F1 Score Assesment of Gaussian Mixture Background Subtraction Algorithms Using the MuHAVi Dataset

Jorge Sepúlveda, Sergio A. Velastín

Universidad de Santiago de Chile, {jorge.sepulvedao,sergio.velastin}@usach.cl

Keywords: Background Subtraction Algorithms, Gaussian Mixture model, MuHAVi, F1-Score.

Abstract

Background subtraction algorithms are mainly used to segment some specific moving objects in an image sequences. Within of the action recognition context, these methods may be proper to generate automatically silhouettes of the human actions. In this way, MuHAVi is a human action dataset which provides a small set of manually annotated silhouettes and a large set of multi-camera raw video. The purpose of this work is to use a segmentation algorithm to generate automatically the whole dataset of silhouettes for the MuHAVi raw video. The F1-score unit measurement is the selection criterion as for the best method to generate such silhouettes. This paper focuses especially on background subtraction methods that create a statistical model of the background, typically using a mixture of Gaussian. The best-evaluated algorithm can then be used to generate automatically a set of silhouettes.

1 Introduction

Recognising a human activity using a computer system is a popular objective in computer vision research. The underlying idea is to try to emulate the mind ability to discover from an image sequence, which human activity is taking place typically from a universe of pre-defined activities such as walking, running, jumping, etc. Extracting the essential information contained within of an image sequence is thus part of the challenges human activities recognition.

A general approach to identifying a human activity in an images sequence could be represented as a processing pipeline comprised of several techniques, methods, or algorithms applied either in image processing or machine learning activities. It starts with data acquisition (e.g. from surveillance or traffic monitoring cameras) and the pre-processing of the images (e.g. temporal and spatial filtering). A next step involves features extraction to find shapes and relevant objects. This stage, such as, characterizes a region that delineates within the image some particular artifact. At this phase, a segmentation process could take place, to separate foreground from its background, where the outcome consists a set of (ideally perfect) segmented silhouettes which then are used by the higher-level stages. The feature extraction represents the input image as a set of characteristic elements grouped in a vector of image properties. A silhouette represents part of an image of interest (typically the body of a person) and is later used to show where in the images, characteristic features can be extracted that might be meaningful for human action recognition (such as optical flow, motion history images, histogram of gradients, visual hulls, etc.). Intuitively, the quality of subsequent human action recognition methods will depend on the quality of the extracted silhouettes i.e., on the quality of the background subtraction method.

A background subtraction algorithm typically operates at the pixel level, labelling every pixel of an image either as foreground (part of the silhouette) or background. There are various approaches for segmentation an object [1]. The simplest methods consist of modeling the background with averages, median, or histograms of the image sequences frames over time. A more elaborated method creates independent statistical models for every pixel so that a pixel can be labeled during a sequence as a foreground or background according to some statistic rules e.g. established by Gaussian or Kernel models.

Several types of metrics measures the background subtraction performance, most of them supported on comparisons. Essentially, the results are compared with a reference (ground-truth), the bigger of a discrepancy between its reference the worst is the evaluation. Furthermore, the computer vision community has produced some datasets[8] to evaluate these methods, i.e., in complex illumination and environmental conditions. In this way, these datasets can offer a base of comparison among different methods. Consequently, MuHAVi[5] is presented as a dataset of common human activities, inspired by CCTV surveillance needs, with a set of manually annotated images, which, at the same time, can be used as ground-truth for making comparative evaluations of the segmentation methods.

Thus, this paper describes the performance evaluation of some background subtraction algorithms using MuHAVi[5] as a comparative dataset. Testing was focused on algorithms based on a Gaussian Mixture Model. The evaluation metric used was the F1-Score, commonly used in information retrieval and computer vision applications. The purpose was to find a comparison number based on this metric which gave optimal operation points for each of these algorithms, the final aim was to be able to generate a full dataset of silhouettes for MuHAVi, using the best algorithm. As the results will be representative of a good, but imperfect segmentation algorithm, researchers in human action recognition might be able to test the robustness of their methods to segmentation errors.

This paper is organized as follows, section two describes

background subtraction methods that use a statistical model of the background, specifically as a mixture of Gaussian components. Section three explains evaluation methods making a distinction between evaluations based on the discrepancy and those on visuals perception. Section four provides an overview of the MuHAVi dataset, section five presents experimental results of the tests conducted using the human action sequences with ground truth images, and section six concludes this work.

2 Background Subtraction methods

This section describes the background subtraction methods, based on Gaussian Mixture Models (*GMM*), used in this work to assess foreground segmentation in the MuHAVI dataset. A detailed description of these specific methods and other more complex algorithms are in the survey presented by Bouwmans[1], where it is also claimed that *GMM* algorithms are one of the most referenced methods in the literature.

In general, a background subtraction process constantly updates a background model of an image sequence. The purpose is to detect any unusual presence within of such sequence. A probability density function models the dynamics of every pixel in the image background. This is based on the assumption that the image background, with no movements in the objects of interest, presents a highly consistent and smooth behavior, which can be described by a statistical model. As a result, any new event should be detected by just identifying the pixels (of the object in movement) which do not fit the generated model.

Figure 1: Example of two frames taken from the MuHAVi sequence that describes human action of kicking (*Kick-Camera3-Person4*). The red spot shows a single pixel at two different times.

GMM models a pixel using a combination of Gaussian distributions, Figures 1, and 2 provide a simplified outline of this Gaussian Mixture operation. A red point in both frames of Figure 1 is highlighted to show the state of a simple pixel in two separate events. The first scene is part of the standing actor (the pixel gets the values of the actor's clothing) on the floor, while in the second, it just lies on the stage floor (background) acquiring the color of the stage floor. Also, Figure 2 shows the histogram of the sequence, for this specific pixel, evidencing two Gaussian distributions, with low and high variances for the background and foreground.

2.1 Gaussian Mixture Models

In a traffic monitoring system, Friedman and Russell[3] presented one of the first proposals of this method. They introduced a model of the background using a fixed number of Gaussian components, modeling a pixel with only three Gaussian: vehicles, shadows, and road (background). However, if a pixel does not fit any of these three components, it is not considered part of the model. Stauffer and Grimson[6] introduced a similar approach, but instead of a fixed number of components, they selected a K number of Gaussian mixtures (K typically between three and five), representing the background by more than one component. This model also incorporates two new adjustable parameters: a learning rate and a background proportion factor. Later, Zivkovic and Heijden[9] approximates the problem from a Bayesian perspective, selecting in real-time a suitable number of mixture Gaussian for a pixel. The algorithm, using a Dirichlet prior distribution, estimates in real-time the parameters of the mixtures and selects the number of Gaussian. Thus, in this case, the number K adapts automatically to each pixel distribution (multi-mode) and a pixel could be represented by just one or more Gaussian components. In an urban traffic monitoring context, Chen[2] improved the model proposed by Zivkovic & Heijden[9], taking into account illumination changes and the shadows cast by moving cars. He proposed an auto-adaptive Gaussian mixture model, introducing a factor to compensate global illumination changes, which also modifies dynamically the learning rate, to attenuate sudden changes in the global illumination. He also added a temporal-spatial filter to decrease noise and vibration problems produced e.g., by the wind on the cameras. Salvadori and Petracca[4] proposed implementing the GMM model using micro-controllers. To reduce memory use and computational cost, they approximate the integer numbers precision to overcome the lack of floating point in low-cost processors. The updating rules of Gaussian components were also modified as approximations that they call the Linear and Staircase models.

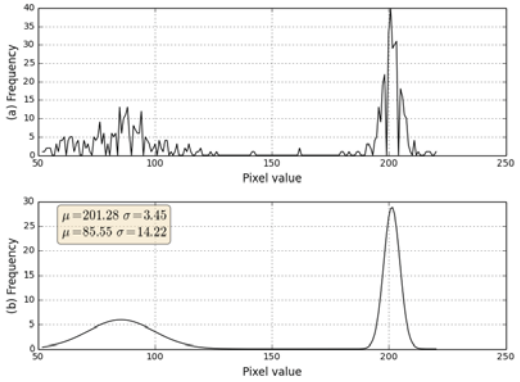


Figure 2: Pixel histogram of the whole sequence (*Kick-Camera3-Person4*) mentioned in Figure 1. Top: actual histogram, Bottom: Gaussian Mixture Model

3 Evaluation Measure

Performance evaluation of background subtraction methods, measure segmentation quality and provides a common criterion to compare the general performance of the algorithms. Ex-

ist several evaluation metrics in the literature, but no general agreement for using a common set of metrics. The measures often mentioned in classification are Precision, Recall, and F1-score, all generated by a mix of true/false positive/negative measures (TP, TN, FP, FN). In the same way, a BMC workshop [7] proposed evaluation criteria for evaluating such algorithms, with static quality and visual perception metrics. The advantage of F1 is that it provides a single measure of quality easier to understand by end-users. It combines the results values of precision and recall. Precision is a proportion measure of the foreground pixels correctly selected. Conversely, recall is a proportion measure of the foreground pixels selected.

The empirical discrepancy is the most common evaluation method based on comparative measures, mostly comparing the binary mask (foreground) and its ground-truth (reference image). A ground-truth image is also a binary mask, which labels a pixel as background or foreground, providing a set of reference images for comparing segmentation results. Figure 3 shows an example of a ground-truth frame (obtained from *WalkTurnBack-Camera3-Person1* sequence) and its segmented silhouette mask. The *True Positive* (TP), in this example, is the amount silhouette pixels (foreground) the system got right, *True Negative* are the pixels of background image correctly selected. *False Positive* is the number of background pixels wrongly selected as a silhouette and *False Negative* are the silhouette pixels wrongly classified as background. The result of these pixel measures populates a contingency table which is the base of many of the metrics used for performance evaluation.

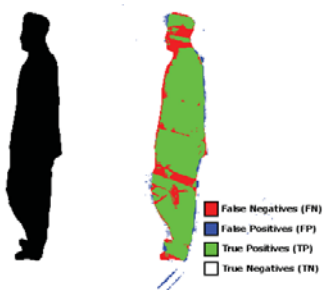


Figure 3: Ground-truth image and its annotated foreground mask taken from human action of walking and turn back sequence (*WalkTurnBack Camera3 Person 1*).

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (3)$$

4 MuHAVi dataset

MuHAVi[5] is a video dataset mainly aimed at testing human action recognition algorithms based on silhouettes. A set of manually annotated silhouettes has been specially prepared for evaluation of these type of methods, providing accurate silhouette data of the video frames. This subset is called MuHAVi-MAS (Manually Annotated Silhouettes). Nevertheless, the manually annotated silhouettes can be also used as ground truth for evaluating segmentation methods (i.e. the purpose of this paper).

All the actions in the dataset were registered with standard CCTV not calibrated cameras, from different observation angles at similar distances from the actors. The stage was also illuminated with night street lights simulating surveillance monitoring in uneven background conditions. Diverse classes of human actions (17) compose this dataset, including common actions such as run, walk, run, punch, kick, (shotgun) collapse/fall and other more complex activities, from a human point of view, e.g., look in a car, drunk walk, jump over a gap, etc. 14 actors perform each action, obtained by eight cameras at the 4 sides and 4 corners of a rectangular stage, shows that figure 4. The dataset has 8x17x14=1904 video segments in total. A subset of 5 actions (ShotGunCollapse, WalkTurnBack, Punch, Kick, Run) constitute the annotated silhouettes (MuHAVi-MAS), which in turn played by 2 actors (Person1, Person4), and registered by 2 cameras (Camera3, Camera4), i.e 5x2x2=20 actions. Clearly, if a good segmentation algorithm is identified, it will open up opportunities for much more extensive work in human action recognition methods.

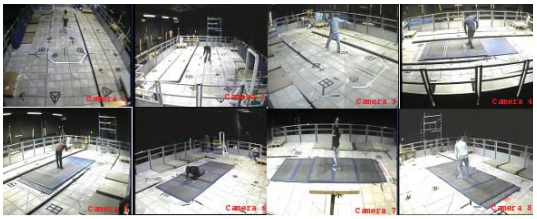


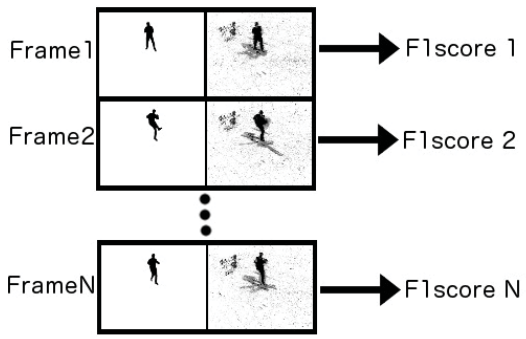
Figure 4: View of the eight cameras provided by MuHAVi.

Figure 5: Example of annotated images provided by MuHAVi

5 Evaluation Methodology

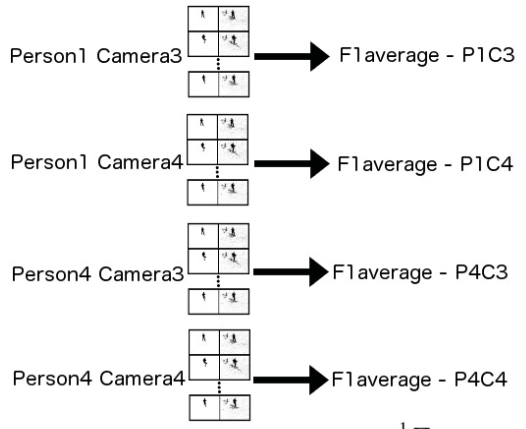
Using the F1-score comparisons helped evaluate and rank the algorithms performance. Doing extensive experiments get a representative (best) F1 measure per algorithm. Thus, comparing a generated silhouette (by the algorithm under evaluation) with its corresponding reference (annotated image) produced a base unit of measure (Figure 6). A final measure, defined as this score number, was conducted taking an average of the base unit for each frame in a specific sequence (a role played, for example, by the actor Person1 registered in Camera3), and then

averaging each one of such averaged sequences (combination of the averages Person1-Camera3, Person1-Camera4, Person4-Camera3, and Person4-Camera4). Finally, the overall score is the average of F1 measures achieved by each action (those with annotated silhouettes) of MuHAVi. Figures 6, 7, and 8 illustrate how the last F1 measure was prepared.



$$F1_{AVERAGE} = \frac{1}{N}\sum_{i=1}^{N} F1_i$$

Figure 6: The F1 score unit base compares each single frame of a sequence with its ground-truth mask.

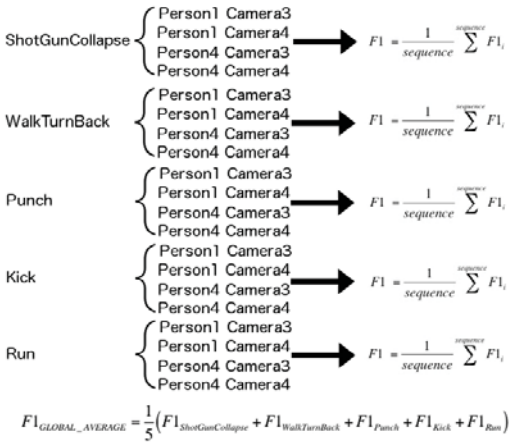


$$F1_{sequence} = \frac{1}{4}\sum F1_{average_i}$$

Figure 7: The average of each sequence per actor and camera is averaged to obtain a F1 score per sequence.

The representative curves of F1-score measurements were plotted on a graph to facilitate comparisons among the algorithms (Figure 9). The different points were collected just modifying a specific configuration parameter of the methods and running again the same algorithm with this new configuration.

The GMM methods depend on a set of parameters, mainly the *learning rate* and *threshold* (Mahalanobis distance). The *learning rate* in all experiments remained in 0.001 following [9], with this value the algorithms take roughly 100 frames to compute Gaussian parameters of each individual pixel. Thereby, an object (silhouette) that is static for around 100 frames becomes part of the background. Moreover, the *threshold* was increased systematically within a range of 21 fixed values. A F1-score curve for a single action was then built combining a fixed parameter (*learning rate*) and incremental variations of the *threshold* (Mahalanobis distance). All MuHAVi sequences with the annotated images (20 actions) repeated this same procedure. Finally, the overall F1-score curve for a given algorithm is the average of the independent F1-score curves of each action sequence experimented.



$$F1_{GLOBAL\_AVERAGE} = \frac{1}{5}\left(F1_{ShotGunCollapse} + F1_{WalkTurnBack} + F1_{Punch} + F1_{Kick} + F1_{Run}\right)$$

Figure 8: The final F1 score representative of an algorithm is the average of F1 measures obtained per each MuHAVi action.

6 Experimental Results

The experiments used four different classes of GMM methods, described in section 2.1. A suitable verification software (written in C++), specially prepared for the experiments, included these four alternative methods. The GMM models were implemented in C++ and based in OpenCV libraries. The Mixture of Gaussian[9] method is available in OpenCV and designated as MOG2, was incorporated in the verification software as a C++ library. The Select Adaptive Gaussian Mixture Model[2] (SAGMM), originally implemented in Matlab, was also ported to a C++ library and used in the same verification software. At last, UCV GMM[4] algorithms included in the verification software, the author provided a library (developed in C language) with two alternative versions: *Linear* and *Staircase*.

Figure 9 shows the overall F1 measurements averages of all GMM algorithms examined. The F1 measures are compared with a threshold value; this is a factor related with the Mahalanobis Distance (MD). As the distance increases the F1 measures get better. But, seems no appreciable F1 improvement (for all algorithms) when MD is greater than 10. A good operation point for every one of these algorithms could be in the range five to ten. In general terms, SAGMM[2] has better overall F1-score with respect the other algorithms. On the one hand, verified by the F1 curve located just above of the other curves, and on the other by a good trade-off among measures of *Precision* and *Recall*, this latter reflected by F1 measures within 0.6 as its MD is more than five.

F1 is a measure of the true positives (silhouette) influenced by false positives (pixels of background interpreted as fore-

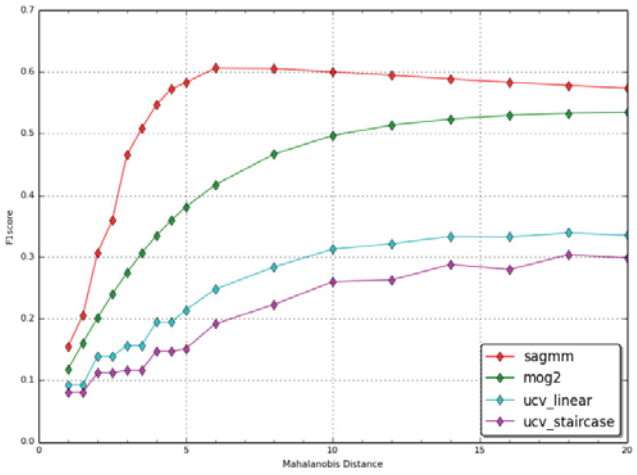


Figure 9: The final F1 score representative per algorithms.

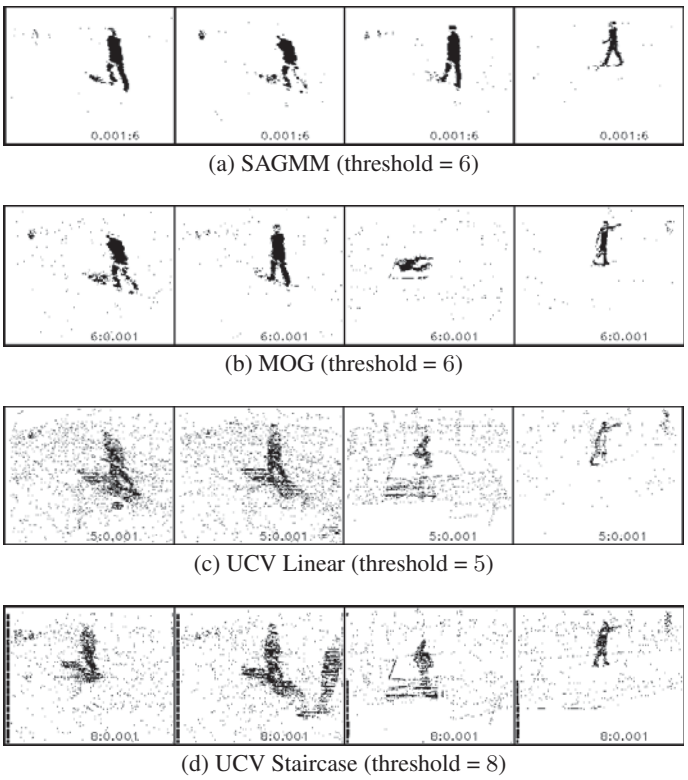
ground) and false negatives (pixels of foreground recognize as background). The F1-score around 0.6 can be interpreted as silhouettes with a low rate of false negatives and a similar rate of false positives, a good trade-off between errors and thus the silhouette could recover. It can also be checked that changes proposed by SAGMM[2] (spatial-temporal and illumination change filters) have improved the overall performance with respect MOG2[9]. Furthermore, MOG also has better performance of two versions of UCV[4] in all range of MD, and UCV[4] linear version gets better of UCV staircase.

Figure 10 shows examples of silhouette mask obtained for every algorithm. These figures expose the lower performance of the UCV algorithms arising from a higher number of false negatives (pixels within silhouette that were not detected). This mainly affects *Recall* which in turn affects the F1 measure negatively. On the other hand, SAGMM and MOG2 have less false negatives confirming the good F1 score measurement for both algorithms.

From a MuHAVi perspective, dissimilar results per actions are shown in Figure 11. The F1-score results have been separated by algorithms. The human actions of *Walking*, *RunStop* and *Kick* are above the general average for algorithms SAGMM and MOG2. However, the *ShotGunCollapse* and *Punch* actions had the worst performance in SAGMM, which explains a slight degradation in the overall F1 from MD more than ten.

7 Conclusions

This paper has presented an evaluation method of the background subtraction algorithms, based on the Gaussian mixture models, from a perspective of F1-score measurement. The overall F1-score is an average of the average measurements collected from each specific sequence of a particular algorithm. This score represents a consolidated number that gives a general idea on how every algorithm works, even though at the same time might hide a better (or worse) performance of a particular sequence. Notwithstanding this, the overall score showed a good assessment of the algorithms, giving a certain selection and configuration criterion for the different algorithms.



(a) SAGMM (threshold = 6)

(b) MOG (threshold = 6)

(c) UCV Linear (threshold = 5)

(d) UCV Staircase (threshold = 8)

Figure 10: Example of silhouette masks obtained for each algorithm.

The evaluation method made comparisons between a foreground mask (silhouette achieved by the method) and the annotated silhouettes (ground-truth) provided by MuHAVi. The experimental results have concluded, examining the F1-score curves that SAGMM has better performance than the MOG2 and UCV methods. However, despite the poorer performance of the UCV algorithms (designed to be embedded on microcontroller cards), their computational speeds are clearly better than SAGMM and MOG2, something that was not evaluated in this paper.

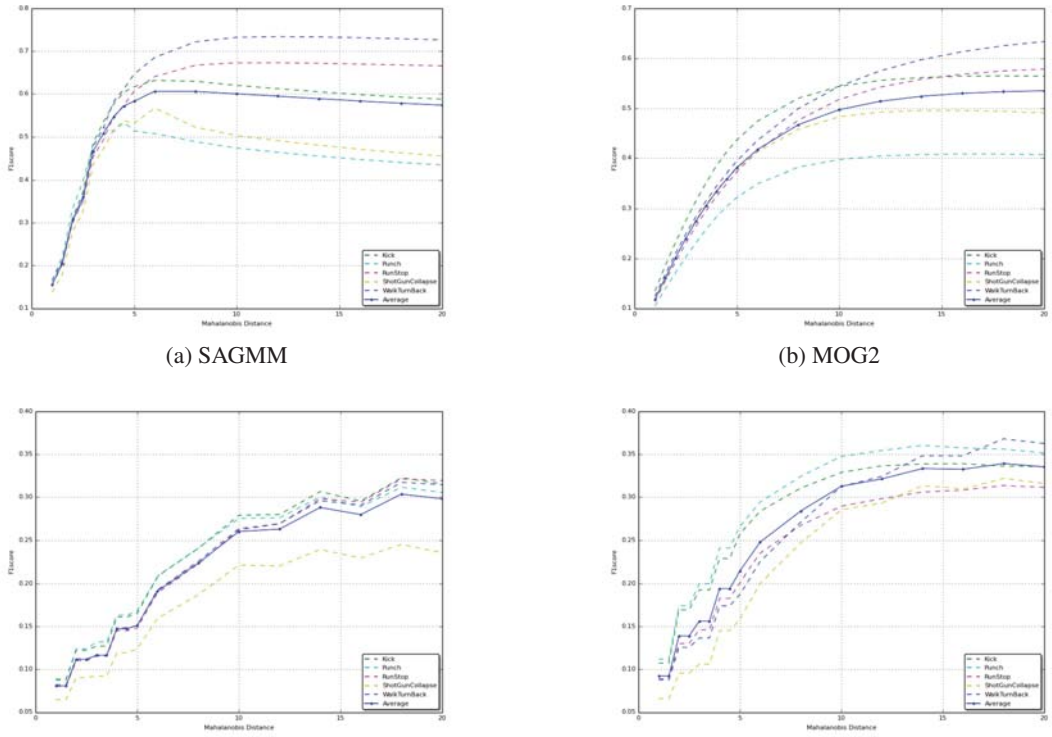
Finally a complete set of silhouettes all MuHAVi sequences were generated using the SAGMM algorithm configured with the parameters found in the experiments and is now available to the research community (as MuHAVi-uncut).

Acknowledgements

The authors gratefully acknowledge the Chilean National Science and Technology Council (Conicyt) for its funding under grant CONICYT-Fondecyt Regular no. 1140209 ("OBSERVE").

References

[1] T. Bouwmans. Recent advanced statistical background modeling for foreground detection: A systematic survey. *Recent Patents on Computer Science*, 4(3), 2011.



(a) SAGMM

(b) MOG2

(c) UCV Staircase

(d) UCV Linear

Figure 11: Performance F1 measures obtained during experiments separated by algorithms and actions. It shows the results as Mahalanobis Distance value was increased.

[2] Z. Chen and T. Ellis. Self-adaptive gaussian mixture model for urban traffic monitoring system. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1769–1776, 2011.

[3] Nir Friedman and Stuart Russell. Image segmentation in video sequences: A probabilistic approach. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, UAI'97, page 175181, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[4] Claudio Salvadori, Dimitrios Makris, Matteo Petracca, Jesus Martinez-del Rincon, and Sergio Velastin. Gaussian mixture background modelling optimisation for microcontrollers. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Charless Fowlkes, Sen Wang, MinHyung Choi, Stephan Mantler, Jrge n Schulze, Daniel Acevedo, Klaus Mueller, and Michael Papka, editors, *Advances in Visual Computing*, volume 7431 of *Lecture Notes in Computer Science*, pages 241–251. Springer Berlin Heidelberg, 2012.

[5] Sanchit Singh, Sergio A. Velastin, and Hossein Ragheb. MuHAVi: a multicamera human action video dataset for the evaluation of action recognition methods. In *Proceedings of the 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, AVSS '10, page 4855, Washington, DC, USA, 2010. IEEE Computer Society.

[6] Chris Stauffer and W. Eric L. Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition*, volume 2, page 22462252, 1999.

[7] Antoine Vacavant, Thierry Chateau, Alexis Wilhelm, and Laurent Lequivre. A benchmark dataset for outdoor Foreground/Background extraction. In Jong-Il Park and Junmo Kim, editors, *Computer Vision - ACCV 2012 Workshops*, volume 7728 of *Lecture Notes in Computer Science*, pages 291–300. Springer Berlin Heidelberg, 2013.

[8] Daniel Weinland, Remi Ronfard, and Edmond Boyer. Free viewpoint action recognition using motion history volumes. In *COMPUTER VISION AND IMAGE UNDERSTANDING*, 2006.

[9] Zoran Zivkovic and Ferdinand van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recogn. Lett.*, 27(7):773780, May 2006.