User Identification and Object Recognition in Clutter Scenes Based on RGB-Depth Analysis

Albert Clapés[1,2], Miguel Reyes[1,2], and Sergio Escalera[1,2]

[1] Dept. Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007, Barcelona, Spain
[2] Computer Vision Center, Campus UAB, Edifici O, 08193, Bellaterra, Spain
{aclapes,mreyes,sescalera}@cvc.uab.es

Abstract. We propose an automatic system for user identification and object recognition based on multi-modal RGB-Depth data analysis. We model a RGBD environment learning a pixel-based background Gaussian distribution. Then, user and object candidate regions are detected and recognized online using robust statistical approaches over RGBD descriptions. Finally, the system saves the historic of user-object assignments, being specially useful for surveillance scenarios. The system has been evaluated on a novel data set containing different indoor/outdoor scenarios, objects, and users, showing accurate recognition and better performance than standard state-of-the-art approaches.

Keywords: Multi-modal RGB-Depth data analysis, User identification, Object Recognition, Visual features, Statistical learning.

1 Introduction

In most monitoring surveillance scenarios a vast majority of video is permanently lost without any useful processing being gained from it. Several automatic approaches related to this topic has been published [1]. These works base on Computer Vision techniques to examine the video streams to determine activities, events, or behaviors that might be considered suspicious and provide an appropriate response when such actions occur. The detection of motion in many current tracking systems relies on the technique of background subtraction. The ability to represent multiple modes for the background values allows some techniques to model motion which is part of the background [2]. However, almost none of the state-of-the-art methods can adapt to quick image variations such as a light turning on or off.

Computer Vision techniques have been studied for decades in the surveillance scenario, and although huge improvements have been performed, still it is difficult to robustly identify users and objects in visual data. Some works have addressed the problem of developing complete vision systems for both object recognition and tracking in order to obtain a rough scene understanding [3]. However, still occlusions and noise can generate false object appearance in the scene.

F.J. Perales, R.B. Fisher, and T.B. Moeslund (Eds.): AMDO 2012, LNCS 7378, pp. 1–11, 2012.
© Springer-Verlag Berlin Heidelberg 2012

With the objective of improving the discriminability of relevant surveillance events in the scenes, some authors use calibrated cameras which are synchronized in order to obtain an approximation of the 3D representation of the scene. Although this approach can be useful in some situations, it requires from a perfect multi-camera synchronization, and a strategic location of each camera that could not be feasible in most real environments. Recently, with the appearance of the Depth maps introduced by the Kinect Microsoft device, a new source of information has emerged. With the use of depth maps, 3D information of the scene from a particular point of view is easily computed, and thus, working with consecutive frames, we obtain RGBDT information, from Red, Green, Blue, Depth, and Time data, respectively. This motivates the use of multi-modal data fusion strategies to benefit from the new data representation. In particular, Girshick and Shotton et al. [4] present one of the greatest advances in the extraction of the human body pose from depth images, that also forms the core of the Kinect human recognition framework. Through this technology are emerging work on reconstruction of dense surfaces and 3D object detection [5].

In this paper, we propose an automatic surveillance system for user identification and object recognition based on multi-modal RGB-Depth data analysis. We model a RGBD environment learning a pixel based background Gaussian distribution. Then, user and object candidate regions are detected and recognized using robust statistical approaches. The system robustly recognize users and update the system in an online way, identifying and detecting new actors in the scene. On the other hand, segmented regions of candidate objects are described, matched, and recognized using view-point 3D descriptions of normal vectors using spatial and depth information, being robust to partial occlusions and local 3D viewpoint rotations. Moreover, 3D object information is online updated as well as new views of the object are detected. Finally, the system saves the historic of user-object pick ups assignments, being specially useful for surveillance scenarios. The system has been evaluated on a novel data set containing different scenarios, objects, and users, showing accurate recognition results.

The rest of the paper is organized as follows: Section 2 presents the system for user identification and object recognition. Section 3 presents the results, and finally, Section 4 concludes the paper.

2 Muti-modal User Identification and Object Recognition

In this section, we present our system for automatic user-object interaction analysis using multi-modal RGBD data. The system is composed by four main modules which are described next. The control automata of the system that calls to the different module functionalities is summarized in Algorithm1. The scheme of the whole system is illustrated in Fig. 1.

2.1 Environment Modeling

Given the frame set $F = \{I, D\}$ containing a RGB image $I \in [0,1]^{h \times w}$ and a depth map $D \in [0,\infty]^{h \times w}$ with the depth value of each pixel obtained by the

Data: $F_{\{1,..,T\}}$
Environment modeling of $F_{\{1,..,T\}}$ using pixel adaptive learning (section 2.1)
while *true* **do**
  Acquire new frame $F_t = \{I_t, D_t\}$ composed by RGB image I and depth map D (section 2.1)
  Segment new regions of F_t based on environment modeling (section 2.1)
  Look for subject/s and identification/s in F_t (section 2.2)
  Look for new objects or object removals in F_t (section 2.3)
  Look for getting/leaving objects in scene (section 2.4)
  User-object association analysis
end

Algorithm 1. Control automata of the RGBD surveillance system

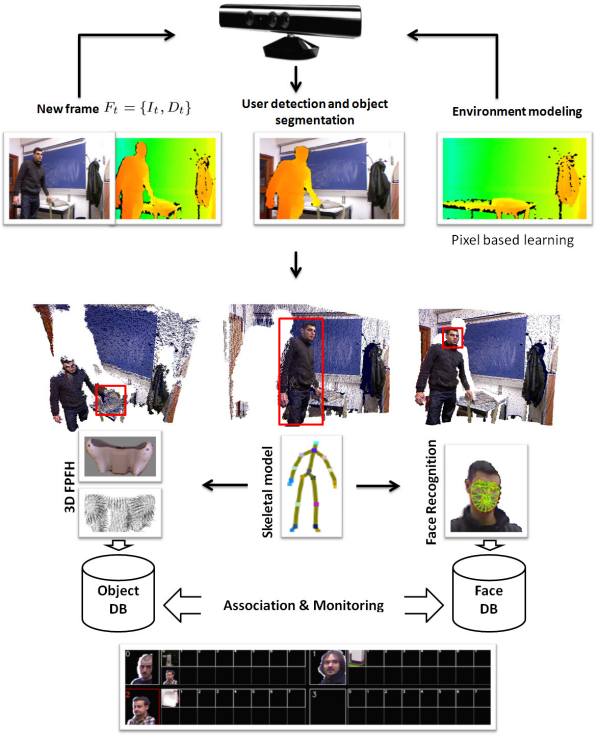


Fig. 1. Muti-modal user identification and object recognition surveillance system

Kinect infrared sensor, an adaptive model is learnt for each pixel. Supposing a RGBD Gaussian distribution for each pixel, the training procedure is performed as,

$$\mu_{\mathbf{X},t} = (1-\alpha)\mu_{\mathbf{X},t-1} + \alpha\left(\frac{D_{\mathbf{X},t}}{\max D_t} \cup I_{\mathbf{X},t}\right), \tag{1}$$

$$\sigma^2_{\mathbf{X},t} = (1-\alpha)\sigma^2_{\mathbf{X},t-1} + \alpha\left(\frac{D_{\mathbf{X},t}}{\max D_t} \cup I_{\mathbf{X},t} - \mu_{\mathbf{X},t}\right)^T\left(\frac{D_{\mathbf{X},t}}{\max D_t} \cup I_{\mathbf{X},t} - \mu_{\mathbf{X},t}\right), \tag{2}$$

where $\mu_{\mathbf{X},t}$ is the mean depth learnt at pixel $\mathbf{x} = (i, j)$ at frame t, α is a training weight of the parameters during learning, $D_{\mathbf{X},t}$ is the depth at pixel $\mathbf{x}$ at frame t, $I_{\mathbf{X},t}$ is the RGB values at pixel $\mathbf{x}$ at frame t, and σ^2 is the covariance. The computation of μ and σ given a fixed α value is performed during a perfect stationary background composed of T frames, so that $t \in [1, .., T]$. Once the background has been modeled, a new change of a pixel in the scene produced by the appearance/disappearance of items is detected as follows,

$$\sigma_{\mathbf{X},T} - \left| \frac{D_{\mathbf{X},t}}{\max D_t} \cup I_{\mathbf{X},t} - \mu_{\mathbf{X},T} \right| > \theta_S, \tag{3}$$

where $|.|$ corresponds to the absolute value and θ_S is an experimentally set background segmentation hypothesis value. At the top of Fig. 1 one can see the background modeling procedure, a new frame F, and the detection of a new item corresponding to a user in the scene.

2.2 User Detection and Identification

Given the segmented image M that contains 1 at those positions satisfying Eq. 3 and 0 otherwise, the procedure for user detection and identification is only applied on the activated pixels of M. The algorithm for user detection and identification is summarized in Algorithm 2. Note that we track each particular user based on its distance to previous detections in time, as well as the counter for the n identifications is treated for each user independently. Moreover, temporal coherence is taken into account by filtering the detections in time based on region density and 3D coordinates, discarding isolated detections and recovering miss-detections, resulting in a reduction of false detections and allowing a continuous detection of objects and users within the sequence.

User Identification Procedure. For the user identification module we propose to use the combination of body color model $\mathcal{C}$ with the face recognition probability $\mathcal{F}$ based on the matching of visual features, defining the following energy functional,

$$E(c_i, u) = \mathcal{C}(H_u, H_i) \cdot \beta + \mathcal{F}(f_u, f_i) \cdot (1 - \beta), \tag{4}$$

where β is a trade-off energy parameter. Energy functional $E \in [0, 1]$ is computed between a new test user $u = \{H_u, f_u\}$ and a candidate user class $c_i = \{H_i, f_i\}$, where H_i is the set of RGB color histograms for user i, and f_i is the set of face descriptions. Given a set of k possible users $C = \{c_1, .., c_k\}$ learnt online by the system, using the energy functional of Eq. 4, the new user candidate u is identified as follows,

$$\begin{array}{ll} i & \text{if} \quad E(c_i, u) > \theta_u, E(c_i, u) > E(c_j, u), \forall j \in [1, k], i \neq j \\ 0 & \text{otherwise} \end{array} \tag{5}$$

```
   Data: M_t, F_t, count, n
 1 if count < n then
 2     a) User detection [4] on D_t for the activated pixels in M
 3     if Detected user then
 4         b) Skeletal model description [4] on the pixels corresponding to the detected user
 5         c) Run Viola & Jones lateral and frontal face detectors on the surrounding areas to
           the detected head joint after background removal
 6         if Detected face then
 7             d) Use Active Shape Model with a set of face landmark to align the detected
               face to the closest data set training sample for each subject based on the mesh
               fitting error
 8             e) Compute user body color histogram excluding face region (section 2.2)
 9             f) Perform user identification (section 2.2)
10             g) Save the partial user identification ID_count to the class of the closest user
               probability, or 0 if none of the possible users achieve a probability threshold θ_u
11             count++
12         else
13             count=0
14         end
15     else
16         count=0
17     end
18 else
19     h) Assign class label to subject based on majority voting of ID or define new user if the
       majority vote is 0 count=0
20 end
```

Algorithm 2. User detection and identification algorithm

In the case that the new user defines a new model (classification label 0), it is used to update the user model C with a new identifier $C = C \cup \{H_u, f_u\}$. In the case that the user has been identified as a previously learnt user, the user model can be updated if the energy E for the classified user is bellow a particular update threshold parameter, so that if $E(c_i, u) < \theta_u$ for the identified user i, then $c_i = \{H_i, f_i\} \cup \{H_u, f_u\}$, subtracting the oldest data to reduce an uncontrolled growing of model information. Next, we describe the computation of the color and face models.

Color Model Computation $\mathcal{C}$. Once a new user is identified in the environment, a predefined number of color histograms is defined, computed, and saved in the histogram set H_i for user i. Each histogram in this set is computed as a 62 bin normalized histogram (30-H and 32S) from HSV color representation (PDF of the HSV data for the subject) for each frame considered to model the user body color model, without considered the region of the subject detected as the face region. Once a new candidate user u is detected by the system, its color model histogram is computed and compared with each learnt possible user i, defining the energy $\mathcal{C}(H_u, H_i)$ of Eq. 4. This energy is based on the Bhattacharyya distance of two histogram distributions $\mathcal{B}(h_u, h_i) = \sqrt{1 - \sum_j \frac{\sqrt{h_u^j \cdot h_i^j}}{\sqrt{\sum_j h_u^j \cdot \sum_j h_i^j}}}$, where h_i^j is the j-th position of one of the histograms of the set H_i. Once this distance is computed among the candidate user u and each histogram in the training set, the m lowest distances for each user class are selected to compute the

mean confidence for that class. Thus, the final color energy term is defined as $\mathcal{C}(H_u, H_i) = \frac{\sum_m 1-\mathcal{B}(h_u,h_m)}{m}$ for the m largest confidences (lowest Bhattacharyya distances) for candidate user i.

Face Model Computation $\mathcal{F}$. Describing in more detail lines 7-10 of Algorithm 2, our steps for face model computation are,

• We perform face alignment after face detection and background removal using Active Shape Model by means of linear transformation of position, rotation, and scale, computed using the mesh fitting changes.

• We use fast SURF point detection and description on the RGB user face f_u and each candidate face f_i for user i.

• We match SURF features between f_u and f_i using nearest neighbor assignments using a k-d tree with Best-bin-first search [6].

• We use RANSAC to discard final outliers based on the difference of the pair of features assignment to the computed linear transformation. Inliers are selected based on linear least squares.

• Using the initial set of v descriptions and the w final selected inliers, we compute a probabilistic membership of user model f_u to face model f_i for class i as follows [7]: Let $P(y|\neg f_i)$ be the probability that the matched features y would arise by accident if the model f_i is not present. We assume the w feature matches arose from v possible features, each of which matches by accident with probability p. Therefore, we can use the cumulative binomial distribution for the probability of an event with probability p occurring at least w times out of v trials $P(y|\neg f_i) = \sum_{j=w}^{v} \binom{v}{j} p^j (1-p)^{v-j}$. To compute $P(f_i|y)$ we use Bayes' theorem $P(f_i|y) = \frac{P(y|f_i)\cdot P(f_i)}{P(y|f_i)\cdot P(f_i) + P(y|\neg f_i)\cdot P(\neg f_i)}$. We approximate $P(y|f_i)$ as 1 as we normally expect to see at least w features present when the model is present. We also approximate $P(\neg f_i)$ with the value 1 as there is a very low prior probability of a model appearing at a particular pose. Therefore, our face energy model $\mathcal{F}$ is computed as $\mathcal{F}(f_u, f_i) = P(f_i|y) \approx \frac{P(f_i)}{P(f_i)+P(y|\neg f_i)}$. As in the case of the color model $\mathcal{C}$, detected faces are used online to update the user model of faces either for the case of a new user or for the case of previously identified user. Figure 2 shows real application examples of the user identification approach based on the face energy $\mathcal{F}$.

2.3 Object Recognition

Each segmented region (connected component) of M which has not been identified as a user is considered as a new object in case where the distance to the camera at those segmented pixels in D are reduced from the modeled background, or as the absence of an object if depth values increase. The case where an object has been removed is straightforward to analyze since we saved the description of the object located at those positions from previous frame description. This means that if a user picks an object, we immediately know looking at the label of the object from the removed location which object it was.

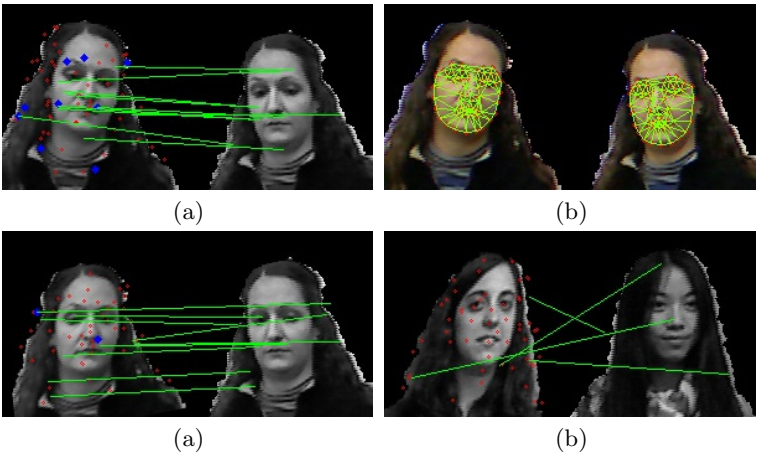


Fig. 2. Face identification analysis. Red dots: SURF candidate keypoints not matched based on descriptor distance. Blue dots: candidate keypoints discarded as outliers using RANSAC based on mesh transformation criteria. Green line: final matches considered for identification. (a) Example of images not aligned after face detection and background removal. Several outliers are detected using RANSAC (blue dots), reducing final identification probability of being the same user category (71.4% of probability in this example). (b) Shows the intermediate results of applying ASM meshes to both faces before alignment. (c) Applying the whole proposed process. Now the probability of identification increases up to 98.4%. (d) An example of alignment and identification for two different categories, with a result of 32.3% of probability.

In the case that a new object is located in a scene by a user, we take advantage of the 3D object information provided by the depth map D to compute a normalized description of that particular 3D view [5]. For this task, we take use of the recently proposed Fast Point Feature Histogram (FPFH) to compute a 3D rotation invariant object description for each particular point of view of an object $\mathcal{P}$ in the scene. A visualization of the descriptors for a set of objects is shown in Fig. 3. This procedure is performed for each new object cluster in M, and the object description is compared to the data set of descriptions saved in memory as in the case of the user color model $\mathcal{C}$. In this case, k-NN are used to classify the new object view as a previous detected object if it achieves majority voting and a threshold value over object threshold θ_o, being also used to update online the data set of object descriptions. In cases where two objects achieve high similarity with the new sample, we update the model and fuse two previous object descriptions. An example of object segmentation and 3D visual description using FPFH is shown in the middle of Fig. 1 for a detected object in the scene.

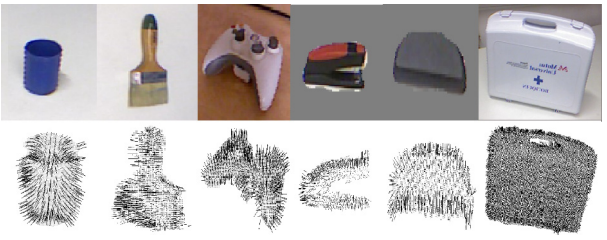


Fig. 3. Views of different objects and descriptions based on the normal components

2.4 User-Object Interaction

The analysis of object-user interaction is based on the definition of pairs of values (user,object) for those new objects that appear in the scene or those users that pick up an object, looking for past memberships in order to activate the required surveillance alerts. Some interface examples are shown in Fig. 5.

3 Results

In order to present the results of the proposed system, first, we discuss the data, methods and parameters, and evaluation measurements of the different experiments.

• **Data.** We defined a novel set of data recorded with the Kinect device. The data set consists of 10 videos of one minute each one in indoor scenes and 5 videos of one minute each one in outdoor scenes. The whole data set contains a total of 23600 semi-supervised labeled frames, containing a total of 8 different subjects and 11 different objects.

• **Methods and Parameters.** The values of our method parameters have been experimentally set via cross-validation. We also compare the proposed system with state-of-the-art methods: SURF and Bag-of-visual-words (BOVW) description, and the effect of background substraction and face alignment for user identification. Finally we also compare with RGB SIFT description in the case of object classification.

• **Evaluation Measurements.** We compute the performance of the system in terms of user detection, user identification, object detection, object classification, user-object association, and theft. For each of these evaluations we measure the number of true positives, false positives, and false negatives.

3.1 Surveillance System Evaluation

The mean global performance of the presented surveillance system is shown in Fig. 4. The Y-axis corresponds to the absolute value of true positives, false

positives, and false negatives for each event category. One can see that we are able to correctly detect most of the events, corresponding to an accuracy upon 90%. Most true positives are detected. False positives are almost non existent except for the case of object detection, where small noisy regions of the image are sporadically detected as small objects. Only few false positives occur in the case of user identification and theft, where an error in the case of object or user

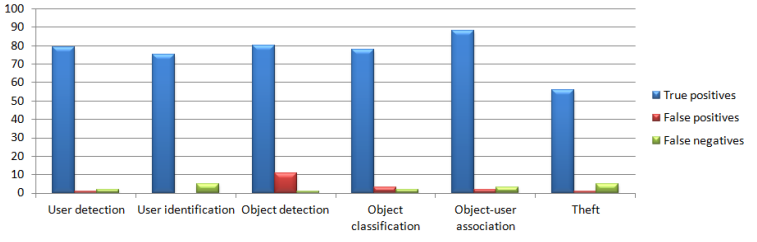
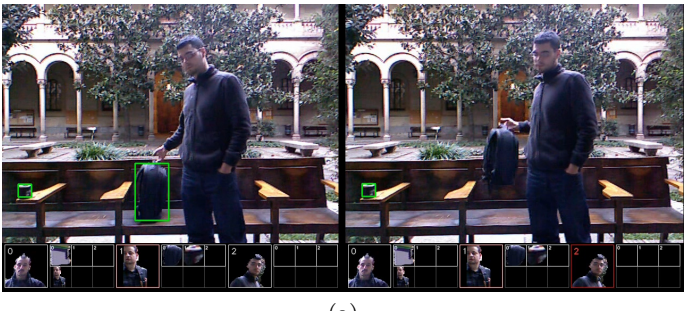
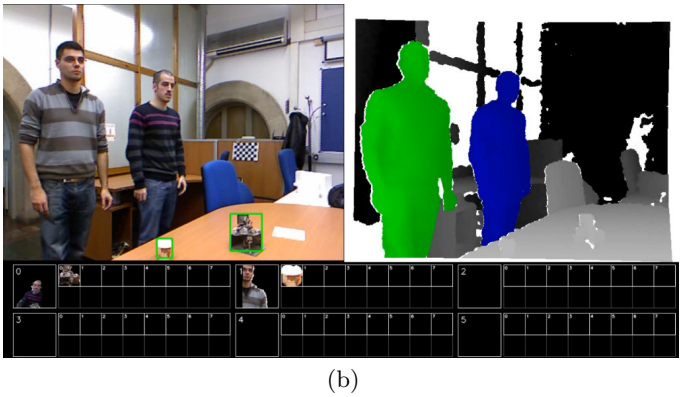


Fig. 4. Mean surveillance system performance



(a)



(b)

Fig. 5. (a) Outdoor scenario: user is identified, theft is recognized, and different objects, included a small cup are detected. (b) Users and object memberships are correctly identified and classified. Different users can be identified simultaneously by the system.

detection/recognition immediately propagates an error in the final theft detection step. Some qualitative results of the execution of the surveillance system are shown in Fig. 5.

3.2 User Identification Comparative

In Table 1 we show the identification accuracy of our method (Statistical Surf) and the standard SURF description using Bag of Visual Words (SURF BOVW) [8] for the user identification module of our system. Moreover, for each of these two configurations, we test the effect of removing background and aligning faces. In particular, A, $\overline{A}$, B, and $\overline{B}$ correspond to aligned, not aligned, with background, and background substraction, respectively. Comparing these approaches on the data set, one can see that removing background not only reduces the posterior complexity of the approach but also improves final identification performance. Aligning the face also increases the performance. Finally, one can see the robustness and better performance of our approach compared to the classical SURF BOVW technique, with a global mean improvement of 20% for the best configuration between both approaches.

Table 1. User identification performance results

SURF BOVW				STATISTICAL SURF			
$B+\overline{A}$	$B+A$	$\overline{B}+\overline{A}$	$\overline{B}+A$	$B+\overline{A}$	$B+A$	$\overline{B}+\overline{A}$	$\overline{B}+A$
33.3%	47.1%	52.8%	74.4%	52.9%	60.9%	76.3%	96.4%

3.3 Object Recognition Comparative

In order to analyze the high discriminative power of the used FPFH descriptor encoding the normal vector distributions of a 3D object view, we compare the obtained recognition results with the standard object description using SIFT on the RGB segmented object region. The results are shown in Table 2. One can see that contrary to the state-of-the-art SIFT descriptor, the 3D-normal vector distributions improve classification results in 12% in the presented experiments.

Table 2. Object recognition performance results

RGB SIFT	DEPTH FPFH
86.2%	98.5%

4 Conclusion

We proposed an automatic system for user identification and object recognition based on multi-modal RGB-Depth data analysis. We modeled a RGBD environment learning a pixel based background Gaussian distribution. Then, user and

object candidate regions were detected and recognized using robust statistical approaches. The system was evaluated on a novel data set containing different indoor and outdoor scenarios, objects, and users, showing accurate recognition results and better performance than classical approaches.

References

1. Lipton, A.J., Fujiyoshi, H.: Moving target classification and tracking from real-time video. In: Proceedings of the Fourth IEEE Workshop on Applications of Computer Vision, WACV 1998, pp. 8–14 (1998)
2. Elgammal, A., Harwood, D., Davis, L.: Non-parametric Model for Background Subtraction. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 751–767. Springer, Heidelberg (2000)
3. Brown, L.M., Senior, A.W., Li Tian, Y., Connell, J., Hampapur, A., Fe Shu, C., Merkl, H., Lu, M.: Performance evaluation of surveillance systems under varying conditions. In: Proceedings of IEEE PETS Workshop, pp. 1–8 (2005)
4. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images (2011)
5. Rusu, R.B.: Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments. Articial Intelligence (KI - Kuenstliche Intelligenz) (2010)
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
7. Lowe, D.G.: Local feature view clustering for 3d object recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 682–688 (2001)
8. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, pp. 1–22 (2004)