An Integrated Tracking Approach to the Assessment of Object Description Models

Telmo Oliveira

Pedro Carvalho

Lucian Ciobanu

Jaime S. Cardoso

Luís Côrte-Real

INESC Porto,
Faculdade de Engenharia da Universidade do Porto,
Campus da FEUP, Rua Dr. Roberto Frias, n 378, 4200 - 465 Porto, Portugal

Abstract

A key issue in video object tracking is the representation of the objects and how effectively it discriminates between different objects and the background. Several techniques have been proposed, but it is missing a generally accepted method. In a video object tracking framework, the appearance model is one of the components with its performance depending on previous processing stages and affecting those that succeed it. We extended the analysis of description methods to better understand the impact in the overall tracking process, by comparing a set of widely used descriptors in a common object tracking solution. This work provides foundations for future tests, contributing to a more informed selection of techniques adequate for a given application context.

1 Introduction

Tracking multiple objects, and in particular humans, presents many challenges, especially if it occurs in non-controlled environments as is the case of most everyday scenarios, were factors such as coverage of large areas, group movement and occlusion must be efficiently addressed. Several techniques have been proposed for object representation, each with its strengths and weaknesses, but a generally accepted method is missing. The robustness of some methods typically implies a higher computational cost; inversely, more lightweight methods may originate higher error rates, making them unsuitable for most scenarios. In a video object tracking (VOT) framework, the appearance model's performance will depend on previous processing stages and affect those that succeed it. Consequently, it was considered logical to assess the performance of appearance description techniques from a tracking solution point of view, analysing their behaviour in the overall tracking. A set of widely used description methods were integrated and compared in a common object tracking solution. Moreover, the foundations for future tests of other descriptors have been prepared.

Color histograms have been widely used in monocular tracking with interesting results, but with worst performance in multi-view scenarios. Hence, different description methods have been researched, with edge or gradient based features receiving significant attention as an alternative. Unlike color histograms, local descriptors are commonly computed at key points of an image, and should be invariant to noise, geometric, photometric and other deformations [9]. While a lower dimension descriptor may have a faster computation time, it is in general less distinctive than their higher-dimension counterpart. The SIFT (Scale Invariant Feature Transform) interest point detector and descriptor [7] is one of the most well know methods to determine local descriptors that are invariant to changes in scale, rotation and translation. The Gradient Location-Orientation Histogram (GLOH) [9] is a variant of SIFT using a log-polar binning structure instead of four quadrants. The Speeded-Up Robust Features (SURF) [3] shares many similarities with SIFT, but with performance gains due to the approach followed on the detection of interest points and on the matching process. The Fast Invariant to Rotation and Scale feature Transform (FIRST) [2] is a recent feature detection and matching technique. It is stated that the algorithm is faster than most traditional approaches, such as SIFT, although yielding less distinctiveness in terms of rotation and scale. This common trade-off between quality and performance is reflected in the number of key points extracted by FIRST (smaller than its counterparts). The bag-of-words (BoW) [6] is a key point based representation, which consists of grouping of similar key points descriptors into a large number of clusters; each cluster is treated as a visual word which in turn is used to build a visual vocabulary. A different approach is followed in the computation of the Histogram of Oriented Gradients (HOG) [5] descriptor, which uses a dense grid of uniformly spaced cells and it has been shown that HOG is insensitive to color variation. A more detailed survey analysing the aforementioned and other different object descriptors can be found at [8, 10].

The rest of this paper is organized as follows: section 2 presents the tracking algorithm, used dataset and the evaluation metrics used to assess our results; the results are (partially) shown on section 3 and our final comments are presented on section 4.

2 Experiments

2.1 Framework

The test environment consists of an implementation of the tracking algorithm proposed in [11]. It is focused on a surveillance scenario and was described as having the capability of real time operation and an acceptable detection and tracking rate in low to medium complexity scenarios. Also, it has features that enable human detection and the separation of objects in a group. Throughout these experiments only the extraction of the object appearance and the matching between two individual instances are model specific. All other stages of the algorithm are common.

For our experiments, we selected the following state-of-the-art description methods: SIFT; SURF; HOG; color histogram; BoW. For the color histogram the bounding box is divided into an upper and lower part with a greater weight assigned to the upper part. Several variations for the appearance models were tested in order to emphasize particular contributions thereof to the performance of the tracking algorithm. For the key point based descriptors the ensemble of descriptors was adopted as the model. A grid-based variation was also considered for the SIFT and SURF appearance models; it is composed of evenly spaced points and intends to provide an alternative to the use of interest point's detection by performing a dense scan. We tested grids with a number of points equal to 1% and 4% of the object's image. For the BoW approach, the dictionaries were trained for SIFT and SURF descriptors using generic images. The experiments were performed in a computer with an Intel(R) Core(TM) i5CPU at 3.20GHz with 8GB of RAM.

2.2 Evaluation

For the experiments we selected sequences of two widely used datasets. Specifically, we used sequences: cam3 (ST1C3) and cam4 (ST1C4) from the ST1-C1 set of the PETS 2006 workshop; OneShopOneWait1 (OSOW1) and OneShopOneWait2 (OSOW2) from the Caviar project. These sequences are representative of monitoring and surveillance scenarios depicting commonly observed problems: group movement; appearance similarity; occlusion. Also, they were captured with dissimilar cameras and offer different perspectives over the scene.

We decided to use two complementary metrics to objectively evaluate the impact of the different representation models in the tracking solution. We chose the hybrid framework [4]), enabling the computation of an error metric for every frame of the sequence, whose output we'll refer to as

Table 1: Overall comparison of the experimented models. For grid based representations, the best results were selected for each type of descriptor.

		Appearance Model Comparison						
	Metric	SIFT	SIFT Grid (1%)	SURF	SURF Grid (1%)	HOG	Histogram	Texture
OSOW1	TRDR	0.68	0.60	0.14	0.71	0.73	**0.79**	0,50
	TSR	0.29	0.43	0.43	**0.86**	0.57	0.57	0,43
	DR	0.68	0.60	0.14	0.71	**0.73**	0.69	0,50
	FAR	0.10	0.11	0.04	**0.02**	0.08	0.14	0,35
	FNR	0.32	**0.04**	0.86	0.29	0.27	0.22	0,50
	FER	2.57	1.57	1.57	**1.00**	1.29	1.29	1,33
OSOW2	TRDR	0.64	0.66	0.20	0.72	0.69	**0.76**	0,58
	TSR	0.20	0.40	0.40	0.40	**0.60**	0.40	0,50
	DR	0.63	0.65	0.20	0.71	0.69	**0.73**	0,57
	FAR	0.23	0.20	0.21	**0.16**	0.28	0.25	0,37
	FNR	0.37	0.35	0.80	**0.29**	0.31	0.27	0,43
	FER	3.38	2.88	4.25	2.56	**1.25**	1.56	2,50

Table 2: Comparison of the processing times for the solution and for individual components of the models: extraction and matching.

	OSOW1			ST1C3		
	SPT (s/f)	DET (s/t)	DMT (s/t)	SPT (s/f)	DET (s/t)	DMT (s/t)
SIFT	0.420	0.175	0.000	0.800	0.559	0.003
SIFT Grid (1%)	0.242	0.092	0.000	0.416	0.239	0.000
SURF	0.033	0.002	0.000	0.055	0.014	0.000
SURF Grid (1%)	0.194	0.074	0.002	0.354	0.202	0.010
Histogram	0.213	0.000	0.000	0.190	0.000	0.000
HOG	0.052	0.007	0.001	0.058	0.010	0.002
Texture	0.084	0.000	0.000	0.177	0.000	0.000

'hybrid metric'. We also use some of the measures proposed in [1] to summarise evaluation results for a complete sequence. Specifically, we chose: Tracker Detection Rate (TRDR); Tracking Success Rate (TSR); Detection Rate (DR); False Alarm Rate (FAR); False negative Rate (FNR).Note that for the first three metrics we wish to obtain high values, while for remaining metrics low values are desirable.

3 Results

Given the large number of experiments conducted, we will only present the most expressive results.

Table 1 summarises the results obtained for the tracking solution using the different methods and variations. For the experiments using grid points, we selected those with the best performance for each model. The results show that SURF with grid points (1% density), HOG and color histogram enabled the best performances, but it is difficult to highlight one of the three. Color histogram enabled a good overall tracking performance, accompanied closely by solutions using a HOG based model; the solution using SURF with grid points also presented a good reliability concerning false alarms and misdetections. Signature based models behaved poorly and were omitted here for simplicity.

Table 2 presents the time measures for the most relevant experiments considering the tracking performance results. We present the following time measures: average frame processing time (SPT), measured in seconds per frame; average descriptor extraction time (DET), which we measured in seconds per track (as in a single 'object track'); average descriptor matching time (DMT), also in seconds per track. We restricted Table 2 to a single CAVIAR and PETS sequence for readability issues, since presenting additional sequences did not necessarily bring additional information.

As expected, color histogram and HOG exhibited a good computational performance. SIFT presented a higher complexity, but the use of a grid of points enabled a significant decrease of the computational time. SURF also presents low computational time values, but the tracking results were poorer. However, it is noteworthy the values obtained for SURF with a grid of points; although they are higher than with the use of interest point detection, they are lower than with SIFT and it was demonstrated in Table 1 that they enable better tracking results.

4 Discussion and Future Work

Given the difference between tracking and standalone image or object matching, it was considered logical the analysis of different description models in a common tracking solution. A set of well known description techniques and different representations were analysed and assessed by measuring the output of the tracking solution in terms of computational performance and accuracy of the results.

It was observed that a dense scan did not provide a significant contribute when using SIFT. In contrast, the performance of the tracking algorithm with SURF descriptors increased significantly by using a dense scan in alternative to interest point detection. With SIFT, the number of identified points is higher, hence grid points may not contribute with relevant information. When using grid points, overall better results were obtained for smaller patch's heights, indicating that grid points may introduce noise in the model.

The colour histogram and HOG descriptors were computed over the object's image, thus a grid of points was not applicable in these cases. Nevertheless, using small object patches conduced to better results, which is coherent with the argument that smaller heights for the object's bounding box (and consequently its image) can limit the background noise added to the models. In particular, the results for color histograms were not surprising considering the scenario: the use of a single camera with a reasonable frame rate enabled smooth transitions from one frame to the next. Furthermore, object patches with small dimension and smooth regions are often an obstacle to the computation of edge or gradient based descriptors.

Future research work will include the integration and analysis of different object descriptors and appearance models. Furthermore, the VOT framework itself is to be subjected of integration of additional techniques. It would be interesting to perform the analysis proposed in this paper in multi-camera scenarios thus obtaining more complete information regarding the description methods robustness in both contexts.

References

[1] F. Bashir and F. Porikli. Performance evaluation of object detection and tracking systems. In *Proceedings of IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), PETS 2006.*, June 2006.

[2] R. Bastos and M. S. Dias. FIRST - Fast Invariant to Rotation and Scale Transform: Invariant image features for augmented reality and computer vision. In *VDM Verlag*, 2009.

[3] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110: 346–359, June 2008.

[4] P. Carvalho, J. S. Cardoso, and L. Corte-Real. Hybrid framework for evaluating video object tracking algorithms. *Electronics Letters*, 46(6):411–412, 2010.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc, of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 886–893. IEEE Computer Society, 2005.

[6] Nowak E, F. Jurie, and Bill Triggs. Sampling strategies for Bag-of-Features image classification. In Aleš Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision - ECCV 2006*, volume 3954, chapter 38, pages 490–503. Springer Berlin Heidelberg, 2006.

[7] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[8] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(10):1615–1630, 2005.

[9] R. Szeliski. *Computer Vision : Algorithms and Applications*. Springer-Verlag New York Inc, 2010.

[10] P. Tissainayagam and D. Suter. Assessing the performance of corner detectors for point feature tracking applications. *Image and Vision Computing*, 22:663–679, 2004.

[11] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1208–1211, Sep 2004.