Metrics to evaluate binary classifiers for video background extraction: weaknesses, strengths, and perspectives

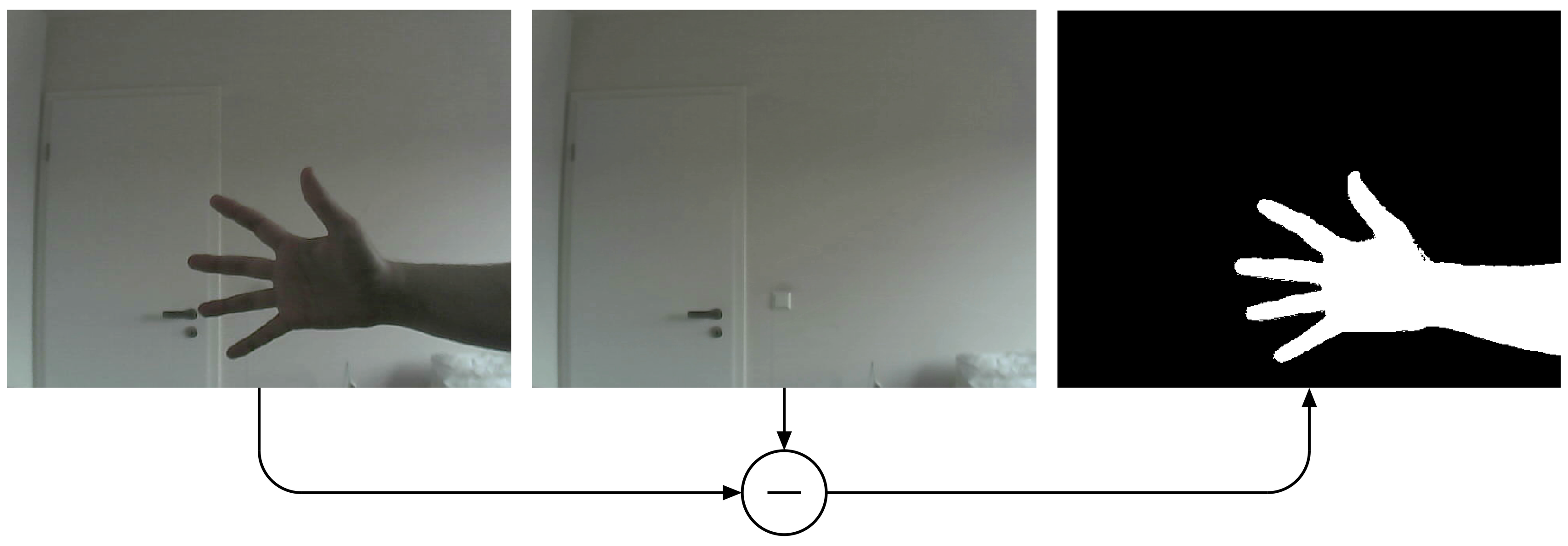
Université de Liège

Benjamin Laugraud, Sébastien Piérard, and Marc Van Droogenbroeck
{BLaugraud, Sebastien.Pierard, M.VanDroogenbroeck}@ulg.ac.be
INTELSIG, Montefiore Institute, University of Liège, Belgium

Introduction

The *background subtraction* is a typical instance of binary classification problems, which predict the membership of an object to a class C. Such a classifier highlights the motion in a video sequence. At each frame, it decides whether the pixels belong to the foreground or to the background of the scene. To make it possible, the biggest challenge is to find a solid way to model the background. Once such a model has been constructed, at each time t and for each pixel x, the observed intensity $I_t(x)$ is compared with the estimated background $B_t(x)$, using a distance d (e.g. Euclidean, Mahalanobis, etc) and a decision threshold τ [1]:

$$C_t(x) = \begin{cases} \text{foreground}\,(+) & \text{if } d(I_t(x),\, B_t(x)) > \tau \\ \text{background}\,(-) & \text{if } d(I_t(x),\, B_t(x)) \leq \tau \end{cases}$$



The limits of the classical evaluation approach

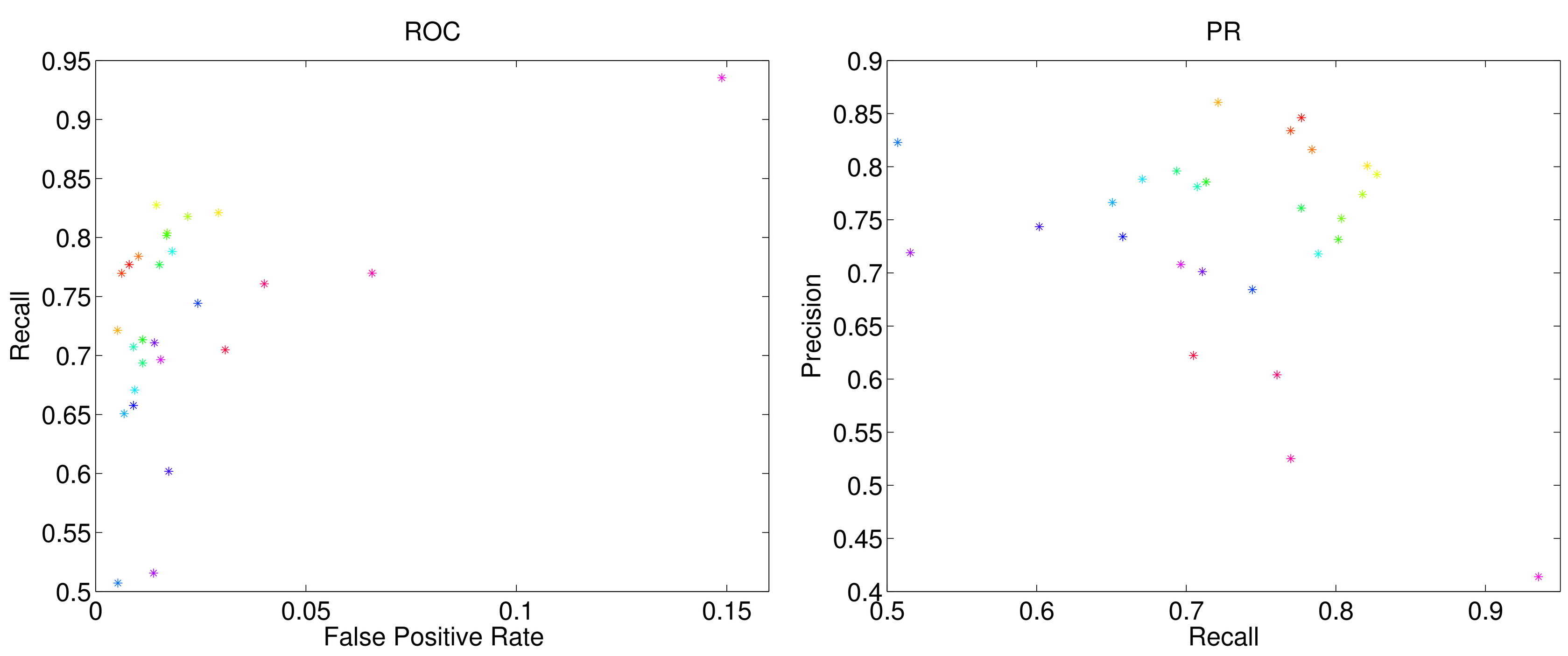
The evaluation of the results of a binary classifier is not trivial. The classical approach consists of measuring how close a classifier is to the best theoretical classifier. To achieve this goal, a *confusion matrix* that offers the basis to build a rigorous evaluation is constructed. Such a matrix contains the number of true and false positive classifications (TP and FP), and the number of true and false negative classifications (TN and FN). Note that this kind of constructions is only possible by using a ground-truth (i.e. the results of the best theoretical classifier) made by a human expert.

Image [2] — Ground-Truth [2] — Example of classification [2]

Considering the confusion matrix, the main limits of the classical approach is to use this information to perform a relevant and a fair measure of the performance of a given classifier.

Performance spaces

Based on TP, FP, TN and FN, one can plot the results of a classifier in spaces such as *Receiver Operating Characteristic* (*ROC*) and *Precision-Recall* (*PR*). These spaces presents the performance of a binary classifier and display its position beside theoretical trade-offs based on different rates. For each space, there is a target point which represents the ideal trade-off, and subsequently, the best theoretical classifier. However, ROC and PR graphs only provide partial information. Indeed, the spaces ignore the information provided by the exact number of foreground and background pixels in the ground-truth. Moreover, they are hard (and might be impossible) to interpret. Therefore, this leads to a subjective interpretation of objective measures.



Let *recall* $(\mathrm{Re}) = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, *false positive rate* $(\mathrm{FPR}) = \mathrm{FP}/(\mathrm{FP}+\mathrm{TN})$ and *precision* $(\mathrm{Pr}) = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$. The figure above plots examples of both ROC and PR graphs from data taken out of the CD.net dataset [3] in October 2013 for 27 distinct methods.

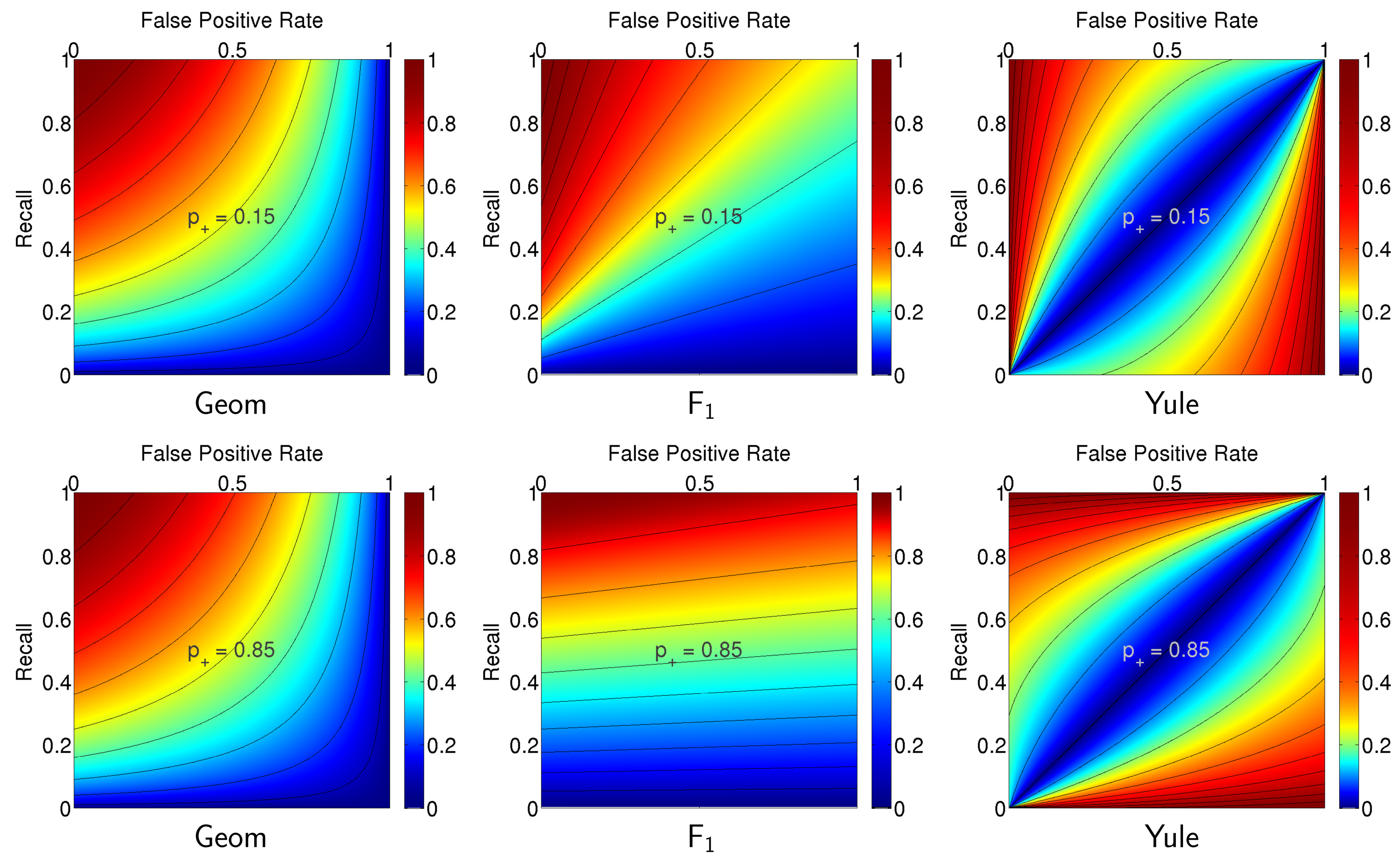
The metrics and their strengths

To overcome the interpretation issue of the performance spaces, some measures called *metrics* and based on combinations of the values of the confusion matrix, have been elaborated. They characterize the performance of a binary classifier in a simple number which can be used as a score. In other words, their goal is to measure the distance of a binary classifier to the best theoretical classifier. For example, in the field of the background subtraction, such metrics serve to rank algorithms and to choose the best one. Authors have presented many of them, but none of these metrics is unanimously adopted.

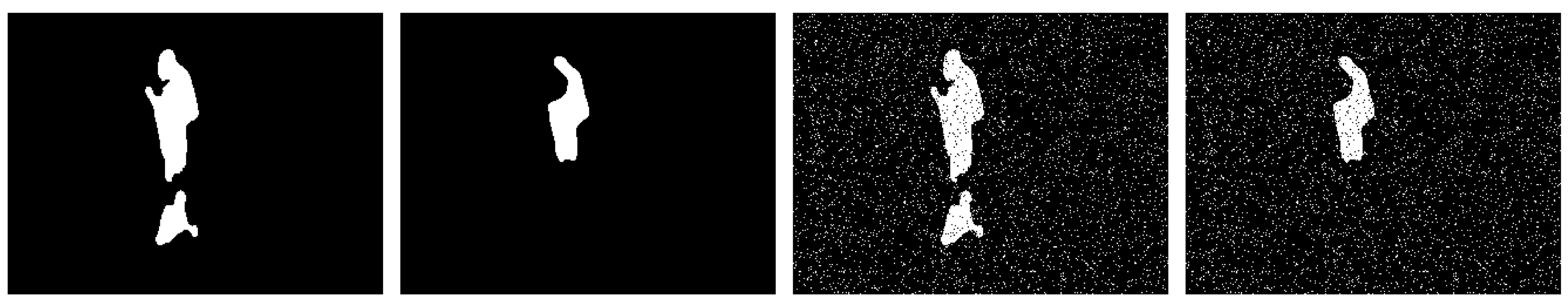
Let *true negative rate* $(\mathrm{TNR}) = \mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$ and *false negative rate* $(\mathrm{FNR}) = \mathrm{FN}/(\mathrm{FN}+\mathrm{TP})$. Some examples of common metrics are the *geometric mean of* Re *and* TNR (*Geom*) $= \sqrt{\mathrm{Re} \times \mathrm{TNR}}$ [4], $F_1 = 2.\mathrm{Pr}.\mathrm{Re}/(\mathrm{Pr}+\mathrm{Re})$ [5] and *Yule* $= |\mathrm{Pr} + \mathrm{TN}/(\mathrm{TN}+\mathrm{FN}) - 1|$ [6].

Weaknesses

- It is possible to create an infinity of metrics. Thus, it is hard to choose one, even though the high-level application relying on background subtraction is known.
- As a matter of fact, it is impossible to recover the confusion matrix based on a single metric. It follows that there is a loss of information when one considers only a single metric to interpret the performance of a background subtraction method. The set of performances leading to the same metric lie on an iso-performance curve [7].
- Combining several metrics is, in practice, delicate. On the one hand, combining highly correlated metrics is useless. On the other hand, combining weakly correlated metrics could be meaningless since those metrics might contradict.
- Let p_+ be the *positive prior* (i.e. the probability for a pixel to be in the foreground according to the ground-truth). It should be stressed that some metrics depend on p_+ (e.g. F_1 and Yule), to the contrary of others (e.g. Geom). The knowledge of the prior used to assess a background subtraction method is necessary to interpret the prior-dependent metrics. The same value for a metric may correspond to a random classifier, or a classifier behaving significantly better than a random classifier, depending on the particular priors.



- Some metrics are sensible to noise [8]. The figure below illustrates this phenomenon with F_1 and the PBAS [9] and Spectral-360 [10] algorithms applied to the 975$^{\text{th}}$ frame of the "corridor" video sequence of the CD.net dataset. When the considered metric is calculated for the two first classifications, it says that Spectral-360 is better than PBAS. However, when some noise is added to the original classifications as a post-processing step, the ranking based on F_1 is inverted.



PBAS $\mathrm{F}_1 = 0.762$ — Spectral-360 $\mathrm{F}_1 = 0.806$ — PBAS + noise $\mathrm{F}_1 = 0.416$ — Spectral-360 + noise $\mathrm{F}_1 = 0.376$

Perspectives

While it is obvious that we have to solve some of the weaknesses listed above, there are some perspectives to explore. Firstly, as the performance is application dependent, we need to predict it in a new context based on the performance measured previously in other contexts. Secondly, it would be interesting to add some semantic in the evaluation process. For example, one could distinguish the positive errors according to their connectivity to the foreground in the ground-truth. Unconnected errors give birth to new *blobs* (i.e. positive areas), while the others introduce noisy contours. According to the target application requirements, different costs could be given to these two types of positive errors, based on the filtering possibilities.

References

[1] Y. Benezeth, P. Jodoin, B. Emile, H. Laurent, and C. Rosenberger. Review and evaluation of commonly-implemented background subtraction algorithms. In *IEEE International Conference on Pattern Recognition (ICPR)*, pages 1–4, December 2008.

[2] O. Barnich and M. Van Droogenbroeck. ViBe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image Processing*, 20(6):1709–1724, June 2011.

[3] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar. changedetection.net: A new change detection benchmark dataset. In *Change Detection Workshop (CDW)*, Providence, Rhode Island, June 2012.

[4] R. Barandela, J. Sánchez, V. García, and E. Rangel. Strategies for learning in class imbalance problems. *Pattern Recognition*, 36(3):849–851, 2003.

[5] A. Baumann, M. Boltz, J. Ebling, M. Koenig, H. Loos, M. Merkel, W. Niem, J. Warzelhan, and J. Yu. A review and comparison of measures for automatic video surveillance systems. *EURASIP Journal on Image and Video Processing*, page 30 pages, 2008.

[6] S. Elhabian, K. El-Sayed, and S. Ahmed. Moving object detection in spatial domain using background removal techniques – State-of-art. *Recent Patents on Computer Science*, 1:32–54, January 2008.

[7] F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, 2001.

[8] P.-M. Jodoin, S. Piérard, Y. Wang, and M. Van Droogenbroeck. Overview and benchmarking of motion detection methods. In T. Bouwmans, F. Porikli, B. Hoferlin, and A. Vacavant, editors, *Background Modeling and Foreground Detection for Video Surveillance*. Chapman and Hall/CRC, 2014. to be published.

[9] M. Hofmann, P. Tiefenbacher, and G. Rigoll. Background segmentation with feedback: The pixel-based adaptive segmenter. In *Change Detection Workshop (CDW)*, Providence, Rhode Island, June 2012.

[10] M. Sedky, C. Chibelushi, and M. Moniri. Image processing: Object segmentation using full-spectrum matching of albedo derived from colour images. Patent application PCT GB2009/002829, 2009.

BEMEKO — Liège, Belgium — November 7th 2013