S. PIÉRARD, A. LEJEUNE, and M. VAN DROOGENBROECK. **A probabilitic pixel-based approach to detect humans in video streams**. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, pages 921-924, May 2011.

A PROBABILISTIC PIXEL-BASED APPROACH TO DETECT HUMANS IN VIDEO STREAMS

S. Piérard , A. Lejeune, M. Van Droogenbroeck

INTELSIG Laboratory, Montefiore Institute, University of Liège, Belgium

ABSTRACT

Human detection in video streams is an important task in many applications including video surveillance. Surprisingly, only few papers have been devoted to this topic.

This paper presents a new approach to detect humans in video streams. Our approach is based on the temporal information present in videos. A background subtraction algorithm is first used to segment the silhouettes of the users and the moving objects. Then a classification process in two steps determines for each connected component if it corresponds to the silhouette of a human or not. During the first step, a probabilistic information is computed for each pixel independently. The information from a subset of pixels is then gathered to predict the class of the observed silhouette.

This paper presents the principles and some results obtained on real silhouettes. It is shown that our approach is efficient for the detection of humans in video streams.

***Index Terms*—** Identification of humans, Image sequence analysis, Image matching, Video surveillance, Video processing

1. INTRODUCTION

The number of cameras used worldwide for video surveillance is huge. These cameras produce large bit-streams that need to be interpreted automatically. In particular, a crucial task in video surveillance applications consists in detecting the presence of humans in the observed scene. This paper proposes a new approach to detect human silhouettes in video streams. In Section 2, we detail our method. Results are provided in Section 3 and Section 4 concludes the paper.

1.1. Previous works on images (not on video streams)

Most existing techniques detect humans in images. Among them, a popular approach is the technique proposed by Dalal and Triggs [1] that uses a set of Histograms of Oriented Gradients (*HOGs*). Image based detection techniques have two main drawbacks. The first one is that they are based on appearance (*that is* on colors and textures), which depends on lightning conditions and is unpredictable in uncontrolled scenes. These techniques manipulate a running window that is moved along the image and, consequently, require to process numerous overlapping windows, at multiple locations and scales, which is also a major drawback. For example, 12800 windows are considered for a 320×240 image in the work of Zhu *et al.* [2]. These authors showed that *HOGs* can be computed in real-time but that it comes at the cost of a performance reduction (this is because, among other things, it is not possible to use the Gaussian mask of [1] anymore). Finally, note that the Dalal and Triggs technique [1] has a miss rate of 10^{-1} for a false positive rate per window of 10^{-4}. This means that, without any further processing, a video surveillance system based on *HOGs* would produce 1.28 false alarm per 320×240 image while missing one person out of 10. This is not acceptable in a practical situation.

Fig. 1. Results of a person detection technique as proposed by Barnich *et al.* [3]. Objects included in rectangular frames are classified as human silhouettes (images taken from [4]).

1.2. Detection of humans in video streams

Barnich *et al.* [3] proposed an alternative approach to detect humans applicable to video streams. They used a background subtraction algorithm to extract the silhouettes of moving objects in the scene. The segmentation map is then split in its connected components to provide a set of silhouettes, as shown in Figure 1.

With silhouettes, we take advantage of the temporal information present in videos to extract them, while avoiding to base the decision on appearance. Moreover, silhouettes can be obtained, not only from color cameras, but also from other sensors like range cameras or laser scanners [5]. Techniques for classifying silhouettes have therefore a broad range of applications.

1.3. Describing silhouettes

In order to classify silhouettes with machine learning algorithms, silhouettes have to be summarized as a fixed amount of information called *attributes*. Popular techniques to compute attributes include the image moments introduced by Hu [6] and the Fourier descriptors [7]. Unfortunately, these techniques have an important drawback: each attribute is global, meaning that it depends on the whole silhouette. Thus, damaged silhouettes have noisy attributes. The solution to decrease the sensitivity to local silhouette modifications consists in cutting silhouettes in a set of smaller regions. Both the work of Barnich *et al.* [3] and this paper describe such techniques.

1.4. Description of Barnich's method

In the technique proposed by Barnich *et al.* [3], silhouettes are decomposed in a set of overlapping elements; this set includes all the largest rectangles that can be wedged into the silhouette (see Figure 2). Once the algorithm has computed the set of all rectangles, to each rectangle is separately given a class label (by means of a machine learning algorithm called "*ExtRaTrees*" [8]) that can be "rectangle belonging to a human silhouette" or "rectangle belonging to a non-human silhouette". Thereby, each rectangle votes for one class of silhouettes and the class with the most votes is assigned to the

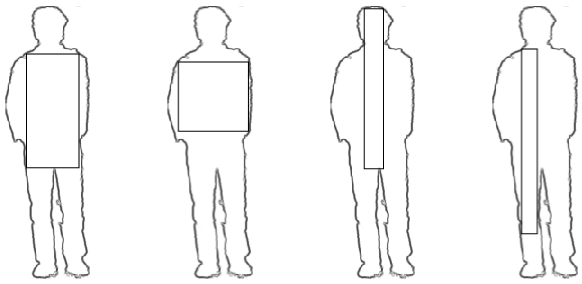

Fig. 2. Largest rectangles included in a silhouette (reproduced from [4]).

silhouette. Recently, Barnich showed in [4] that this approach yields better results than those based on Hu's moments.

From our experience we noticed that the results obtained by Barnich *et al.* are sensitive to the proportion of human silhouettes in the Learning Set (denoted *LS* hereafter). When one gives equal importance to correctly classify human and non-human silhouettes, it is imperative to keep an equal amount of human and non-human silhouettes in *LS*. In addition, we established that with non global attributes, the classes "human rectangle" and "non-human rectangle" are not separable. It turns out that the *ExtRaTrees* estimate probabilities, and that an unbalanced learning set introduces a bias in the estimate (this is further discussed in Section 2.3). In the following, we propose a new method that goes beyond the estimation of a probability to belong to a human silhouette.

2. OUR METHOD

Like in [3], we divide the silhouette in a set of elements. But, instead of assigning them to a class, we evaluate the probability that they originate from a human silhouette. In addition, we remove the voting scheme to replace the whole process by a two steps mechanism. First, we compute attributes for each pixel and associate a probability per pixel using *ExtRaTrees* (see Figure 3). Then, the set of computed probabilities is interpreted to predict the class of the corresponding silhouette (see Section 2.4).

2.1. Towards a pixel-based approach

The reason to replace the set of largest rectangles chosen by Barnich *et al.* by pixels is threefold:

1. Pixel related attributes permit to evaluate a local probability map (see Figure 3).
2. Probability maps can help to improve the segmentation mask, for example by locating parts of the silhouette which originates from shadows.
3. Attributes that characterize a pixel are computed on a neighborhood of the pixel. The choice of these attributes implies an implicit choice of the neighborhood. To date, literature gives no indication on the best size of the neighborhood to be considered. Pixel-based methods are helpful for the selection of an optimal neighborhood.

2.2. The attributes

It is hard to determine a prior best suited set of attributes. Therefore, we reuse the proved decomposition of silhouettes into the set of the largest included rectangles to build a robust characterization because this set of rectangles has proved to be resilient to noise. However, we derive information from this set to allocate it to pixels directly.

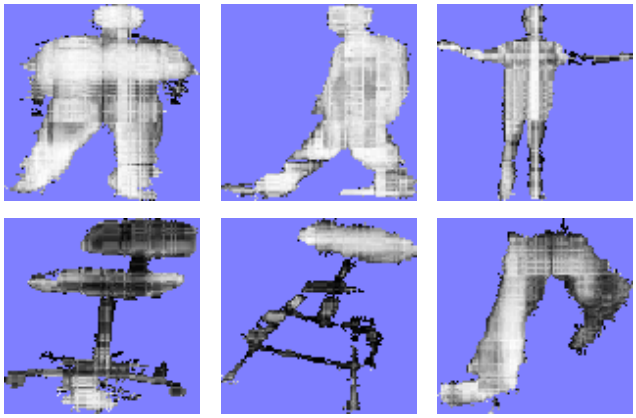

Fig. 3. Examples of probability maps. The upper row shows the estimated probabilities for 3 human silhouettes, and the lower one shows the probabilities for 3 non-human silhouettes. Dark and bright values respectively denote low or high probabilities for a pixel to be part of a human silhouette.

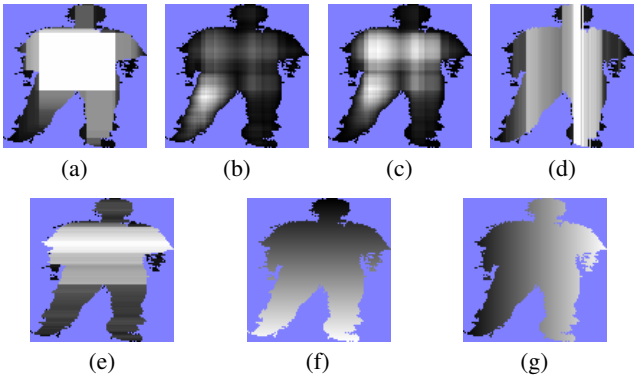

Fig. 4. Examples of attributes that are used for characterizing a pixel: (a) largest rectangle area; (b) number of rectangles; (c) sum of the areas of all rectangles; (d) largest height; (e) largest width; and (f-g) the position relative to the silhouette center of gravity.

To each pixel contained in a silhouette, we evaluate 10 attributes from the set of all largest rectangles containing that pixel. We also consider 2 additional location attributes to encode the relative location of the pixel with respect to the center of gravity of the silhouette (see Figure 4 for a subset of these 12 attributes). Note that we don't have to measure the relevance of an attribute since the *ExtRaTrees* select the most useful attributes automatically. In addition, to avoid learning the typical size of human and non-human silhouettes, we resize and stretch all silhouettes to fit a 100×100 pixels wide bounding box before computing the attributes.

2.3. Estimation of probabilities

Let η be the number of attributes used to describe a pixel. Each pixel is mapped onto a point of an η-dimensional space $\mathcal{S}$. Now, consider all possible pixels of a human silhouette. In the space $\mathcal{S}$, their statistical distribution follows a Probability Density Function,

denoted pdf hereafter, $\rho_+(\cdot)$. Likewise, $\rho_-(\cdot)$ is the pdf associated to pixels of the non-human class. Assume that the two classes occur at the same frequency. In this case, the probability of a pixel x to be part of a human silhouette (as shown in Figure 3) is given by

$$p_+(x) = \frac{\rho_+(x)}{\rho_+(x) + \rho_-(x)}.$$

The classification technique of *ExtRaTrees* compute several trees that can be combined to estimate $p_+(x)$. Let $\Pi_+(x)$ be the proportion of trees voting for the human class. We propose the following estimator for $p_+(x)$

$$\widehat{p_+(x)} = \frac{n_- \Pi_+(x)}{n_+ + (n_- - n_+)\Pi_+(x)}$$

where n_+ and n_- are respectively the total amount of human and non-human pixels in the *LS*. For commodity, we assign the labels $+1$ and -1 respectively to the human and non-human classes. Following Bayes' decision rule, we select a class for a pixel x according to

$$\mathrm{y}(x) = \mathrm{sign}\left(\widehat{p_+(x)} - \frac{1}{2}\right) = \mathrm{sign}\left(\Pi_+(x) - \frac{n_+}{n_+ + n_-}\right).$$

The probability for this decision to be correct is given by $\mathrm{r}(x) = \frac{1}{2} + \left|\widehat{p_+(x)} - \frac{1}{2}\right|$. Barnich did not consider the impact of n_+ and n_- and used $\mathrm{y}(x) = \mathrm{sign}(\Pi_+(x) - 0.5)$ instead.

2.4. The second step: classification of silhouettes

The advantage of probability maps with respect to a simple majority vote is that we can further enforce high probabilities with appropriate weights. Therefore we introduce the pixel dependent weights $w(x) \geq 0$. Let us denote the set of pixels used to characterize a silhouette s by $\Psi(s)$. We propose the following weighted decision rule for a silhouette s

$$\mathrm{y}(s) = \mathrm{sign}\left(\frac{1}{|\Psi(s)|}\sum_{x\in\Psi(s)} \mathrm{y}(x)\, w(x) + b\right),$$

where b is a parameter that permits to move on the ROC curve. The last question is how we determine the weights. One could define the weighting function $w(x)$ in terms of $\mathrm{r}(x)$. However, in the following subsection, we show that the determination of $\mathrm{y}(s)$ is a linear classification problem, and that the optimal separating hyperplane is related to $w(\cdot)$. Subsequently, $w(\cdot)$ can be learned automatically.

2.4.1. *Silhouettes classification as a linear classification problem*

Let T be the number of trees used to compute $\Pi_+(\cdot)$. There are $T+1$ possible values for $\Pi_+(\cdot)$ since $\Pi_+(\cdot) \in \left\{\frac{0}{T}, \frac{1}{T}, \ldots, \frac{T-1}{T}, \frac{T}{T}\right\}$. This is also true for $\mathrm{y}(\cdot)$, $\mathrm{r}(\cdot)$, and $w(\cdot)$ since they depend only on $\Pi_+(\cdot)$.

Let us denote δ the Kronecker delta, and $\bullet$ the scalar product. All the available information about a silhouette s can be represented as a vector S with $T+1$ dimensions which gives the proportion of pixels for each value of $\Pi_+(\cdot)$:

$$S(j) = \frac{1}{|\Psi(s)|}\sum_{x\in\Psi(s)} \delta(j, T\Pi_+(x)) \quad \forall j \in \{0, 1, \ldots, T\}.$$

If we define a weighting vector W as $W(T\Pi_+(x)) = \mathrm{y}(x)\, w(x)$, then $\mathrm{y}(s) = \mathrm{sign}(S \bullet W + b)$. In other words, the optimal weighting function $w(\cdot)$ can be found by solving a linear classification problem.

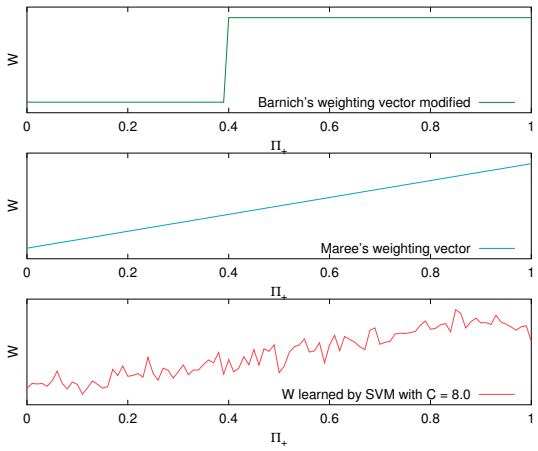

Fig. 5. Three weighting vectors. The learning set contains 39% of human silhouettes.

2.4.2. *Three weighting functions*

In this paper, we compare the results obtained with three different weighting functions, shown in Figure 5. Note that the same ROC curves are obtained for W and $\alpha W + \beta W_1$ where $W_1(\cdot) = 1$, if $\alpha > 0$. This explains why the vertical axes of Figure 5 are not graduated.

The weighting function of Barnich. [3] Barnich used a simple voting scheme by giving the same weight to all the elements, but he didn't take into account that $n_+ \neq n_-$ in the learning set. However, we believe that his intention was to use

$$w(x) = 1 \Leftrightarrow W(j) = \mathrm{sign}\left(\frac{j}{T} - \frac{n_+}{n_+ + n_-}\right)$$

This is referred to as "Barnich's modified weighting vector".

The weighting function of Marée. [9] Marée was also faced with the problem of taking a decision about the class of a composite object based on the classification results of several elementary objects, but in another context. He proposed to use a linearly increasing weighting vector: $W(j) = j$.

An automatically learned weighting function. The two previous weighting vectors were just two possible choices. So, we decided to use a machine learning algorithm to learn the weighting vector from *LS*. The algorithm we tried is the linear C-SVMs. The parameter C was chosen by cross-validation (5 folds) on *LS* to maximize accuracy. We found that the optimal value for C is 8, for an accuracy about 99.56%.

3. MATERIAL, METHODOLOGY AND RESULTS

3.1. Databases

We used the same learning and testing databases as Barnich *et al.* [3]. The *LS* and Test Set *TS* have statistical characteristics that slightly differ; this can be shown using the visual signature of databases introduced in [10]. This is common in practice and informative because it allows to evaluate the generalization ability.

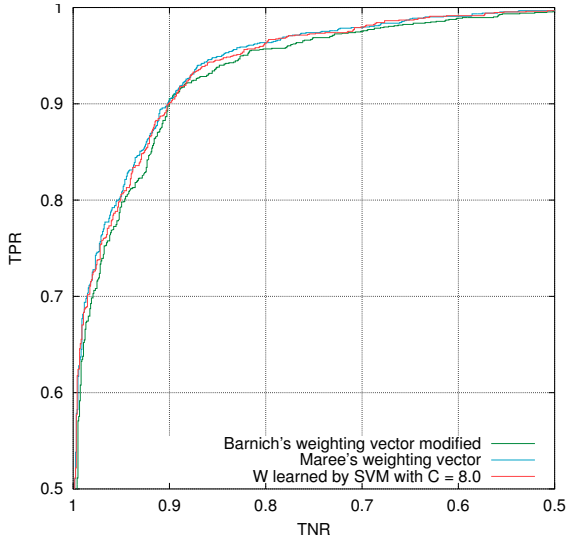

Fig. 6. ROC curves obtained on the Test Set (*TS*) for the three weighting vectors shown in Figure 5. As mentioned, *TS* and *LS* have different statistical characteristics. These curves therefore present the results in the generalization case.

Another common difficulty is the presence of partly occluded human silhouettes in the databases. Unfortunately the used databases contain such silhouettes. Therefore we must interpret the results with some caution.

3.2. Methodology: pixel selection

We use a random pixel selection process for three different tasks: (i) 100 pixels are selected in each silhouette for learning the probability estimator; (ii) only 100 pixels are selected for classifying a silhouette (for efficiency reasons); and, in the same way, (iii) we limit the selection to 100 pixels to build the vector S for learning the weighting vector.

This random selection is acceptable because (1) we have no prior knowledge about the most relevant pixels, and (2) because a random subset of pixels is supposed to follow the same pdf as the full set.

3.3. Results

Unfortunately, no comparison with Barnich's results can be provided, because no quantitative results were reported in [3].

By cross-validation on *LS*, we get an accuracy of 99.56%. This means that, if the observed scene contains a single moving person or object, everything will be detected thanks to the background subtraction algorithm (to the contrary of *HOGs*), but on average one silhouette will be misclassified every 200 frames.

Figure 6 presents the results obtained on *TS* in generalization. Correct classification rates around 90% can be obtained for both human and non-human silhouettes. It is interesting to note that the ROC curves obtained with the three weighting vectors drawn in Figure 5 are similar. This means that better weighting vectors are not helpful for generalization. However, we think our results could be improved by studying how to get both a more robust set of attributes and a better learning database (*eg* more diversified).

4. CONCLUSIONS

This paper presents a new system for the detection of humans, suited for video streams. A background subtraction algorithm first extract silhouettes of moving objects. Then, silhouettes are classified into two classes: human and non-human. Classification is achieved in a two steps process. A probabilistic information is calculated for each pixel independently during the first step. Then this information is used to predict the class of each silhouette. Results show that our approach is effective for the detection of humans in video streams.

Acknowledgments

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

5. REFERENCES

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

[2] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, USA, June 2006, vol. 2, pp. 1491–1498.

[3] 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)*. September 2006, vol. 4179 of *Lecture Notes on Computer Science*, pp. 734–745, Springer.

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

[5] S. Piérard, V. Pierlot, O. Barnich, M. Van Droogenbroeck, and J. Verly, "A platform for the fast interpretation of movements and localization of users in 3D applications driven by a range camera," in *3DTV Conference*, Tampere, Finland, June 2010.

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

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

[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. thesis, University of Liège, Belgium, February 2005.

[10] 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.