Machine Vision and Applications manuscript No.
(will be inserted by the editor)

Analysis of Object Description Methods in a Video Object Tracking Environment

Pedro Carvalho · Telmo Oliveira · Lucian Ciobanu · Filipe Gaspar · Luís F. Teixeira · Rafael Bastos · Jaime S. Cardoso · Miguel S. Dias · Luís Côrte-Real

Received: date / Accepted: date

Abstract A key issue in video object tracking is the representation of the objects and how effectively it discriminates between different objects. Several techniques have been proposed, but without a generally accepted method. While analysis and comparisons of these individual methods have been presented in the literature, their evaluation as part of a global solution has been overlooked. The appearance model for the objects is a component of a video object tracking framework, depending on previous processing stages and affect-

The authors would like to thank the Fundação para a Ciência e a Tecnologia (FCT) - Portugal - and the European Commission, for financing this work through the grants SFRH/BD/31259/2006, SFRH/BD/73667/2010 and Fundo Social Europeu (FSE). Part of the work was also developed in the context of Project QREN 7900 LUL (Living Usability Lab), a co-promotion R&D project funded by European Structural Funds for Portugal (FEDER) through COMPETE as part of the National Strategic Reference Framework (QREN), and managed by Agência de Inovação (ADI).

Pedro Carvalho, Telmo Oliveira, Lucian Ciobanu, Jaime S. Cardoso and Luís Côrte-Real
INESC TEC (formerly INESC Porto) and Faculdade de Engenharia, Universidade do Porto
Campus da FEUP, Rua Dr. Roberto Frias, n 378, 4200 - 465 Porto, Portugal
Tel.: +351-222094000
Fax: +351-222094050
E-mail: pedro.carvalho@inescporto.pt

Filipe Gaspar
ADETTI-IUL / ISCTE-Lisbon University Institute,
Av. das Forças Armadas, Edifício ISCTE, 1600-082 Lisboa, Portugal

Luís F. Teixeira
Faculdade de Engenharia da Universidade do Porto,
Campus da FEUP, Rua Dr. Roberto Frias, no. 378, 4200 - 465 Porto, Portugal

Rafael Bastos
Vision-Box & ADETTI-IUL / ISCTE-Lisbon University Institute,
Rua Casal do Canas no. 2, Zona Industrial de Alfragide, 2790-204, Carnaxide, Portugal

Miguel S. Dias
Microsoft Language Development Center & ISCTE-Lisbon University Institute,
Av. das Forças Armadas, Edifício ISCTE, 1600-082 Lisboa, Portugal

ing those that succeed it. As a result, these interdependencies should be taken into account when analysing the performance of the object description techniques. We propose an integrated analysis of object descriptors and appearance models through their comparison in a common object tracking solution. The goal is to contribute to a better understanding of object description methods and their impact on the tracking process. Our contributions are threefold: propose a novel descriptor evaluation and characterisation paradigm; perform the first integrated analysis of state-of-the-art description methods in a scenario of people tracking; put forward some ideas for appearance models to use in this context. This work provides foundations for future tests and the proposed assessment approach contributes to the informed selection of techniques more adequately for a given tracking application context.

Keywords Computer Vision · Descriptors · Appearance Models · Tracking Assessment · Video Object Tracking

1 Introduction

The automatic tracking of objects has been gaining importance in recent years. However, everyday situations still present complex problems within the scope of research activities. There are innumerous research efforts targeting different aspects of the problem and a vast number of published work in international journals, conferences and workshops. Tracking multiple objects, and in particular humans, is a difficult problem presenting many challenges, especially if it occurs in non-controlled environments as in everyday scenarios. In these situations, algorithms or tracking systems must deal with factors such as coverage of large areas, group movement, partial or total occlusion, shape deformation, fast changes in direction, illumination variations and shadows, among others.

The main steps of a video object tracking (VOT) framework can be summarised as: object detection, often based on background/foreground segmentation; object description, where different cues (e.g., appearance, shape) are extracted to uniquely characterise each object; definition of correspondences for each tracked object throughout the video sequence. Consequently, the representation of objects and how effectively it discriminates between different instances is a key issue [16]. Several techniques have been proposed, each with its strengths and weaknesses, but without a generally accepted method.

The appearance model receives and provides information from and to other modules of a VOT framework. As a result, its performance will depend on previous processing stages and will affect those that succeed it. For example, by estimating the position of an object in the next frame it is possible to reduce the uncertainty that the appearance model must resolve; lower discriminative capabilities of the appearance model may cause track drift or even track loss (inability to detect an object being tracked or erroneous identity assignment to the tracked object).

The standalone comparison of descriptors has been the subject of several studies [24, 25, 7, 38, 31] and is not within the scope of this paper to make an

exhaustive review or benchmarking. Rather, recognising the importance of these techniques to visual tracking we found necessary, and a logic evolution, to also perform this analysis from a tracking solution point of view. This will enable a better understanding of the impact on the overall tracking. To validate our proposal, we compare a set of widely used description methods in a common object tracking solution.

Video object tracking strategies are still highly related to the application scenario. Even though the concepts and strategy proposed in this paper can be applied to different tracking scenarios, we present a proof of concept by assessing state-of-the-art descriptors in a people tracking solution. Furthermore, we also put forward some ideas for appearance models.

The remaining of this paper is organised as follows. A brief overview of state-of-the-art description methods and their use in video object tracking algorithms is presented in Section 2.1; in Section 2.2 we present a brief summary of the results from a standalone comparison of the descriptors. Our proposal is described in Section 3. It includes the main concepts, suggested metrics and datasets. In Section 4 we describe a set of experiments intended to serve as proof of concept for the proposed assessment approach. The results and conclusions are presented in Sections 5 and 6 respectively.

2 Object Descriptors

2.1 Literature review

Ideally, an object's description should enable its discrimination from other objects during the tracking process. However, this is hard to achieve in a real scenario. Towards this goal, two main approaches can be found on the literature: (1) based on the analysis of histrogram responses; (2) based on the extraction and matching of local feature descriptors. Following the formulation that objects' appearance and shape, within an image, can be described by its distribution of colour, intensity gradients, or edge directions, histogram analysis has been widely studied. Interesting results have been reported using colour appearance and colour histograms in single camera tracking, but poor performance was achieved in multi-view scenarios. As a result, increasing attention has been given to the research of edge or gradient based features as an alternative or complement to colour based descriptors [1].

Local feature descriptors have been widely used in different areas including video object tracking and image retrieval. These descriptors are commonly computed at key points of an image, which are salient patches that contain rich local information about the image [17]. High repeatability of the key points is desirable since it expresses the reliability of a detector for finding the same physical key points under different viewing conditions. The feature vector used to represent the neighbourhood of the key point should be robust (invariant) to noise, detection displacements, and geometric and photometric deformations. It has been recognised that the dimension of the descriptors has

a direct impact on the computation time [4]. Although descriptors of lower dimensions are desirable for fast key point matching, they are in general less distinctive than their high-dimensional counterparts.

The SIFT (Scale Invariant Feature Transform) key point detector and descriptor proposed by Lowe [23] is one of the most well known methods to determine local descriptors that are invariant to changes in scale, rotation and translation. The detector generates a great number of key points compared to other detectors, but the extraction process tends to be slower. Furthermore, the high dimensionality of SIFT descriptors has significant impact on the matching step. Ke and Sukthankar [19] applied Principal Component Analysis (PCA) to decrease the dimension of the vector. Although the resulting vector enabled a faster matching, Mikolajczyk and Schmid [25] proved that it is less distinctive than SIFT and the PCA slows down the feature computation. Gradient location-orientation histogram (GLOH) [34] is another variant of SIFT using a log-polar binning structure instead of four quadrants. In the study by Mikolajczyk and Schmid [25], GLOH slightly outperforms SIFT, but the use of PCA has a negative impact on the computational weight [19]. In [4], Bay et al. described a fast scale and rotation invariant key point detector and descriptor which they named SURF (Speeded-Up Robust Features). It shares many similarities with SIFT, but with performance gains due to the approach followed on the detection of key points and on the matching process. FIRST (Fast Invariant to Rotation and Scale feature Transform), proposed by Bastos et al. [3] in the context of Augmented Reality and Computer Vision, relies on the detection of local maxima Shi and Tomasi corners to define key points location. Each feature's intrinsic scale factor is determined by selecting the multiplication factor that maximizes the luminance average around the key point, using integral images for fast indexation. Similar to SIFT, FIRST uses an orientation histogram to make the descriptor rotation invariant. However, instead of recording the histogram itself, the final patch is rotated by the maximum bin value found. Since FIRST uses 15x15 non-normalized patches, matching is performed through Normalized Cross Correlation, which formulation is largely simplified by describing features through a DoG (Difference of Gaussians). In order to accelerate the matching step on large feature databases, DoG patches are clustered based on its sum of all pixels sign (positive or negative). Each patch is evaluated in 6 distinct regions (3 vertical regions and 3 horizontal regions) resulting in a 6 bits string which enables fast cluster creation and matching problem reduction. Reported results indicate that FIRST algorithm is faster than SURF, although yielding slightly less repeatability.

Methods using key point selection and local descriptors have produced interesting results. However, in some cases, such as their application to objects with large smooth regions, the number of selected points may be insufficient for a successful matching process [1].

Dalal et al. [12] proposed a new human detector based on a grid of Histograms of Oriented Gradient (HOG). Unlike descriptors such as SIFT, these are computed on a dense grid of uniformly spaced cells. It has been shown that HOG is insensitive to colour variation. The HOG descriptor has a high

dimensionality, thus requiring a large amount of memory storage [18]. Also, it has difficulties with the effective representation of objects or backgrounds with large smooth regions since the contours are indistinctive. Usually HOG presents high sensitivity to rotation transformations.

Most of the base invariant descriptors presented, being local based key point descriptors such as SIFT, SURF and FIRST, or being based on histogram analysis such as colour histograms and HOG, fall into the same problem formulation: the need to identify corresponding properties between different images taken from the same scene under different conditions, such as presence of affine transformations (rotation, scale), changes in noise, image blurring, and luminance conditions variation. This ability is usually denoted as repeatability.

The above mentioned techniques share some shortcomings. It has been recognised that in image patches of reduced dimension or with little texture it may not be possible to extract an effective description of the objects [1]. Moreover, these descriptors disregard spatial information, which has been identified as an important feature to increase the object identification rate [28,37,42,33]. Histogram approaches are typically faster than common invariant descriptors, but ignore spatial and shape information. Han et al. [16] proposed the combination of colour histograms and HOG descriptors arguing that the computation is efficient and the combination of these two features can effectively represent an object because they complement each other.

Jiang et al. [18] used the HOG people detector to initialise the bounding box (BB) of pedestrians and colour histograms to describe each object. Rather, than extracting a single histogram, the BB was divided into a lower and upper part. A colour histogram was computed for each part, thus incorporating spatial information. During the tracking process, different weights were used for the upper and lower parts based on the argument that the lower part is more likely to become occluded. Tang and Tao [35] used SIFT descriptors in conjunction with a graph-based model to increase the robustness to occlusion. However, the matching efficiency tended to deteriorate with the increase of the number of objects or the complexity of the model (number of key points).

Other methods have been proposed to represent objects namely human objects. In [40], the authors propose modelling the human shape using skeleton decomposition. The human shape is subdivided iteratively into elementary disks. Following the same approach, in [41] it was presented an inter-frame interpolation method. The shape model was updated with partial changes to the skeleton decomposition model, which was built based on the interpolation of the input and output frames. The authors claimed that better performances could be achieved with only a small complexity overhead.

The mean-shift algorithm has been widely used for visual tracking [6,22, 45] and colour-based mean-shift has been identified as a fast and effective algorithm for tracking colour blobs, but sensitive to large "distracters". Zhou et al. [45] used mean-shift in combination with the SIFT technique to improve tracking. The SIFT features were used to match a region of interest across frames and mean-shift applied for a similarity search through the use of colour

histograms. Shahed et al. [32] tracked a deformable object using rectangles with a minimum overlap and colour information extracted for each box. In [1] SIFT and SURF were tested in a cascade of region descriptors.

2.2 Overview of Individual Benchmarkings

Many different studies have performed the independent evaluation of descriptor methods [24, 25]. In Bastos et al. [3] a framework simulating the complete pipeline of images 3D viewing and projection was proposed following the same evaluation metrics. As a result, it was possible to generate and render images under different viewing variations, namely scale and rotation, as well under different perturbations, specifically, luminance, blurriness and random noise changes. The authors also analysed the extraction and matching times per key point for the descriptors. The studies with this benchmarking framework showed that SIFT is typically superior to FIRST and SURF in terms of repeatability and distinctiveness. Exceptions occurred in the presence of noise and luminance variations, which suggests that the DoH (Determinant of Hessian) adopted in SURF is less sensitive to these perturbations than key points extracted in SIFT via DoG (Difference of Gaussians). Considering perspective transformations tested, the computation of Sobel derivatives to find key points' dominant orientation in SIFT and FIRST proved more robust than the Haar wavelet responses used in SURF. Since FIRST key points do not have a scale-space representation, the repeatability across scale changes is considerable smaller than in SIFT, even though is similar to SURF. Regarding extraction and matching times per key point, SURF outperformed SIFT in both measures, while FIRST presented the best computational performance. In terms of quality, SIFT key points are highly distinctive and accurate, while SURF is the most error-prone algorithm, but has the highest matching rate due to the large number of key point extracted.

The authors also performed the tests for colour histogram and HOG algorithms, which are not based on local key points. Due to the nature of these methods, the evaluation was made by measuring the average chi-squared (χ^2) distance between the histogram created from an original image, and the histogram created from the transformed or perturbed image. Colour histogram proved to be more robust than HOG to all invariance tests, but at a higher computational cost.

A noticeable result of these standalone tests was the need for different evaluation approaches due to the nature of the methods.

3 The Proposed Integrated Analysis Strategy

The main concepts of the proposed assessment approach are described in Section 3.1. While the proposed concepts are horizontal to different tracking contexts, in this paper we focus on people tracking scenarios. In Sections 3.2

and 3.3 we describe a set of suggested metrics and datasets applicable to this context.

3.1 A New Assessment Paradigm

In the specific context of video object tracking, the performance of the appearance model and the corresponding descriptor technique is not independent of other modules. Rather, it is influenced by the preceding modules and affects those that succeed it. In Figure 1 it is represented the conceptual architecture of a tracking solution[1]. This architecture focuses on common tracking modules. Upon receiving images of a stream, some pre-processing can be performed, for example to reduce noise in the image or change the colour space. An initial step of the tracking solution consists of the detection of the objects of interest, in this case people, and the initialisation of the corresponding models (in our particular case the focus is on the appearance model). Once the objects are detected in the current frame it is necessary to define correspondences with previous detected objects; this process makes use of the information stored in the appearance model. The final step consists of the update of the models with the new observed information. Modules for higher level processing using the output of tracking may exist, but they are not relevant to this proposal.

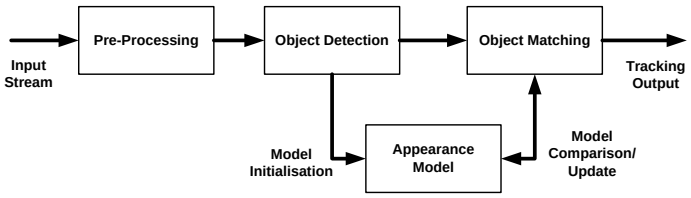


Fig. 1 Conceptual generic architecture of a video object tracking solution.

In this paper we propose a new paradigm for the assessment of object description techniques for video object tracking, with the individual description methods being integrated in a common tracking solution and the evaluation performed upon the tracking results. This approach takes into account the interdependencies among the modules of the system and is performed from the tracking point of view, i.e., it uses the tracking output. To our knowledge, such approach has never been attempted.

The underlying framework must be prepared to accommodate the integration and test of different appearance models. Throughout the experiments only the extraction of the object appearance and the matching between two individual instances should be model specific. All other stages of the algorithm must be common. The appearance models are to be encompassed in modules

[1] The reader may note some similarities with the architecture proposed by Moeslund [26].

with a common and generic API (application programming interface). Each of these modules should filter and store the required information and be responsible for performing all model specific information, namely initialisation, comparison and update.

3.2 Evaluation Metrics

The evaluation of tracking results is by itself a research problem with several open issues, such as: what factors to evaluate?; what information is available?; what metrics to use?

The criteria used in the evaluation of the tracking algorithms should be appropriate to the application scenario. This has resulted in the use of different features (e.g., as objects' trajectory, silhouette or assigned identifier) and application of different metrics (e.g., trajectory root mean square error, detected and reference region overlap, identity consistency). Proposals for evaluation frameworks and metrics already exist, but they haven't been generally adopted by the research community [5,30,15,2,27,13,8]. Some of these approaches focus on evaluation without comparing with a reference (commonly known as ground truth (GT)), but the results provided typically lack sufficient discriminative information. Consequently, evaluations based on comparisons with GT are commonly favoured and most test sequences are not accompanied by this information.

To objectively evaluate the impact of the different representation models in the tracking solution we propose two complementary strategies, intended to provide greater flexibility. The first consists of a set of metrics proposed by Bashir et al. [2] to summarise evaluation results for a complete sequence. Specifically, the object paradigm was used, which implies a previous alignment of reference and detected tracks. These will be referred to as 'object metrics'. The use of these metrics was motivated by its use, or of slight variations, in other papers of the literature such as [14,20,5,39]. From the proposed set of metrics in [2] we selected:

$$\text{Tracker Detection Rate (TRDR)} = \frac{TP}{TG} \tag{1}$$

$$\text{Detection Rate (DR)} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{False Alarm Rate (FAR)} = \frac{FP}{TP + FP} \tag{3}$$

TG is the total number of ground truth tracks; true positives (TP) consists of the number of tracked objects mapped to a GT track; similarly, false positives (FP) are the number of tracked objects without a correspondence in the GT and false negatives (FN) the number of GT track without a tracked object mapped to it. The tracker detection rate (TRDR) and false alarm rate (FAR) characterise the tracking performance of the algorithm, while the detection

rate (DR) measures the sensitivity of the algorithm. The fragmentation error rate (FER) was also considered to indicate how often a track label changes; it consists of the average number of detected tracks per GT track (tracks paired with a GT track). Ideally, the FER value should be one, with larger values reflecting poor tracking and trajectory maintenance. Note that for the first two metrics, TRDR and DR, we wish to obtain high values, while for metric FAR low values are desirable. The computation of these metrics requires the existence of reference information for the sequences. Specifically, the metrics are calculated using information of the bounding boxes enclosing each object. This requirement limits the test sequences that can be used.

The second evaluation strategy consists of the hybrid framework described in [9,10], which enables the computation of an error metric for every frame of the sequence, i.e., it can also capture the temporal evolution of the error. Moreover, it has been demonstrated [8–10] that this framework: (1) enables the use of different types of GT; (2) this information is not required for the complete sequence. The output of this framework will be referred to as 'hybrid metric'.

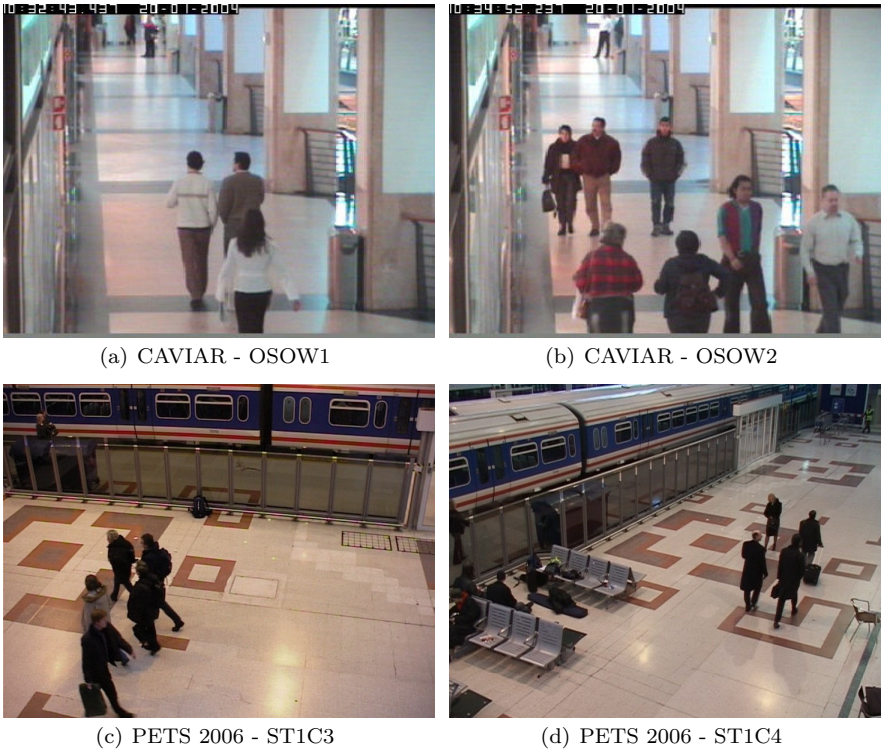
Due to the described properties of these metrics, we suggest their use, particularly in a context of people tracking. Nevertheless, other metrics can be selected and, for different application contexts, that can be a requirement. The important point is that the evaluation is done using the tracking output. Moreover, different application scenarios may require the maximisation/minimisaton of a specific metric or subset of metrics.

Additionally, we propose the computation of several time measures in each experiment to assess the impact of the descriptor and appearance representation in the algorithm execution and the corresponding computational weight. These consist of: the average frame processing time, which we will refer to as SPT (sequence processing time), in seconds per frame; the average descriptor extraction time (DET), in seconds per track; the average descriptor matching time (DMT), in seconds per track. For each track, the values for DET and DMT were averaged over the number of iterations performed to define object correspondences.

3.3 Dataset

The evaluation should be performed using datasets representative of the scenario and accessible to researchers, thus favouring replication and comparison of results. For our particular demonstration scenario we selected sequences of two widely used datasets: CAVIAR project [11]; PETS 2006 [29] workshop. Specifically, we used the sequences: OneShopOneWait1 (OSOW1) and OneShopOneWait2 (OSOW2) from the CAVIAR project; ST1-C1/cam3 (ST1C3) and ST1-C1/cam4 (ST1C4) from the PETS 2006 workshop. These sequences are representative of monitoring and surveillance scenarios depicting commonly observed problems: group movement; shape deformation; appearance similarity; occlusion; object crossing. Also, they were captured with dissimilar

cameras and offer different perspectives over the scene. Figure 2 depicts illustrative frames of the sequences. The sequences of the CAVIAR project have half-resolution of the PAL (Phase Alternating Line) standard (384x288 pixels, 25 frames per second) and were compressed using MPEG-2; the resolution of the PETS sequences are PAL standard (768x576 pixels, 25 frames per second) and were compressed as JPEG image sequences (approx. 90% quality).



(a) CAVIAR - OSOW1 (b) CAVIAR - OSOW2

(c) PETS 2006 - ST1C3 (d) PETS 2006 - ST1C4

Fig. 2 Illustrative frames of the dataset used in the experiments.

The sequences have associated different types of ground truth (GT). The CAVIAR project provided reference information, in the form of bounding boxes (BB), for every sequence of the dataset; this was provided in an XML file following the CVML (Computer Vision Markup Language) syntax [21]. This BB GT information was also mapped to the corresponding image mask [10]. Additionally, for the sequences OSOW1 and OSOW2 reference segmentations were manually generated for a subset of the frames. Sequences ST1C3 and ST1C4 are not provided with reference information applicable to this context (provided GT consists of a list of events), thus reference segmentations were manually produced for a subset of frames.

4 Proof of Concept

In the previous section we described a new paradigm for assessing object description techniques in a video object tracking context. In this section we aim to provide a proof of concept and demonstrate the benefits of this proposal.

Tracking algorithms are intrinsically related to a given context. Hence, different tracking solutions can be conceived. It is not our intention to restrict the tracking solution. Rather, we selected a state-of-the-art people tracking algorithm, presented below, which is well described in the literature.

4.1 Tracking Test Environment

The selected testing environment consists of an implementation of the tracking algorithm proposed by Zhao and Nevatia [44,43]. The algorithm was intended for surveillance or human monitoring scenarios and was described as having: capability of real time operation; acceptable detection and tracking rate in low to medium complexity scenarios. Furthermore, it has features that enable human detection and the separation of individuals in a group.

In the original proposed solution, a shape model - an ellipsoid - is used to approximate the human shape, in the detection of people and in their discrimination when moving together or with partial occlusion. The shape model helps to minimise some problems introduced by segmentation, such as object split or boundary detection errors. Once an object, more precisely a person, is detected, an instance of the shape model is assigned and an appearance template is initialised; it consists of colour information. In subsequent frames, the shape model is used to indicate the possible area of the image corresponding to the person after a position estimation performed through the use of a Kalman filter. A scan is conducted in the uncertainty region centred on the estimated position, with a dissimilarity measure obtained in each iteration using information from the input image and segmentation, and from the appearance model.

The test framework has been prepared to accommodate the integration and test of different appearance models according to the description done in section 3.1. Object matching is performed by searching for the most similar instance in an uncertainty region. In each iteration, the appearance model instance receives the images, shape and position information; it computes and compares the object descriptors. The output is the corresponding dissimilarity measure that will be used by the tracking algorithm to define the best match; once it is found, information is passed to the module for model update. Video segmentation was performed using the algorithm proposed in [36]. The experiments were performed on a computer with an Intel(R) Core(TM) i5 CPU at 3.20GHz with 8GB of RAM.

4.2 Appearance Models

For our experiments, we have selected the following state-of-the-art description methods, using available implementations: SIFT; SURF; FIRST; HOG; colour histogram. For colour histogram, we used the proposal described in [18], where the bounding box is divided into an upper and lower part with a greater weight assigned to the upper part, thus adding more relevant spatial information to the model. For the local descriptors, the set of description vectors extracted from the region corresponding to an object was taken as model.

Variations for the appearance models were also tested in order to emphasise particular contributions thereof to the performance of the tracking algorithm, e.g., accuracy and consistency of human detection, computational complexity of appearance extraction and matching, adequate discrimination of humans when occlusions occur. Specifically:

- *Grid Points*: Commonly, SIFT, SURF and FIRST descriptors use a key point detector. In tracking, due to the reduced dimensions of the objects or their smoothness, the number of detected key points may be small. As an alternative to the automatic detection we also tested the use of a dense scan, where the points for the computation of the descriptors are indicated by an evenly spaced grid. We tested grids with a number of points equal to 1% and 4% of the object's image.

- *Bounding Box Scaling*: Typically, models are computed in the object's region defined by its bounding box. This region tends to include information from the background or even from other objects (Figure 3). We argue that the top and bottom of the bounding box (for a human in the upright position these are associated with the head and feet) are more likely to be subject to noise (e.g., from the segmentation) and more background points may be wrongly included into the object's model. Hence, we experimented with the variation of the size of the bounding box where the models are determined. Specifically, the bounding box's height was scaled down around the centre to a final BB height that is a percentage (90%, 75% and 50%) of the original BB. This concept is illustrated in Figure 3. We will refer to these variations as object patch, or simply BB, accompanied by a percentage, e.g., 'for a 90% object patch' or '90% BB', where the percentage is related to the full BB height in each instance.

Concluding, a total number of 49 appearance models were tested: 4 for colour histogram, 12 for SIFT, 12 for SURF, 4 for HOG, 16 for FIRST and the baseline (texture model proposed in [44]). These were applied to the 4 sequences (Section 3.3), adding up to 196 experiments.

(a) 100% BB (b) 90% BB (c) 75% BB

Fig. 3 Example of the height variation of the bounding box. In the left image, full bounding box is used; the bounding box is scaled down from left to right. (sample frame of the CAVIAR project [11])

4.3 Measuring Models Dissimilarity

Due to the nature of the description techniques, two different dissimilarity measures where used. Note that these measures are model specific, but are hidden from the overall tracking solution. For both colour histogram and HOG models, the straightforward and well-known normalised χ^2 distance was used as dissimilarity measure. For the key point based descriptors SIFT, SURF and FIRST, the ensemble of descriptors was adopted as the model; the descriptors were computed in an object's patch and in a possible instance of the object in the next frame, and a matching process was formulated. For two sets with N_1 and N_2 descriptors, and K matches, a dissimilarity measure is calculated. The simple use of the distances between matched descriptors can easily induce errors. As an example, consider the situation depicted in Figure 4 where two local descriptor based models are compared.

For object #1, six descriptors were computed (N_1), while for object #2 only four descriptors were calculated (N_2). From these, only two pairs were identified (K). Due to the small number of matched descriptors, one could consider these to be different objects. However, if only the distance of the two pairs is used, the dissimilarity may be low and the objects matched; this is particularly true if the two individual distances $(d1, d2)$ are small. Hence, the number of unmatched descriptors must also be taken into account. Toward this objective, different measures may be defined; we chose the formulation

$$D = \sum_{i=1}^{K} d_i + P_{max} \left[\max(N_1, N_2) - K \right] \tag{4}$$

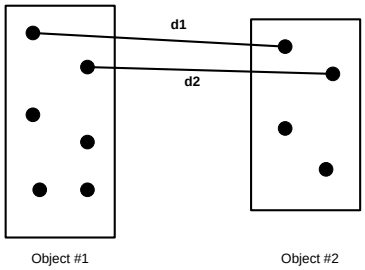


Fig. 4 Dissimilarity for local descriptor based models.

where d_i is the distance for the i^{th} match; the second component of the sum was introduced to penalise non-matched descriptors, where P_{max} is a constant greater than the maximum possible value for the distance between two descriptors. The measure is normalised by P_{max} in each iteration. In each case, the best match corresponds to the pair that minimises the dissimilarity.

5 Results

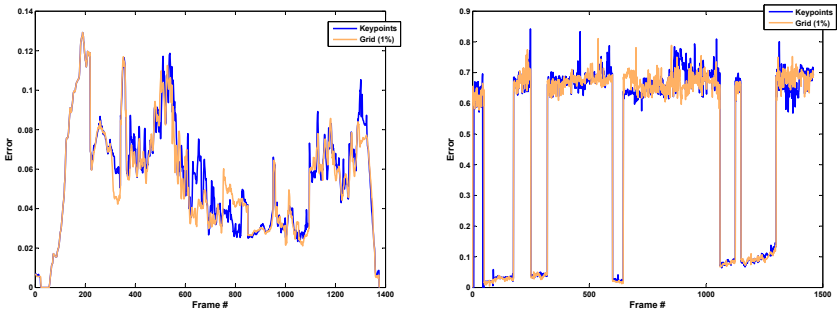
Given the large number of experiments conducted, we will adopt a phased approach intended to make the presentation of the results clearer. First, the local descriptor base models are compared with both key and grid point strategies. It follows the comparison of the histogram base descriptors. Finally, an overall comparison of all the models is performed.

With regards to the hybrid metric (Section 3.2), note that only a reduced number of frames with reference information exists for the PETS's sequence; this may prevent the identification of events if the corresponding frames are not encompassed in a sequence segment with frames containing reference information [10]. Since this metric tends to be 'noisier' due to the sparse GT (about 3% of the full sequence) and given the possible superimposition of several results, we chose to present only the hybrid metric for the PETS' sequences when justified by a relevant behaviour of the assessed methods.

In Table 1, it is compared the widely used SIFT and SURF descriptors, as well as the more recent FIRST descriptor. The results correspond to the object metrics (Section 3.2) for sequences OSOW1 and OSOW2 of the CAVIAR dataset; different heights of the object patches (in percentage of the object's full bounding box height) were considered (Section 4.2). Specifically we analyse the key point detectors and grid points in the computation of the models. Furthermore, given the characteristics and existing implementation of FIRST, we tested the use of different patch sizes in the calculation of these descriptor. The best results are highlighted per sequence and BB variation (i.e., per table column). Note that for the first two metrics, TRDR and DR, we wish to obtain high values, while for metric FAR low values are desirable. In the case of the FER the best result is 1, with increasing values traducing a performance

degradation. These experiments were also assessed with the hybrid metrics for the full dataset and the corresponding results are depicted in Figure 5 and Figure 6 for SIFT and SURF respectively.

From the observation of Table 1 a major conclusion is the best tracking results obtained with SURF using a grid of points in alternative to the key point detector. While the performance of the tracking algorithm with SURF descriptors increased by using a dense scan, with SIFT no significant advantage was observed. These individual behaviours can also be observed in Figure 5 and Figure 6 respectively. Given a superimposition of the values and for clarity of the graphs, only the values for full object's patch are used in this analysis. For SIFT, (Figure 5), little changes are visible between the use of key and grid points; when applied to the PETSs' sequences; slight differences are visible around frame 1500 (Figure 5(c) and Figure 5(d)) and frame 900 (for Figure 5(d)) where the error value is zero due to the misdetection of objects.



(a) CAVIAR - OSOW1 (b) CAVIAR - OSOW2

(c) PETS - ST1C3 (d) PETS - ST1C4

Fig. 5 Tracking error, using the hybrid metric, for SIFT descriptors with both key points and a grid of points using the full object patch.

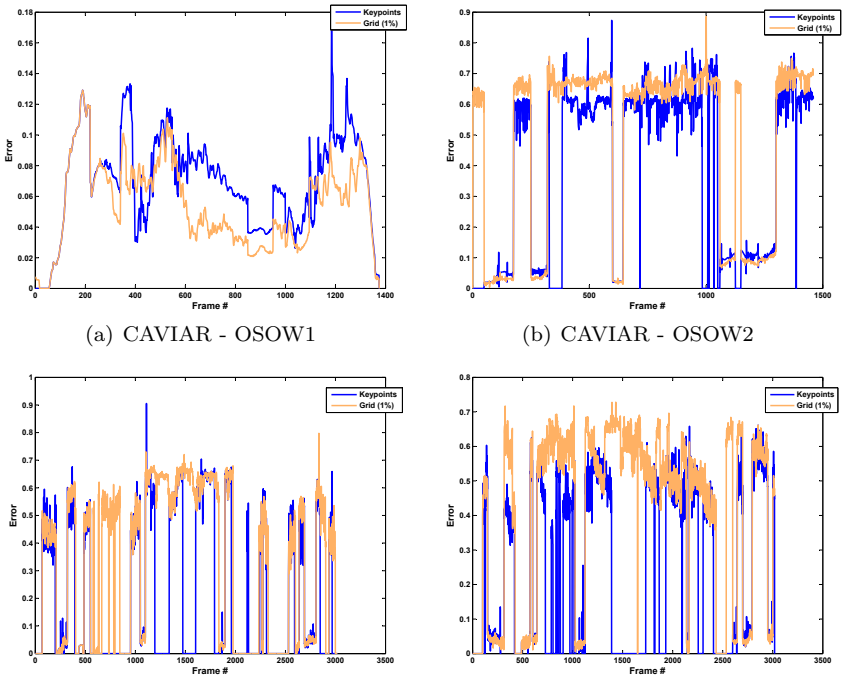
In Figure 7, we compare SIFT and SURF through the hybrid metric using the best methods for each descriptor: SIFT with key points; SURF with grid points. In both cases, full object patch was used. The results for the CAVIAR

Table 1 Assessment of tracking performance using object metrics. It summarises results for the use of a sparse (key points) or dense (grid) scan in the computation of SIFT, SURF and FIRST descriptors, and for different heights of the object patch.

		OSOW1				**OSOW2**			
		Bounding Box Height (%)				Bounding Box Height (%)			
	Metric	100	90	75	50	100	90	75	50
SIFT	TRDR	0.68	0.58	0.61	0.44	0.64	0.64	0.58	0.59
	DR	0.68	0.57	0.61	0.44	0.63	0.62	0.57	0.57
	FAR	0.10	0.12	0.12	0.26	0.23	0.25	0.32	0.24
	FER	2.57	2.71	3.43	4.71	3.38	4.12	4.50	6.33
SIFT Grid 1%	TRDR	0.56	0.51	0.60	0.53	0.64	0.66	0.66	**0.66**
	DR	0.55	0.51	0.60	0.53	0.63	0.65	0.64	**0.65**
	FAR	0.19	0.27	0.11	0.26	0.24	0.20	0.23	0.21
	FER	2.00	1.86	1.57	1.86	2.78	2.88	3.56	3.13
SIFT Grid 4%	TRDR	0.45	0.47	0.63	0.43	0.56	0.56	0.62	**0.66**
	DR	0.45	0.47	0.63	0.43	0.55	0.56	0.61	**0.65**
	FAR	0.34	0.36	0.15	0.38	0.39	0.36	0.34	0.29
	FER	1.43	1.71	1.71	1.57	3.11	2.50	2.67	2.78
SURF	TRDR	0.14	0.14	0.14	0.09	0.20	0.17	0.17	0.10
	DR	0.14	0.14	0.14	0.09	0.20	0.17	0.17	0.10
	FAR	0.08	**0.04**	0.12	0.52	0.21	0.23	0.22	0.50
	FER	1.57	1.57	1.86	1.57	4.25	4.25	4.13	4.33
SURF Grid 1%	TRDR	0.71	**0.72**	0.70	0.71	**0.68**	**0.70**	**0.72**	0.65
	DR	0.71	**0.71**	0.70	0.71	**0.66**	**0.68**	**0.71**	**0.65**
	FAR	**0.02**	0.09	0.10	**0.03**	0.21	0.18	0.16	0.19
	FER	**1.00**	1.29	**1.14**	1.57	2.25	1.88	2.56	3.13
SURF Grid 4%	TRDR	**0.78**	0.54	**0.72**	**0.75**	0.64	0.42	0.63	0.64
	DR	**0.75**	0.53	**0.72**	**0.75**	0.63	0.42	0.62	0.64
	FAR	0.12	0.36	0.09	0.05	0.37	0.59	0.38	0.36
	FER	1.71	1.14	1.43	**1.43**	**1.38**	**1.14**	**1.38**	2.00
FIRST 15x15 patch	TRDR	0.09	0.09	0.11	0.10	0.18	0.18	0.11	0.04
	DR	0.09	0.09	0.11	0.10	0.18	0.18	0.11	0.04
	FAR	0.15	0.13	**0.06**	0.18	**0.06**	0.09	**0.05**	**0.07**
	FER	2.00	1.33	1.00	1.67	4.33	3.50	3.67	**1.50**
FIRST 9x9 patch	TRDR	0.11	0.10	0.10	0.10	0.23	0.23	0.21	0.21
	DR	0.11	0.10	0.10	0.10	0.23	0.23	0.21	0.21
	FAR	0.09	0.19	0.20	0.15	0.08	**0.06**	0.21	0.15
	FER	1.33	**1.00**	1.00	1.00	2.67	2.50	3.00	5.00
FIRST Grid 1% 9x9 patch	TRDR	0.08	0.07	0.08	0.07	0.22	0.23	0.18	0.20
	DR	0.08	0.07	0.08	0.07	0.22	0.23	0.18	0.20
	FAR	0.39	0.43	0.35	0.38	0.09	0.11	0.12	0.12
	FER	2.00	3.00	1.33	1.00	2.33	2.50	3.33	2.67
FIRST Grid 1% 6x6 patch	TRDR	0.12	0.05	0.10	0.06	0.22	0.21	0.21	0.21
	DR	0.12	0.05	0.10	0.06	0.22	0.21	0.21	0.21
	FAR	0.09	0.58	0.33	0.55	0.07	0.12	0.15	0.11
	FER	1.33	**1.00**	1.33	1.00	2.00	1.83	2.33	2.50

sequences (Figure 7(a) and Figure 7(b)) are very similar, indicating SURF as a valid alternative to the heavier SIFT for tracking. For sequence ST1C3 (Figure 7(c)), the results are also alike, with the SIFT based solution failing to detect any object around frame 1700 and 2100. For sequence ST1C4 (Figure 7(d)), tracking with both methods does not detect objects around frame 1700, but with the SURF based solution failing for a longer period.

Unlike in the individual assessments, the tracking results using the FIRST descriptor in the appearance model were not competitive with regards to SIFT and SURF (see Table 1). Similar to SIFT, the use of a grid of points did not provide significant advantages. From our experiments, it is clear that a reduc-



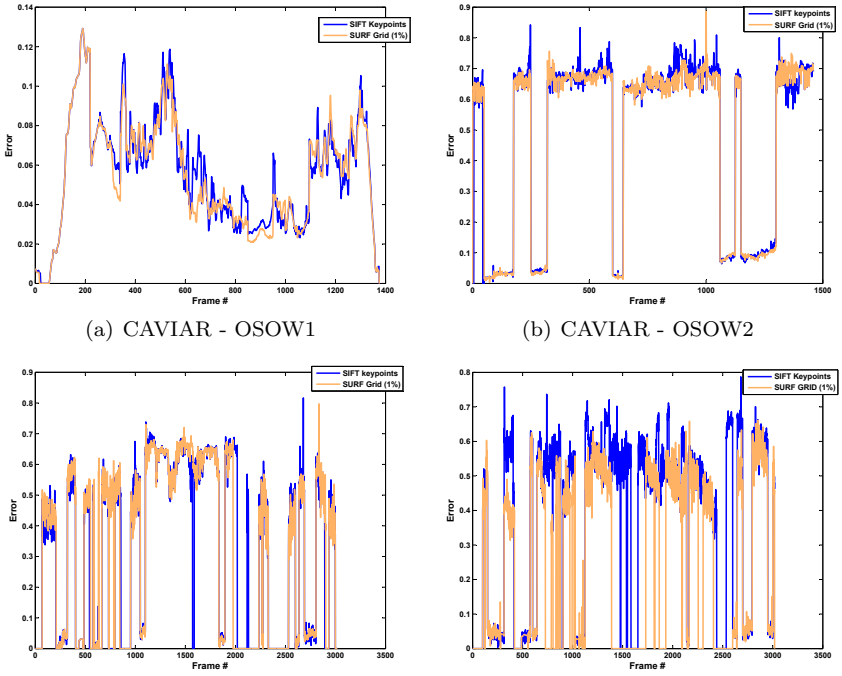
(a) CAVIAR - OSOW1 (b) CAVIAR - OSOW2

(c) PETS - ST1C3 (d) PETS - ST1C4

Fig. 6 Tracking error, using the hybrid metric, for SURF descriptors with both key points and a grid of points using the full object patch.

tion of the default patch size (region around a point used in the computation of the description vector) can lead to better results. It is noteworthy that modification of the patch size was a preliminary contribution of these experiments to the FIRST implementation.

The object metrics' values for the tracking results using colour histogram and HOG based models are summarised in Table 2. These descriptors are computed over the object's patch, thus a grid of points is not applicable in this case. It can be observed that HOG presented the best results with the exception of the fragmentation error; nevertheless, the differences of the methods with regards to this measure is small. Moreover, the results show that variations of the object patches are more noticeable when using the colour histogram model; while small object patches (50% of the bounding box height) tend to slightly deteriorate the performance, the use of the full bounding box does not lead to better results either. This is coherent with the argument that smaller heights for the object patch (and consequently its image) can limit the background noise added to the models, without forgetting the known limitations of these descriptors within small, possibly homogeneous, regions. The results for full



(a) CAVIAR - OSOW1 (b) CAVIAR - OSOW2

(c) PETS - ST1C3 (d) PETS - ST1C4

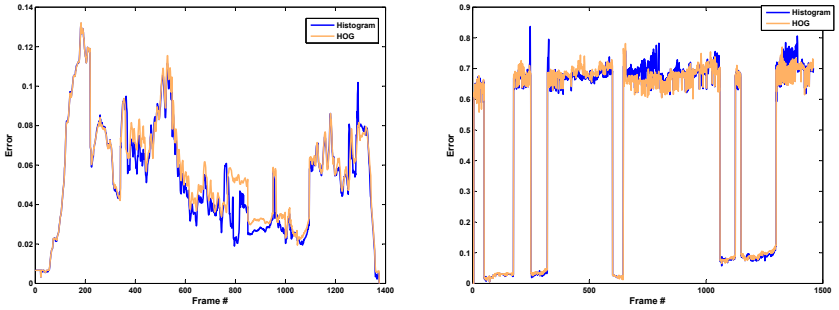
Fig. 7 Comparative analysis of using SIFT with key point detection and SURF with a grid of points (number of points equal to 1% of the object's patch pixels) over the test sequences.

Table 2 Assessment of tracking performance using object metrics. It summarises tracking results for the used HOG and histogram based models.

		OSOW1				**OSOW2**			
		Bounding Box Height (%)				Bounding Box Height (%)			
	Metric	100	90	75	50	100	90	75	50
Histogram	TRDR	0.56	0.73	0.72	0.54	0.68	0.69	0.66	0.67
	DR	0.56	0.73	0.72	0.54	0.67	0.69	0.66	0.67
	FAR	0.32	**0.08**	**0.08**	0.33	0.27	0.28	**0.28**	0.30
	FER	**1.00**	**1.29**	**1.29**	1.29	**1.29**	**1.25**	**1.22**	**1.11**
HOG	TRDR	**0.77**	**0.75**	**0.79**	**0.60**	**0.74**	**0.72**	**0.72**	**0.76**
	DR	**0.76**	**0.74**	**0.78**	**0.59**	**0.73**	**0.72**	**0.70**	**0.73**
	FAR	**0.09**	0.13	0.14	**0.27**	**0.24**	**0.25**	**0.28**	**0.25**
	FER	1.29	1.57	**1.29**	**1.14**	1.87	1.62	1.50	1.56

BB conveyed by the hybrid metric are represented in Figure 8 and show that no method clearly surpassed the other in terms of tracking performance.

Table 3 summarises the results of all experiments. With regards to the BB variations we selected the best results for each model. The best results per metric are highlighted. It is noticeable that the assessment approach proposed in this paper enables the comparison of the different methods and variations. It can be observed that HOG, color histogram and SURF with grid points en-



(a) CAVIAR - OSOW1 (b) CAVIAR - OSOW2

Fig. 8 Comparative analysis of colour histogram and HOG based appearance models for the CAVIAR sequences.

abled the best tracking performances, but without a clear dominant technique. Furthermore, with these methods, the tracking solution performed better than with the originally proposed model (using texture [44]).

Table 4 presents the time measures described in Section 3.2 for the most relevant experiments corresponding to the presented tracking performance results. As expected, tracking using colour histogram and HOG exhibited a good computational performance. Even though the time for descriptor extraction and matching for HOG was slightly higher than with color histogram, the overall processing time was smaller due to less iterations in object matching; this is a clear example of the interdependencies between the different modules of the tracking solution. As expected, SIFT implied a high computational complexity, the highest for our experiments. An interesting observation is the significant decrease of the computational time through the use of a grid of points and without a degradation of the tracking results. SURF enabled low computational time values, but as already stated the tracking results were poorer. However, it is noteworthy the time values obtained for SURF with a grid of points; although they are higher than with the use of key point detection, they remain smaller than with SIFT and it was demonstrated in Table 3 that they enable better tracking results.

6 Conclusions

In this paper, we proposed a new paradigm for the assessment of appearance models and corresponding description techniques. Recognising the difference between tracking and standalone image or object matching, it was considered logic to analyse the different description models from a tracking point of view. In this novel approach, different description techniques or models are assessed through their integration in a common tracking solution with the assessment performed on the tracking output. While the current proposal has people track-

Table 3 Overall comparison of the experimented models. For grid based representations, the best results are highlighted for each type of descriptor.

		Appearance Model Comparison												
	Metric	SIFT	SIFT Grid (1%)	SIFT Grid (4%)	SURF	SURF Grid (1%)	SURF Grid (4%)	FIRST (15x15)	FIRST (9x9)	FIRST Grid (9x9)	FIRST Grid (6x6)	Histogram	HOG	Texture
OSOW1	TRDR	0.68	0.60	0.63	0.14	0.71	0.78	0.11	0.11	0.08	0.12	0.73	**0.79**	0,50
	DR	0.68	0.60	0.63	0.14	0.71	0.75	0.11	0.11	0.08	0.12	0.73	**0.78**	0,50
	FAR	0.10	0.11	0.15	0.04	**0.02**	0.12	0.06	0.09	0.35	0.09	0.08	0.14	0,35
	FER	2.57	1.57	1.71	1.57	**1.00**	1.71	1.00	1.33	1.33	1.33	1.29	1.29	1,33
OSOW2	TRDR	0.64	0.66	0.66	0.20	0.72	0.64	0.18	0.23	0.23	0.22	0.69	**0.76**	0,58
	DR	0.63	0.65	0.65	0.20	0.71	0.63	0.18	0.23	0.23	0.22	0.69	**0.73**	0,57
	FAR	0.23	0.20	0.29	0.21	**0.16**	0.37	0.06	0.06	0.04	0.07	0.28	0.25	0,37
	FER	3.38	2.88	2.78	4.25	2.56	1.38	4.33	2.50	2.50	2.00	**1.25**	1.56	2,50

Table 4 Comparison of the processing times for the overall solution and for individual components of the models: extraction and matching.

	OSOW1			OSOW2			ST1C3			ST1C4		
	SPT (secs/ frame)	DET (secs/ track)	DMT (secs/ track)	SPT (secs/ frame)	DET (secs/ track)	DMT (secs/ track)	SPT (secs/ frame)	DET (secs/ track)	DMT (secs/ track)	SPT (secs/ frame)	DET (secs/ track)	DMT (secs/ track)
SIFT	0.420	0.175	0.000	0.678	0.126	0.000	0.800	0.559	0.003	1.150	0.420	0.001
SIFT Grid (1%)	0.242	0.092	0.000	0.427	0.070	0.000	0.416	0.239	0.000	0.565	0.175	0.000
SURF	0.033	0.002	0.000	0.061	0.002	0.000	0.055	0.014	0.000	0.092	0.012	0.000
SURF Grid (1%)	0.194	0.074	0.002	0.316	0.036	0.001	0.354	0.202	0.010	0.509	0.151	0.007
FIRST (15x15)	0.272	0.229	0.002	0.548	0.113	0.000	0.749	0.756	0.001	1.389	0.728	0.002
FIRST Grid (9x9)	0.277	0.212	0.001	0.602	0.149	0.000	0.851	0.733	0.001	1.447	0.716	0.001
Histogram	0.213	0.000	0.000	0.351	0.000	0.000	0.190	0.000	0.000	0.292	0.000	0.000
HOG	0.052	0.007	0.001	0.090	0.008	0.001	0.058	0.010	0.002	0.090	0.009	0.001
Texture	0.084	0.000	0.000	0.125	0.000	0.000	0.177	0.000	0.000	0.199	0.000	0.000

ing as application scenario, the underlying concepts can be seamlessly applied to other tracking contexts. In addition to the assessment paradigm description, a set of metrics are suggested towards the goal of a complete and flexible evaluation.

For proof of concept, we assessed the computational performance and accuracy of the results for a set of well known description techniques and different representations. Furthermore, some ideas for appearance models were put forward and the first application and analysis of the FIRST descriptors to video object tracking was made. From this work it naturally derived the set up of foundations for the future test of other techniques. Although the choice of the algorithm can be argued considering the vast number of published work, such discussion has no effect in the current context since all representations methods are assessed in a common framework under the same conditions; it is the impact of the individual models that we wish to evaluate.

A major result of the experiments conducted as proof of concept is the ability of a uniform assessment of heterogeneous techniques. Moreover, the results depicted different behaviours from those observed in individual evaluations. For example, HOG and SURF with grid points performed better than colour histogram and SIFT respectively. These results support the proposal of this paper.

Tracking with the SIFT base models achieved competitive accuracy, but at the expense of a higher computational weight. While the use of a dense scan in alternative to key points proved negligible in terms of robustness, it enabled a significant reduction of the computational performance - similar results were achieved in less time. An opposite behaviour was exhibited by SURF since its computational performance was significantly better than SIFT, where the tracking results were degraded due to the reduced dimensions and homogeneity of the objects, which caused a small number of detected key points. The use of dense scan greatly improved the tracking results surpassing, in several cases, the other methods. This is indicative that using a dense scan over the object's patch can offer a more powerful representation than the sparse key points. In small image patches with possibly smooth regions the number of detected key points may be very small or even null. By using a grid of specified points this problem can be minimised enabling better tracking results. It is noteworthy that, even with the use of a grid of points, tracking using SURF still exhibited a better computational performance than with SIFT. While SIFT and FIRST formulations rely on 1^{st} order derivatives to discard low contrast key points located on eigenvalues to create a robust descriptor, SURF finds key points on 2^{nd} order mixed derivatives using DoH, which is sensitive to several perturbations such as noise. These differences may define why only SURF clearly improves TRDR using a regular grid of key points, since alternative approaches describe more specific image regions, thus failing in a regular grid.

The results obtained with FIRST illustrate advantages of the proposed assessment paradigm. While individual evaluations showed that the FIRST descriptors are faster than SIFT and SURF, the results of these experiments show a different behavior due to interdependencies of the modules. Tracks are

only considered if they remain stable for a minimum number of frames; the difficulties in establishing object correspondences between consecutive frames reflects in an increased uncertainty (conveyed by the TRDR and DR measures) of the overall process implying the initialization of more (temporary) tracks and a greater number of iterations during the matching step. Regarding the several configurations tested in FIRST, none of them led to drastic changes. Reducing the patch size slightly improved the results, while reducing the BB causes an opposite behavior. This motivates a more in-depth study of the parameterization influence on the overall tracking.

As expected from the literature, tracking using colour histograms and HOG based appearance models exhibited good performances. In particular, the results for colour 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. Nevertheless, the integrated assessment of these techniques indicated a superiority of the HOG base models in both tracking robustness and overall processing time.

There was a decision to use sequences widely known to the research community, focusing on surveillance scenarios, and a solution designed for people tracking. However, the reasoning underlying this assessment proposal can and should be extended to other scenarios and description methods. Given the additional challenges placed by multi-camera scenarios, e.g., differences in scale and color, it would be interesting to perform the analysis proposed in this paper in multi-camera scenarios. This can result in more complete information regarding the description methods robustness and the feasibility of their use for both single and multiple cameras. Even though interest results were obtained from our proof of concept, a major outcome is the setup for future experiments not only at level of the appearance model, but also in the other modules of a tracking solution, e.g., object detection.

References

1. Alahi, A., Vandergheynst, P., Bierlaire, M., Kunt, M.: Cascade of descriptors to detect and track objects across any network of cameras. Computer Vision and Image Understanding **114**, 624–640 (2010)
2. Bashir, F., Porikli, F.: Performance evaluation of object detection and tracking systems. In: Proceedings of IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), PETS 2006. (2006)
3. Bastos, R., Dias, M.S.: FIRST - Fast Invariant to Rotation and Scale Transform: Invariant image features for augmented reality and computer vision. In: VDM Verlag (2009)
4. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). Computer Vision and Image Understanding **110**, 346–359 (2008)
5. Black, J., Ellis, T., Rosin, P.: A novel method for video tracking performance evaluation. In: In Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), pp. 125–132 (2003)
6. Bradski, G.R.: Computer vision face tracking for use in a perceptual user interface. Intel Technology Journal (Q2) (1998)
7. Brown, M., Lowe, D.: Invariant features from interest point groups. In: In British Machine Vision Conference, pp. 656–665 (2002)

8. Cardoso, J.S., Carvalho, P., Teixeira, L.F., Corte-Real, L.: Partition-distance methods for assessing spatial segmentations of images and videos. Computer Vision and Image Understanding **113**(7), 811–823 (2009)
9. Carvalho, P., Cardoso, J.S., Corte-Real, L.: Hybrid framework for evaluating video object tracking algorithms. Electronics Letters **46**(6), 411–412 (2010). URL http://www.inescporto.pt/ jsc/publications/journals/2010PCarvalhoIET.pdf
10. Carvalho, P., Cardoso, J.S., Corte-Real, L.: Filling the gap in quality assessment of video object tracking. Image and Vision Computing **30**(9), 630–640 (2012). DOI 10.1016/j.imavis.2012.06.002
11. Caviar: Ec-funded-caviar-project, i. 2001-37540 (2004). URL http://homepages.inf.ed.ac.uk/rbf/CAVIAR/
12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01, CVPR '05, pp. 886–893. IEEE Computer Society, Washington, DC, USA (2005)
13. Denman, S., Fookes, C., Sridharan, S., Lakemond, R.: Dynamic performance measures for object tracking systems. In: Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS '09, pp. 541–546. IEEE Computer Society, Washington, DC, USA (2009)
14. Ellis, T.: Performance metrics and methods for tracking in surveillance. In: 3rd IEEE International Workshop on Performance Evaluation of Tracking and Surveillance PETS'2002. Copenhagen, Denmark (2002)
15. Çiğdem Eroğlu Erdem, Sankur, B., Tekalp, A.M.: Performance measures for video object segmentation and tracking. IEEE Transactions on Image Processing **13**(7), 937–951 (2004)
16. Han, Z., Ye, Q., Jiao, J.: Combined feature evaluation for adaptive visual object tracking. Computer Vision and Image Understanding **115**, 69–80 (2011)
17. Jiang, Y.G., Yang, J., Ngo, C.W., Hauptmann, A.G.: Representations of Keypoint-Based semantic concept detection: A comprehensive study. Multimedia, IEEE Transactions on **12**(1), 42–53 (2009)
18. Jiang, Z., Huynh, D.Q., Moran, W., Challa, S., Spadaccini, N.: Multiple pedestrian tracking using colour and motion models. Digital Image Computing: Techniques and Applications **0**, 328–334 (2010)
19. Ke, Y., Sukthankar, R.: PCA-SIFT: a more distinctive representation for local image descriptors. In: Proceedings of the 2004 IEEE computer society conference on Computer vision and pattern recognition, CVPR'04, pp. 506–513. IEEE Computer Society, Washington, DC, USA (2004)
20. Lazarevic-McManus, N., Renno, J.R., Makris, D., Jones, G.A.: An object-based comparative methodology for motion detection based on the F-Measure. Computer Vision and Image Understanding **111**(1), 74–85 (2008)
21. List, T., Fisher, R.B.: CVML - an XML-based Computer Vision Markup Language. In: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 1 - Volume 01, ICPR '04, pp. 789–792. IEEE Computer Society, Washington, DC, USA (2004)
22. Liu, H., Yu, Z., Zha, H., Zou, Y., Zhang, L.: Robust human tracking based on multi-cue integration and mean-shift. Pattern Recognition Letters **30** (2009)
23. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
24. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. International Journal of Computer Vision **60**, 63–86 (2004)
25. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence **2**(10), 1615–1630 (2005)
26. Moeslund, T.B., Granum, E.: A survey of computer vision-based human motion capture. Computer Vision and Image Understanding **81**, 231–268 (2001)
27. Nghiem, A.T., Bremond, F., Thonnat, M., Valentin, V.: Etiseo, performance evaluation for video surveillance systems. In: Proceedings of the 2007 IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 476–481. IEEE Computer Society, Washington, DC, USA (2007)

28. Opelt, A., Pinz, A., Zisserman, A.: Incremental learning of object detectors using a visual shape alphabet. Computer Vision and Pattern Recognition, IEEE Computer Society Conference on **1**, 3–10 (2006)
29. PETS: IEEE international workshop on performance evaluation of tracking and surveillance 2006 (2006). URL http://www.cvg.rdg.ac.uk/PETS2006/index.html
30. Schlogl, T., Beleznai, C., Winter, M., Bischof, H.: Performance evaluation metrics for motion detection and tracking. In: ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 4, pp. 519–522. IEEE Computer Society, Washington, DC, USA (2004)
31. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. International Journal of Computer Vision **37**, 151–172 (2000)
32. Shahed, S.M.N., Ho, J., Yang, M.H.: Online visual tracking with histograms and articulating blocks. Computer Vision and Image Understanding **114**(8), 901–914 (2010)
33. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their localization in images. In: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 - Volume 01, pp. 370–377. IEEE Computer Society, Washington, DC, USA (2005)
34. Szeliski, R.: Computer Vision : Algorithms and Applications. Springer-Verlag New York Inc (2010)
35. Tang, F., Tao, H.: Object tracking with dynamic feature graph. In: ICCCN '05: Proceedings of the 14th International Conference on Computer Communications and Networks, pp. 25–32 (2005)
36. Teixeira, L.F., Cardoso, J.S., Corte-Real, L.: Object segmentation using background modelling and cascaded change detection. Journal of Multimedia (JMM) **2**, 55–65 (2007)
37. Tell, D., Carlsson, S.: Combining appearance and topology for wide baseline matching. In: Proceedings of the 7th European Conference on Computer Vision-Part I, ECCV '02, pp. 68–81. Springer-Verlag, London, UK (2002)
38. Tissainayagam, P., Suter, D.: Assessing the performance of corner detectors for point feature tracking applications. Image and Vision Computing **22**, 663–679 (2004)
39. Venetianer, P.L., Deng, H.: Performance evaluation of an intelligent video surveillance system - a case study. Computer Vision and Image Understanding **114**, 1292–1302 (2010)
40. Vizireanu, D.N.: Generalizations of binary morphological shape decomposition. Journal of Electronic Imaging **16**, 013,002 (2007)
41. Vizireanu, N., Halunga, S., Marghescu, G.: Morphological skeleton decomposition interframe interpolation method. Journal of Electronic Imaging **19**, 023,018 (2010)
42. Wu, L., Hu, Y., Li, M., Yu, N., Hua, X.S.: Scale-invariant visual language modeling for object categorization. IEEE Transactions on Multimedia **11**, 286–294 (2009)
43. Zhao, T.: Model-based segmentation and tracking of multiple humans in complex situations. Ph.D. thesis, Faculty of the Graduate School of the University of Southern California (2004)
44. Zhao, T., Nevatia, R.: Tracking multiple humans in complex situations. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(9), 1208–1211 (2004)
45. Zhou, H., Yuan, Y., Shi, C.: Object tracking using SIFT features and mean shift. Computer Vision and Image Understanding **113**, 345–352 (2009)