3D information is valuable for the detection of humans in video streams

Sébastien Piérard
INTELSIG Laboratory
Montefiore Institute
University of Liège, Belgium
Email : Sebastien.Pierard@ulg.ac.be

Antoine Lejeune
INTELSIG Laboratory
Montefiore Institute
University of Liège, Belgium
Email : Antoine.Lejeune@doct.ulg.ac.be

Marc Van Droogenbroeck
INTELSIG Laboratory
Montefiore Institute
University of Liège, Belgium
Email : M.VanDroogenbroeck@ulg.ac.be

***Abstract*—In this paper, we propose a technique based on 3D information (also called depth or range) for the detection of humans. First, a background subtraction technique operates to detect the silhouettes of humans and objects moving in the scene. Then, a machine learning algorithm is used to predict if a silhouette annotated with depth matches a human silhouette or not. The complete method is designed to cope with defects introduced during the segmentation step.**

Results, obtained on computer generated data, show that 3D depth data is a valuable information for detecting humans in that it improves over techniques based on binary silhouettes. In our experiments, we have reached an accuracy of 99.9% thanks to the depth information.

***Index Terms*—3D, Range, Depth, People detection, Video analysis, Video surveillance**

I. Introduction

Human detection in video scenes is a crucial task for a large variety of applications including video surveillance. So far, methods to detect humans were developed for grayscale or color images because of their large availability. Nowadays, 3D range sensors become affordable. Therefore we have considered the use of 3D cameras (also called depth cameras) for detecting humans in videos, particularly because techniques based on grayscale or color images have to be improved.

In this paper, we study the impact of adding 3D information by adapting a technique proposed by Barnich *et al.* [1], which is based on geometric information. After a short overview of related works, we introduce, discuss and modify Barnich's method in Section II. Section III explains how we build attributes that include 3D depth information in our method. In Section IV, we describe the (synthetic) data used in the experiments and present our results. Section V concludes the paper.

A. Approaches based on silhouettes

There are two main approaches to detect humans depending on whether temporal coherence between successive images is considered or not. When the temporal coherence is ignored, each image is processed separately. A popular approach to detect humans in static images, based on *Histograms of Oriented Gradients*, was proposed by Dalal and Triggs [2]. Generally, such methods have two weaknesses. These techniques are based on appearance which largely depends on lightning conditions and are unpredictable in uncontrolled scenes. In addition, they require to process a large number of detection windows. This impacts both on the processing speed and on performance. Indeed, these techniques can not focus on moving regions and a large number of detection windows are therefore located in the background. As a result, it is mandatory to keep a low rate of false positives, which unfortunately implies a limitation on the rate of true positives.

A better approach consists in analyzing successive frames to select the foreground only. Background subtraction algorithms use this temporal coherence to extract the silhouettes of moving objects and persons in the scene. This approach has been followed by Barnich [1] and Diaz de Leon [3]: each silhouette is classified as human or non-human (see Figure 1). They based the decision on the geometric information present in silhouettes, and ignored any information related to appearance.

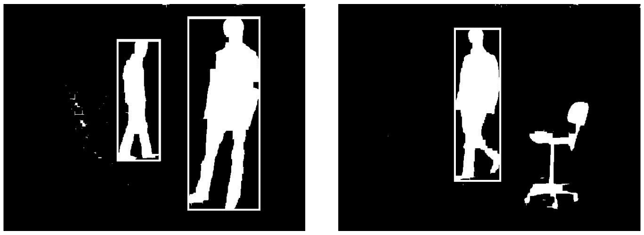


Figure 1. Results of a person detection technique as proposed by Barnich *et al.* [1]. Objects included in rectangular boxes are classified as human silhouettes (images taken from [4]).

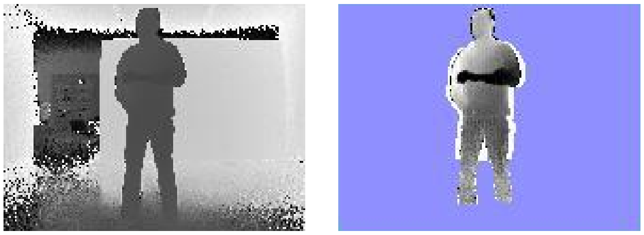


Figure 2. Background subtraction algorithms can be applied on depth maps. On the left hand side, a depth map acquired by a range camera (*PMD[vision]19k*). On the right hand side, the result of the background subtraction algorithm [5] after a distance normalization.

It should be noted that background subtraction algorithms can be applied to depth maps (see Figure 2). For example, the state of the art algorithm *ViBe* [5] has been used in [6]. Techniques based on silhouettes are thus applicable to 3D video streams without any modifications. However, as shown in this paper, the use of depth improves the detection performance.

B. Describing silhouettes

Once silhouettes are extracted from the video stream, one has to classify them. In order to use machine learning algorithms, silhouettes must be summarized in a fixed amount of information called *attributes*. Popular techniques to compute attributes include image moments introduced by Hu [7] and Fourier descriptors (used for example in [3]). In these methods, each attribute is global, meaning

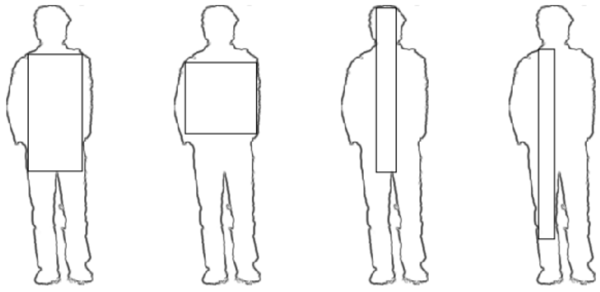


Figure 3. Largest rectangles included in a silhouette (reproduced from [4]).

that it depends on the whole silhouette. Thus, when silhouettes present flaws, all the attributes are corrupted.

To evaluate the importance of 3D information, it is preferable to limit the scope of noise to the local neighborhood. In Barnich *et al.* [1], silhouettes are split in a set of smaller regions to decrease the influence of silhouette defects. As expected, Barnich showed in [4] that, in practice, this approach yields to better results than those based on Hu's moments.

II. Our detection method

A. Description of Barnich's method

Our technique adapts and extends the algorithm proposed by Barnich *et al.* [1]. In that algorithm, silhouettes are decomposed in the set of all the largest rectangles that can be wedged inside the silhouette (see Figure 3) and that are aligned with the image axes.

In a first step, each rectangle is given one out the two following labels: the $+$ label if the rectangle belongs to a human silhouette, the $-$ label if not. This labelization results from a machine learning algorithm named *ExtRaTrees* [8] which consists in a set of decision trees. Let $\Pi_+(r)$ be the proportion of trees classifying a rectangle r as human. Barnich *et al.* used the following decision rule to assign a class $\mathrm{y}(r)$ to a rectangle r

$$\mathrm{y}(r) = \mathrm{sign}\left(\Pi_+(r) - \frac{1}{2}\right). \tag{1}$$

Then, in a second stage, the class with the most votes is assigned to the silhouette s. If we denote $\Psi(s)$ the set of rectangles that are used to take a decision about the class of the silhouette (see Section II-B), this rule can be expressed as

$$\mathrm{y}(s) = \mathrm{sign}\left(\sum_{r\in\Psi(s)} \mathrm{y}(r)\, w(r)\right), \tag{2}$$

where the function $w(\cdot)$ (which is supposed to be positive) gives a weight to each rectangle. Barnich *et al.* chose $w(r) = 1$. Marée [9] was also faced with the classification problem of a composite object based on the classification results of several elementary objects, but in another context. He proposed to use $w(r) = \left|\Pi_+(r) - \frac{1}{2}\right|$. In fact, the choices of Barnich *et al.* and Marée are arbitrary, and the decision rule (1) is just one possible rule. In fact, $\mathrm{y}(r)\, w(r)$ can be learned automatically with a linear classifier such as the linear Support Vector Machine methods. We will see in Section IV-B that these different ways to weight rectangles lead to similar results, although the weighting function learned automatically has a slight advantage.

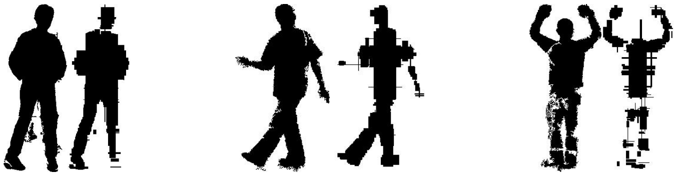


Figure 4. The effect of randomly selecting 100 maximal rectangles. This figure shows three original human silhouettes (on the left hand side), and the silhouettes reconstructed with a random subset of 100 maximal rectangles (on the right hand side).

B. The selection of rectangles

To analyze a silhouette s, it is impractical to consider the complete set of the largest included rectangles for the following reasons.

When learning the model $r \rightarrow \Pi_+(r)$, an equal number of rectangles must be used for each silhouette to avoid a bias in the decision rule. In [4], Barnich showed that the best results are obtained for a number of selected rectangles comprised between 50 and 200. In this work, we use 100 rectangles per silhouette.

A subset of rectangles should also be used to predict the class of silhouettes, for efficiency reasons. Figure 4 shows the effect of a random selection limited to 100 rectangles.

Recently, Barnich showed in [4] that a random selection of rectangles gives better results than deterministic strategies such as the selection of the largest rectangles or the maximization of the covered area. Thus, in this paper, $\Psi(s)$ is a random subset of the set of all the maximal rectangles contained inside a silhouette s.

III. The attributes

In this section, we describe how to transform Barnich's method to include a depth map.

A. Desired invariants

We require that the attributes used to describe rectangles have several invariance properties:

1) The method has to be invariant to the location of users in the observed scene (the variety of poses and orientations are integrated in the learning set). We assume that users remain at a reasonable distance from the camera. Under this assumption, the motion of the users results in a scaling and a translation of the silhouettes in the image plane, and in an offset of the 3D (depth) information.
2) The method has to be insensitive to an horizontal flip of the analyzed silhouette. This is required because a flip does not change the class of the silhouette.
3) Lastly, as described in Section IV-A, we use synthetic data to assess the performance of the method. The size of the 3D models we use to generate the set of non-human silhouettes is not expressed in physical units and thus cannot be compared with the real size of a human. Therefore attributes have to be insensitive to a modification of the depth scale.

B. Notations

We denote the width and height of an object o (which is a rectangle or a silhouette) by w_o and h_o respectively. The values $x(p)$, $y(p)$ denote the coordinates of a pixel p in the image plane and $d(p)$ its associated depth information. $\mathcal{M}_{v,w}(o)$ represents the weighted mean of the values taken by a function v over o; weights are given by a function w. Using these notations, we have

$$\mathcal{M}_{v,w}\left(o\right)=\frac{\sum_{p\in o}v(p)w(p)}{\sum_{p\in o}w(p)}.$$

C. *The attributes*

To describe a rectangle, we use eight attributes. The first four ones characterize the silhouette, and are not related to 3D information:

$$\left\{\frac{\left|\mathcal{M}_{x,1}\left(r\right)-\mathcal{M}_{x,1}\left(s\right)\right|}{w_s},\frac{\mathcal{M}_{y,1}\left(r\right)-\mathcal{M}_{y,1}\left(s\right)}{h_s},\frac{w_r}{w_s},\frac{h_r}{h_s}\right\}$$

These attributes are the same as the ones used in [1]. Barnich *et al.* also used an additional attribute that is the proportion of pixels covered by a sole rectangle. However, this attribute is not scale-invariant and therefore it was discarded.

The last four attributes characterize the depth:

$$\left\{\frac{\left|\mathcal{M}_{x,d}\left(r\right)-\mathcal{M}_{x,1}\left(r\right)\right|}{w_r},\frac{\mathcal{M}_{y,d}\left(r\right)-\mathcal{M}_{y,1}\left(r\right)}{h_r},\right.$$

$$\left.\mathcal{M}_{d,1}\left(r\right),\sqrt{\frac{1}{w_rh_r-1}\sum_{p\in r}\left(d\left(p\right)-\mathcal{M}_{d,1}\left(r\right)\right)^2}\right\}.$$

It should be noted that these four last attributes have not been designed to be insensitive to noise on the distance estimate. As mentioned before, to assess our method, we did not use data obtained from a range camera but, instead, we decided to use a database of computer generated silhouettes (see Section IV-A). Indeed, computer generated silhouettes are free of noise. However, depth map obtained from depth cameras are often very noisy. This means that the last four attributes will probably have to be reconsidered and adapted for real applications.

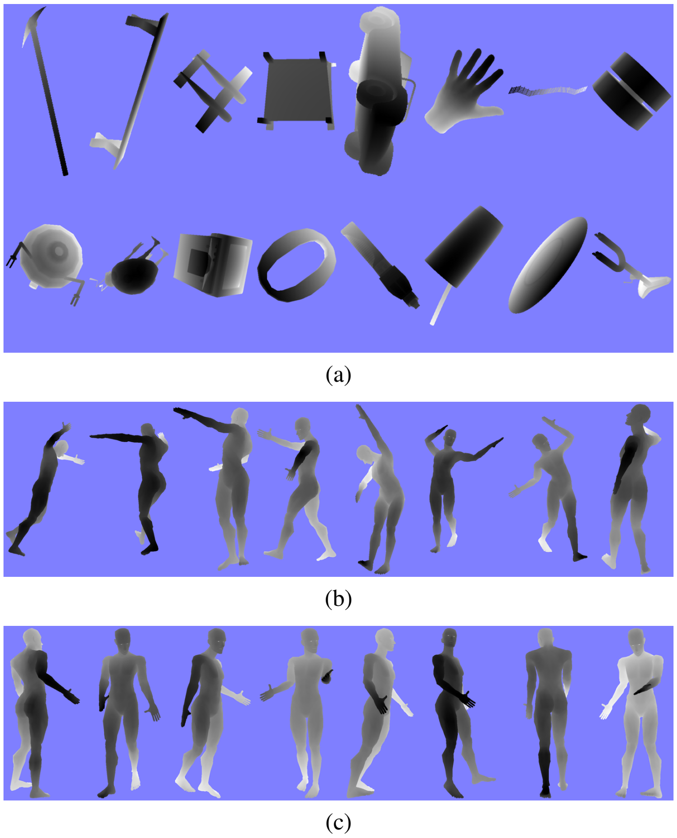
IV. Data sets and results

A. *Data sets*

To evaluate the contribution of the depth data, we need databases of human and non-human silhouettes annotated with depth. To our knowledge, there are no publicly available databases providing such real silhouettes. Because building such a database is an intensive job by itself, we decided to build our own databases filled with synthetic views of objects and humans. Silhouettes and depth images are derived from these views.

We created a set of $10,000$ non-human silhouettes using the NTU 3D model database [10]. The virtual camera is placed randomly around the object, at a distance such that the object is fully comprised in the image. Silhouettes composed of several connected components are discarded. It should be noted that the NTU database contains models of human-like objects. However, these silhouettes are taken from a totally arbitrary viewpoint and do not correspond to a standing (vertical) person. Examples of produced silhouettes are shown in Figure 5(a).

Human silhouettes have also been generated automatically. We used the avatar provided with the open source software *MakeHuman*[1], version 0.9. The pose of the avatar is controlled by a set of parameters. Realistic poses are obtained using the technique described in [11]. The virtual camera looks towards the avatar, and is placed randomly around him, in an horizontal plane. As we assume that, in a real application, the results are sensitive to the set of chosen human poses, we created two different sets of $10,000$ human silhouettes: one set with a high pose variability and one with silhouettes closer to the one of a walker. They correspond to the sets $\mathcal{B}$ and $\mathcal{C}$ of [11] and are shown in Figures 5 (b) and (c) respectively.


(a)

(b)

(c)

Figure 5. Examples of non-human and human synthetic silhouettes annotated with depth. (a) Non-human silhouettes. (b) Human silhouettes with a weakly constrained set of poses. (c) Human silhouettes with a strongly constrained set of poses. Dark and bright values respectively denote the nearest and farthest points.

B. *Results*

We provide the results for two particular experiments: one experiment where human poses are strongly constrained, and an experiment where the arms are completely free to move (this is a most challenging situation because the diversity of silhouettes in the class to be recognized is higher). In both experiments, the learning set, denoted *LS*, and the testing set, denoted *TS*, are equally distributed, and contain 5000 human silhouettes and 5000 non-human silhouettes. The selection of silhouettes in *LS* and *TS* is random; in addition, we ensure that the silhouettes in *TS* are different from silhouettes contained in *LS*.

Detection Error Trade-off (DET) graphs, plotting false positive rate versus false negative rate, are drawn in Figures 7 and 6. Two families of curves are displayed; they correspond to classification results for variations on weighting functions for silhouettes including or not depth (3D) attributes. These graphs present results similar to Receiver Operating Characteristic (ROC) curves, but the axes are logarithmic to enlighten the high recognition rates. Four conclusions can be drawn:

1) adding depth to silhouettes always improves the performance;
2) experiments with strongly constrained human silhouettes in the learning set (which results in a lower diversity) exhibit a better performance;

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

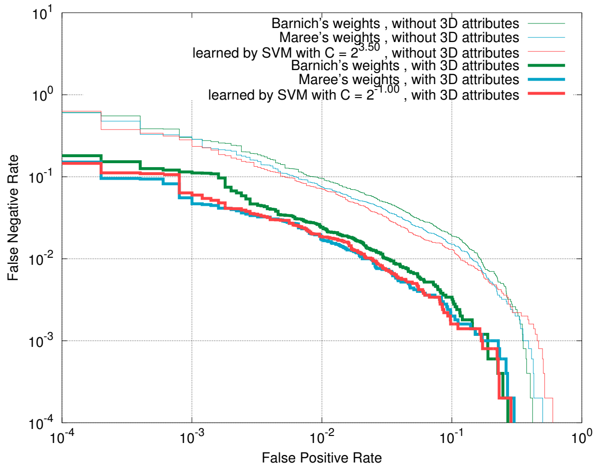


Figure 6. Results for a weakly constrained set of human poses.

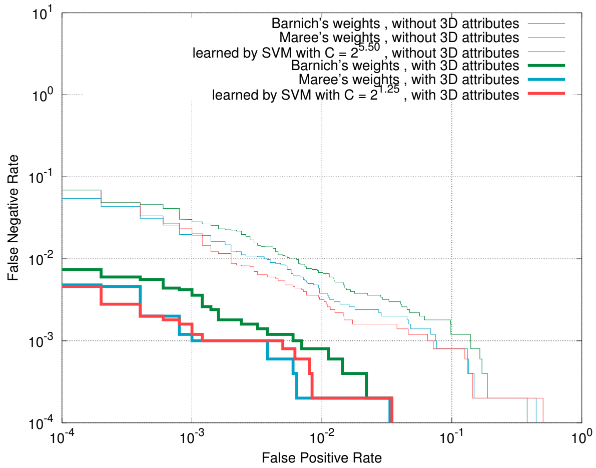


Figure 7. Results for a strongly constrained set of human poses.

3) when the human silhouettes used for learning are strongly constrained (like for a walker), true negative and true positive rates up to 99.9% can be reached;
4) several weighting functions were considered for the rectangles, as mentioned in Section II. The performance of these functions are almost similar, although the weighting function learned automatically provides a slight increase of accuracy.

V. Conclusion and future work

In this paper, we propose a technique that considers silhouettes annotated with depth information for the detection of humans in video sequences. Our technique extends Barnich's method to accommodate to 3D data and uses a machine learning algorithm to predict if a silhouette annotated with depth corresponds to a human silhouette or not.

Results obtained on synthetic silhouettes show that 3D depth data is relevant and useful for detecting humans in a scene; an accuracy of 99.9% has been reached on synthetic data. It remains to apply the method on real data and to propose attributes capable to deal with noise on depth signals.

Acknowledgments

S. Piérard has a grant funded by the FRIA, Belgium.

References

[1] O. Barnich, S. Jodogne, and M. Van Droogenbroeck, "Robust analysis of silhouettes by morphological size distributions," in *Advanced Concepts for Intelligent Vision Systems (ACIVS 2006)*, ser. Lecture Notes on Computer Science, vol. 4179. Springer, September 2006, pp. 734–745.

[2] N. Dalal and B. Triggs, "Hog, histograms of oriented gradients for human detection," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, San Diego, USA, June 2005, pp. 886–893.

[3] R. D. de León and L. Sucar, "Human silhouette recognition with Fourier descriptors," in *IEEE International Conference on Pattern Recognition (ICPR)*, vol. 3, Barcelona, Spain, September 2000, pp. 709–712.

[4] O. Barnich, "Motion detection and human recognition in video sequences," Ph.D. dissertation, University of Liège, Belgium, September 2010.

[5] O. Barnich and M. Van Droogenbroeck, "ViBe: a powerful random technique to estimate the background in video sequences," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2009, pp. 945–948.

[6] S. Piérard, J. Leens, and M. Van Droogenbroeck, "Real-time processing of depth and color video streams to improve the reliability of depth maps," in *Proceedings of 3D Stereo MEDIA*, Liège, Belgium, November 2009.

[7] M. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, pp. 179–187, 1962.

[8] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, April 2006.

[9] R. Marée, "Classification automatique d'images par arbres de décision," Ph.D. dissertation, University of Liège, Belgium, February 2005.

[10] D.-Y. Chen, X.-P. Tian, Y.-T. Shen, and M. Ouhyoung, "On visual similarity based 3D model retrieval," *Computer Graphics Forum*, vol. 22, no. 3, pp. 223–232, September 2003.

[11] S. Piérard and M. Van Droogenbroeck, "A technique for building databases of annotated and realistic human silhouettes based on an avatar," in *Workshop on Circuits, Systems and Signal Processing (ProRISC)*, Veldhoven, The Netherlands, November 2009, pp. 243–246.