Performance Evaluation of Text Detection and Tracking in Video

Vasant Manohar[1], Padmanabhan Soundararajan[1], Matthew Boonstra[1], Harish Raju[2], Dmitry Goldgof[1], Rangachar Kasturi[1], and John Garofolo[3]

[1] University of South Florida, Tampa, FL
{vmanohar, psoundar, boonstra, goldgof, r1k}@cse.usf.edu
[2] Advanced Interfaces Inc., State College, PA
hraju@advancedinterfaces.com
[3] National Institute of Standards and Technology, Gaithersburg, MD
john.garofolo@nist.gov

Abstract. Text detection and tracking is an important step in a video content analysis system as it brings important semantic clues which is a vital supplemental source of index information. While there has been a significant amount of research done on video text detection and tracking, there are very few works on performance evaluation of such systems. Evaluations of this nature have not been attempted because of the extensive effort required to establish a reliable ground truth even for a moderate video dataset. However, such ventures are gaining importance now.

In this paper, we propose a generic method for evaluation of object detection and tracking systems in video domains where ground truth objects can be bounded by simple geometric shapes (polygons, ellipses). Two comprehensive measures, one each for detection and tracking, are proposed and substantiated to capture different aspects of the task in a single score. We choose text detection and tracking tasks to show the effectiveness of our evaluation framework. Results are presented from evaluations of existing algorithms using real world data and the metrics are shown to be effective in measuring the total accuracy of these detection and tracking algorithms.

1 Introduction

Text embedded in video frames often carries important information such as time, place, name, topics and other relevant information. These semantic cues can be used in video indexing and video content understanding. To extract textual information from video, which is often referred to as *Video Optical Character Recognition*, the first essential step is to detect and track the text region in the video sequence. There have been several published efforts addressing the problem of text area detection in video [1]. Performance metrics assume significance in the presence of such numerous systems with high claims on accuracy and robustness.

Empirical evaluation is highly challenging, due to the fundamental difficulty in establishing a valid "ground truth" or "gold standard". This is the process of establishing the "ideal output" for what *exactly* the algorithm is expected to generate. The secondary challenge with quantitative validation is assessing the relative

H. Bunke and A.L. Spitz (Eds.): DAS 2006, LNCS 3872, pp. 576–587, 2006.

importance of different types of errors. In this work, we adopt a comprehensive evaluation framework that carefully examines and finds solutions to each of these factors.

Earlier works on empirical evaluation of object detection and tracking [2–8], either present a single measure that concentrates on a particular aspect of the task or a suite of measures that look at different aspects. While the former approach cannot capture the performance of the system in its entirety, the latter results in a multitude of scores which makes it difficult to make a relative comparison between any two systems.

Similarly, while evaluating tracking systems, earlier approaches either concentrate on the spatial aspect of the task, i.e., assess correctness in terms of number of trackers and locations in frames [5, 7] or the temporal aspect which emphasizes on maintaining a consistent identity over long periods of time [3]. In the very recent works of [4, 8], a spatio-temporal approach towards the evaluation of tracking systems is adopted. However, these approaches do not provide the flexibility to adapt the relative importance of each of these individual aspects. Finally, majority of these undertakings make little effort in actually comparing the performance of existing algorithms on real world applications using their proposed measures.

In this paper, we apply two comprehensive measures that we have used in our evaluations of computer vision algorithms for face detection and tracking in boardroom meetings videos [9]. While the detection measure looks just at the spatial aspect, we approach with a spatio-temporal concept for the tracking measure. By adopting a thresholded approach to evaluation (See Secs 3.1 and 3.2), the relative significance of the individual aspects of the task can be modified. In the end, text detection and tracking is picked as a prototype task for evaluation and select algorithm performances are evaluated on a reasonable corpus.

The remainder of the paper is organized in the following manner. Section 2 briefs the ground truth annotation process which as explained earlier is a vital part of any evaluation. Section 3 briefs the detection and tracking metrics deployed in this evaluation. Section 4 explains the one-to-one mapping which is integral to this evaluation. Section 5.1 details the experimental results describing the behavior of the measures for different types of detection and tracking errors. Section 5.2 discusses and evaluates the results of three text detection algorithms and a tracking algorithm on a data set containing video clips from broadcast news segments. We conclude and summarize the findings in Section 6.

2 Ground Truth Annotations

Having a consistent and reliable ground truth is imperative to carrying out a scientific evaluation. There are many different ways to create a reference annotation for evaluation purposes. Domain characteristics such as spatial resolution of objects, temporal persistence of objects, object movement in the video and scene transitions decide the way annotation is carried out so that the marking is consistent and reliable. Also, in selecting a specific annotation approach, one has to

keep in mind that object detection and tracking is essentially a lower level task in a video understanding system, the output of which is passed on to a higher level system which extracts semantic meaning from it. For instance, the result of a text detection and tracking system is often used by an *OCR* system that pulls out textual information producing indexible keywords. Thus, the objective of performance evaluation is to identify a system that best generates an output which can be fed to any reasonable OCR system to obtain satisfying recognition results. To achieve this, the reference annotations should be such an ideal output.

In this paper, the method used for ground truthing is one in which text regions are bounded by a rectangular box with features of the region used as guides for marking the limits of the edges. If the features are occluded, which is often the case, the markings are approximated. Unique IDs are assigned to individual text objects and are consistently maintained over subsequent frames.

There are many free and commercially available tools which can be used for ground truthing videos such as Anvil, VideoAnnex, ViPER [10] and many others. In our case, we used ViPER [1] (Video Performance Evaluation Resource), a ground truth authoring tool developed by the University of Maryland.

Fig 1 shows a sample annotation using ViPER for text in a broadcast news segment.

Fig. 1. Sample annotation of text in broadcast news using rectangular boxes. Textual features such as readability, size and font are used as guides for marking the edges of the box. Internally, a unique Object ID is maintained for each of the text objects shown which helps in measuring the performance of tracking. Courtesy: CNN News.

2.1 Annotation Guidelines

In order to reduce the intra-annotator variability (the same annotator marking the boxes inconsistently at different times) and inter-annotator variability (mismatch between different annotators), a clear and exhaustive set of guidelines

[1] http://viper-toolkit.sourceforge.net

is established. These were strictly and diligently adhered to while creating the reference annotations. Further, considerable effort was directed in developing a ground truth that is rich with details. Hence, each text block is associated with a set of attributes which characterizes the region both from an evaluational and informational point of view. This section explains the set of guidelines and additional flags used in this evaluation for the text annotation task.

Every new text area is marked with a box when it appears in the video. Moving and scaling the selection box tracks the text as it moves in succeeding frames. This process is done at the line level (with offsets specified for word boundaries) until the text disappears from the frame.

There are two types of text:

- Graphic text is anything overlaid onto the picture. Example, the "abc" logo in Fig 1.
- Scene text is anything in the background/foreground of what is actually being filmed. Example, all text regions on the newspaper in Fig 1.

Text readability consists of three levels. Completely unreadable text is signified by READABILITY = 0 (green boxes in Fig 1) and is defined as text in which no character is identifiable. Partially readable text is given READABILITY = 1 (blue boxes in Fig 1) and contains characters that are both identifiable and non-identifiable. Clearly readable text is assigned READABILITY = 2 (red boxes in Fig 1) and is used for text in which all letters are identifiable.

The OCCLUSION attribute is set to TRUE when the text is cut off by the bounds of the frame or by another object. The LOGO attribute is set to TRUE when the text region being marked is a company logo imprinted in stylish fonts. Example, the texts "The Washington Post" and "abc" in Fig 1.

Of all the objects of interest in video, text is particularly difficult to be uniformly bound. For this reason, text regions are marked meticulously based on a comprehensive set of rules, namely,

- All text within a selected block must contain the same readability level and type.
- Blocks of text must contain the same size and font. Two allowances are given to this rule. A different font or size may be included in the case of a unique single character and the font color may vary among text in a group.
- The bounding box should be tight to the extent that there is no space between the box and text. The maximum distance from the box to the edge of bounded text may not exceed half the height of the characters when Readability = 2 (clearly readable). When Readability = 0 or 1 the box should be kept tight but does not require separate blocks for partial lines in a paragraph.
- Text boxes may not overlap other text boxes unless the characters themselves are specifically transposed atop one another.

The additional set of attributes described above is used in deciding whether a particular text region should be evaluated. The specific settings for evaluating a text region used in this evaluation are - TEXT-TYPE = Graphic, READABILITY = 2, OCCLUSION = FALSE and LOGO = FALSE.

All other regions are treated as "Don't Care" where the system output is neither penalized for missing nor given credit for detecting. It has to be noted that each of these attributes can be selectively specified to be included in evaluation through the scoring tool that we have developed.

2.2 Annotation Quality

It has been well appreciated in the research community that when manual labeling is involved, it is important to evaluate the consistency of labeling empirically. This becomes extremely critical when the marking involves subjective concepts like object bounds and readability. For this reason, 10% of the entire corpus was doubly annotated and checked for quality using the evaluation measures. Using the thresholded approach described in Secs 3.1 and 3.2, we found that at 60% spatial threshold, the average SFDA and the average ATA scores for the doubly annotated corpus were 0.97 and 0.90 respectively. This process assures that the reference annotations are reliable which is essential for genuine evaluations.

The threshold for a given application is derived from spatial disagreements between the annotators in the 10% double annotated data. The motivation behind this is to eliminate the error in the scores induced due to ground truth inconsistencies in terms of spatial alignment. Also, such an approach of arriving at the spatial threshold reflects the difficulties in how humans perceive the task. It has to be noted that though we get a good performance at 60% spatial threshold on the double annotations, we run the actual evaluations at a threshold of 10%. By adopting this method, systems are less penalized for spatial alignment errors.

3 Performance Measures

The performance measures that were used in the evaluation were proposed and discussed in detail in [9]. The performance measures are based primarily on area calculations of the spatial overlap between the ground truth objects and the system output. To generate the best score for an algorithm's performance, a one-to-one mapping is performed between the ground truth and system output objects such that the metric scores are maximized. All of the measure scores are normalized to a scale from 0, the worst performance, to 1, the best performance.

Secs 3.1 and 3.2 discuss the frame based detection measure and the sequence based tracking measure respectively, while Sec 4 briefs the one-to-one matching strategy.

The following are the notations used in the remainder of the paper,

- G_i denotes the i^{th} ground truth object and $G_i^{(t)}$ denotes the i^{th} ground truth object in t^{th} frame.
- D_i denotes the i^{th} detected object and $D_i^{(t)}$ denotes the i^{th} detected object in t^{th} frame.
- $N_G^{(t)}$ and $N_D^{(t)}$ denote the number of ground truth objects and the number of detected objects in frame t respectively.

- N_G and N_D denote the number of unique ground truth objects and the number of unique detected objects in the given sequence respectively. Uniqueness is defined by object IDs.
- N_{frames} is the number of frames in the sequence.
- N^i_{frames}, depending on the context, is the number of frames the ground truth object (G_i) or the detected object (D_i) existed in the sequence.
- N_{mapped} is the number of mapped ground truth and detected objects in a frame or the whole sequence depending on the context (detection/tracking).

3.1 Detection – Frame Based Evaluation

The Sequence Frame Detection Accuracy **(SFDA)** is a frame-level measure that penalizes for fragmentations in the spatial dimension while accounting for number of objects detected, missed detects, false alarms and spatial alignment of system output and ground truth objects.

The frame-based detection measure **(FDA)** which was used is defined as,

$$FDA(t) = \frac{\text{Overlap\_Ratio}}{\left[\frac{N_G^{(t)}+N_D^{(t)}}{2}\right]} \tag{1}$$

$$\text{where, Overlap\_Ratio} = \sum_{i=1}^{N_{mapped}} \frac{|G_i^{(t)} \bigcap D_i^{(t)}|}{|G_i^{(t)} \bigcup D_i^{(t)}|} \tag{2}$$

Here, the N_{mapped} is the number of mapped objects, where the mapping is done between objects which have the best spatial overlap in the given frame t.

To calculate the Sequence Frame Detection Accuracy (SFDA), the FDA scores from each frame are summed together and normalized by the total number of frames which either has a ground truth or a detected object. This normalization accounts for both missed detections and false alarms. This formula can be expressed as,

$$SFDA = \frac{\sum_{t=1}^{t=N_{frames}} FDA(t)}{\sum_{t=1}^{t=N_{frames}} \exists(N_G^{(t)}\ OR\ N_D^{(t)})} \tag{3}$$

Relaxing Spatial Alignment. For many systems, it would be sufficient to just detect the presence of an object in a frame, and not be concerned with the spatial accuracy of detection. To evaluate such systems, we employed a thresholded approach to evaluation of detection. Here, the detected object is given full credit even when it overlaps just a portion of the ground truth. OLP_DET is the spatial overlap threshold.

$$\text{Overlap\_Ratio\_Thresholded} = \sum_{i=1}^{N_{mapped}} \frac{Ovlp\_Thres(G_i^{(t)}, D_i^{(t)})}{|G_i^{(t)} \cup D_i^{(t)}|} \tag{4}$$

where,

$$Ovlp\_Thres(G_i^{(t)}, D_i^{(t)}) = \begin{cases} |G_i^{(t)} \cup D_i^{(t)}|, & \text{if } \frac{|G_i^{(t)} \cap D_i^{(t)}|}{|G_i^{(t)}|} \geq OLP\_DET \\ |G_i^{(t)} \cap D_i^{(t)}|, & otherwise \end{cases}$$

3.2 Tracking – Sequence Based Evaluation

For the tracking evaluation, we use the Average Tracking Accuracy **(ATA)** measure. This measure is a spatio-temporal measure which penalizes fragmentation in both the temporal and spatial dimensions. It also accounts for number of objects detected and tracked, missed objects, and false alarms. A one-to-one mapping between the ground truth objects and the system output is determined by computing the measure over all combinations of system output objects and ground truth objects and using an optimization strategy to maximize the overall score for the sequence.

We first determine the Sequence Track Detection Accuracy **(STDA)** which, is the performance of tracking on all ground truth objects. The $STDA$ is calculated as,

$$STDA = \sum_{i=1}^{N_{mapped}} \frac{\sum_{t=1}^{N_{frames}} \left[\frac{|G_i^{(t)} \cap D_i^{(t)}|}{|G_i^{(t)} \cup D_i^{(t)}|} \right]}{N_{(G_i \cup D_i \neq \emptyset)}} \tag{5}$$

Finally, the Average Tracking Accuracy **(ATA)** is the STDA score normalized by the number of objects in the sequence. It is defined as,

$$ATA = \frac{STDA}{\left[\frac{N_G + N_D}{2} \right]} \tag{6}$$

In cases when it is desirable to measure the tracking aspect of the algorithm and not be concerned with the detection accuracy, we can relax the detection penalty by using an area thresholded approach similar to the technique described in Sec 3.1.

4 Matching Strategies

From Eqs 2 and 5, it is clear that both the detection and the tracking measures distinguish between individual objects at the frame and at the sequence level respectively. The maximal scoring is obtained for the *optimal* ground-truth and system output pairs. Potential strategies to solve this assignment problem are the weighted bi-partite graph matching [11] and the Hungarian algorithm [12].

	DT_1	DT_2	...	DT_M
GT_1	x			
GT_2				x
$\vdots$				
GT_N		x		

Assume that there are N ground truth objects and M detected objects. A brute force algorithm would have an exponential complexity, a result of having to try out all possible combination of matches (n!). However, this is a standard optimization problem and there are standard techniques to get the optimal match. The matching is generated with the constraint that the sum of the chosen function of the matched pairs is minimized or maximized as the case may be. In

usual assignment problems, the number of objects in both cases are equal, i.e, when $N = M$. However, this is not a requirement and unequal number of objects can also be matched.

There are many variations of the basic Hungarian strategy most of which exploit constraints from specific problem domains they deal with. The algorithm has a series of steps which is followed iteratively and has a polynomial time complexity, specifically some implementations have $O(N^3)$. Faster implementations have been known to exist and have the current best bound to be at $O(N^2 logN + NM)$ [13]. In our case, we take advantage of the fact that the matrix is mostly sparse by implementing a hash function for mapping sub-inputs from the whole set of inputs.

5 Results and Analysis

5.1 Analytical Observations

For an object detection and tracking task the errors that can affect the metric scores can be due to a single or a combination of the following errors, namely, spatial inaccuracy, temporal inaccuracy, missed detects and false alarms. In our earlier work, we presented a detailed analysis of the influence of missed detects and false alarms on the metric scores. In a text detection and tracking scenario, most likely the outputs will be used to drive the text recognition module to extract the transcriptions before deriving semantic information. To this extent, we focus on the effects of spatial and temporal inaccuracies as these are as important as the other errors. For the purpose of completeness, we also present the analytical equations that drive the metrics in the case of missed detects and false alarms. We have developed an evaluation tool which reports each of the above components as auxiliary measures. These can be used for debugging purposes by algorithm developers to identify strengths and weaknesses of an approach and also for determining the operating point for their algorithm.

Effect of Spatial Inaccuracy. Assume a ground truth, G_i, and a corresponding detected object, D_i, of the same size ($|G_i|$). Fig 2 shows the effect of percentage overlap of ground truth on the overlap ratio. The main motivation behind taking the ratio of the spatial intersection of the two bounding boxes with their spatial union instead of the ground truth object size is to penalize bigger detected objects with the same spatial overlap with the ground truth.

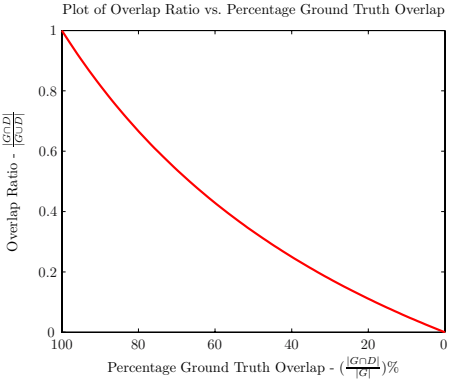

Fig. 2. Plot of overlap ratio vs. percentage ground truth overlap between a ground truth object and a detected object of the same size

For a ground truth, G_i and a detected object, D_i, that overlaps $x_i\%$ of G_i, we can derive "Overlap_Ratio" in Eq 2 for G_i as a function of x_i.

$$\text{Overlap\_Ratio(i)} = \frac{x_i}{1 + \frac{|D_i|}{|G_i|} - x_i} \tag{7}$$

Effect of Temporal Inaccuracy. There are two kinds of temporal inaccuracies that induce errors in the tracking task, namely,

- Incorrect object ID propagation in time. In this case, there can still be perfect detection.
- Missed object frames during tracking. In this case, it is treated as missed detects at the frame level by the detection measure.

Assuming that there are no false alarms and perfect spatial accuracy for the detected objects, we can analytically characterize the SFDA and the ATA measures for temporal inaccuracies as shown in Eqs 8 and 9.

$$\text{SFDA} = \frac{\sum_{i=1}^{N_D} N^i_{det\_frames}}{\frac{\sum_{i=1}^{N_D} N^i_{det\_frames} + \sum_{j=1}^{N_G} N^j_{frames}}{2}} \tag{8}$$

where, $N^i_{det\_frames}$ is the number of frames the output box D_i ideally detected the ground truth G_i irrespective of identification.

$$\text{ATA} = \frac{\sum_{i=1}^{N_{mapped}} \frac{N^i_{trk\_frames}}{N^i_{frames}}}{\left[\frac{N_G + N_D}{2}\right]} \tag{9}$$

where, $N^i_{trk\_frames}$ is the number of frames the output box D_i ideally detected and tracked (identified) the ground truth G_i.

Effect of Missed Detects. Given an ideal detection and tracking for the remaining objects in the sequence, we can characterize the SFDA and the ATA measures for missed detects as shown in Eqs 10 and 11.

$$\text{SFDA} = \frac{\sum_{i=1}^{N_D} N^i_{frames}}{\frac{\sum_{i=1}^{N_D} N^i_{frames} + \sum_{j=1}^{N_G} N^j_{frames}}{2}} \tag{10}$$

$$\text{ATA} = \frac{N_D}{\left[\frac{N_G + N_D}{2}\right]} \tag{11}$$

It can be seen that there will be a uniform degradation of the ATA score while the SFDA score will exhibit a non-uniform behavior. Clearly, the SFDA score is influenced by temporally predominant objects (existing in more frames) in the sequence, while the ATA score is independent of the frame persistence of objects.

Effect of False Alarms. Having looked at the effect of missed detects on the SFDA and the ATA, it is fairly straightforward to imagine the effect of false alarms on the measure scores. Given an ideal detection and tracking for all the objects in the sequence, we can analytically characterize the SFDA and the ATA measures for false alarms as shown in Eqs 12 and 13.

$$\text{SFDA} = \frac{\sum_{j=1}^{N_G} N_{frames}^{j}}{\frac{\sum_{i=1}^{N_D} N_{frames}^{i} + \sum_{j=1}^{N_G} N_{frames}^{j}}{2}} \tag{12}$$

$$\text{ATA} = \frac{N_G}{\left[\frac{N_G + N_D}{2}\right]} \tag{13}$$

Just as missing a predominantly occurring object decreases the SFDA score by a higher extent, introducing an object in a large number of frames affects the SFDA score more. However, the ATA score is affected by the number of unique objects (different object IDs) inserted into the sequence.

5.2 Text Detection and Tracking Evaluation

In this section, we describe the framework that we use in our evaluation of text detection and tracking algorithms. We evaluated three text detection algorithms and a text tracking algorithm using the measures discussed. The algorithm outputs were obtained from the original authors and thus can be safely assumed that the reported outputs are for the optimal parameter settings of the algorithm without any implementation errors. For anonymity purposes, these algorithms will be referred to as Algo 1, Algo 2 and Algo 3. The source video was in MPEG-2 standard in NTSC format encoded at 29.97 frames per second at 704x480 resolution (Aspect Ratio – 4:3).

The algorithms were trained on 50 clips, each averaging about 3 minutes (approx. 5400 frames) and tested on 30 clips, whose average length was the same as that of the training data. The ground truth was provided to algorithm developers for the 50 clips to facilitate training of algorithm parameters.

Fig 3 shows the SFDA scores of the three text detection algorithms on the 30 test clips. It also reports the SFDA scores thresholded at 10% spatial overlap, missed detects, and false alarms associated with each sequence. By adopting a thresholded approach, we alleviate the effect of errors caused due to spatial anomalies. Thus, the errors in the thresholded SFDA scores are primarily due to missed detects and false alarms. One can observe a strong correlation between the SFDA scores and the missed detects/false alarms. All three algorithms have reasonably low missed detection and false alarm rates with Algo 1 being the lowest in majority cases. As a result of thresholding, the average increase in scores for Algo 1, Algo 2 and Algo 3 is 29.56%, 63.54% and 24.55% respectively. This shows that Algo 1 and Algo 3 have good localization accuracy which is important if these outputs are to be used by a text recognition system.

Fig 4 shows the ATA scores of a text tracking system on the test set. Additionally, ATA scores thresholded at 10% spatial overlap, missed detects, and false alarms associated with each sequence are reported.

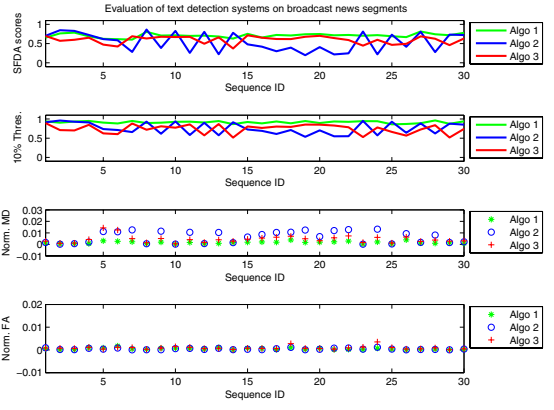

Fig. 3. Evaluation results of three text detection systems. Missed Detects (MD) and False Alarms (FA) are normalized with respect to total number of evaluation frames.

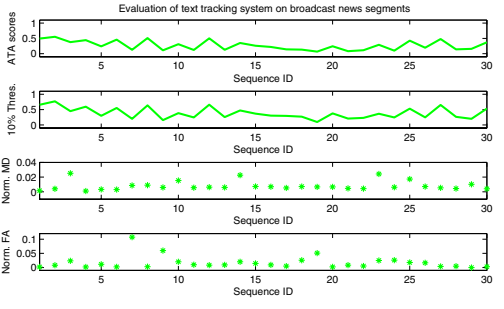

Fig. 4. Evaluation results of a text tracking algorithm. Missed Detects are normalized with respect to total number of unique ground truth objects in the sequence and False Alarms are normalized with respect to total number of unique system output objects in the sequence.

It can be observed from the results that the tracking algorithms are not as accurate as the detection algorithms. This is a direct result of inconsistent ID tracks. Fig 4 also shows that induction of *random* false alarms is detrimental to the performance of the tracking system. Through more analysis using the auxiliary measures discussed in Sec 5.1, we observed that these false alarms generally do not persist for more than a couple of frames. This gives an idea that trackers should perhaps look for evidence in the spatio-temporal space before declaring an object's presence.

6 Conclusions

A comprehensive approach to evaluation of object detection and tracking algorithms is presented for video domains where an object bounding approach to

ground truth annotation is followed. An area based metric, that depends on spatial overlap between ground truth objects and system output objects to generate the score, is used in the case of an object bounding annotation. For the detection task, the SFDA metric captures both the detection capabilities (number of objects detected) and the goodness of detection (spatial accuracy). Similarly, for the tracking task, both the tracking capabilities (number of objects detected and tracked) and the goodness of tracking (spatial and temporal accuracy) are taken into account by the ATA metric. Evaluation results of text detection and tracking systems on broadcast news segments show the effectiveness of the metrics in capturing their performance. Results show that the state-of-the-art is fairly mature in the detection of clear, readable text that is overlaid on video. It can also be seen that text tracking systems suffer from irregular identification and insertion of sporadic false alarms.

References

1. Jung, K., Kim, K.I., Jain, A.K.: Text information extraction in images and video: a survey. Pattern Recognition **37** (2004) 977–997
2. Antani, S., Crandall, D., Narasimhamurthy, A., Mariano, V.Y., Kasturi, R.: Evaluation of Methods for Detection and Localization of Text in Video. In: Proceedings in International Workshop on Document Analysis Systems. (2000) 507–514
3. Black, J., Ellis, T.J., Rosin, P.: A Novel Method for Video Tracking Performance Evaluation. In: Proceedings of IEEE PETS Workshop. (2003)
4. Brown, L.M., Senior, A.W., Tian, Y., Connell, J., Hampapur, A., Shu, C., Merkl, H., Lu, M.: Performance Evaluation of Surveillance Systems Under Varying Conditions. In: Proceedings of IEEE PETS Workshop. (2005)
5. Collins, R., Zhou, X., Teh, S.: An Open Source Tracking Testbed and Evaluation Web Site. In: Proceedings of IEEE PETS Workshop. (2005)
6. Hua, X., Wenyin, L., Zhang, H.: Automatic Performance Evaluation for Video Text Detection. In: Proc. International Conference on Document Analysis and Recognition. (2001) 545–550
7. Nascimento, J., Marques, J.: New Performance Evaluation Metrics for Object Detection Algorithms. In: Proceedings of IEEE PETS Workshop. (2004)
8. Smith, K., Gatica-Perez, D., Odobez, J., Ba, S.: Evaluating Multi-Object Tracking. In: Proceedings of IEEE Empirical Evaluation Methods in Computer Vision Workshop. (2005)
9. Manohar, V., Soundararajan, P., Raju, H., Goldgof, D., Kasturi, R., Garofolo, J.: Performance Evaluation of Object Detection and Tracking in Video. In: Proceedings of Asian Conference on Computer Vision. (2006) 151–161
10. Doermann, D., Mihalcik, D.: Tools and Techniques for Video Performance Evaluation. In: ICPR. Volume 4. (2000) 167–170
11. Papadimitriou, C.H., Steiglitz, K.: Combinatorial optimization: algorithms and complexity. Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1982)
12. Munkres, J.R.: Algorithms for the Assignment and Transportation Problems. J. SIAM **5** (1957) 32–38
13. Fredman, M.L., Tarjan, R.E.: Fibonacci Heaps and their uses in Improved Network Optimization Algorithms. Journal of ACM **34** (1987) 596–615