs successfully in a particular image database. Experiments strongly suggest that the background subtraction approach is better for mice images, mostly in cases with distinguished contrast. However, is also need to focus attention on basic features, for instance, C57 mouse is a black coat animal with short hairs, which at a young age their epidermal structures are quite visible. From the gesture recognition results, supervised learning approach achieve better results for both illumination conditions, in specific support vector machines. Future research should focus on different image segmentation techniques. We believe that another focus of attention is to explore unsupervised approaches.

Acknowledgment This work has received financial support from UCDB, FUNDECT, FINEP and CPP. Some of the authors have received scholarships from CNPQ and CAPES.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

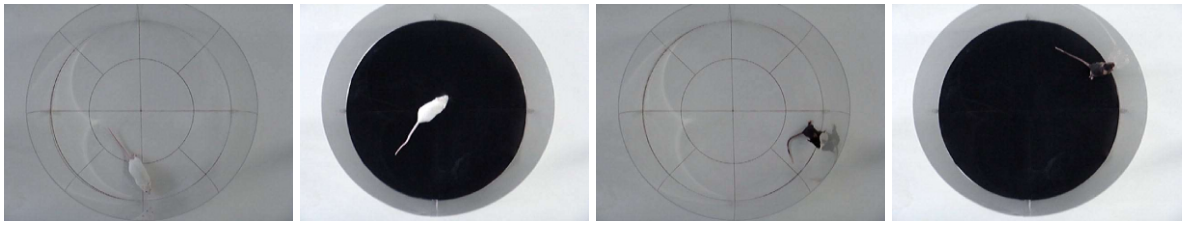

Figure 1. Examples of the two mouse behaviors: spatial locomotion and vertical exploration. From left to right, the first two images are mice in spatial locomotion, while the other two are mice in vertical exploration.

Figure 2. Ten different poses of 6 different signers taken of outdoor and indoor environments, showing the difference in illuminations and backgrounds. We are interested in evaluating the robusteness when signers wear accessories, such as sunglasses, lens and cap.

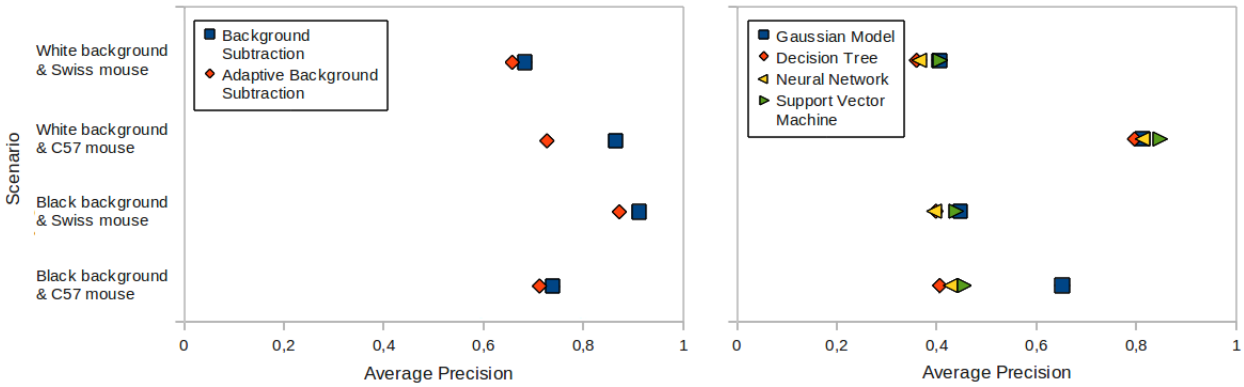

Figure 3. Mouse Behavior database evaluation influenced by combinations of background and coat color. Swiss are white haired animals and C57 black haired animals. In this experiment just artificial illumination is evaluated because experiments shall occurs in indoor environments.

[2] S.-C. S. Cheung and C. Kamath. Robust background subtraction with foreground validation for urban traffic video. *EURASIP J. Appl. Signal Process.*, 2005:2330–2340, January 2005.

[3] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, and O. Hasegawa. A sys-

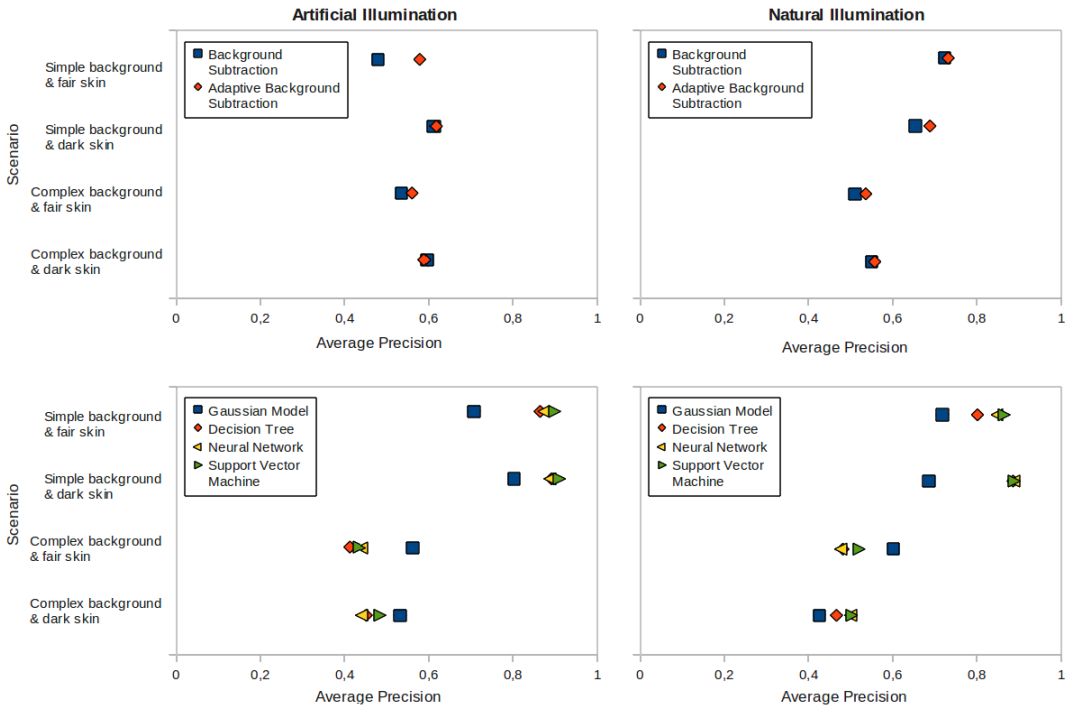

Figure 4. Evaluates combinations of backgrounds and skin tones of the Brazilian Sign Language database. We test both approaches under two illumination conditions: artificial and natural.

tem for video surveillance and monitoring. Technical Report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, May 2000.

[4] R. Cucchiara, A. Prati, and R. Vezzani. Real-time motion segmentation from moving cameras. *Real-Time Imaging*, 10(3):127–143, 2004.

[5] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2nd edition edition, 2001.

[6] J. Heikkilä and O. Silvén. A real-time system for monitoring of cyclists and pedestrians. In *Proceedings of the Second IEEE Workshop on Visual Surveillance*, VS '99, pages 74–, Washington, DC, USA, 1999. IEEE Computer Society.

[7] M. H. Khan, I. Kypraios, and U. Khan. A robust background subtraction algorithm for motion based video scene segmentation in embedded platforms. In *Proceedings of the 7th International Conference on Frontiers of Information Technology*, FIT '09, pages 31:1–31:6, New York, NY, USA, 2009. ACM.

[8] A. McIvor. Background subtraction techniques. In *Proceedings of Image and Vision Computing*, 2000.

[9] G. R. S. Murthy and R. S. Jadon. A review of vision based hand gesture recognition. *International Journal of Information Technology and Knowledge Management*, 2(2):405–410, July-December 2009.

[10] H. Pistori, V. V. Viana Aguiar Odakura, J. a. B. Oliveira Monteiro, W. N. Gonçalves, A. R. Roel, J. de Andrade Silva, and B. B. Machado. Mice and larvae tracking using a particle filter with an auto-adjustable observation model. *Pattern Recognition Letters*, 31:337–346, March 2010.

[11] J. C. Platt. *Fast training of support vector machines using sequential minimal optimization*, pages 185–208. MIT Press, Cambridge, MA, USA, 1999.

[12] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, March 1986.

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning internal representations by error propagation*, pages 318–362. MIT Press, Cambridge, MA, USA, 1986.

[14] M. Sezgin and B. Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, 13(1):146–168, 2004.

[15] J.-C. Terrillon, H. Fukamachi, S. Akamatsu, and M. N. Shirazi. Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 54–63, Washington, DC, USA, 2000. IEEE Computer Society.

[16] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data [in Russian]*. Nauka, Moscow, Russia, 1979. (English translation: Springer Verlag, New York, 1982).

[17] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, NY, 1998.