A New Evaluation Approach for Video Processing Algorithms

Anh-Tuan Nghiem, François Bremond, Monique Thonnat, Ma Ruihua

A New Evaluation Approach for Video Processing Algorithms

A.T NGHIEM, F. BREMOND, M. THONNAT and R. MA
Project Orion
INRIA-Sophia Antipolis - France
{atnghiem,Francois.Bremond,Monique.Thonnat,Ruihua.Ma}@sophia.inria.fr

Abstract

We present a new evaluation methodology to better evaluate video processing performance. Recent evaluation methods [10], [9], [11] depend heavily on the benchmark dataset. The result may be different if we change the testing video sequences. The difference is mainly due to the video sequence content which usually includes many video processing problems (illumination changes, weak contrast etc.) at different difficulty levels. Hence it is difficult to extrapolate the evaluation result on new sequences.

In this paper, we propose an evaluation methodology that help to reuse the evaluation result. We try to isolate each video processing problem and define quantitative measures to compute the difficulty level of a video relatively to the given problem. The maximum difficulty level of the videos at which the algorithm is performing good enough is defined as the upper bound of the algorithm capacity for handling the problem. To illustrate this methodology, we present metrics that evaluate the algorithm performance relatively to the problems of handling weakly contrasted objects and shadows.

1. Introduction

In this paper we propose a new methodology to evaluate video processing programs in order to obtain evaluation result that can be generalized to new video sequences. With the development of video surveillance systems, many algorithms are proposed to automate the processing of video flows. To select the most appropriate algorithm given a characterization of the scene, the performance evaluation stage becomes crucial. Usually, to evaluate video processing algorithms, a set of video sequences is collected together with the ground truth characterizing the tasks to be performed. The performance of one algorithm indicates how good it processes these specific video sequences. Although adopted by many projects, this approach contains two main limitations.

First, the evaluation results depend on testing sequences. In other words, these results may change dramatically with a new set of video sequences. The reason is that each video sequence comprises various problems at various difficulty levels and the final results are affected by all these factors. With a new video sequence, there is a new combination of problems. Thus, the algorithm performance on this sequence is unpredictable.

Secondly, a video processing algorithm is usually designed to work in specific conditions (outdoor, indoor scene, containing fast/slow illumination changes etc.). However there is no quantitative measure to compute the difficulty level of a video sequence relatively to a given problem. Therefore we do not know up to which difficulty level of the video, the algorithm can still achieve good result.

This paper introduces an approach that partly solves these issues. By defining problem-specific metrics, we can measure the algorithm capacity for solving each problem separately. Then, this capacity can be generalized to estimate the algorithm possibility of success on other sequences. For instance, if the difficulty level of a new sequence on one problem is higher than the capacity of a given algorithm, this algorithm may not work well on this sequence. Based on this approach, we present the metrics that evaluate the algorithm performance relatively to the problems of handing weakly contrasted objects and shadows.

2. Related works

There are many individual works on the evaluation of some aspects of video surveillance systems. For instance, [10] characterizes object detection algorithms using the metrics concerning correct detection, detection failures, number of splits, merges and matching area.[9] introduces metrics to measure the similarity between two trajectories to evaluate the tracking tasks. [7] presents a framework using the pseudo synthetic videos to evaluate video tracking performance. [11] uses the metrics like true positive, false positive, true negative on benchmarking data set to evaluate the performance of different shadow detection algorithms.

For an overview of the individual works as well as current workshops and projects on performance evaluation, voir [8]. Nevertheless, these works have little influence on the research community because they do not constitute a whole evaluation platform enabling a new algorithm to be evaluated. Moreover, their data set is not big enough to achieve reliable evaluation results.

Therefore, to answer the need of having a publicly available set of annotated video sequences, many projects (CAVIAR [1]) and workshops (PETS [6], VisualSurveillance) have been created. These research programs provide video sequences at various subjective "difficulty levels" together with associated ground truth. Nevertheless, because people participating to the workshops such as PETS often choose the testing sequences arbitrarily and evaluate their algorithm using the metrics defined by themselves, the performance comparison with other algorithms remains difficult. Other programs such as [3], [5], [2] try to overcome this problem by providing standard metrics and dataset to evaluate the performance of different algorithms. However they still suffer from several issues. Firstly, the "difficulty level" of video sequences is estimated manually by experts. For the same sequence, two experts may assign two different difficulty levels. Secondly, one video sequence may contain several problems at different difficulty levels. All these problems contribute to the "difficulty level" of the input data of the algorithm (e.g. the video for the object detection algorithm). Therefore, there are plenty ways of combining problems that produce the same difficulty level. Consequently, the ranking of one algorithm on two sequences at the same difficulty level may be different because the algorithm is efficient relatively to one particular problem. Thirdly the extrapolation of the evaluation results for a new video is nearly impossible. The performance of one algorithm on a new sequence is estimated through its performance on the most "similar" sequence in the testing set. The selection of "the most similar" sequence and the prediction of the performance based on the "closeness" of two sequence are often subjective and inaccurate. Finally The evaluation process does not enable to gain some insight into each video processing algorithm. In particular, the evaluation process does not determine the necessary works remaining to be done on the algorithm (which improvement is the most crucial) to achieve satisfactory performance given some environment conditions.

3. ETISEO, a performance evaluation program

ETISEO, one of the latest evaluation programs, has tried to address these issues. One of the main objectives of ETISEO is to "acquire precise knowledge of vision algorithms". In other words, ETISEO tries to underline the "dependencies between algorithms and their conditions of use". At the end of the project "strengths and weaknesses of algorithms as well as unsolved problems should be highlighted" [4].

ETISEO tries to address each video processing problem separately, by defining accurately the problem. For instance, we should handle shadows within at least three different problems: (1) shadows at different intensity levels (i.e. weakly or strongly contrasted shadows) with uniform non color background, (2) shadows at the same intensity level with different types of background images in terms of color and texture and (3) shadows with different illumination sources in terms of source position and wavelengths.

Firstly, for each problem, it collects the video sequences illustrating only the current problem. The video sequences should illustrate the problem at different difficulty levels. For instance, for the problem of shadows and intensity levels, we should select video sequences containing shadows at different intensity levels (more or less contrasted). On these selected sequences, the the appropriate part of the ground truth is filtered and extracted to isolate video processing problems. For instance, for the detection task, we can evaluate the algorithm performance relatively to the problem of handling occluded objects by considering only the ground truth related to the occluded objects.

Secondly, for a given task (object detection, tracking, object classification and event recognition) ETISEO defines a sufficient number of metrics to measure and characterize the algorithm performance on various aspects. For instance, in ETISEO there are 7 metrics for the task of object detection.

Thirdly, ETISEO computes the reference data which corresponds to the expected output of the algorithm to be evaluated relatively to a given video processing task. The reference data are computed from the ground truth provided by human operators and can be improved to better correspond to the expected results. For instance, instead of evaluating the mobile object positions from the ground truth (2D-points), we can use 3D-point reference data to measure the computation of 3D object position.

Finally, ETISEO provides a unique automatic evaluation tool to accurately analyze how a given algorithm address a given problem.

In ETISEO, for each video sequence, there are three types of associate data. The first one is the ground truth (e.g. object bounding box, object class, event etc.) given by human operators at each level of the four video processing tasks. The second one is the general annotation on the video sequences concerning video processing problems (e.g. weak shadows) or concerning recording conditions (e.g. weather conditions such as sunny day). The final information is the camera calibration and contextual information about the empty scene describing the topology of the scene (e.g. zone of interest)

Table 1. ETI-VS1-BE-18-C4: ETISEO evaluation on object detection task

Participant	8	1	11	13	22	12
Precision	0.69	0.79	0.49	0.39	0.30	0.98
Sensitivity	0.71	0.42	0.48	0.42	0.40	0.16
F-Score	0.7	0.55	0.48	0.41	0.34	0.27

Table 2. ETI-VS1-BE-18-C4: ETISEO evaluation on tracking task

Participant	11	1	13	8	12
Tracking	0.47	0.45	0.42	0.4	0.19

All the video sequences of ETISEO (about 40 sequences) are selected and classified according to the problems they illustrate. These sequences have been processed by 16 international teams participating to the evaluation program in two phases. This work reports on the first phase of the evaluation results.

ETISEO's first phase also has faced two main limitations. Because the participants do the experiment themselves, they often have different assumptions. For instance, several participants do not detect the objects that do not move for a certain period of time. As a consequence, for some sequences, the algorithm results cannot be compared with each other. Table 1 and 2 show the evaluation results of the object detection and tracking tasks on the ETI-VS1-BE-18-C4 sequence. In the ETISEO point of view, we can observe that there is no coherence between these two tasks, one algorithm having good performance on object detection can perform poorly in the tracking task. However, these results are quite different from those of the proposed evaluation described in section 5.

Another limitation is that ETISEO does not define quantitative methods to measure the difficulty level of the videos illustrating a given video processing problem. For instance, ETISEO uses the terms "normal" or "dark" to describe the intensity levels of video sequences. Therefore, the selection of video sequences in ETISEO according to their difficulty levels is subjective and not precise enough. Furthermore, this subjective judgment also makes arbitrary the comparison between the new sequences with the tested ones.

Working in the ETISEO program, we have extended its methodology to propose a new approach of evaluation.

4. Proposed evaluation

Similar to ETISEO, we address each video processing problem separately. The steps of our methodology are as follow:

- Define a measure to compute the difficulty levels of the input data (e.g. video sequence) relatively to the current video processing problem, for instance weakly contrasted objects for the object detection task.
- Define metrics to evaluate the algorithm performance.
- Select video sequences illustrating the problem and the associate reference data to perform the evaluation.
- Evaluate algorithm performance on these sequences using the defined metrics.
- For each algorithm, determine the highest difficulty level where the algorithm can still achieve an acceptable performance. This value is defined as the algorithm capacity for addressing the current problem.

With this new approach, we still cannot predict the performance of an algorithm for a new sequence. We can only estimate the upper bound of the algorithm performance relatively to a specific video processing problem. The reason is that the algorithm performance on a new sequence also depends on other problems like small object size, illumination change etc. Thus the main objective of this methodology is to state that the performance may be unsatisfactory in case where the difficulty level relatively to one particular problem is greater than the algorithm capacity. In other words, for a given algorithm, we cannot determine its sufficient conditions of success but we can estimate the necessary ones.

To implement this approach, we need six elements: the algorithm output results of several participants, the video processing problem, the performance evaluation metrics, the input data measure, the reference data and the selected video sequences illustrating the problem. Concerning the video processing algorithms to be evaluated, it is important to define the parameters requiring a training state during the configuration to process the selected videos. In this paper, we consider that the algorithm developers were able to tune their algorithms and to provide results which are satisfactory and representative enough. Therefore, to apply this methodology, two main elements remain: the performance evaluation metric and the input data measure. In the following section, to illustrate the evaluation methodology, we describe the implementation of our methodology for two video processing problems: handling weakly contrasted objects and handling objects mixed with shadows.

4.1. Handling weakly contrasted objects

Problem description: Usually, the performance of video processing algorithms is proportional to the contrast level between mobile objects to be detected and the

background image. The lower the contrast of the object, the worse the performance of video processing algorithms. Therefore we would like to determine the contrast level where a given algorithm can still obtain an acceptable performance.

Measure of a blob contrast level: Because most of the video understanding algorithms are dedicated to the recognition of human activities, in our experiment, we have chosen blobs corresponding to persons as testing data. Most of the time, a person blob (the image region corresponding to a person), in terms of contrast level, is composed of three main regions distributed horizontally: head, body (covered by shirt, jacket, etc.) and legs. Therefore if we divide the blob horizontally into small strips, we hope that the contrast level inside one strip is homogeneous. Based on this idea, we propose the following procedure to determine the blob contrast:

- Divide the blob into a set of horizontal strips
- Calculate the contrast level of each strip
- The set of the contrast levels of all strips constitutes the contrast level of the blob (by removing all duplicate values)

Measure of a strip contrast level We apply the following procedure to compute the strip contrast:

- Divide the strip vertically into sub-regions.
- For each sub-region, compute the average contrast level of all pixels inside the sub-region.
- The contrast level of the strip is the maximum contrast level of all sub-regions inside the strip.

Measure of a pixel contrast level: Given both the current image which contains the mobile objects to be detected and the background image of the scene, the contrast of one object pixel is calculated using the following formula:

$$\frac{|R_b - R_f| + |G_b - G_f| + |B_b - B_f|}{255 \times 3}$$

In which:

(R_f, G_f, B_f) is the color value of one object pixel in the RGB color space.

(R_b, G_b, B_b) is the color value of the corresponding background pixel.

Measure of the algorithm capacity for dealing with weakly contrasted objects: The video interpretation system consists of several consecutive tasks (e.g. object detection, object classification, object tracking etc). The effect of weak contrast problem propagates from the lowest level task (the object detection) to the higher ones. Therefore we can evaluate the system capacity with respect to this problem at many points. However, the performance at one particular task does not necessarily reflect the performance of the whole system. For instance, on one sequence, a system may achieve good result in object detection but have difficulties in tracking objects. Therefore we would like to evaluate the system capacity in handling weakly contrasted objects at every possible tasks. To perform the evaluation, the best is to select the video sequences which do not contain other problems (at high levels) such as object crossing or occlusion. Unfortunately ETISEO videos contain often more than one problems. To isolate weakly contrasted objects, we have to extract one or several clips from a sequence. Because in the first phase of ETISEO program, too few participants have submitted output results on object classification and event recognition, we have chosen to evaluate the system performance for only two tasks: object detection and object tracking.

For the object detection task, the system capacity is computed using the detection error rate at each object contrast level. To determine this value, we have changed the evaluation space. Instead of evaluating algorithms using objects (i.e. blobs) having several contrast levels, we consider homogeneous regions with only one contrast level. To transform blobs into homogeneous regions, we assume that in each blob, the regions having the same contrast level are homogeneous. Therefore, if the set of contrast levels of a blob is composed of m levels, this blob will lead to m homogeneous regions in the new evaluation space.

With this transformation, from a set of n blobs we obtain m homogeneous regions representative of different contrast levels. At a given contrast level, if the total number of the regions is a and the number of the regions that an algorithm can detect is x, then the error rate (i.e. misdetected regions or false negative rate) of this algorithm on the testing data at the current contrast level is $1 - \frac{x}{a}$.

Then, we define the capacity of an object detection algorithm for dealing with weakly contrasted objects as the lowest contrast level at which the error rate of this algorithm is smaller than a certain threshold. One may claim that considering only the error rate (false negative) can only lead to a partial evaluation. However, we suppose that the algorithm has been tuned to handle all types of problems in the video sequences and that algorithm will be evaluated considering all these aspects. This evaluation is only used to assess the sensitivity of the algorithm on one specific problem.

For the object tracking task, the system can track an object if and only if in most of the frames, the system detects this object correctly. It means that all the object regions at various contrast levels should be detected. Therefore, for the weakly contrasted objects, we define the difficulty level of a sequence for the object tracking task as the minimum contrast level of the mobile objects in this sequence. Then the performance of the tracker at this difficulty level can be measured using classical metrics, for instance, the metric defined in ETISEO program (described in the section 5). The capacity of a tracker for dealing with weakly contrasted objects is defined as the lowest contrast level of the sequence with which the performance of the tracker is higher than a certain threshold.

With this definition, it is difficult to collect appropriate

Table 3. Analysis of contrast levels

Number of regions	Contrast level						
Sequence	0	1	2	3	4	5	6
BE-18-C1	0	19	46	29	13	5	5
BE-18-C4	0	15	16	6	11	12	14
MO-7-C1	0	3	30	30	30	30	1

Table 4. Error rate on BE-18-C1

Error rate	Contrast level						
BE-18-C1							
Participant	0	1	2	3	4	5	6
1	0	0	0	0	0	0	0
22	0	0.21	0.11	0.03	0.15	0.2	0.2
12	0	0.79	0.35	0.1	0.08	0	0.4
13	0	0	0.63	0.1	0.08	0	0.2
8	0	0.89	0.41	0.17	0.15	0.20	0.20
11	0	0.95	0.89	0.76	0.54	0.8	1
BE-18-C4							
1	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0
12	0	0.2	0.06	0.08	0.15	0.08	0
15	0	0.5	0.19	0.17	0.23	0.17	0.07
8	0	0.8	0.38	0.17	0.15	0.25	0.33
13	0	0.9	0.63	0.58	0.08	0.17	0.33
MO-7-C1							
11	0	0	0	0	0	0	0
1	0	0	0	0.2	0.1	0	0
9	0	0	0.13	0.13	0.07	0	0
8	0	0.67	0.07	0.07	0.7	0.07	0

data because the selected sequences should contain enough mobile objects with minimum contrast levels varying between 0 and 20. However, with the existing data in ETISEO program, in some cases, we can still deduce the tracking performance of an algorithm on a new sequence based on the evaluation results performed on a previous sequence. For instance, if an algorithm fails to track the objects in one particular testing sequence, this algorithm is likely to fail with more difficult sequences (i.e. with higher difficulty levels).

4.2. Handling objects mixed with shadows

Problem of handling shadows: When an object appears in the scene containing a strong light source like the sun or a lamp, the object is often detected together with its shadow. Algorithms often have difficulties in distinguishing the mobile object from its shadow because the contrast between shadows and the background is quite high. Therefore, often full or parts of the shadow is mixed with the object. Because shadow detection algorithms use the chromatisity and the texture of the background and objects to detect shadows, in this paper, we propose to assess the algorithm performance against the shadows under natural light at different intensity levels (more or less contrasted) in challenging situations with uniform non color background.

Measure of a shadow contrast levels: Unlike person blob, the shadow contrast levels change according to the direction of light source. Therefore, we should compute the shadow contrast using both vertical and horizontal strips. For example, if we divide one shadow into vertical strips, the set of contrast levels of these strips can be $\{5, 7, 6, 4\}$. If we divide that shadow into horizontal strips, the set of contrast levels of these strips can be $\{2, 7, 6, 3\}$. Then, the contrast of the shadow is defined as the union of these two sets: $\{2, 3, 5, 7, 6, 4\}$.

Capacity of the algorithm for dealing with shadows: Usually, shadow detection algorithms [11], often construct a shadow model and apply machine learning methods to determine the model parameters which is appropriate for a specific scene. Depending on the type of algorithms, two situations can be challenging: a large range between the maximum and the minimum contrast levels of the shadows or strongly contrasted shadows. We focus on the first challenge because this situation is relevant to a larger variety of algorithms. Thus, for each algorithm, we would like to know the maximum range of the contrast levels that this algorithm can handle.

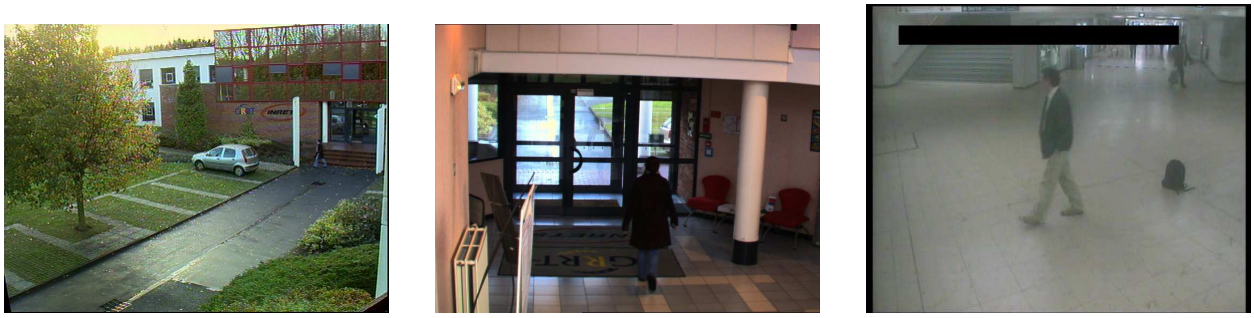
For handling weakly contrasted objects, we have tried to evaluate the whole system at every possible task. However the problem of objects mixed with shadows has small impact on the object tracking performance, except in case where objects are crossing each others. As we want to isolate video processing problems, we propose to evaluate the effect of the shadow contrast only at the object detection task.

5. Experimental results

This section describes the experiments we have realized to evaluate algorithm sensitivity on handling problems involving weakly contrasted objects and shadows.

5.1. Handling weakly contrasted objects

Video sequence selection: The chosen sequences should contain mobile objects (corresponding to isolated individuals) at different contrast levels. In addition, there should be no strong shadow and no illumination change so that the evaluation process is not influenced by other video processing problems. On the other hand, the selected sequences should not contain other problems such as occlusion or small object size. Finally, these selected sequences



(a) BE-18-C1 (b) BE-18-C4 (c) MO-7-C1

Figure 1. Sample images from the testing sequences

should be processed by a sufficient number of algorithms within the ETISEO project.

According to these criteria, we have chosen three clips in three video sequences. The first clips starts at frame 336 and ends at frame 404 from BE-18-C1 sequence (ETI-VS1-BE-18-C1 sequence in ETISEO). The second period starts at frame 90, ends at frame 105 from BE-18-C4 sequence (ETI-VS1-BE-18-C4 sequence in ETISEO). The final period starts at frame 5321, ends at frame 5350 from MO-7-C1 sequence (ETI-VS1-MO-7-C1 sequence in ETISEO). In the first clip, the size of the mobile object is quite small in comparison with those of the last two sequences. We have chosen this clip because we want to evaluate the algorithm performance at low contrast levels.

For simplicity, from now on, we will use the sequence name to refer to the selected clip in that sequence.

Evaluation results: In our experiment, there are 20 levels of contrast. The maximum contrast level (19) corresponds to the contrast between a completely black pixel (RGB(0,0,0)) and a completely white one (RGB(255,255,255)). Beside that, the height of each strip and the length of each block inside a strip are 10 pixels.

In ETISEO project, to ensure the fairness of evaluation, the algorithm performance of each participant is assigned with an anonymous number and we do not know which number belongs to which participant. Hence, in this section we will use these numbers to identify the participant algorithms.

The numbers of the participants having processed the BE-18-C1, the BE-18-C4 and the MO-7-C1 sequences are 6, 7 and 4 respectively.

Table 3 shows the number of regions at different contrast levels for three sequences. In these tables, the columns show the number of homogeneous regions at a given contrast level and the rows show the distributions of homogeneous regions of the testing sequences. At some contrast levels there are too few regions to get a reliable evaluation results. Therefore we will ignore these levels when evaluating the algorithm performance. From these table, we notice that there are more regions at lower contrast levels in the BE-18-C1 sequence than the others. By applying our methodology, we can observe that the selection of sequences illustrating the weak contrast problem becomes easier because we can obtain a quantitative description of the selected sequences. For the object detection task, table 4 illustrates the evaluation results on three sequences. In this table, the columns correspond to the contrast levels of the object regions that the algorithms have to detect. The rows correspond to the performance (i.e. error rate) of each algorithm for the different contrast levels. From the results we can highlight the general trend that the error rate is high (close to 1) at low contrast levels and it reduces gradually down to zero at the higher ones.

Table 5. Detection capacity

Capacity	Participant							
Sequence	1	8	9	11	12	13	15	22
BE-18-C1	0	2	-	5	2	3	-	0
BE-18-C4	0	2	-	0	0	4	2	0
MO-7-C1	0	2	0	0	-	-	-	-

For the sequence BE-18-C1, the participant 1 can handle very well the problem of low contrasted objects. The behavior of the algorithm of participant 13 is not coherent with the general trend. It can recognize object at low contrast level but its performance is poor at the high levels. A deep analysis shows that, this algorithm detects poorly small regions even though they are at high contrast levels. For the BE-18-C4 and MO-7-C1 sequences, the algorithms having good results on the previous sequence still maintain good performance. Nevertheless, the algorithm 11 does not have any error on these two sequences. By looking in details at the results of this algorithm, we realize that the size of the mobile objects in the first sequence are too small for algorithm 11.

Table 6. Tracking performance

Tracking	Participant							
Sequence	1	8	9	11	12	13	15	22
BE-18-C1	1	0.59	-	0.08	0.69	0.71	-	0.84
BE-18-C4	1	0.44	-	1	1	0.69	0.5	-
MO-7-C1	1	0.9	0.97	1	-	-	-	-

The object detection capacity of tested algorithms for each sequence is described in table 5. In our experiment, we take the threshold of error rate equal to 0.5 to compute the algorithm capacity for handling the weakly contrasted object problem. This capacity corresponds to a contrast level and means that the algorithm cannot handle 50% of regions at this contrast level. From these three tables, we can observe that the capacity of each algorithm does not change across the last two video sequences. This is an important result showing that the evaluation results are the same for different videos at similar difficulty levels. The difference between BE-18-C1 and the last two sequences, especially for the participant 11, is mainly due to the small size of the person in the first sequence. Therefore, for new sequences, depending on the size of mobile objects, we could consider the evaluation result on the first or on the last two sequences as the upper bound of the algorithm capacity.

For the object tracking task, in our experiment, we use a metric defined in ETISEO program. This metric measures the percentage of time an object in the reference data (RD) has been observed and tracked (C) with a consistent ID over the tracking period. The mobile object is considered to be observed if the distance between the bounding boxes of reference data and the algorithm (computed using the Dice coefficient: $(2 \times card(RD \bigcap C))/(card(RD) + card(C))$) is smaller than 0.7. The formula of this metric is as follow:

$$T_{Tracked} = \frac{1}{NB_{RefData}} \sum_{RefData} \frac{card(RD \cap C)}{card(RD)}$$

Where $card$ corresponds to the number of elements in a set. Table 6 shows the object tracking evaluation results on the three sequences. The algorithm ranking is nearly the same as in the experiment of object detection task except the ranking of participant 13 in the ETI-VS1-BE-18-C4 sequence. For participant 13, even though the object detection module achieves the worst performance, the tracking output of this participant is of better quality than that of participants 8 and 15. It means that the tracking algorithm of participant 13 is more robust to object detection failures.

From this experiment we conclude that the evaluation results at a specific task does not always reflect the performance of the whole system in dealing in particular with weakly contrasted mobile objects and in general with video processing problems.

Table 7. Error rate of shadow detection algorithms

	Shadow contrast level			
Participant	8	9	10	11
19	0	0.03	0	0
8	0.05	0.14	0.21	0.21
11	0.52	0.46	0.35	0.43
13	1	0.89	0.73	1
12	1	0.99	0.98	1

For object tracking, we define the algorithm capacity as the threshold of tracking performance equal to 1. Then only the algorithm of participant 1 can handle the mobile objects in the ETI-VS1-BE-18-C1. As the lowest contrast level of the tracked object in this sequence is 1, we state that in case of sequences containing small mobile objects, the algorithm of the participant 1 can track the objects with the lowest contrast level at least equal to 1.

For the sequences with big object size, the algorithms of participants 1, 11, 12 can track the mobile object in the ETI-VS1-BE-18-C4 sequence. Because the lowest contrast level of the tracked object in this sequence is 1, we state that these algorithms can track the objects with the lowest contrast level at least equal to 1. Thus as on the ETI-VS1-MO-7-C1, the tracked objects have the lowest contrast level equal to 2, we can verify on table 6 that the algorithms 1 and 11 achieve good tracking performance. Therefore the evaluation results obtained on the ETI-VS-BE-18-C4 sequence can be extrapolated to the other sequence.

5.2. Handling objects mixed with shadows

Video selection: As described in the ETISEO section, there are many types of shadows. In this section we propose to test the algorithm performance against the shadows at different intensity levels (more or less contrasted) with uniform non color background. Hence, in the dataset of ETISEO program, we have selected 74 shadow regions in the ETI-VS1-RD-16-C4 sequence. This sequence has been processed by the algorithms of 5 participants.

Evaluation results: To compute the contrast levels of the shadow regions we have taken the same parameters used in the previous experiment: there are 20 levels of contrast, the height of the strip is 10 pixel high and the size of the sub-regions inside a strip is 10x10 pixels.

In our experiment the shadows are strongly contrasted and all the contrast levels are within the range of 8 to 11.

The evaluation results are illustrated in table 7. From the

table we observe that the algorithms 12 and 13 do not have a mechanism to detect shadows. Therefore they consider nearly all shadows as mobile objects. For the remaining algorithms, the algorithms 19 have a perfect performance in handling shadows. The algorithm 8 still makes mistakes for strongly contrasted shadow regions (error rate: 0.21) and the algorithm 11 gets the worst performance among the three.

If we take the threshold of error rate equal to 0.5 as the capacity of handling shadows, we observe that algorithms 8 and 19 handle well the shadows with the contrast level in the range of 8 to 11. Therefore, the range of the shadow handling mechanism of these algorithms is at least more than 4 contrast levels. In the contrary, the error rate of the algorithm 11 at contrast level 8 is higher than the acceptable threshold (0.52). Hence, the range of contrast levels of the shadow regions that can be handled is 3 (from level 9 to level 11).

6. Conclusion

In this paper, we propose a new evaluation methodology that helps to generalize the evaluation results performed on selected videos to new video sequences. More precisely, we address each video processing problem separately and estimate the upper bound of algorithm capacity in solving the given problem. If this value is smaller than the difficulty level of new sequences, we can conclude that the algorithm cannot achieve acceptable performance on these sequences. To illustrate the new evaluation methodology, we present two metrics to address the problems of handling weakly contrasted objects and handling objects mixed with shadow. The preliminary results show that, with this methodology, we can extrapolate the evaluation results for new sequences.

There are three main limitations to the proposed evaluation methodology. First, this is a challenging task to select videos illustrating only one video processing problem and illustrating this problem at different difficulty levels. However once the videos have been selected they can be reused for any type of algorithms. Second the evaluation results can be **partially** extrapolated on new videos. This evaluation methodology only determines the upper bound of the algorithm capacity for solving one problem. Usually, as videos illustrate several video processing problems, the difference between the upper bound of the algorithm capacity and the real performance on videos containing more than one problem can be important. Third, for a given algorithm, the same set of parameters can be tuned to handle different problems. If two problems require two different ways of changing parameters, the difference between the upper bound and the real performance could be considerable. To limit this issue, the algorithms have been tuned on videos containing a mixture of problems and tested on sub-parts (of these videos) illustrating only one problem at a time.

In the future we plan to propose new evaluation metrics on more video processing problems and tasks to validate the generalizing power of this evaluation methodology. We are also planning to compute the dependencies between the parameter sets necessary for handling specific problems. Knowing these dependencies, we will be able to estimate the reliability of the computation of the algorithm capacity upper bound.

Acknowledgement: We would like to thank the ETISEO team (in particular SILOGIC) for providing us the ETISEO data and support to accomplish this work.

References

[1] http://homepages.inf.ed.ac.uk/rbf/caviar/. *CAVIAR: Context Aware Vision using Image-based Active Recognition*.

[2] http://www.clear-evaluation.org/. *CLEAR: Classification of Events, Activities and Relationships - Evaluation Campaign and Workshop*.

[3] http://www.ic-arda.org/infoexploit/vace/index.html. *VACE: Video Analysis and Content Extraction*.

[4] http://www.silogic.fr/etiseo. *ETISEO: Video understanding Evaluation*.

[5] http://www-dsp.elet.polimi.it/avss2005/creds.pdf. *CREDS: Call for Real-Time Event Detection Solutions (CREDS) for Enhanced Security and Safety in Public Transportation*, 2005.

[6] http://www.pets2006.net/. *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2006.

[7] J. Black, T. Ellis, and P.Rosin. A novel method for video tracking performance evaluation. *Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pages 125–132, 2003.

[8] B. GEORIS. Program supervision techques for easy configuration of video understanding systems. *PhD Thesis - Universit Catholique de Louvain*, 2006.

[9] C. J.Needham and R. Boyle. Performance evaluation metrics and statistics for positional tracker evaluation. *Computer Vision Systems Third International Conference, ICVS*, pages 278–289, 2003.

[10] J. Nascimento and J. Marques. Performance evaluation of object detection algorithms for video surveillance. *IEEE Transactions on Multimedia*, 8:761–774, 2006.

[11] A. Prati, I. Mikic, M. Trivedi, and R. Cucchiara. Comparative analysis of moving shadow detection algorithms. *Image and Vision Computing Journal (special issue on Visual Surveillance)*, 2003.