A Benchmark Dataset for Outdoor Foreground/Background Extraction

Antoine Vacavant[1,2], Thierry Chateau[3], Alexis Wilhelm[3], and Laurent Lequièvre[3]

[1] Clermont Université, Université d'Auvergne, ISIT, BP10448, F-63000 Clermont-Ferrand
[2] CNRS, UMR6284, BP10448, F-63000 Clermont-Ferrand
[3] Pascal Institute, Blaise Pascal University, CNRS, UMR6602, Clermont-Ferrand

Abstract. Most of video-surveillance based applications use a foreground extraction algorithm to detect interest objects from videos provided by static cameras. This paper presents a benchmark dataset and evaluation process built from both synthetic and real videos, used in the BMC workshop (Background Models Challenge). This dataset focuses on outdoor situations with weather variations such as wind, sun or rain. Moreover, we propose some evaluation criteria and an associated free software to compute them from several challenging testing videos. The evaluation process has been applied for several state of the art algorithms like gaussian mixture models or codebooks.

1 Introduction

The ability to detect objects in videos is an important issue for a number of computer vision applications like intrusion detection, object tracking, people counting, *etc.* In the case of a static camera, a foreground extraction algorithm is a popular operation to point out objects of interest in the video sequence. Although modeling background seems simple, challenging situations occur in classic outdoor environments such as variation of illumination conditions or local appearance modifications resulting to wind or rain. In order to handle such situations, many background/foreground adaptive models have been proposed in the last fifteen years. An important issue is to provide a way to evaluate and compare most popular models according to standard criteria.

Although the evaluation of background subtraction algorithms (BSA) is an important issue, the impact of relevant papers that handle with both benchmarks and annotated dataset is limited [1, 10]. Moreover, many authors that propose a novel approach use [11] as a gold-standard, but rarely compare their method with recent related work. This paper proposes a set of both synthetic and real video and several performance evaluation criteria in order to evaluate and rank background/foreground algorithms. Popular methods are then evaluated and ranked according to these criteria.

The next section (Section 2) presents the annotated datasets we have proposed for the BMC (Background Models Challenge), composed of 20 synthetic videos

and 9 real videos. We also define the quality metrics available in the benchmark, and computable with a free software (BMCW). In Section 3, we conduct a complete evaluation of six classic background subtraction algorithms of the literature, thanks to the benchmark of BMC.

2 Datasets and Evaluation Criteria

2.1 Learning and Evaluation Videos

In the contest BMC (Background Models Challenge) [4], we have proposed a complete benchmark composed of both synthetic and real videos. They are divided into two distinct sets of sequences: learning and evaluation.

The benchmark is first composed of 20 urban video sequences rendered with the SiVIC simulator [4]. With this tool, we are also able to render the associate ground truth, frame by frame, for each video (at 25 fps). Two scenes are used for the benchmark:

1. a street;
2. a rotary.

For each scene, we propose 5 event types:

1. cloudy, without acquisition noise;
2. cloudy, with noise;
3. sunny, with noise;
4. foggy, with noise;
5. wind, with noise.

For each configuration, we have two possible use-cases:

1. 10 seconds without objects, then moving objects during 50 seconds;
2. 20 seconds without event, then event (*e.g.* sun uprising or fog) during 20 seconds, finally 20 seconds without event.

The *learning* set is composed of the 10 synthetic videos representing the use-case 1. Each video is numbered according to presented event type (from 1 to 5), the scene number (1 or 2), and the use-case (1 or 2). For example, the video 311 of our benchmark describes a sunny street, under the use-case 1 (see Figure 1). In the learning phase of the BMC contest, authors use these sequences in order to set the parameters of their BSA, thanks to the ground truth of each image that is available, and to a software of computation of quality criteria (see next section).

The *Evaluation* set first contains the 10 synthetic videos with use-case 2. In Figure 1, the video 422, presenting a foggy rotary under use-case 2, is depicted. This set is also composed of real videos acquired from static cameras in video-surveillance contexts (see Figure 2). This dataset has been built in order test

[4] http://bmc.univ-bpclermont.fr

Fig. 1. Examples of synthetic videos and their associated ground truth in our dataset. Left: scene 1, configuration 3, use-case 1 (learning phase). Right: scene 2, configuration 4, use-case 2 (evaluation phase)

the algorithms reliability during time and in difficult situations such as outdoor scenes. So, real long videos (about one hour and up to four hours) are available, and they may present long time change in luminosity with small density of objects in time compared to previous synthetic ones. This dataset allows to test the influence of some difficulties encountered during the object extraction phase. Those difficulties have been sorted according to:

1. the ground type (bitumen, ballast or ground);
2. the presence of vegetation (trees for instance);
3. casted shadows;
4. the presence of a continuous car flow near to the surveillance zone;
5. the general climatic conditions (sunny, rainy and snowy conditions);
6. fast light changes in the scene;
7. the presence of big objects.

Fig. 2. Examples of real videos and their associated ground truth in our dataset (evaluation phase)

For each of these videos have been manually segmented some representative frames that can be used to evaluate a BSA. In the evaluation phase of the BMC contest, no ground truth image is available, and authors should test their BSA with the parameters they have set in the learning phase.

2.2 Quality Assessment of a Background Subtraction Algorithm

In our benchmark, several criteria have been considered, and represents different kinds of quality of a BSA.

Static Quality Metrics Let S be the set of n images computed thanks to a given BSA, and G be the ground truth video sequence. For a given frame i, we denote by TP_i and FP_i the true and false positive detections, and by TN_i and FN_i the true and false negative ones. We first propose to compute the F-measure, defined by:

$$F = \frac{1}{n} \sum_{i=1}^{n} 2 \frac{Prec_i \times Rec_i}{Prec_i + Rec_i}, \tag{1}$$

with

$$Rec_i(P) = TP_i/(TP_i + FN_i) \; ; \qquad Prec_i(P) = TP_i/(TP_i + FP_i) \tag{2}$$

$$Rec_i(N) = TN_i/(TN_i + FP_i) \; ; \qquad Prec_i(N) = TN_i/(TN_i + FN_i) \tag{3}$$

$$Rec_i = (1/2)(Rec_i(P) + Rec_i(P)) \; ; \quad Prec_i = (1/2)(Prec_i(P) + Prec_i(P)). \tag{4}$$

We also compute the PSNR (Peak Signal-Noise Ratio), defined by:

$$PSNR = \frac{1}{n} \sum_{i=1}^{n} 10 \log_{10} \frac{m}{\sum_{j=1}^{m} ||S_i(j) - G_i(j)||^2} \tag{5}$$

where $S_i(j)$ is the jth pixel of image i (of size m) in the sequence S (with length n). These two criteria should permit to compare the raw behavior of each algorithm for moving object segmentation.

Application Quality Metrics We also consider the problem of background subtraction in a visual and perceptual way. To do so, we use the gray-scale images of the input and ground truth sequences (see Figure 3) to compute the perceptual measure SSIM (Structural SIMilarity), given by [14]:

$$SSIM(S, G) = \frac{1}{n} \sum_{i=1}^{n} \frac{(2\mu_{S_i}\mu_{G_i} + c_1)(2cov_{S_iG_i} + c_2)}{(\mu_{S_i}^2 + \mu_{G_i}^2 + c_1)(\sigma_{S_i}^2 + \sigma_{G_i}^2 + c_2)}, \tag{6}$$

where μ_{S_i}, μ_{G_i} are the means, $\sigma_{S_i}, \sigma_{G_i}$ the standard deviations, and $cov_{S_iG_i}$ the covariance of S_i and G_i. In our benchmark, we set $c_1 = (k_1 \times L)^2$ and $c_2 = (k_2 \times L)^2$, where L is the size of the dimension of the signal processed (that is, $L = 255$ for gray-scale images), $k_1 = 0.01$ and $k_2 = 0.03$ (which are the most used values in the literature).

We finally use the D-Score [8], which consists in considering localization of errors according to real object position. As Baddeleys distance, it is a similarity measure for binary images based on distance transform. To compute this measure we only consider mistakes in BSA results. Each error cost depends on the distance with the nearest corresponding pixel in the ground-truth. As a matter of fact, for object recognition, short or long range errors in segmentation step are less important than medium range error, because pixels on medium range impact greatly on object's shape. Hence, the penalty applied to medium range errors is heavier than the one applied to those in a short or large range, as shown on Figure 4.

Fig. 3. To compute the SSIM, we need the intensities of pixels, in the ground truth sequence G (Left), and in the sequence computed by a BSA (Right)

Fig. 4. Examples of computation of the D-Score. From Left to Right: a ground-truth image; cost map based on a DT; example of long ranges errors, leading to a D-Score of 0.003; omissions with medium range errors, with D-Score: 0.058

More precisely, the D-Score is computed by using:

$$D{-}score(S_i(j)) = \exp\left((-\log_2\left(2.DT(S_i(j)) - 5/2\right)^2\right) \tag{7}$$

where $DT(S_i(j))$ is given by the minimal distance between the pixel $S_i(j)$ and the nearest reference point (by any distance transformation algorithm). With such a function, we punish errors with a tolerance of 3 pixels from the ground-truth, because these local errors do not really affect the recognition process. For the same reason, we allow the errors that occur at more than a 10 pixels distance. Details about such metric can be found in [8]. Few local/far errors will produce a near zero D-Score. On the contrary, medium range errors will produce high D-Score. A good D-Score has to tend to 0.

3 Results and Analysis

3.1 Material and methods

In this article, we will present the quality measures presented in the previous section for the methods depicted in Table 1. Most of those approaches are available thanks to the OpenCV library [5]. The parameters were tuned with a stochastic gradient descent to maximize the F-measure for the sequences of the learning phase.

We present the values of all the quality criteria exposed in the previous section, for the evaluation set of videos. Criteria are calculated thanks to the

[5] http://opencv.org/

Table 1. The methods tested in this article, with their associated references

Name	Description
NA	Naive approach, where pixels differing from the first image of the sequence (under a given threshold) are considered as foreground ($threshold = 22$).
GMM 1	Gaussian mixture models from [5, 11], improved by [6] for a faster learning phase.
GMM 2	Gaussian mixture models improved with [12, 13] to select the correct number of components of the GMM ($history\ size = 355$, $background\ ratio = 16$).
BC	Bayesian classification processed on feature statistics [9] ($L = 256$, $N_1 = 9$, $N_2 = 15$, $L^c = 128$, $N_1^c = 25$, $N_2^c = 25$, no holes, 1 morphing step, $\alpha_1 = 0.0422409$, $\alpha_2 = 0.0111677$, $\alpha_3 = 0.109716$, $\delta = 1.0068$, $T = 0.437219$, $min\ area = 5.61266$).
CB	Codewords and Codebooks framework [7].
VM	VuMeter [3], which uses histograms of occurences to model the background ($\alpha = 0.00795629$ and $threshold = 0.027915$).

BMC Wizard (BMCW, see a screenshot in Figure 5), which can be downloaded from the BMC website [6].

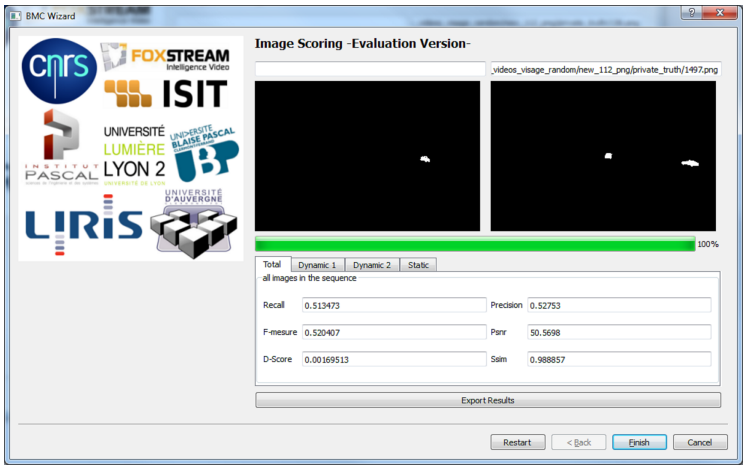


Fig. 5. The BMC Wizard, a free software to compute criteria of our benchmark

3.2 Results

Figures 6 to 10 show the global performance of each method for each evaluated score.

[6] http://bmc.univ-bpclermont.fr/?q=node/7

Fig. 6. F-measure for each method

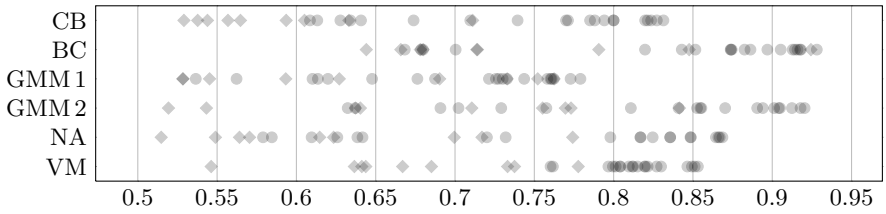


Fig. 7. D-score for each method

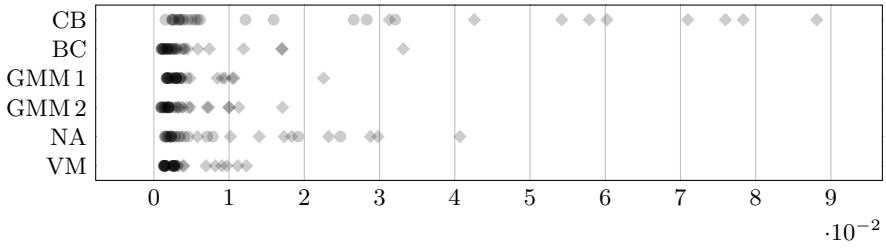


Fig. 8. PSNR for each method

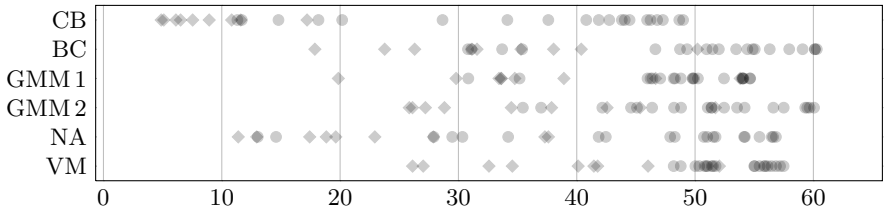


Fig. 9. SSIM for each method

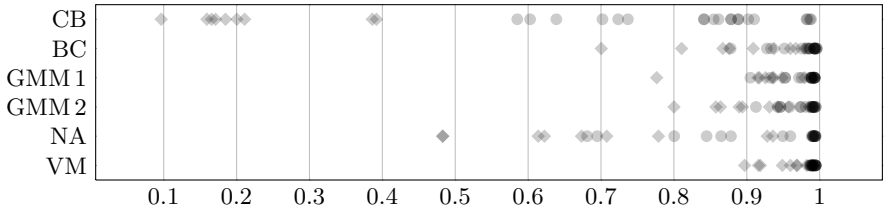
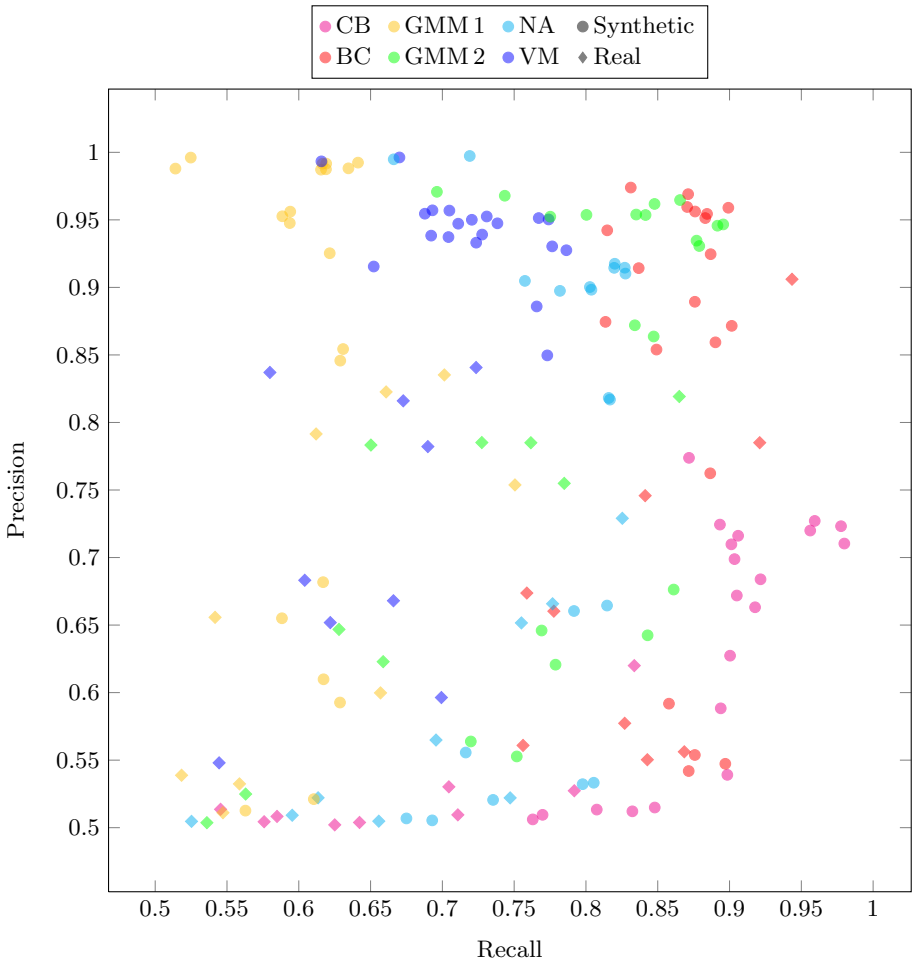


Fig. 10. Precision and Recall for each method



Tables 1 to 29, from the supplementary material of this article, show the performance of each method for each sequence:

- Learning phase:
  - Street: tables 1 to 5;
  - Rotary: tables 6 to 10.
- Evaluation phase:
  - Street: tables 11 to 15;
  - Rotary: tables 16 to 20;
  - Real applications: tables 21 to 29.

3.3 Analysis

From a statistical point of view (Figure 6), we can notice that the best method of our tests is BC, since its F-measure has the shortest range of values, with highest values (from 0.65 to 0.93 approximately). The case of the VM method is interesting because its F-measure is focused around the interval $[0.8; 0.85]$. These observation can be confirmed by Figure 10, where BC and VM have the greatest numbers of points coming close the $(1, 1)$ point. GMM1 has also a similar behaviour, around the 0.75 value, and a very good precision. GMM2 has a point of focus around the 0.9 value, but has also a wide interval of F-measures. The CB approach returns a very wide range of values, which could be induced by the high variability of the parameters of the method. Figure 10 informs us that the real videos of our benchamrk are not correctly processed by CB, impacting a global bad results. This phenomenon can also be observed for the NA, in a more negative way.

As illustrated in Figure 8, the PSNR gives us equivalent general informations about the tested BSA. We can also notice an increasing feeling of non-control of the results of CB and NA. Points of focus are also observable for VM ($[50; 60]$) and GMM1 ($[45; 55]$).

From a structural point of view, the values of SSIM and D-score lead to similar conclusions: CB and NA are not constant, and not efficient on the whole benchmark. Its seems even better to choose NA (SSIM greater than 0.4) instead of CB (SSIM can be around 0.1 or 0.2).

4 Conclusion

In this article, we have proposed to test the benchmark proposed in the BMC contest, with six classic background subtraction algorithms of the literature. Thanks to the measures we have computed, we can determine several qualities of the tested methods.

We would like to propose an other contest in 2013, with maybe more real videos, containing complex contexts. The BMC website is an interesting way to keep our benchmark available to researchers who want to test their algorithm.

References

1. Y. Benezeth, P-M. Jodoin, B. Emile, H. Laurent, and C. Rosenberger. Review and evaluation of commonly-implemented background subtraction algorithms. In Proc. of *IEEE Int. Conf. on Pat. Rec.*, 2008.
2. Y. Dhome, N. Tronson, A. Vacavant, T. Chateau, C. Gabard, Y. Goyat, and D. Gruyer. A benchmark for background subtraction algorithms in monocular vision: a comparative study. In Proc. of *IEEE Int. Conf. on Image Proc. Theory, Tools and App.*, 2010.
3. Y. Goyat, T. Chateau, L. Malaterre and L. Trassoudaine. Vehicle trajectories evaluation by static video sensors. In Proc. of *IEEE Int. Conf. on Intel. Transp. Sys.*, 2006.
4. D. Gruyer, C. Royere, N. du Lac, G. Michel and J.-M. Blosseville. SiVIC and RTMaps, interconnected platforms for the conception and the evaluation of driving assistance systems. In Proc. of *World Cong. and Exh. on Intel. Trans. Sys. and Serv.*, 2006.
5. E. Hayman and J.-O. Eklundh. Statistical background subtraction for amobile observer. In Proc. of *Int. Conf. on Comp. Vis.*, 2003.
6. P. Kaewtrakulpong and R. Bowden. An improved adaptive background mixture model for realtime tracking with shadow detection. In Proc. of *Eur. Work. on Adv. Video Based Surv. Sys.*, 2001.
7. K. Kim, T. H. Chalidabhongse, D. Harwood and L. Davis. Real-time foreground-background segmentation using codebook model. *Real-time Imag.*, **11**(3):167–256, 2005.
8. C. Lallier, E. Renaud, L. Robinault, L. Tougne. In Proc. of *IEEE Int. Conf. on Adv. Video and Signal-based Surv.*, 2011.
9. L. Li, W. Huang, I. Y. H. Gu, and Q. Tian. Foreground Object Detection from Videos Containing Complex Back- ground. In Proc. of *ACM Multimedia*, 2003.
10. A. Prati, I. Mikic, M. Trivedi, and R. Cucchiara. Detecting moving shadows: Algorithms and evaluation. *IEEE Trans. on PAMI*, **25**(7):918–923, 2003.
11. C. Stauffer and W. E. L. Grimson. Adaptative background mixture models for a real-time tracking. In Proc. of *IEEE Int. Conf. on Comp. Vision and Pat. Rec.*, 1999.
12. Z. Zivkovic. Improved adaptive Gaussian mixture model for background subtraction. In Proc. of *IEEE Int. Conf. on Pat. Rec.*, 2004.
13. Z. Zivkovic and F. v. d. Heijden. Efficient adaptive density estimapion per image pixel for the task of background subtraction. *Pat. Rec. Let.*, **27**(7):773–780,2006
14. Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. on IP*, **13**(4):600–612, 2004.