Object Tracking Guided by Segmentation Reliability Measures and Local Features

Cristian M. Orellana[1] and Marcos D. Zuniga[2(✉)]

[1] Department of Computer Science, Universidad Técnica Federico Santa María, Av. España 1680, Valparaíso, Chile
[2] Electronics Department, Universidad Técnica Federico Santa María, Av. España 1680, Valparaíso, Chile
marcos.zuniga@usm.cl

Abstract. Real world applications need to cope with unreliable data sources that affect negatively the performance of visual systems, adding error to the whole process. Existing solutions focus their efforts on decreasing the probability of making errors, but if an error occurs, there is no mechanism to deal with it. This work focuses in dealing with this problem by modelling the quality of the segmentation phase in order to apply control mechanisms to mitigate negative effects in later stages. Our control mechanism is based on determining the reliability of local features to discard the less reliables. Local features are characterized using colour, texture, and an illumination reliability model to quantify the quality of illumination. The use of local features enables us to deal with partial occlusion problems by determining the global object position via local features consensus. Experiments were performed, showing promising results in object position estimation under poor illumination conditions.

Keywords: Multi-target tracking · Feature tracking · Local descriptors · Background subtraction · Surveillance tracking · Reliability measures · Quality segmentation

1 Introduction

Real problems often lack on the possibility of obtaining manual initialisation for properly obtaining a reliable first model of an object. Many tracking algorithms require a robust initial object model to perform tracking, often obtained with manual procedures [1,2]. These methods often fail in dealing with problems as severe illumination changes or lack of contrast, or perform expensive procedures to keep the coherence of tracking in these complex situations. Also, these tracking approaches are focused on moving camera applications, so they neglect the utilisation of background subtraction to determine the regions of interest in the scene.

J. Braz et al. (Eds.): VISIGRAPP 2016, CCIS 693, pp. 534–554, 2017.
DOI: 10.1007/978-3-319-64870-5_26

A wide variety of applications can be solved utilising a fixed camera setup (e.g. video-surveillance, health-care at distance, behaviour analysis, traffic monitoring). This kind of setup allows the consideration of inexpensively utilising background subtraction approaches to detect potential regions of interest in the scene. This work focuses on this kind of applications, focusing in solving the problem of robust tracking of multiple unknown (uninitialised) objects, independently of the scene illumination conditions, in real-time. Then, tracking is performed without manual intervention.

Segmentation is commonly the early stage of any vision system, prior to tracking and higher level analysis stages, where regions of interest are extracted from the video sequence. Background subtraction approaches present several issues as: low contrast, poor illumination, gradual and sudden illumination changes, superfluous movement, shadows, among others [3]. Any error emerging from this stage would be propagated to the subsequent stages. A way to deal with these issues is to determine the *quality of the segmentation process* in order to activate control mechanisms to mitigate those errors on later stages.

Assuming that we do not know the model of objects present in the scene, we initially use a bounding box representation extracted from segmented blobs using background subtraction methods. This representation is general enough to track any object in real-time, and serves as the initial region of interest for applying more complex object models. Nevertheless, as the segmented blobs are obtained from background subtraction, they are sensitive to changes in contrast and illumination. This sensitivity affects the object tracking process incorporating noise (in terms of false positive and negative) to the system.

In order to control the effect of noisy information in tracking, we propose a local feature tracking approach, which reinforces the tracking of the bounding box associated to the object. We extract a contrast map from segmentation, to obtain reliability measures which allow us to characterise the local features in terms of illumination and contrast conditions. The local descriptors are obtained from a multi-criteria approach, considering colour (through HSV histograms), structural (through a binary descriptor), and segmentation region (through foreground mask and contrast maps) features. Then, the most reliably tracked local features are utilised, together with the tracked bounding box and the foreground information associated to the tracked object in the current frame, to adjust the estimation of the bounding box in the current frame.

This paper is organised as follows. First, Sect. 2 presents the state-of-the-art in order to clearly establish the contribution of the proposed approach. Then, Sect. 3 performs a complete description of the approach. Next, Sect. 4 presents the results obtained on several benchmark videos. Finally, Sect. 5 presents the conclusion and future work.

2 State of the Art

In the context of segmentation quality measures, the most recent approach is presented in [4]. The authors propose a metric to quantify the segmentation

quality for remote sensing segmentation, in terms of over-segmentation and under-segmentation. In order to detect under or over-segmentation, they use a similarity function to evaluate the quality of the segmentation. A good segmentation is obtained if a segment is well separated from its neighbouring segments. Errors can occur, like splitting a segment in two similar segments (over-segmentation) or merging two distinct segments (under-segmentation). Using the similarity function, the authors are able to measure over-segmentation and under-segmentation for each segment in the image. That information then is utilised to improve the segmentation applying the corresponding mechanisms to the erroneous segment (e.g. splitting a segment with under-segmentation problem).

In [5] the authors make a review of video segmentation quality. They identify that quality measurements can be object-based (individually) or globally (as meaning of overall segmentation). These measurements can also be classified as *relative*, when the segmentation mask is compared with ground-truth or as *stand-alone*, when the evaluation is made without using a reference image. Other classifications are subjective evaluation using human judgement or objective evaluation, using a set of a priori expected properties. For our scope, we are interested on a *individual stand-alone objective* quality measurement. In the same article, the features describing this kind of measures are intra-object metrics such as shape regularity, spatial uniformity, temporal stability and motion uniformity; or inter-object metrics like local contrast or neighbouring objects feature difference. The authors propose measures for each two classes of content, the stable content and the moving content. The first one is temporally stable and has regular shape, while the second one has strong and uniform motion. These measures take into account the characteristics of each content to make an unique quality value for the object.

In [6] the authors proposed three disparity metrics: local bound contrast, temporal color histogram difference and motion difference along object boundary. The local bound contrast is focused on determining the quality of the bounds by

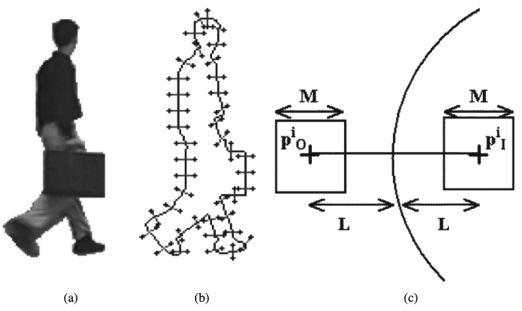

Fig. 1. Spatial color contrast along boundary metric from [6]. (a) image: Object detected, (b) image: Boundary with normal lines, (c) image: A zoom-in of a normal line where each cross represents a pixel inside (P_I) or outside the object (P_O).

comparing internal features (inside of the object) with external features (outside of the object). The next image depicts this metric:

To determine the quality of the boundary, a pixel P_I from the object is compared with a pixel of its neighbourhood P_O, both at distance L of the boundary. The comparison considers the average color in the square of size M, centered in the pixel P_* as shown in the Fig. 1(c). In this sense, good quality segmentation is achieved when there is a high difference between internal and external features. Special care must be taken with the meaning of the value, because a good boundary can be represented by a high quality value, but a high quality value does not necessarily mean a good quality boundary. The second metric tries to measure the temporal stability of color histogram distribution by comparing current object histogram with a smoothed version generated as an average of k previous histograms. A good temporal color stability is obtained if both histograms are similar. The third metric models the quality of the movement by estimating how the points P_* change from one frame to another. The movement metric considers the difference of motion vectors from both points (P_i and P_O) and a reliability factor defined as the precision of the estimation compare the measurement and the color consistency of the points in the square. The authors proposed a combined metric to determine the quality of the object segmentation. As well, they can determine if a particular segment of the boundary has poor quality using a combination of local bound contrast and motion metrics. If the combined value is higher than a predefined threshold, the related segment is considered as low quality. This threshold is obtained as a factor of the standard deviation of the mean object quality.

In the context of, local descriptor-based trackers, some similar approaches are presented in the literature. In [7] a reliable appearance model (RAM) that uses local descriptor (HOG) to learn the object shape and histogram is proposed. This appearance model effectively incorporate color and edge information as discriminative features. However, it is necessary to get a reliable first model to perform the training of the Adaboost learner, leaving this approach as semi-automatic, as well as many other approaches [1,2,8–10].

In [8] the authors proposed a weighted histogram that gives a higher weight to foreground pixel in order to make target features more prominent. The weighted component is based on the pixel's degree of belonging to the foreground. The way of producing the weighted histogram is very similar to our weighted histogram from Eq. (4), but it does not incorporate the reliability of illumination $R_i(y)$, that defines how illumination affect color-based features.

The authors in [9] combine a local descriptor (SIFT) with a global representation (PCA). In contrast to classical PCA, where pixels are weighted uniformly, they add a higher weight to pixels close to SIFT descriptor's position. The tracking phase depends on how reliable are the descriptors matching. This reliability is obtained based on how well the descriptor has been matched previously. Also the amount of reliable descriptors is used to determine if the occlusion is present in the frame. There are three modes of tracking, (1) if there are enough descriptor matched and they are reliable, then the tracking is perform by approximating the affine matrix that described the movement of the previous frame's descriptors

with the current descriptors. (2) if there are reliable matched descriptor but they are scarce, a translation model (position and velocity) is calculated instead. (3) is there no reliable matches, previous information is used to estimate the object's movement. In our case, the reliability of the descriptors comes from the reliability map, but the idea of use previous information when there is no reliable match of the descriptors remains. Another tracker that uses reliability is presented in [11]. In this case, the reliability is based on self-incorporated object detector (that is trained off-line). In order to get a good tracking performance, it is necessary to weight properly the information of tracking history and the classifier, otherwise drifting problems may arise.

Fragtrack is proposed in [10]. It uses local patches to avoid partial occlusion problems. If a patch is occluded, other patches can be used to predict the bounding box position (they assume that at least 25% of patches are visible). Each of this patches has associated a histogram and the relative position of its bounding box. The estimation of the bounding box in the next frame is done by a voting scheme. Each patch's histogram is searched in a neighbourhood and votes for a possible position of the bounding box. So, the estimated bounding box's position is whose has more votes. As the method rely heavily on the use of histogram, they use integral matching to perform real time tracking. This also allows search in different scales at without increasing so much the computational cost.

We summarise the contributions of the proposed approach as:

- A reliability model for background subtraction methods (or methods with similar behaviour: background modelling, comparing current frame with background model and applying a threshold to classify pixels into foreground or background). This is a pixel-level reliability model, which we refer as *reliability map*.
- An illumination reliability model to quantify the effects of illumination on color-based features.
- A way to convert a *reliability map* to *attribute-level reliability*. The attributes depend on the object representation. In our case, we will use a 2D bounding box and local features as object representation.
- A multi-target tracking approach incorporating attribute-level reliability measures for weighting the contribution of detected local features to the object model. The idea is to prevent the incorporation of information that could negatively affect the estimation of the object model, and focus on the most reliable information to reduce the effect of noise.

3 Reliable Local Feature Tracking

The proposed tracking approach is depicted in Fig. 2.

For each new frame of the video sequence, a background subtraction algorithm is applied for obtaining the foreground mask, the reliability map (see Sect. 3.1, for details), and the regions of interest (ROI), represented as a set of

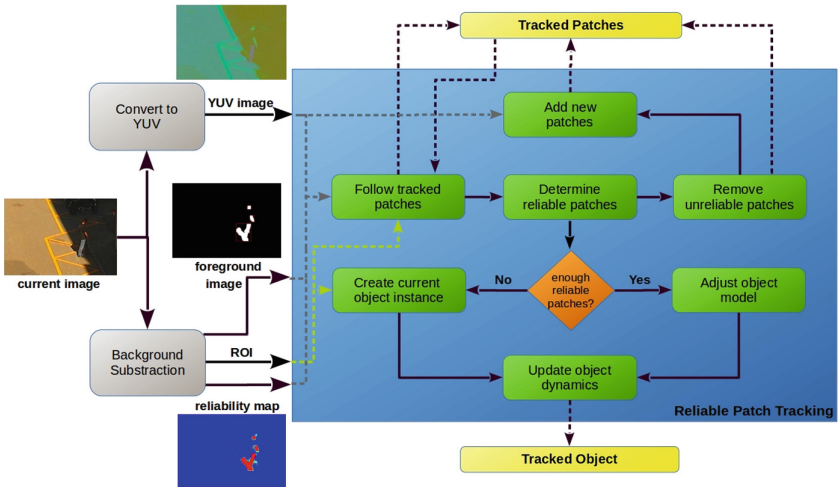

Fig. 2. General schema of the proposed tracking approach.

bounding boxes, using a connected components algorithms. Also, the new frame is converted to YUV color space.

For the first frame where a new object appears (new bounding box not associated to any other previously tracked object), a set of tracked patches is initialised, according to the procedure described in Sect. 3.2.

For the next frames, a ROI (or merge of partial ROIs), determined with a Multi-Hypothesis Tracking (MHT) algorithm [12], is associated to the object as input to the robust patch tracking approach, and the following procedure is applied:

1. If a patch is considered unreliable in terms of positioning. Then, an optimal association to the patch is searched in the current frame considering the information of the ROI displacement and dimension change, compared to the previously associated ROI. This optimal association is determined using a global reliability measure, which integrates temporal coherence, structural, colour, and contrast measures (see Sect. 3.4). If a set of patches has been reliably tracked from previous frames, this information is utilised to determine the displacement of all the patches for the current frame, according to the procedure detailed in Sect. 3.3.
2. Then, according to the global reliability measure calculated at the previous step, the highest reliability patches can be classified as *highly reliable*, the patches with low reliability are classified as *unreliable* and marked for elimination (see Sect. 3.3, for details).
3. Next, *unreliable* patches are eliminated and new patches are added in positions not properly covered by the remaining tracked patches. The construction of these patches follows the same procedure as the patch initialisation phase (Sect. 3.2).

4. If a significant number of patches is classified as reliable, they are utilised for adjusting the estimation of the object model bounding box for the current frame. If this number is not significant, the object model bounding box is obtained from the input ROI and the estimated bounding box from the object model dynamics (see Sect. 3.5, for details).
5. Finally, the dynamics object model is updated with the current object model bounding box (see Sect. 3.5, for details). Bottom image of Fig. 3 depicts the result of the tracking process.

Fig. 3. Top figure shows the current frame. Center figure depicts the reliability map, with a thermal map, where high reliability is red. Bottom figure shows the result of the tracking process; red boxes represent the bounding boxes from segmentation, the blue box represents the estimated bounding box of the tracked object, the dots represent the tracked patches coloured according to reliability in thermal scale, and blue segments represent the object trajectory. (Color figure online)

3.1 Reliability Map from Background Subtraction

The key factor for a good tracking is how distinguishable is the object of interest from its surroundings. If we are working in a background subtraction scheme, we are going to interpret the *surrounding* of the object as the background model and *how distinguishable is* as the degree of difference between the current image and the background model. If we have a significant difference, we have certain margin of error on defining the threshold and the segmentation algorithm will still be able to perform a good classification. Nevertheless, if that difference is low, we have to accurately define the threshold value to avoid a misclassification. In this sense, the last example is less *reliable*, because it is more prone to make a wrong classification.

Based on the previous idea, we propose a method that can model the reliability of any background subtraction technique through the following steps:

1. Generate a pixel-level difference value D between current image pixels and background model pixels.
2. Define a range $[inf, sup]$ for the difference value D. We are interested in generating a reliability image representation with different degrees of reliability. If we consider all the range, sometimes it can generate a binary image (just low and high reliability) that is not useful for our interest. This range can be

defined as the neighbourhood of the threshold value, for example using the range $[inf = \alpha \times Threshold,\ sup = \beta \times Threshold]$, with $0 < \alpha < 1$ and $1 < \beta$.

3. Apply the scaling function from Eq. (1), to every difference value D generated in Step 1, to convert difference values into reliability measures:

$$S(D) = \begin{cases} 0\% \ \ if & D < inf \\ f(x) \ \ if & inf \leq D \leq sup \\ 100\% \ if & D > sup \end{cases}, \tag{1}$$

where D is the differnece value, inf and sup are values defined in Step 2 and $f(x)$ is a increasing function (we use a linear function).

At the end of these steps we generate a pixel-level representation of the reliability which we named as *reliability map*. This map is internally represented as a grayscale image, but for proper visualisation we transform it into thermal scale, as shown in Fig. 4.

Usually, several post-processing functions are applied to the segmentation mask in order to reduce the noise. This operation also should be applied to the reliability map to maintain the coherence of its representation with the foreground mask. Figure 5 is an example of applying morphology operations to the foreground image and the reliability map (considering gray-scale morphological operators).

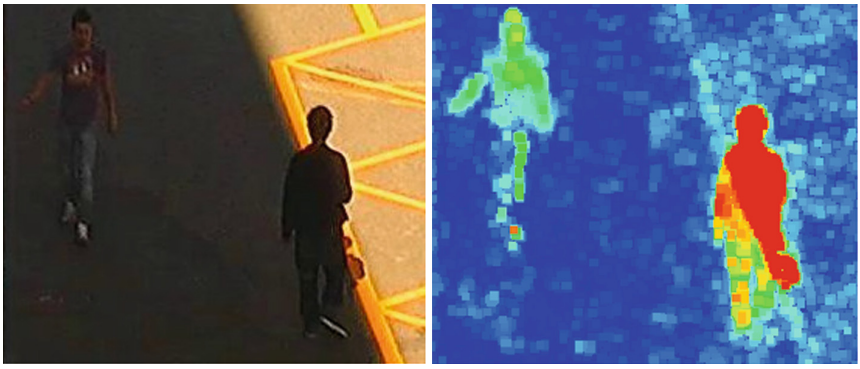

Fig. 4. Reliability map visualization. Left image: current image frame, right image: thermal scale reliability map. Blue color means a low difference between modeled background and current frame. Red color means a high difference. (Color figure online)

We illustrate how this method works using naive background subtraction [13]: This model performs difference of current image with a background subtraction image (image without any object interest). Our implementation uses the sum of square differences as distance value before applying the classification threshold.

The sum of square difference, shown in the Eq. (2), is a common metric to measure the distance between current pixel and background pixel in a RGB color space:

$$D = (R_{bg} - R_i)^2 + (G_{bg} - G_i)^2 + (B_{bg} - B_i)^2, \tag{2}$$

where subindex $(\cdot)_i$ refers to current image pixel and $(\cdot)_{bg}$ refers to background pixel.

The classification is performed applying the threshold value as in the Eq. (3):

$$fg_{mask} = \begin{cases} foreground\ if\ D > \tau^2 \\ background\ if\ D < \tau^2 \end{cases}, \tag{3}$$

where τ is the classification threshold.

Applying the proposed scheme to this method using a range of $[0.1 \times \tau^2,\ 2.0 \times \tau^2]$ with $\tau = 13$ and using a morphology window of size 7×7, we can obtain image shown in Fig. 6.

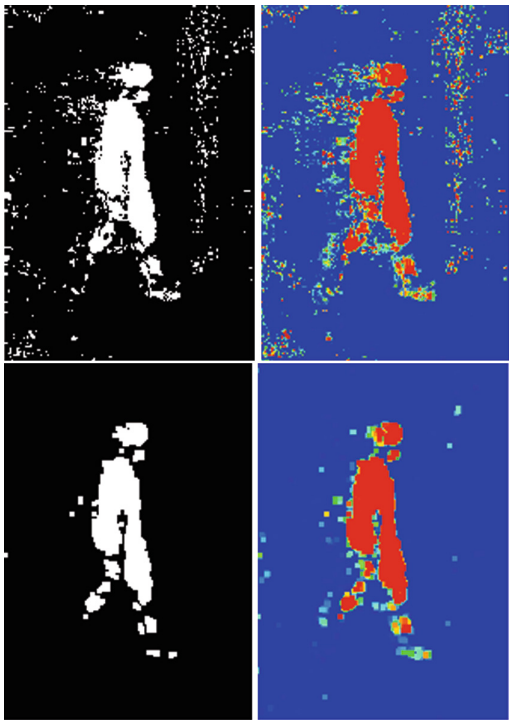

Fig. 5. Example of applying morphology operations to foreground mask and reliability map. The top images show the foreground mask and the reliability map with noise. The bottom images show the results after applying the morphological operation (binary morphology for foreground mask and gray-scale morphology for reliability map).

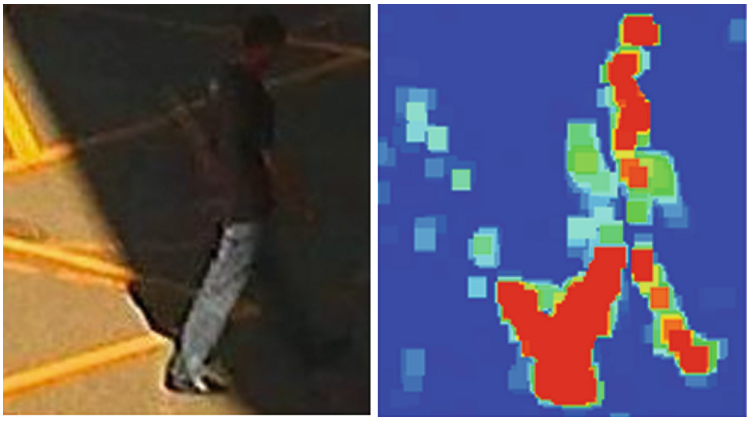

Fig. 6. Reliability map using naive background subtraction. Left image: current image, right image: reliability map from naive background subtraction.

3.2 Patch Initialisation Phase

The first step is to find patches of size $patchSize \times patchSize$ in the contour of the object (defined by the foreground mask) in such way that any two patches do not overlap between each other. Then, the strongest point inside of the patch, obtained by FAST algorithm [14] from the Y-channel of the current frame converted to YUV color space, is added as a new patch position if no other existing patch is near this position.

Then, each candidate patch stores the following information:

- The central patch position (x, y).
- The 512 bits FREAK descriptor [15], generated using the reliability map, representing the structural information of the patch.
- A normalised colour histogram, using chroma channels U and V from the YUV current frame, considering only pixels belonging to the foreground mask in the analysed patch. Considering $H_{UV}(i, j)$ as the bin of a 2D histogram of the UV channels, with $i, j \in [0..BinsNumber]$, The Eq. (4) represents the way this histogram is calculated.

$$H_{UV}(i,j) = \frac{\sum_{p \in Q} F(p) R_m(p) R_i\left(Y(p)\right)}{\sum_{p \in P} F(p) R_m(p) R_i\left(Y(p)\right)}, \tag{4}$$

with

$$Q = \left\{ p \in P : \left\lfloor \frac{U(p)}{binSize} \right\rfloor = i \ \wedge \ \left\lfloor \frac{V(p)}{binSize} \right\rfloor = j \right\}, \tag{5}$$

where $Y(p)$, $U(p)$, and $V(p)$ correspond to the channel level in [0..255] in pixel position p of the current frame in YUV color space, P is the set of pixel positions

inside the analysed patch, and Q is the set of patch positions, where values $U(p)$ and $V(p)$ fall inside the bin $H_{UV}(i,j)$. For each pixel a weighted value is added, where: $F(p) = 1$ if the pixel p corresponds to the foreground, and 0 otherwise; $R_m(p) \in [0;1]$ is the reliability map value in position p, where a value of 1 corresponds to maximum contrast reliability (see Sect. 3.4, for details); and $R_i(Y(p))$ corresponds to the illumination reliability, accounting the pertinence of colour information given different illumination levels, according to the gray-scale level in channel $Y \in [0..255]$ at pixel position p. The reliability measure R_i considers maximum reliability near 128 value (medium illumination) and decays to 0 near the extremes of the interval. Eq. (6) formulates this reliability and Fig. 7 depicts the reliability function.

$$R_i(Y) = \begin{cases} 0 & \text{if } Y \leq 128 - \gamma \\ \frac{Y+\gamma-128}{\beta} & \text{if } 128 - \gamma < Y < 128 - \alpha \\ 1 & \text{if } 128 - \alpha \leq Y \leq 128 + \alpha \\ \frac{128+\gamma-Y}{\beta} & \text{if } 128 + \alpha < Y < 128 + \gamma \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where α and β are predefined parameters, and $\gamma = \alpha + \beta$.

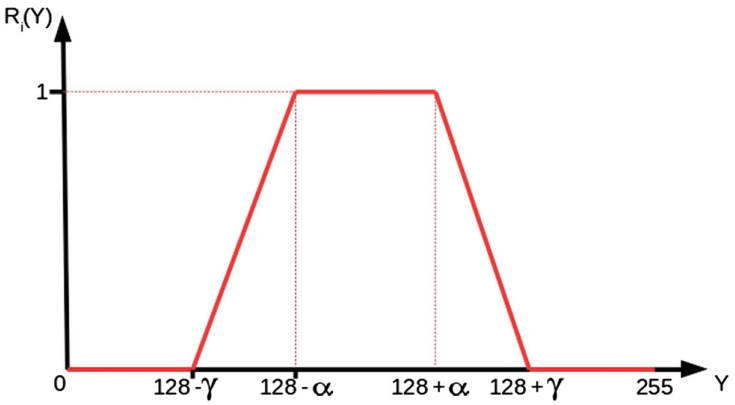

Fig. 7. Illumination reliability function.

– A colour histogram reliability measure accounting for the reliability of colour information (Eq. (7)).

$$R_{colour} = \left(\sum_{p \in P} F(p) R_m(p) R_i(Y(p)) \right) / N_{pix}, \tag{7}$$

where N_{pix} is the number of foreground pixels in the patch.

– A normalised gray-scale histogram of $NumBins$ bins, accumulating channel Y of the current image in YUV color space, for those pixels inside the patch which belong to the foreground.

All this information is utilised to properly characterise the patch, in order to match with potential patches in future frames. These patches then initialise patch tracking buffers for future processing.

3.3 Patch Tracking Phase

Given a set of patches S from the previous frame, the patch tracking process follows the process described bellow:

– Consider S_H as the set of tracked patches considered as *highly reliable* from the previously processed frame. A reliably tracked frame is a frame of high reliability, which has a coherent movement with the mobile object and high contrast, colour, and structural accumulated reliabilities (as described in Sect. 3.4). Then, these patches are considered able to estimate the behaviour of less reliable patches near to them. For this reason, tracking becomes more exhaustive for these patches, but in a reduced region. Then, the reliable patches are tracked in the following way:
  1. Displacement vector (dx, dy) is determined from the displacement vector inferred from their associated patch tracking buffer.
  2. Search window is determined from the accumulated difference (x_d, y_d) between the accumulated object center movement vector with the accumulated movement vector of the patch, considering all the patches in the tracking buffer. The window is centered in $(x_W, y_W) = (x_p + dx, y_p + dy)$, where (x_p, y_p) is the position of the patch in the previous frame.
  3. Then, the patch position with minimal global distance D_{global} to the previous patch is associated to the current reliable patch position, following the Eq. (8).

$$(x^*, y^*) = \arg \max_{(x,y) \in W_H} D_{global}(p_t(x, y), p_{t-1}), \tag{8}$$

with

$$W_H = \{(x, y) : |x - x_W| \leq x_d \wedge |y - y_W| \leq y_d\}, \tag{9}$$

where $p_t(x, y)$ is the current patch at position (x, y), and p_{t-1} is the patch at previous frame. The distance measure D_{global} globally calculates the patch distance, considering the structural, colour, segmentation and gray-scale information. This measure is described in detail, in Sect. 3.4.

– If the patch buffer has been built just in the previous frame (previous initialisation step) or the patch is not *highly reliable*, the positioning of the patch is determined in the following way:
  1. If set S_H size is adequate, the displacement vector (dx, dy) for the patch is determined from the displacement vectors of *highly reliable* patches, each weighted by the position of the *highly reliable* patch to the analysed patch in the previous frame and the R_{global} reliability measure.

2. The window is determined in a similar way as for *highly reliable* patches, but, as the patch is less reliable, it would normally have a bigger search window. For this reason, FAST algorithm is applied to the search window for candidate positions.
3. Then, maximal reliability patch is determined in a similar way as in Eq. (8), but from the set of FAST points detected on the window.

– Then, according to the global reliability measure R_{global}, the tracked patches are classified as *highly reliable* if they pass a high threshold T_H. Patches with reliability below a low threshold T_U are classified as unreliable and eliminated.
– As the object can be represented by less patches, new patches are added in positions not properly covered by the remaining tracked patches, using the same procedure described in Sect. 3.2.

3.4 Patch Distance and Reliability Measures

To match two patches, the distance between them in terms of their different attributes must be calculated. We propose the distance measure D_{global}, described in Eq. (10).

$$D_{global} = \frac{w_{st}D_{st} + w_{fg}D_{fg} + w_{co}D_{co} + w_{gs}D_{gs}}{w_{st} + w_{fg} + w_{co} + w_{gs}}, \tag{10}$$

with

$$D_{st}(p_1, p_2) = \frac{\|Freak[p_1]; Freak[p_2]\|_H}{512}, \tag{11}$$

$$D_{fg}(p_1, p_2) = \frac{|\#FG[p_1] - \#FG[p_2]|}{\max(\#FG[p_1], \#FG[p_2])}, \tag{12}$$

$$D_{co}(p_1, p_2) = D_{rcol}(p_1, p_2)\, \|H_{UV}[p_1]; H_{UV}[p_2]\|_B\,, \tag{13}$$

$$D_{rcol}(p_1, p_2) = |R_{colour}(p_1) - R_{colour}(p_2)|, \tag{14}$$

$$D_{gs}(p1, p2) = \|H_Y[p1], H_Y[p2]\|_B\,, \tag{15}$$

where $\|\cdot;\cdot\|_H$ is the distance of Hamming for binary descriptors, and $\|\cdot;\cdot\|_B$ is the Bhattacharyya distance [16] for histograms. $Freak[p]$ corresponds to the FREAK descriptor, $\#FG[p]$ is the number of foreground pixels, $H_{UV}[p]$ is the colour histogram, and $H_Y[p]$ is the gray-scale histogram, of patch p. $D_{rcol}(\cdot\,,\cdot\,)$ accounts for the difference in R_{colour}, considering that histograms are more comparable under similar conditions in terms of illumination and contrast reliability.

It has been previously discussed that we need a measure to account for the reliability of the tracked patches in the scene, in order to determine the usefulness of the patch information on contributing to a more robust object tracking. This reliability measure is R_{global}, described in Eq. (16), considering a tracked patch

buffer $B_p = \{p_1, .., p_N\}$, where p_1 is the current patch, and N is the buffer size, and the object bounding box buffer $B_I = \{I_1, .., I_N\}$, where I_j is the bounding box in buffer position j.

$$R_{global}(B_p) = \frac{R_{pos}(B_p) + R_c(B_p) + R_g(B_p)}{3}, \tag{16}$$

with

$$R_{pos}(B_p) = \frac{\sum_{i=1}^{N-1} (N-i) \left\| c[p_i] - c[p_{i+1}]; c[I_i] - c[I_{i+1}] \right\|_M}{\sum_{i=1}^{N-1} (N-i)}, \tag{17}$$

$$\left\| c1; c2 \right\|_M = |x[c1] - x[c2]| + |y[c1] - y[c2]|, \tag{18}$$

$$R_c(B_p) = \frac{\sum_{i=1}^{N} (N-i+1) C(x[p_i], y[p_i])}{\sum_{i=1}^{N} (N-i+1)}, \tag{19}$$

$$C(x_p, y_p) = \frac{\sum_{x=x_p-\frac{L}{2}}^{x_p+\frac{L}{2}} \sum_{y=y_p-\frac{L}{2}}^{y_p+\frac{L}{2}} G(x-x_p, y-y_p) FG(x,y) R_m(x,y)}{\sum_{x=x_p-\frac{L}{2}}^{x_p+\frac{L}{2}} \sum_{y=y_p-\frac{L}{2}}^{y_p+\frac{L}{2}} G(x-x_p, y-y_p) FG(x,y)}, \tag{20}$$

$$R_g(B_p) = 1 - \frac{\sum_{i=1}^{N-1} (N-i) D_{global}(p_i, p_{i+1})}{\sum_{i=1}^{N-1} (N-i)}. \tag{21}$$

The three components of R_{global} are calculated weighting by the novelty of the information. R_pos is the position coherence reliability, which takes into account the displacement coherence between the history of the patch (measured as the displacement vector of the patch centers $c[p_i] - c[p_{i+1}]$) and the history of the central position of the object model bounding box ($c[I_i] - c[I_{i+1}]|$), using the Manhattan distance between displacement vectors at the different frames. R_c accumulates the contrast reliability measure $C(x, y)$, which accumulates the values of the reliability map R_m, weighted by a Gaussian function G centred at (x, y) and only accumulating foreground pixels (considering $FG(x, y)$ as the foreground image, with value 1 for foreground pixels and 0 for background). R_g accumulates the reliability on the similarity of the patches in the buffer.

3.5 Adjustment of Object Model

Finally, if the current input bounding box is significantly different in dimensions compared to the previous frame or several reliable patches present a low contrast reliability for the current frame and a relevant change on patch mean illumination from previous frame (inferred from Y channel), the bounding box is recalculated based on the information provided by the remaining reliable patches. The displacement of each bound of the bounding box (Left, Right, Bottom, Top) is obtained from the weighted mean of patches displacement from the previous frame, weighted by the distance to the bound and the reliability of the patches.

If no reliable patches are available, the bounding box projected from the object dynamics model is considered as input. We utilise a dynamics model similar to Kalman Filter [12]. If the current input bounding box is similar in size to the previous frame, this bounding box is considered as the object model for the current frame. Then, the dynamics model is updated with the current object model.

4 Experimental Validation

The visual coherence of the estimation has been first tested in three short sequences of diverse contrast. The results are shown in Fig. 8.

A qualitative test has been performed using a section of the sequence Light_Video032 of dataset Alov-300[1]. In this sequence a man walks from a good to a poor illumination region (from left to right). Figure 9 depicts the evolution of the patches through the last five frames:

As we can see in the Fig. 9, each patch can track its next position even with structural deformation (e.g. row num. 0) or some changes in color (e.g. row num. 4) due to illumination changes.

The performace of the approach is compared with a reduced version (without local patches nor reliability measurements). A compacted view of the sequence is shown in the Fig. 10, considering key frames for the analysis.

Initially both methods perform similar, but when the object enters into the poor illumination region, our approach outperforms the basic method. However, at the end of the sequence, when the object is completely immerse in the shadow region, both methods are unable to keep tracking the object. Neverthless, when reliable local information is available, our approach is able to correctly infer the global position of the object and unrealible local features enable the approach to properly characterize regions with poor illumination conditions. This characterization will allow us in the future to properly handle poor illumination situations when they are detected. Figure 11 presents the evolution of the illumiantion conditions for the tested sequence.

To quantify the improvement of the approach, two videos of changing contrast situations have been tested. Both videos have ground-truth segmentation, in order to obtain the ideal track of the analysed objects. The first video consists

[1] http://www.alov300.org/.

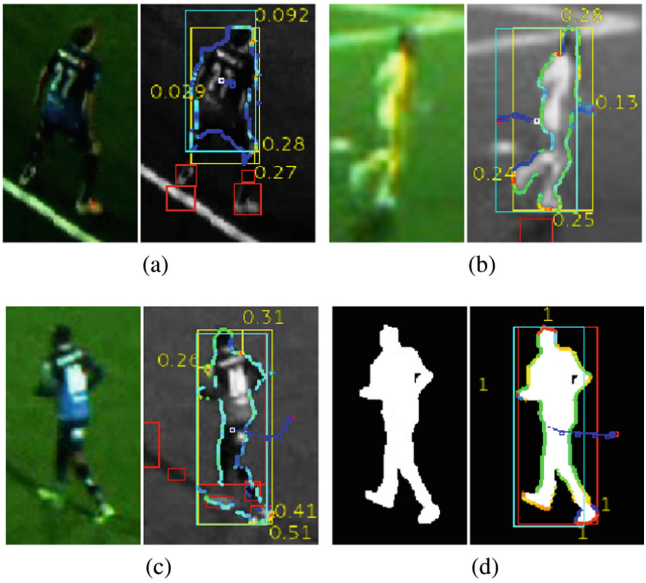

Fig. 8. Resulting tracking for three soccer player sequences with different levels of contrast. Figures (a), (b), and (c) show the result for low, medium, and high contrast situations, respectively. Figure (d) is a control case for ground-truth segmentation. The segmentation blob bounding boxes are colored red, the merged bounding box for the object hypothesis colored yellow, and the estimated bounding box from the dynamics model colored cyan. The central object position trajectory is depicted with blue squares. (Color figure online)

in a single football player sequence (27 frames), where a player goes from a light to a dark zone of the pitch. This video is a zoomed short sequence extracted from the Alfheim Stadium dataset[2]. The second video consists in a sequence (51 frames) where a rodent is exploring a confined space with better illumination in the center. The sequence is part of a set of sequences provided by the Interdisciplinary Center of Neuroscience of Valparaiso[3]. This sequences are intended to study the behavior of the degu, a rodent which commonly presents the Alzheimer disease.

The experiment consists in performing object tracking using the new dynamics model with and without considering the proposed reliability measures, and compare the obtained tracks with the ideal tracks obtained from the ground-truth segmentation. The results were summarised in Table 1.

The results for the first experiment are exemplified in Fig. 12. Figure 12(b) and (c) show the core motivation of this work: the effect of considering different measures for tracked attributes allows a finer control of the trade off between

[2] Open dataset extracted from Alfheim Stadium, the home arena for TromsøIL (Norway). Available from: http://home.ifi.uio.no/paalh/dataset/alfheim/.

[3] Interdisciplinary Center of Neuroscience of Valparaiso, Chile. http://cinv.uv.cl/en/.

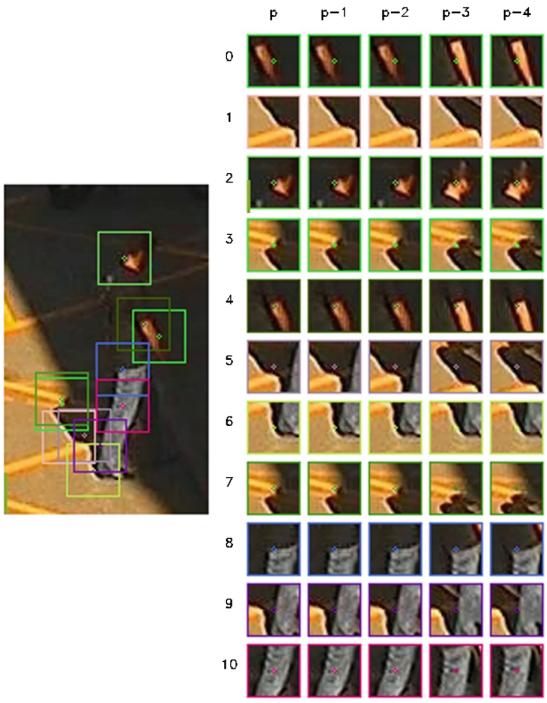

Fig. 9. The local patches are shown in this figure. The patches are sorted from current patches in column p, to older patches (columns $p - i$). (Color figure online)

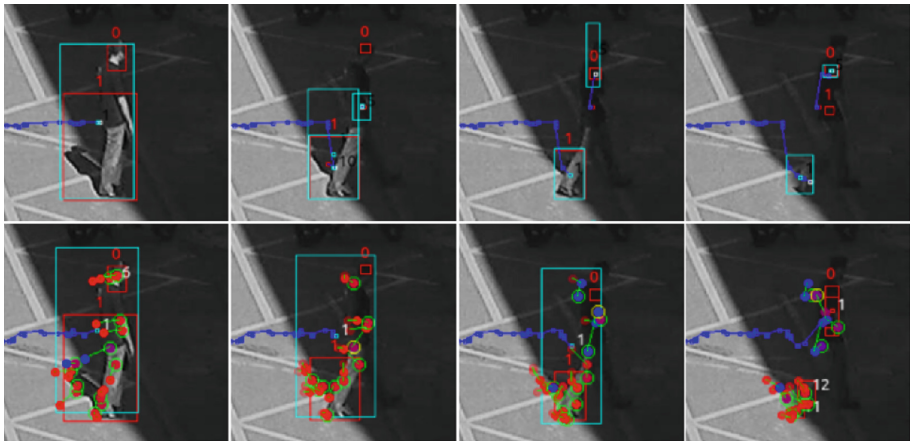

Fig. 10. Comparison of the approach using local patches and reliabilty measurements. First row shows the performance without utilisation of local patches nor reliabilty measurements. Second row depicts our approach. In the second row, red circles represent matched patches and blue circles represent new patches. (Color figure online)

Table 1. Results for evaluation sequences with respect to ground-truth sequences. The column Imp.% is the percent of improvement utilising the proposed approach.

Sequence	Distance (pixels)		
	No rel.	Patch rel.	Imp.%
Football (T = 15)	602.2	579.5	3.8%
Football (T = 20)	640.7	570.8	10.9%
Rodent (T = 10)	600.4	581.6	3.1%
Rodent (T = 15)	506.7	491.4	3.0%
Rodent (T = 20)	1086.8	1011.5	6.9%
Rodent (T = 25)	1071.1	1023.0	4.5%

the estimated state and the measurement in the update process. In the example, the patch tracking algorithm was able to properly weight unreliable data to not affect considerably the dynamics model, and the legs of the player were not lost (Fig. 12(c)).

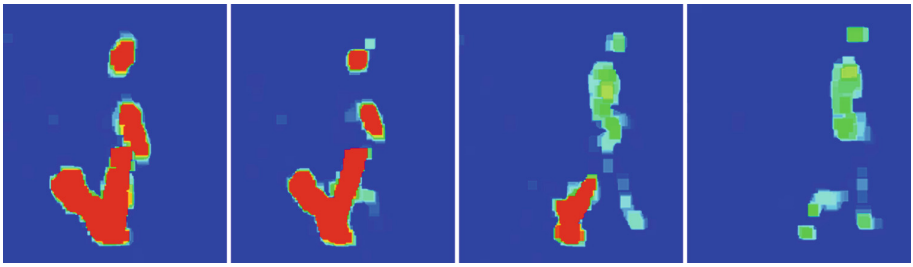

Fig. 11. Reliability map of the object from video Light_Video032. The reliability of the object is properly related to illumination, thus it can be used to detect bad segmentation in order to activate control mechanisms.

For the second experiment, the challenge is to follow a rodent of quick acceleration changes and not homogeneous illumination conditions. Also, poor segmentation occurs due to the sudden changes of speed. The sequence was tested for different segmentation thresholds ($T \in \{10, 15, 20, 25\}$). From these results, we are able to state that a more robust tracking can be achieved utilising the patch reliability measure, with an improvement higher than a 3% in precision. Examples of these results are depicted in Fig. 13.

All the results generated by qualitative and quantitative experiments can be found in: http://profesores.elo.utfsm.cl/~mzuniga/videos/.

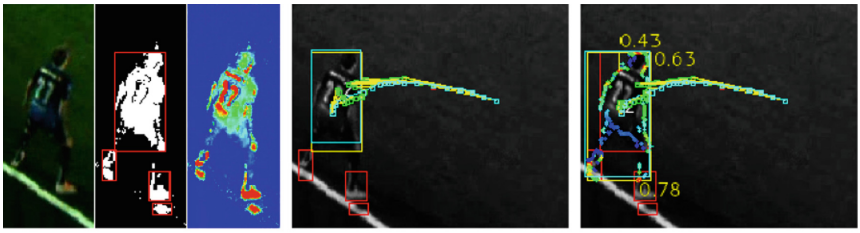

Fig. 12. Example of the effect on utilising the patch reliability on the tracking process ($T = 20$). Figure (a), from left to right, shows the current, segmentation, and contrast map images, respectively. Figure (b) shows the tracking result without considering the patch reliability measures (every reliability is set to 1). Figure (c) shows the result of using the patch reliability measure. Note the difference in tracking bounding box, where the feet of the player are more properly incorporated to the object. The boxes are colored the same way as previous images. The central object position trajectory is depicted with green squares, the ground-truth positions in cyan squares, and the distance between them is represented with a yellow line. (Color figure online)

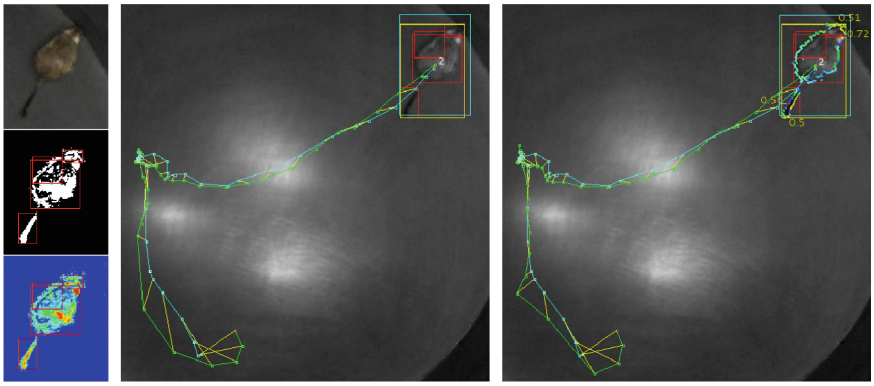

Fig. 13. Example of the effect on utilising the patch reliability on the tracking process ($T = 25$). Figure (a), from top to bottom, shows the current, segmentation, and contrast map images, respectively. Figures (b) and (c) show the tracking result not considering and considering the patch reliability measures, respectively.

5 Conclusions

For addressing real world applications, computer vision techniques must properly handle noisy data. In this direction, we have proposed a new tracking schema considering local features and reliability measures which have shown promising results for improving the dynamics updating process of the tracking phase. The reliability measures were utilised to control the uncertainty in the obtained information, through a direct interpretation of the criteria utilised by the

segmentation phase to determine the foreground regions. In this sense, this approach can be applied to other segmentation algorithms to improve the tracking phase in the same way.

In particular, the proposed global patch reliability measure, considering a diverse range of features, has shown one of the many possible ways of integrating segmentation phase data to object modelling. In the present work, no a priori knowledge has been considered about the objects to be tracked. The integration of the data from the segmentation phase with more complex object models can also improve the tracking phase, by better determining the objects of interest for a context or application. At the same time, these reliability measures can help these object models to better determine their parameters, subject to noisy measurements.

The preliminary evaluation obtained promising results both in robust tracking and quick processing. Nevertheless, extensive testing is required for fully validating the approach.

This work can be extended in several ways: the approach can be tested for different types of detectors of interest points and local feature detectors. Also, the algorithm can be tested for different background subtraction approaches. However an extensive parameter sensitivity evaluation is still needed. As local features are utilised for partial occlusion, this approach could be naturally extended to deal with dynamic occlusion situations. As previously mentioned, one of the most important extension of this work will be the development of more sophisticated control mechanisms in case of local patches describing the presence of poor illumination conditions.

Acknowledgements. This research has been supported, in part, by Fondecyt Project 11121383, Chile.

References

1. Kalal, Z., Matas, J., Mikolajczyk, K.: Tracking learning detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 1409–1422 (2011)
2. Yang, F., Lu, H., Yang, M.: Robust superpixel tracking. IEEE Trans. Image Process. **23**, 1639–1651 (2014)
3. Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower: principles and practice of background maintenance. In: Proceedings of the International Conference on Computer Vision (ICCV 1999), pp. 255–261 (1999). doi:10.1109/ICCV.1999.791228
4. Troya-Galvis, A., Gancarski, P., Passat, N., Berti-Equille, L.: Unsupervised quantification of under- and over-segmentation for object-based remote sensing image analysis. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **8**, 1936–1945 (2015)
5. Correia, P.L., Pereira, F.: Objective evaluation of video segmentation quality. IEEE Trans. Image Process. **12**, 186–200 (2003)
6. Erdem, Ç.E., Sankur, B., et al.: Performance measures for video object segmentation and tracking. IEEE Trans. Image Process. **13**, 937–951 (2004)

7. Lee, S., Horio, K.: Human tracking using particle filter with reliable appearance model. In: 2013 Proceedings of SICE Annual Conference (SICE), pp. 1418–1424 (2013)
8. Wang, L., Yan, H., Wu, H.Y., Pan, C.: Forward-backward mean-shift for visual tracking with local-background-weighted histogram. IEEE Trans. Intell. Transp. Syst. **14**, 1480–1489 (2013)
9. Sun, L., Liu, G.: Visual object tracking based on combination of local description and global representation. IEEE Trans. Circ. Syst. Video Technol. **21**, 408–420 (2011)
10. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 798–805 (2006)
11. Breitenstein, M., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Robust tracking-by-detection using a detector confidence particle filter. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1515–1522 (2009)
12. Zuniga, M.D., Bremond, F., Thonnat, M.: Real-time reliability measure driven multi-hypothesis tracking using 2D and 3D features. EURASIP J. Adv. Signal Process. **2011**, 142 (2011). doi:10.1186/1687-6180-2011-142
13. McIvor, A.: Background subtraction techniques. In: Proceedings of the Conference on Image and Vision Computing (IVCNZ 2000), Hamilton, New Zealand, pp. 147–153 (2000)
14. Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV 2006), vol. 1, pp. 430–443 (2006)
15. Alahi, A., Ortiz, R., Vandergheynst, P.: Freak: fast retina keypoint. In: Procedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012), pp. 510–517 (2012)
16. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by probability distributions. Bull. Calcutta Math. Soc. **35**, 99–110 (1943)