2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance

PETS2010 and PETS2009 Evaluation of Results using Individual Ground Truthed Single Views

A. Ellis and J. Ferryman
Computational Vision Group
School of Systems Engineering
University of Reading
Whiteknights, Reading, RG6 6AY UK
{a.l.ellis|j.m.ferryman}@reading.ac.uk

Abstract

This paper presents the results of the crowd image analysis challenge of the PETS2010 workshop. The evaluation was carried out using a selection of the metrics developed in the Video Analysis and Content Extraction (VACE) program and the CLassification of Events, Activities, and Relationships (CLEAR) consortium. The PETS 2010 evaluation was performed using new ground truthing created from each independant 2D view. In addition, the performance of the submissions to the PETS 2009 and Winter-PETS 2009 were evaluated and included in the results. The evaluation highlights the detection and tracking performance of the authors' systems in areas such as precision, accuracy and robustness.

1. Introduction

This paper discusses the objective evaluation of the submitted results by contributing authors of the PETS 2010, PETS 2009 and Winter-PETS 2009 workshops on the challenges defined on the PETS2009 crowd dataset [14] . The theme of the PETS 2010 workshop was multi-sensor event recognition in crowded public areas. As part of this workshop a challenge was set to evaluate an approach to one or more of people counting, density estimation, tracking, flow estimation and event recognition, and to report results based on annotated datasets made available on the workshop website [1]. In this paper the focus is tracking and people counting challenges due to the fact that the majority of the submitted evaluations and papers were dedicated to these tasks. In the remainder of this paper, the dataset and the ground truth annotation details are presented in Section 2. A brief description of the evaluation methodology follows in Section 3, and analytic discussion of the overall performances is provided in Section 4. Concluding remarks are given in Section 5.

2. Datasets and Ground Truth

2.1. Datasets

Three datasets were recorded for the workshop at Whiteknights Campus, University of Reading, UK. Further details of these datasets may be found in Ferryman and Shahrokni [14]. The datasets comprise of multi-sensor sequences containing crowd scenarios with increasing scene complexity. Dataset S1 concerns person count and density estimation. Dataset S2 addresses people tracking. Dataset S3 involves flow analysis and event recognition. In this paper the first two datasets are the focus.

2.2. Ground Truth

The ground truth for a subsampled set of frames was obtained for each sequence with the average sampling frequency being 1 frame in every 3 frames. The ground truth for people counting was generated by manually counting people in the specified regions, and those that crossed the entry and exit lines at each sampled frame. For the PETS 2010 workshop each of the seven independent 2D camera views (views 1,3,4,5,6,7,8) were ground truthed using the Video Performance Evaluation Resource (ViPER) ground truth tool [2]. This provided the necessary bounding boxes and their identifying key and location for the new evaluations. The original ground truth annotation tool for the tracking challenges presented in the PETS 2009 and Winter-PETS 2009, simultaneously defined bounding boxes in all views corresponding to a person, by locating its 3D position on a discrete grid. Errors in calibration due to the approximation of the ground surface as a plane, in addition to radial distortion, and the spatial resolution of the annotation grid defined on the ground plane, were an intrinsic part of this annotation, and were discussed in a previous workshop pa-

978-0-7695-4264-5/10 $26.00 © 2010 IEEE
DOI 10.1109/AVSS.2010.89

per [13].

3. Evaluation Methodology

The evaluation was based on the framework by Kasturi *et al.* [16], which is a well established protocol for performance evaluation of object detection and tracking in video sequences. These metrics are formally used by the Video Analysis and Content Extraction (VACE) program and the CLassification of Events, Activities, and Relationships (CLEAR) consortium. As part of the PETS 2010 workshop authors of the representative algorithms submitted their results in XML format using the PETS 2010 published XML Schema available at [1]. These results were evaluated using the following metrics:

Notation

- G_i^t denotes i^{th} ground-truth object in frame t; G_i denotes the i^{th} ground-truth object at the sequence level; N_{frames} is the number of frames in the sequence
- D_i^t denotes the i^{th} detected object in frame t; D_i denotes the ith detected object at the sequence level
- 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
- N_{frames}^i refers to the number of frames where either ground-truth object (G_i) or the detected object (D_i) existed in the sequence
- N_{mapped} refers to sequence level detected object and ground truth pairs, N_{mapped}^t refers to frame t mapped ground truth and detected object pairs
- m_t represents the missed detection count, (fp_t) is the false positive count, c_m and c_f represent respectively the cost functions for missed detects and false positives, and $c_s = log_{10}ID - SWITCHES_t$

3.1. Sequence Frame Detection Accuracy (SFDA)

SFDA uses the number of objects detected, the number of missed detections, the number of falsely identified objects, and the calculation of the spatial alignment between the algorithm's output for detected objects and that of the ground truthed objects. It is derived from a Frame Detection Accuracy (FDA) measure. The FDA is calculated using a ratio of the spatial intersection and union of an output object and mapped ground truthed objects

$$OverLapRatio = \sum_{i=1}^{N_{mapped}^t} \frac{|G_i^t \cap D_i^t|}{|G_i^t \cup D_i^t|} \quad (1)$$

$$FDA(t) = \frac{OverlapRatio}{\left[\frac{N_G^t + N_D^t}{2}\right]} \quad (2)$$

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

For this study although the annotation of the ground truth was challenging, as described in Section 2, an overlap threshold of 100 percent for the intersection over union scores, was used.

3.2. Average Tracking Accuracy (ATA)

ATA is obtained from the Sequence Track Detection Accuracy (STDA). The STDA is a measure of the tracking performance over all of the objects in the sequence and from this ATA is defined as the sequence track detection accuracy per object. The mapping between ground truth objects and detected objects is performed so as to maximise the measure score. This metric is implemented with a hash function due to the fact that the track correspondence matrix to be mapped is reasonably sparse.

$$STDA = \sum_{i=1}^{N_{mapped}} \frac{\sum_{i=1}^{N_{frames}} \left[\frac{|G_i^t \cap D_i^t|}{|G_i^t \cup D_i^t|}\right]}{N_{(G_i \cup D_i \neq 0)}} \quad (4)$$

$$ATA = \frac{STDA}{\left[\frac{N_G + N_D}{2}\right]} \quad (5)$$

For both detection and tracking metrics in the following descriptions the accuracy metrics provide a measure of the correctness of the detections or tracks. The precision metrics provide the measure of, in the instance where there has been a correct detection or track, how close to the ground truth that detection or track may be.

3.3. Multiple Object Detection Accuracy (MODA)

MODA is an accuracy measure that uses the number of missed detections and the number of falsely identified objects. Cost functions to allow weighting to either of these errors are included, however for the sake of both PETS 2009 evaluations they were equally set to 1.

$$MODA = 1 - \frac{c_m(m_t) + c_f(f_{p_t})}{N_G^t} \quad (6)$$

3.4. Multiple Object Detection Precision (MODP)

MODP gives the precision of the detection in a given frame. Again, with this metric, an overlap ratio is calculated as previously defined in (1), and, in addition to a count of the number of mapped objects, the MODP is defined as:

$$MODP(t) = \frac{OverLapRatio}{N^t_{mapped}} \quad (7)$$

3.5. Multiple Object Tracking Accuracy (MOTA)

MOTA uses the number of missed detections, the falsely identified objects, and the switches in an algorithm's output track for a given ground truth track. These switches are calculated from the number of identity mismatches in a frame, from the mapped objects in its preceding frame.

$$MOTA = 1 - \frac{\sum_{t=1}^{N_{frames}} (c_m(m_t) + c_f(f_{p_t}) + c_s)}{\sum_{t=1}^{N_{frames}} N^t_G} \quad (8)$$

3.6. Multiple Object Tracking Precision (MOTP)

MOTP is calculated from the spatio-temporal overlap between the ground truthed tracks and the algorithm's output tracks.

$$MOTP = \frac{\sum_{i=1}^{N_{mapped}} \sum_{t=1}^{N^t_{frames}} \left[\frac{|G^t_i \cap D^t_i|}{|G^t_i \cup D^t_i|} \right]}{\sum_{t=1}^{N_frames} N^t_{mapped}} \quad (9)$$

In addition to the evaluation of tracking, a simple comparison of the people count per region, against a ground truth count per region for the sampled frames, produced the average percentage error in counting per region, for each sequence.

4. Evaluation Results

An analysis of the overall performance of the submitted results from the benchmark datasets, using these illustrated metrics, is described in this section. The submitted results were diverse in terms of the sequences and views used and therefore it was not possible to draw general comparisons and conclusions about their performance. Nevertheless, the evaluations presented in this section can lead to a helpful insight about the effectiveness of different methodologies. Both the people counting and tracking challenges were considered.

4.1. People Counting

Figure 1 provides the evaluation of the counting people per region task. Note that the y axis on this graph represents the average error in number of people per frame, where the lower the value, the better the performance per frame. Table 1 gives the corresponding publication reference, for each label, for Figure 1.

Label	Author(s)
Chan	[9]
Sharma	[18]
Albiol	[4]
Choudri	[10]
Alahi	[3]
Conte	[12]
Pätzold	[17]

Table 1. Labels and publication references for Figure 1

Label	Author(s)
Breitenstein	[8]
Sharma	[18]
Yang	[19]
Arsic	[5]
Bolme_ASEF	[7]
Bolme_Cascade	[7]
Bolme_Parts	[7]
Alahi0greedy	[3]
Alahi0lasso	[3]
Ge	[15]
Conte	[11]
BerclazLp	[6]

Table 2. Labels and publication references for Figures 3, 4 and 5

A wide variety of methods have been proposed and tested in this category and from Figure 1 it can been seen that the majority of the methods and their variants have consistent and comparable performance. The algorithms proposed by Alboil *et al.* [4] and Conte *et al.* [12] performed robustly throughout each time sequence. Several methods such as Alahi *et al.* [3], Chan *et al.* [9] and Choudri *et al.* [10] also performed well on the more challenging sequence 14-17. Further details of the variant of each method can be found in their companion workshop paper.

4.2. Tracking

The most tested dataset of the two PETS workshops in 2010 remains as S2.L1, at time sequence 12.34, for the first camera view. Figure 3 shows how the individual algorithms performed according to various VACE and CLEAR metrics on a single representative camera view. Table 2 gives the corresponding publication reference, for each label, for Figures 3, 4 and 5.

Note that in the case of these metrics, higher values indicate better performance. It is clear that for this sequence, us-

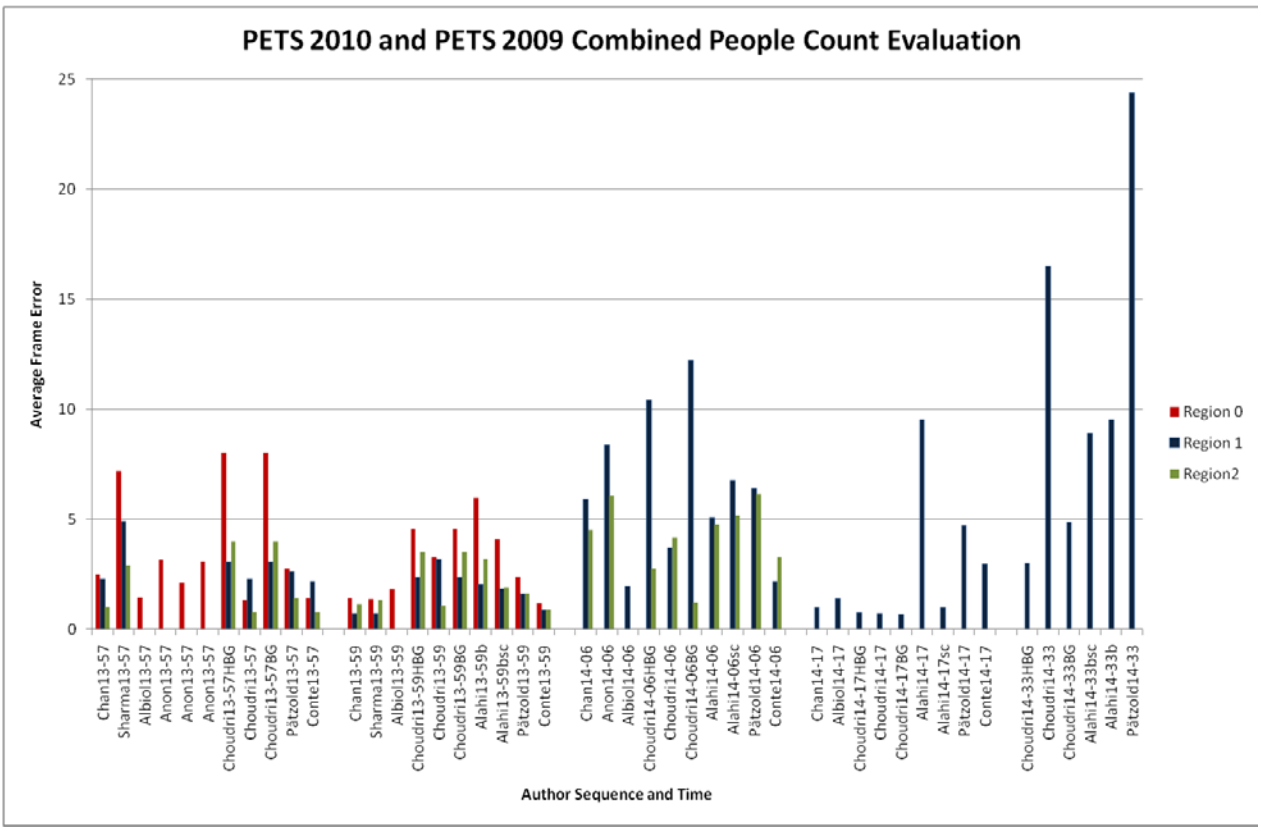

Figure 1. Counting People in Regions.

ing MODA and MOTA as a measure, the systems described by Breitenstein *et al.* [8], Berclaz *et al.* [6], and Conte *et al.* [11] performed strongly at multiple object detection and tracking, with Yang *et al.*'s [19] outperforming all others. Ge and Sikora's [15] detection accuracy (MODA) and Alahi *et al.*'s [3] tracking accuracy (MOTA) results also suggested a robust performance for these particular areas. For precision in this task, using both MODP and MOTP, the system described by Breitenstein *et al.* [8] performed the strongest. Measuring object detection accuracy by frame and sequence using SODA and SFDA metrics, the systems described by Yang *et al.* [19], Breitenstein *et al.* [8], and Conte *et al.* [11] outperformed others.

Figure 4 shows the median of each metric value, for all the computed views, excluding View 2 which was not provided in the dataset, from each author. Again, the performance measures highlight the algorithms provided by Ge and Sikora [15] and Berclaz *et al.* [6] for multiple object detection accuracy and tracking. From this figure it can be seen that although there are variations per metric per author, the results for MOTP, MODP, SFDA, and SODA indicated a general consensus of accuracy.

To estimate the consistency of the metrics themselves another evaluation is illustrated here. Figure 2 showed, for each view, the median value of each metric for all authors. It highlights the relationships between two pairs of metrics. SODA and SFDA present extremely similar median values across all camera views, as do MODP and MOTP. These relationships are emphasised further in the evaluation of the results in Figure 3, where the relationship between each metric measurement pairing (both SODA+SFDA and MODP+MOTP) can be clearly seen. In addition, the tracking accuracy metric values appear to be strongly related to the detection accuracy metric values. The figure suggests that camera view one was the simplest task for the participating authors and camera view six presented a challenge.

As the final evaluation, a view for each metric which corresponds to the median value of the metric for all authors, was used. The results are shown in Figure 5.

From this figure a fair overall performance comparison of each algorithm and their variant forms can be inferred. Due to its robustness to outliers, this visualisation gives a clear indication of how different algorithms perform relative to each other.

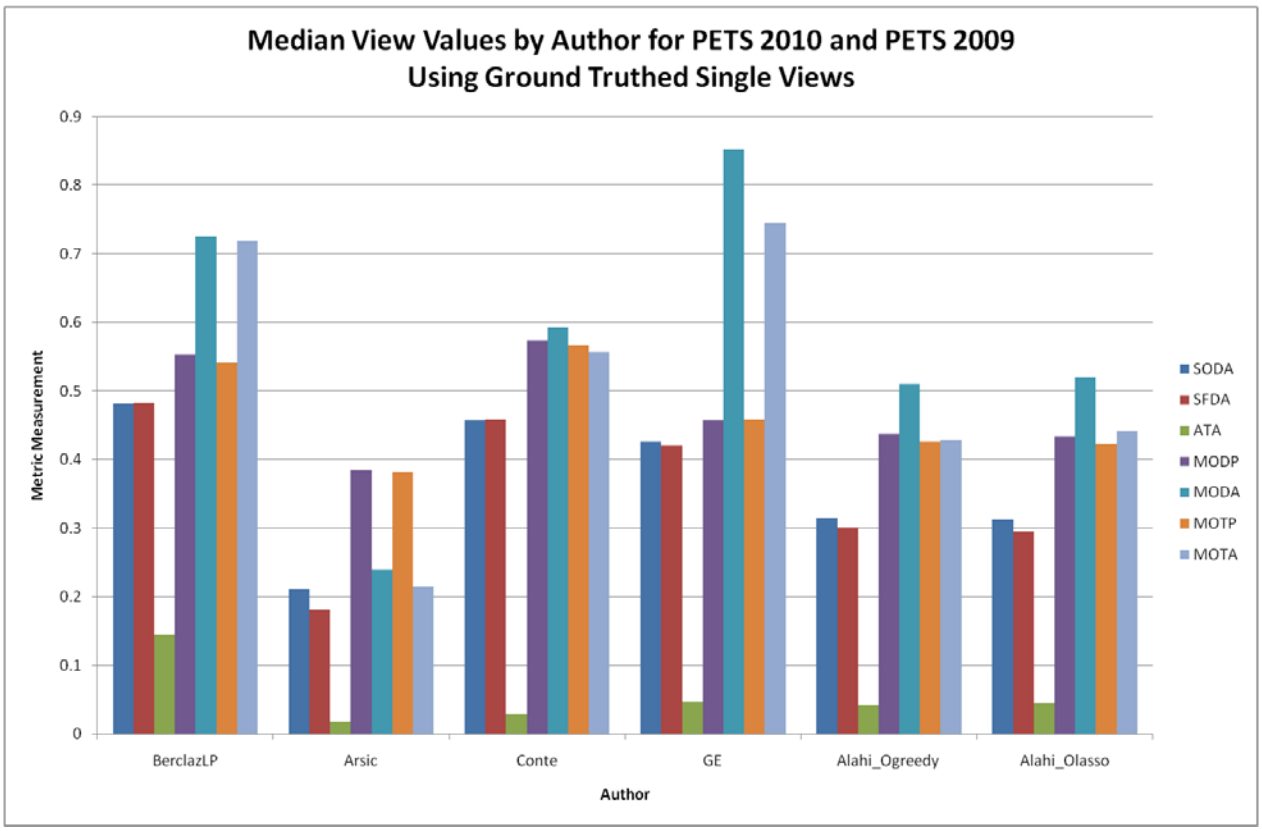

Figure 2. Median Metric Values Among All Authors Per View

5. Conclusion

It is essential that authors are able to objectively evaluate their detection and tracking algorithms with standardised metrics. The ability to compare results, with others, whether anonymous or not, provides a realistic and encouraging research technique towards advanced, robust, real-time visual systems. In addition the latest results highlight the need for careful consideration regarding the ground truthing of data sets and the subsequent evaluation. Whilst the performance of some tracking systems may be shown as lacking using the ground truth produced from a 3D ground truthing tool, using ground truth based on individual 2D viewpoints highlights their robust performance with single camera views. In addition, the use of these metrics and this study provides a mechanism to highlight the strengths of the individual systems, such as accuracy, precision and robustness. It may be used for future decisions for systems placement. For example, those that require a high degree of precision may benefit from techniques described by authors whose systems performed well using precision metrics.

References

[1] PETS Schema. *"http://www.cvg.rdg.ac.uk/PETS2010/"*. 1, 2

[2] ViPER GT. *"http://viper-toolkit.sourceforge.net/"*. 1

[3] A. Alahi, L. Jacques, Y. Boursier, and P. Vandergheynst. Sparsity-driven people localization algorithm: Evaluation in crowded scenes environments. In *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*, pages 1 –8, 7-9 2009. 3, 4

[4] A. Albiol, M. J. Silla, A. Albiol, and J. M. Mossi. Video analysis using corners motion analysis. In *Performance Evaluation of Tracking and Surveillance, 2009 Eleventh IEEE International Workshop on*, pages 31–37, 2009. 3

[5] D. Arsic, A. Lyutskanov, G. Rigoll, and B. Kwolek. Multi camera person tracking applying a graph-cuts based foreground segmentation in a homography framework. In *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*, pages 1 –8, 7-9 2009. 3

[6] J. Berclaz, F. Fleuret, and P. Fua. Multiple object tracking using flow linear programming. In *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*, pages 1 –8, 7-9 2009. 3, 4

[7] D. Bolme, Y. M. Lui, B. Draper, and J. Beveridge. Simple real-time human detection using a single correlation filter. In *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*, pages 1 –8, 7-9 2009. 3

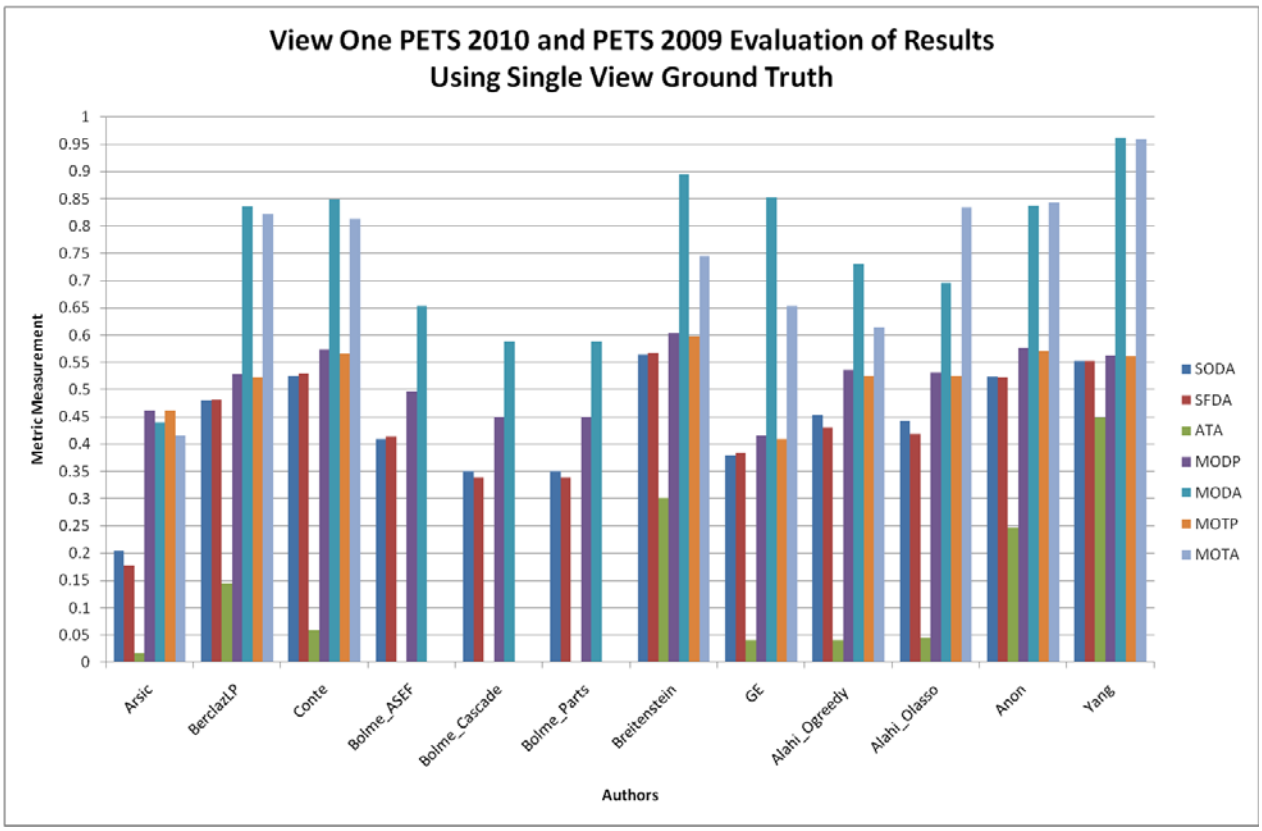

Figure 3. Performance of Authors' Systems Per Metric, Camera View 1, Dataset:S2.L1, Time Sequence: 12.34.

[8] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. van Gool. Markovian tracking-by-detection from a single, uncalibrated camera. In *Performance Evaluation of Tracking and Surveillance, 2009 Eleventh IEEE International Workshop on*, pages 71–78, 2009. 3, 4

[9] A. B. Chan, M. Morrow, and N. Vasconcelos. Analysis of crowded scenes using holistic properties. In *Performance Evaluation of Tracking and Surveillance, 2009 Eleventh IEEE International Workshop on*, pages 101–108, 2009. 3

[10] S. Choudri, J. Ferryman, and A. Badii. Robust background model for pixel based people counting using a single uncalibrated camera. In *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*, pages 1 –8, 7-9 2009. 3

[11] D. Conte, P. Foggia, G. Percannella, and M. Vento. Performance evaluation of a people tracking system on the pets video database. In *Performance Evaluation of Tracking and Surveillance, 2010 Thirteenth IEEE International Workshop on*, 2010. 3, 4

[12] D. Conte, P. Foggio, G. Percannella, and M. Vento. A method based on the indirect approach for counting people in crowded scenes. In *Performance Evaluation of Tracking and Surveillance, 2010 Thirteenth IEEE International Workshop on*, 2010. 3

[13] A. Ellis, A. Shahrokni, and J. Ferryman. Pets2009 and winter-pets 2009 results: A combined evaluation. In *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*, pages 1 –8, 7-9 2009. 2

[14] J. Ferryman and A. Shahrokni. Pets2009: Dataset and challenge. In *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*, pages 1 –6, 7-9 2009. 1

[15] W. Ge and R. Collins. Evaluation of sampling-based pedestrian detection for crowd counting. In *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*, pages 1 –7, 7-9 2009. 3, 4

[16] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):319–336, Feb. 2009. 2

[17] M. Paetzold and T. Sikora. Counting people in crowded environments by fusion of shape and motion information. In *Performance Evaluation of Tracking and Surveillance, 2010 Thirteenth IEEE International Workshop on*, 2010. 3

[18] P. K. Sharma, C. Huang, and R. Nevatia. Evaluation of people tracking, counting and density estimation in crowded environments. In *Performance Evaluation of Tracking and*

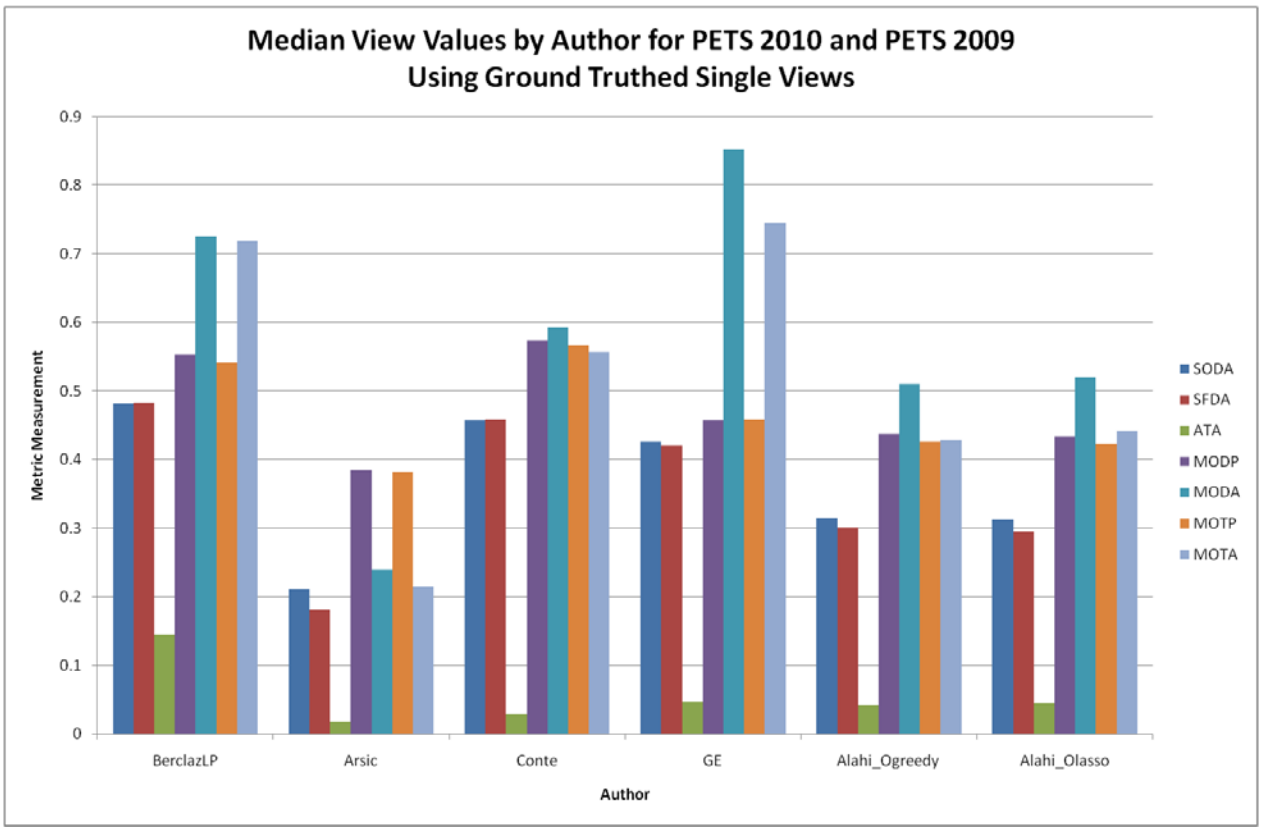

Figure 4. Median Metric Measurement Across All Views for Different Authors, Dataset:S2.L1, Time Sequence: 12.34.

Surveillance, 2009 Eleventh IEEE International Workshop on, pages 39–46, 2009. 3

[19] J. Yang, Z. Shi, P. Vela, and J. Teizer. Probabilistic multiple people tracking through complex situations. In *Performance Evaluation of Tracking and Surveillance, 2009 Eleventh IEEE International Workshop on*, pages 79–86, 2009. 3, 4

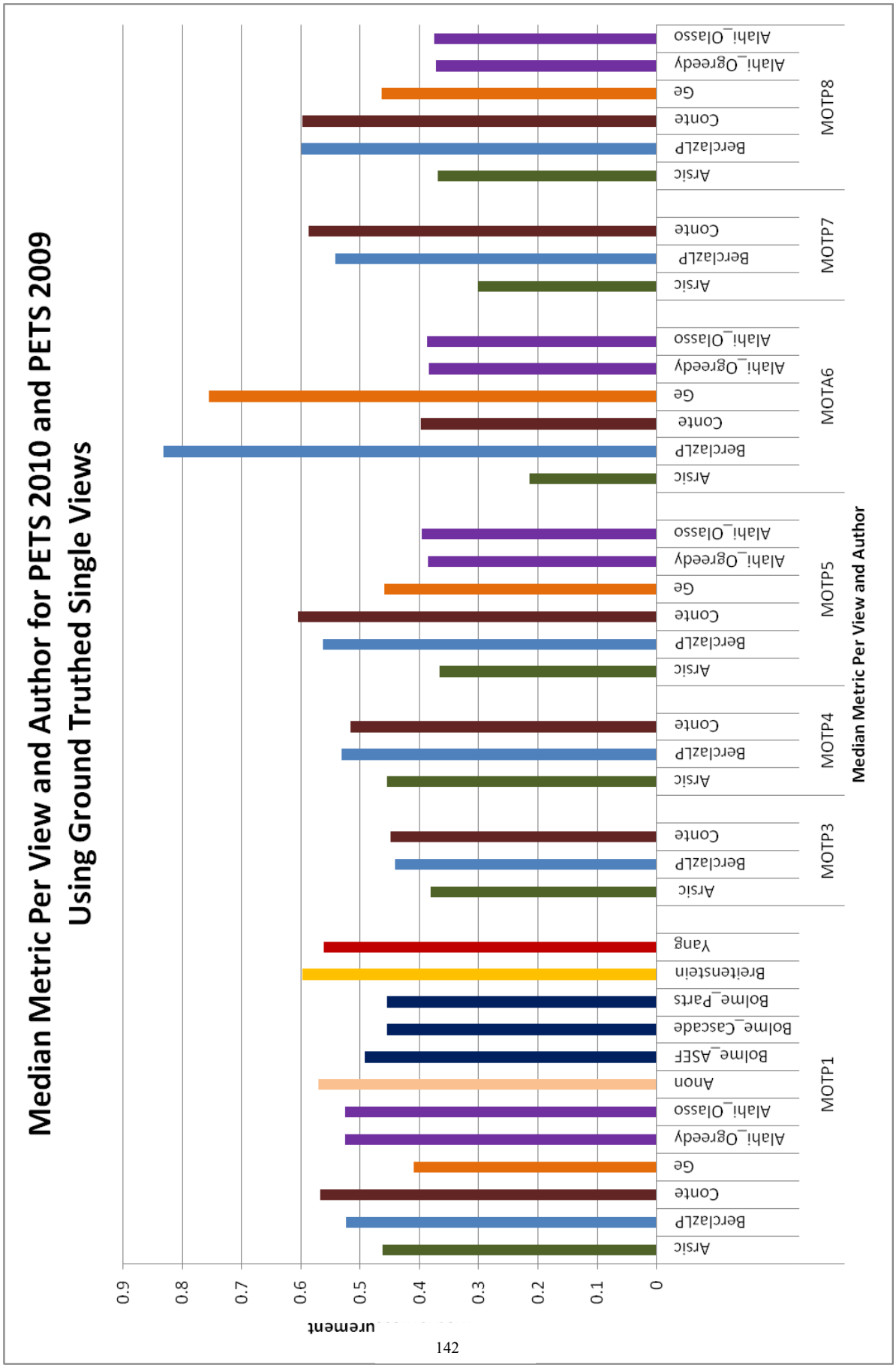

Figure 5. Median Metric Per View and Author.