# Object Detection and Tracking from Fixed and Mobile Platforms

Giovanni B. Garibotto[1] and Francesco Buemi[2(✉)]

[1] Vision technology consultant, Genova, Italy
giovanni.garibotto@gmail.com
[2] Aitek S.p.A., Genova, Italy
francesco.buemi@aitek.it

**Abstract.** Computer Vision technology plays a fundamental role in Video Surveillance applications with the possibility to detect different categories of objects (human beings, faces, vehicles, car plates etc) in a regular stream of video recorded by surveillance cameras. Moreover, the detection process must be validated for a sufficiently long time interval (by tracking), to provide more instances of the same object/subject, and increase the rate of successful recognition/identification (including the possibility of human supervision). The paper address the problem of object detection and tracking and the proposed solution is based on visual appearance model learning during the tracking process. Simplified HOG-like texture features are used, to achieve computationally effective solutions to be applied in practical applications of video analytics. A contrast gradient normalization solution has been adopted, with adaptive threshold estimation, to increase tracking capability along the video flow. Performance of the tracking processing chain is evaluated using the public available TLD dataset [1], to achieve quantitative and comparable data.

**Keywords:** Object detection · Visual tracking · Appearance model learning · Adaptive processing

## 1 Introduction

Object detection and tracking represents a fundamental processing step in any implementation of video analysis. In security applications it is used to collect multiple views of the same object as a support to classification and recognition. It represents a topic of great interest in the vision community with many publications and research studies. A recent tutorial paper [2] describes the state of the art of this technology and the most promising lines of research under investigation.

Our research has been developed within the context of industrial video surveillance to improve the performance of Aitek Video Analytics solution [3] with the primary objective to deal with real every-day life constraints, like poor video quality and the availability of low processing power in embedded camera applications. We intend to face the problem of generic object detection and tracking from fixed cameras (with possible calibration tools) as well as from mobile platforms, including PTZ sensors


V. Murino and E. Puppo (Eds.): ICIAP 2015, Part II, LNCS 9280, pp. 631–642, 2015.
DOI: 10.1007/978-3-319-23234-8_58

and surveillance cameras installed on-board of security vehicles and police cars. The driving line of our research study comes from a security application of supervised detection and tracking of humans and vehicles from video recorded sequences. The final objective is hiding-blurring some selected targets, to prevent their identification in the output video stream (privacy constraints). This operation is performed in a supervisory mode (man-in-the-loop), with a selection of the initial position, and the bounding box, of the target object (it may be the face, the full body of a person, a car or a plate in the scene). The tracking process should be able to follow the target, trying to manage most of its variability including light, scale, translation and 3D pose. The initial information alone is often not sufficient to track the target for long, which means increasing the cost of human supervisor involvement, to select again the target for a next step of the tracking process. The best solution will be a fully autonomous process, able to learn different appearance models along the processing chain and keep track of the selected target as long as possible without human intervention. These are typical specifications of most advanced tracking solutions in the computer vision community, where tracking performance has significantly improved in the last decade.

Among the wide scientific literature on the subject we refer to the most recent results obtained in [4], marketed as "Predator", which exhibits superior performance level as compared to other tracking-by-detection methods. The TLD framework combines together a detector and an optical flow tracker (based on Lucas Kanade technique [5]). The purpose of the detector is to prevent the tracker from drifting away from the object as well as to recover tracking after temporary occlusions. Updates of the learning classifier are accepted only if the discovered image patch is similar to the initial object box, which represents the only prior information about the object.

In the paper the tracking task is addressed as a search problem, at the object level, where the appearance model of the object is searched in the following frames of the video sequence, to recover a consistent trajectory with time and space constraints. The main objective is to demonstrate that a model matching approach, where the appearance model of the object is continuously updated along the tracking process, can be successful in object tracking with competitive performance w.r.t. the most advanced solutions. The choice of a continuous model update and matching process, instead of classical optical flow analysis [6], is motivated by the noisy nature of most commonly used video surveillance data, the possible occurrence of sharp motion displacements between consecutive frames, and the irregularity of intensity image differences at pixel level, mainly due to motion blurring. In [7] HOG based feature descriptors have been already used to deal with large displacement motion, but the research objective, in that case, was to recover an estimation of the continuous optical flow.

In our approach, after an early detection phase of the candidate object (a rectangular patch in the image), there is a continuous loop of tracking search (i.e. the identification of the most likely object patch in the following frames of the video sequence), followed by a learning step, with a continuous update of the target model to be used in the detection and re-initialization phase.

In section 2 we discuss the appearance model which is based on a modification of the widely used texture HOG features. Section 3 describes the forward-backward tracking process and the adopted acceptance-rejection criteria. The process of object appearance model updating is referred in section 4. In spite of the apparent simplicity of the process, the achieved results are quite encouraging, as briefly discussed in section 5, using some of the video sequences collected in [1]. Then, a critical discussion is devoted to the sensitivity to processing parameters and the stability and robustness of the results.

## 2 HOG Features and the Matching Process

The descriptive features of the both tracking and detection models are selected as a collection of histograms of oriented gradients, HOG, following the basic approach as proposed in [8], with the introduction of some minor modifications. The normalization proposed in [8] was found not appropriate when dealing with a very limited number of learning examples, as it happens in the implementation of an on-line tracking process. As such, we have modified the HOG feature representation as follows:

- The object model is divided in a predefined number of cells (NC); the cell-size (Cx, Cy) has a squared aspect ratio (Cx = Cy) and is not fixed a-priori; rather, it depends on the overall size of the selected sample patch (scale factor of the initial model). The resulting appearance-model size (Nx*Cx, Ny*Cy) may differ by a few pixels from the initial selection, due to cell-size quantization effects.
- Cells are not grouped in blocks, and there is no block overlapping; the reason of this choice is to improve specialization capability over the generalization properties of standard HOG models, since we are not planning to use an SVM classifier. Moreover, the 4 corner cells are not considered in the feature representation, to reduce background interference in the tracking process. This choice does not affect the performance even in case of box-like shapes like cars or trucks, since the corner cells do not contain useful information due to projected shape onto the image plane.
- 1-D local histograms of gradient directions are computed for each cell, following the standard scheme of [8]. As usual, each pixel in the cell casts a weighted vote for an orientation, within a pre-defined range of discrete angles. The weight is the gradient magnitude itself, above a suitable noise threshold (as discussed in the following subsection). In our implementation we are using signed-gradient representation, with 8 direction channels in the local 3x3 neighbourhood, for maximum computational efficiency. No histogram normalization is performed at cell-level,
- The local histograms of the different cells are collected together in lexicographic order to obtain a vector of features. This vector is normalized to the overall sum of all features to achieve histogram-like properties (unitary integral value) and the similarity matching between the image patch hypothesis and the running appearance model is computed using a standard Bhattacharyya distance [www.opencv.org].

### 2.1 Gradient Contrast Normalization

Gradient based features are strongly affected by the contrast variability of the video sequence. For that reason it is always necessary to perform some kind of gradient normalization. In our implementation the spatial gradient function is computed on the luminance component of the video signal, as briefly summarized in the following.

***Local Contrast Gradient.*** Spatial derivatives are computed using a Sobel mask in a 3 x 3 neighbourhood and they are normalized to the average grey intensity (in the same local window). In this way it is possible to achieve a better contrast normalization across the whole image frame. At this level, signed-edge orientations are also computed, on the same local window (3 x 3) in the quantized range of 8 directions (0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°). Tracking performance is not much affected by this approximation, as shown in the experimental results of section 5.

***Dynamic Range Expansion.*** An additional normalization is performed on the computed contrast gradient in order to fill the available dynamic range of the signal (0-255 level) for each frame in the video sequence. As such, even relevant contrast variations during time can be successfully managed by the tracking process.

***Adaptive Threshold Estimation.*** Instead of using a fixed threshold of the contrast-gradient magnitude, an adaptive value is computed to track signal changes along the video flow. The histogram distribution of the image gradient magnitude can be roughly approximated by an exponential decreasing function, and a good estimate of a noise threshold can be computed as the median value of this function.

## 3 Tracking as a Search Problem

Due to the presence of noise and possible large displacements of the object trajectory, a classical optical flow approach [6], based on time differences was found not suitable. Moreover, it is necessary to perform a robust and early detection of any deviation from the right object tracking, to avoid unrecoverable drifts.

The adopted solution is a *forward-backward* implementation of a HOG-feature tracker. The best estimate of the target position in the other frame of the video sequence is computed by searching, in a pre-defined search space, centred on the previous target position, the target patch which exhibits the best matching score with the current object features. Starting from object $O_{t-1}$ (at time t-1) we obtain an estimate of the target position $O_t$ (in the frame at time t). The same search procedure is repeated in the backward direction from target $O_t$ to achieve an estimate $\hat{O}_{t-1}$ (in the past frame at time t-1). The matching overlap measure IoU (intersection over Union) is computed between the starting object $O_{t-1}$ and the back-projected object $\hat{O}_{t-1}$. Any drop of the IoU measure is a clear hint of failure which determines the activation of a local detector to re-instantiate the tracking process. As mentioned before, the similarity measure between two HOG feature vectors $\{H_v(n)\}$ and $\{H_w(n)\}$ is the Bhattacharyya distance dB and is computed as:

$$dB = ( 1 - cb) \tag{1}$$

where

$$cb = [\sum_{n \in (0, NF)} (H_v(n) * H_w(n))^{1/2}] / (\sum_{n \in (0, NF)} H_v(n) * \sum_{n \in (0, NF)} H_w(n))^{1/2} \quad (2)$$

NF is the length of the overall feature vector. This distance measure is defined in the range ( 0 – 1) where zero means full match and 0.5 is already representative of large dissimilarities between the two descriptive vectors. To achieve a processing speed-up, each tracking step is performed in a hierarchical way.

- A first ***Search space bucketing*** is performed, by segmenting the search space into non-overlapping blocks having the same size of the target cells (Cx, Cy) at the current tracking step (reference scale factor). Local 1-D histogram features are computed for each bucket and are stored in a working memory {W}. The best-matching target position (minimum distance (1)) is computed by a sliding-window scanning of the search space, with integer steps (Cx, Cy). The computation of the normalized HOG descriptive feature vectors is based on the stored values {W} with significant computational savings. During the tracking process this search area around the previous target object is extended of a few cells in all directions (3 cells in the experimental tests of section 5).
- ***A refinement step*** is repeated in the neighbourhood of the best match with decreasing sampling step (half size at each step until unit shift), by tracking always the minimum score target. During this phase HOG features are fully computed on the contrast-gradient map for each new window shift (without using the stored values in {W}), although some kind of interpolation might be possible [9], with further potential improvements.
- Finally, a ***scale search*** is performed, around the minimum-score target, to check if a zoomed version of it (± 1 in the cell size) may exhibit a better match (i.e. a lower distance score) with the previous object model. In this way it is possible to track expansion/compression effects of the target-object in the video stream. This approximation is quite acceptable due to the short time distance between consecutive frames.

### 3.1 Acceptance/Rejection Criteria

In any visual tracking implementation it is extremely important to manage the risk of drift from the real trajectory. An early and reliable missed-target detection is needed, to stop the tracking process and activate a new boot-strap detection procedure on a wider search area of the image. In our approach the following parameters are used to classify a positive matching candidate:

- The IoU measure between objects $O_{t-1}$ and $\hat{O}_{t-1}$ in forward-backward tracking is a already strong and reliable measure of tracking failure. In case of success it should be very close to 1. Smaller values below 98% are already representative of possible drifts which cannot be corrected by a simple displacement of the predicted object coordinates.
- A grey level image patch is collected for each output box of the tracked object in the frame sequence. It is obtained as a projection of the image pixels onto a small size box (typically 20 x 20 pixels) which is normalized to the maximum dynamic

range and is masked to reduce the peripheral border effects of the background. A Normalized Cross Correlation (NCC) measure is computed between such image box and the reference image patch which has been acquired from the initial target object at the beginning of the tracking process. When the NCC value falls below a confidence level (in the experimental results such threshold was 0.9) the tracking process is considered to fail and a detection search is issued. Since the starting image patch is quickly evolving and modifying during the video flow (due to light changes and 2D - 3D shape variations) and the NCC is a rigid similarity measure, a buffer of image patches is stored and updated for each successful tracking (high IoU of the forward-backward estimates), to keep track of the object evolution.

- Additional structural constraints are used like the distance to the border of the image (the selected processing region) and the tracking displacement from frame to frame of the target trajectory; it must be smaller than a threshold maximum value, which is context dependent (either fixed or mobile platforms).

If the previous constraints are not satisfied, a detection search has been implemented, to manage temporarily occlusions or disappearance of the target and re-initialize the tracking process. Ultimately, if the previous conditions persist, tracking is definitely stopped and a new start-object-detection process is issued (manual or automatic procedures)

## 4 Object Detection by HOG Model Appearance Updating

The use of an adaptive appearance model, which evolves during the tracking process, has proved to be the right choice in many applications of visual tracking [10]. In our approach, due to the choice of the descriptive HOG feature vector {Hv}, such model updating is computed as

$$Av_i = \lambda\, Hv + (1 - \lambda)\, Av_{i-1} \tag{4}$$

where $\{Av_i\}$ represents the appearance model at step (i); the parameter $\lambda$ represents a learning-rate of the new validated tracking features {Hv}. $\lambda = 0.5$ has been selected in all experimental results referred in section 5. This short-term estimate of the appearance model gives more weight to the last validated instance of the tracked object and it ensures a good continuity of the tracking process.

As mentioned before, the detection process is activated only when the continuous forward-backward tracking fails. Beside the appearance model update, as in (4), the average size of the target is estimated during the phase of successful tracking, as well as the predicted area of activity. This information provides a useful constraint for the search space and scale of the detection. The detector implementation follows a scheme quite similar to the one-step tracking process as before.

- The search region is defined around the target missing area, and will be expanded progressively in the following video frames, up to a maximum selected size. The range of search scales (cell size) are defined according to the tracking history

- For each selected scale a search region bucketing is performed, with non-overlapping blocks having the same size of the target cells (Cx, Cy) and the histogram features are stored.
- The best target hypothesis is computed again in two consecutive phases: the first one by a sliding window scanning of the search region with integer steps (Cx, Cy) on the stored feature cells, and the second refinement step up to the pixel level.

Once the minimum distance target has been found at multiple scales, a suitable acceptance-rejection criterion is needed. Since we are using positive models only, we must establish a threshold of the distance measure (1). In all examples of section 5 a lower threshold of 0.2 has been used. In our reference video privacy application the presence of the human supervisor plays an important role to fix any possible matching error and reduce the risk of error propagation.

### 4.1 Boot-Strap and First Model Selection

The first instance of the object-target in the scene has particular relevance for the tracking process, since it provides the necessary information to build the first instance of the appearance model. In the experimental results of section 5 the initial object patch is always taken from the ground-truth list of the dataset [1]. In video security applications the automatic detection of the first object instance in the scene is often based on a selected object category (human body, head shape, car-plate, size/type of vehicle, etc.).

In this domain human detection has become a quite consolidated line of research with many contributions mainly based on the use of generalized HOG models [8], [9] [11]. When dealing with fixed surveillance cameras it is possible to take advantage of foreground-background detection to focus the attention on the foreground blobs only, like in security applications of Video Analytics technology [3]. Moreover, using camera calibration constraints, there is an additional possibility to reduce the search space, by prediction of the object size and scale [12].

## 5 Experimental Results

This section is devoted to the evaluation of our approach, by comparing our experimental results with reference solutions from the scientific literature. From dataset [1] we have selected 3 test video sequences, namely *david*, *jumping* and *car*, which seem more related to our reference application and quite representative of challenging motion conditions for a tracking task.

### 5.1 Dataset and Performance Evaluation

For each video sequence, the dataset contains a list of ground-truth box coordinates, as well as additional lists of bounding box results, obtained by the different algorithms

which have been tested in that challenge [1]. The comparative results referred in table 1 are based on such data lists, considering only the best solution in that dataset, namely TLD1.0 [1].

**Table 1.** Comparison of results of our system (AI-T) with the list of bounding boxes named TLD1.0 as provided in the dataset [1]

| Sequence | | TP | P | R | F_M | ACLE | IoU |
|---|---|---|---|---|---|---|---|
| *david* 761 frames | | | | | | | |
| | TLD1.0 | 761 | 1,00 | 0,999 | 0,999 | 2,75 | 0,75 |
| | AI-T | 688 | 0,91 | 0,90 | 0.90 | 7,32 | 0,67 |
| *car* 945 frames | | | | | | | |
| | TLD1.0 | 832 | 0,98 | 0,97 | 0,97 | 13,70 | 0,66 |
| | AI-T | 784 | 0,83 | 0,91 | 0,87 | 17,02 | 0,64 |
| *jumping* 313 frames | | | | | | | |
| | TLD1.0 | 311 | 0,99 | 0,99 | 0,99 | 3,84 | 0,73 |
| | AI-T | 297 | 0,95 | 0,95 | 0,95 | 7,53 | 0,59 |

Performance evaluation is based on the computation of the PASCAL overlap measure IoU (Intersection over Union) considering the bounding boxes of the tracking result and the ground-truth. In the following, a tracking result is considered successful (true positive TP) when IoU > 0.5, a quite challenging goal as compared to the lower value (0.25) which is often adopted in other recent papers. The overall system performance is evaluated using standard precision P, recall R and F-measure statistics. Precision P is the rate of valid (IoU successful) boxes, among all target predictions; recall R is the rate of correct detections over the number of object occurrences that should have been detected. The value ACLE (Average centre Location Estimation) was used in [10] to measure the deviation of the target box with respect to the ground truth. IoU stands for the average overlap measure along the tracking sequence.

### 5.2 Detailed Performance Analysis

The results obtained for the first sequence (*david*: 761 frames length) are shown in table 1 as well as in fig 1; this video exhibits significant contrast variations from the initial frames (in the dark side of the room) to the last ones (the subject moving in full light). During the video sequence there is a consistent pose variation of the face, from frontal to lateral view and return. A total number of 52 cells has been selected for the target, with 8-bin signed gradient directions, for a descriptive vector of 416 elements (corner cells removed). During visual tracking the cell size was varying from a minimum of (6 x 6) pixels to a maximum value of (10 x 10) pixels. The appearance model has been updated (on-line learning) 413 times. The average value of IoU along the full sequence has been 0.67, and the number of correctly tracked frames, TP (IoU > 0.5), has been 688, with an F-Measure 0.90. Fig. 2 shows two different frames of the sequence (at the beginning and towards the end) with the bounding boxes of the ground truth data (blue color) and the tracking result (red color).

a) b)

**Fig. 1.** Video sequence "david"; ground-truth box are displayed in (blue) and tracking result (red) a) frame n.92 b) frame n. 512

The same experimental test has been performed with the "car" sequence (945 frames length). This is a low-contrast video sequence, with partial and temporary occlusions of the vehicle by waving trees. The selected target model is made of 78 cells, the length of the vector feature is 592, and the cell dimensions remain quite unchanged through the full tracking sequence, with an average size of (7 x 7) pixels.

During the tracking process the appearance model has been updated 73 times. A sample of the tracking results is referred in fig.2, with an average score of IoU = 0.64, as computed along the whole sequence. The number of correctly tracked frames was TP = 784 (IoU > 0.5) with an F-Measure of 0.87.

The results in fig 3 are referred to the "jumping" sequence (313 frames). In this video the object-face trajectory is quite irregular due to the up-and-down motion of the subject. Many frames are also affected by motion blur, which makes more difficult to perform model matching. Anyway, the constraints applied to the learning-updating process of the appearance model, are sufficient to achieve satisfactory tracking results: the number of appearance model updating has been 63; the average IoU has been 0.59 along the whole sequence, and the number of corrected tracking frames has been 297, with an F-Measure 0.95.

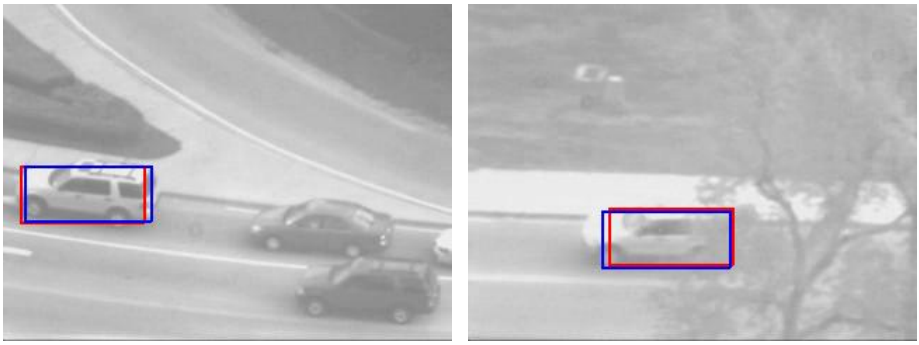

**Fig. 2.** Video sequence "car"; ground-truth box is displayed in blue color and tracking result in red; a) frame n.26 b) frame n.568

**Fig. 3.** Video sequence "jumping"; ground-truth box (blue) and tracking result (red) a) frame n.150 b) frame n.300, c) IoU profile along the video sequence

Finally, the "motocross" sequence has been considered, as shown in fig. 4. The image contrast is quite good through all video frames in the sequence, with strong irregularities in the visual trajectory due to the relative motion of both the on-board camera sensor and the target-motocross. In this case we could not use the ground truth of the dataset, because it was limited to a small subset of the target (the driver shoulders).

By selecting a bounding box around the entire motocross object, it was possible to track almost all visual instances of the target during the video sequence (more than 3000 frames) with a total of 65 updating steps of the appearance model. The selected structure of the target object was made of 72 cells, with variable size from a minimum of (6 x 6) pixels to a maximum of (12 x 12) pixels for close-up views.

### 5.2.1 Discussion

A limitation of our approach is that only positive examples are used during the appearance model updating. As such it is particularly important to select reliable samples to feed the learning process. Some additional experimental tests have been performed, to evaluate the effects caused by small deviations from the ground-truth box, in term of spatial translation and scale factor, during the phase of the initial object selection and the corresponding appearance model. In general such deviations have no significant impact in the continuity of the tracking process. A small increase of the scale factor (object patch size) was always beneficial to capture wider properties of the object. On the contrary a significant shift on the image plane of the initial object detection may produce a wrong starting model which propagates along the full tracking chain. The most critical part of the process is the classification of a tracking result as a positive sample to contribute to the appearance model update. Misclassification results may lead to unstable learning and an early rejection of the tracking process.

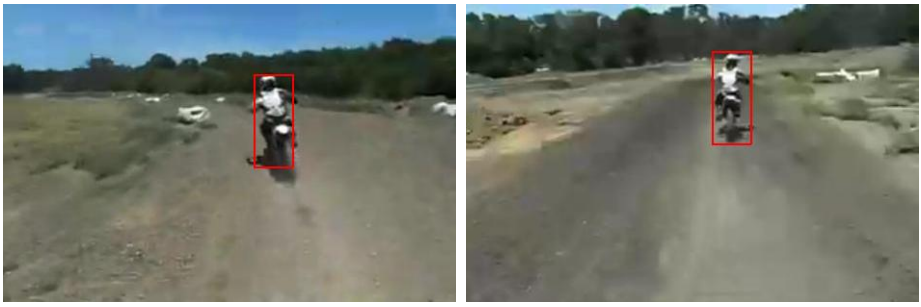

**Fig. 4.** Video sequence "motocross"; a) frame n.234 b) frame n.667

The current AI-T version is a research component still under development, and it is not yet optimized for optimal performance, so that we cannot provide precise figures in terms of processing speed. Regarding the computational cost of our approach we may consider the two most demanding components: the complexity of the target features and the search space (the number of search instances in the scene). The simplified version of the HOG features (small size of the descriptive vector) and the hierarchical search at different scale (cell size and pixel size) allow to achieve satisfactory results of near real-time on small-size video (640 x 480) on a 2.5 Ghz Intel i5 processor. In the referred experimental tests a few thousand vector distance measures are required, as compared to the hundred thousand windows needed by a standard sliding window approach at multiple scales.

In the paper we have shown results of single target tracking, being that the available scenario in the TLD dataset. Actually, our approach has been already successfully tested with multiple-target tracking in the security field, where an industrial application of the AiVu technology [3] has been developed to detect and hide recognizable human faces and car-plates from a standard surveillance video stream, for privacy protection.

## 6 Conclusions

The paper describes an approach to object tracking using adaptive appearance models based on HOG features, with performance results which are comparable with the most successful solutions recently published in the scientific literature. The main contributions of our research are the use of a modified version of the HOG feature descriptor, with a global normalization at target level, and the use of contrast gradient normalization and adaptive threshold. The forward-backward tracking scheme represents a simple and computationally very effective solution to achieve stable tracking results in a set of challenging video sequences.

## References


1. Kalal, Z., Matas, J., Mikolajczyk, K.: P-N learning: bootstrapping binary classifiers by structural constraints. In: 23th IEEE Conf. on Computer Vision and Pattern Recognition CVPR (2010)
2. Smeulders, A.W.M., Chu, D.M., Cucchiara, R., Calderara, S., Delghan, A., Shah, M.: Visual Tracking: an Experimental Survey. IEEE Trans. On Pattern Analysis and Machine Intelligence, PAMI (2013)
3. AiVu Video Analytics Technology. www.aitek.it
4. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Trans. on Pattern Analysis and Machine Intelligence **34**(7), 1409–1422 (2012)
5. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the International Joint Conference on ArtificialIntelligence, pp. 674–679 (1981)
6. Shi, J., Tomasi, C.: Good features to track. In: IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 1994, Seattle, June 1994
7. Brox, T., Malik, J., Flow, L.D.O.: Descriptor Matching in Variational Motion Estimation. PAMI **33**(3), 500–513 (2011)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. of CVPR 2005
9. Dollar, P., Belongie, S., Perona, P.: The fastest pedestrian detector in the west. In: Proceedings of the British Machine Vision Conference, pp. 68.1–68.11. BMVA Press, September 2010
10. Babenko, B., Yang, M.-H., Belongie, S.: Visual tracking with online multiple instance learning. In: CVpPR (2009)
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI **32**(9), 1627–1645 (2010)
12. Zhao, T., Nevatia, R.: Tracking multiple humans in complex situations. IEEE PAMI **26**(9), 1208–1221 (2004)