A ROBUST BACKGROUND SUBTRACTION ALGORITHM USING THE A $\Sigma - \Delta$ ESTIMATION

Applied to the Visual Analysis of Human Motion

Juan Carlos León, Fabio Martínez and Eduardo Romero
CimaLab, Universidad Nacional de Colombia, Bogotá, Colombia

Keywords: Background Subtraction, Motion Analysis, $\Sigma\Delta$ Estimation.

Abstract: This paper introduces a novel method for segmenting the human silhouette in video sequences, based on a local version of the classical $\Sigma\Delta$ filter. A main difference of our approach is that the filter is not pixel-wise oriented, but rather region wise adjusted by using scaled estimations of both the pixel intensity and the horizontal (vertical) gradient, i.e., a multiresolution wavelet decomposition using Haar functions. The classical $\Sigma\Delta$ filter is independently applied to each component of the obtained feature vector, previously normalized and a single scalar value is associated to the pixel by averaging the feature vector components. The background is estimated by setting a threshold in a histogram constructed with these integrated values, attempting to maximize the interclass variance. This strategy was evaluated in a set of 6 videos, taken from the Human Eva data set. Results show that the proposed algorithm provides a better segmentation of the human silhouette, specially in the limbs, which are critical for human movement analysis.

1 INTRODUCTION

Visual analysis of human motion implies the detection, follow up and characterization of relevant patterns in a sequence of images. Usually the main features to detect are the position and alignment of the human body parts (human pose). While visual markers can be employed for this task(Kirtley, 2005), the result is usually a simplified model of the human body. Most detection methods use a background estimation as preprocessing step, attempting to eliminate pixels with no temporal change.

Background subtraction methods use a sequence of images ($\{I_i\}_{i=1:t}$) to build a model of the static scene (M_i), and establish a rule to set a pixel value in I_i as either background or foreground.

A main contribution of the present paper was to adapt the classical $\Sigma\Delta$ pixel wise estimation to a local version of the filter, which is much more robust to local variations and tracks better the image object edges. The basic idea was to approach the pixel information with a multiresolution decomposition, conserving the edge features in the gradient estimations while the low frequency characteristics regularize the numerical difference, i.e., a classical wavelet approximation. The obtained Haar coefficients are used independently in a classical $\Sigma - \Delta$ estimation, averaged and used to construct an histogram in which an optimal threshold maximizes the interclass variance. This paper is organized as follows: Section 2 introduces the $\Sigma - \Delta$ operator, and the proposed extension, section 3 demonstrates the effectiveness of the method, finally section 4 concludes with a discussion of the proposed method and possible future works.

2 MATERIALS AND METHODS

Among the background subtraction techniques, the $\Sigma - \Delta$ operator represents a family of background subtraction methods, well known for their computational efficiency and capability to work without any prior knowledge of the scene, even in no controlled illumination conditions.

While this operator offers a baseline for background subtraction in human movement analysis, it is still limited regarding its accuracy and robustness to noise. As observed in figure 1, relevant parts of the human figure, as the shins and lower arms, are missing. Noise is present on the image, especially while the model converges to a good approximation of the background.

These limitations can be attributed to the selected

Carlos León J., Martínez F. and Romero E..
A ROBUST BACKGROUND SUBTRACTION ALGORITHM USING THE A Σ-âĹĘ ESTIMATION - Applied to the Visual Analysis of Human Motion.
DOI: 10.5220/0003868204840489
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2012), pages 484-489
ISBN: 978-989-8565-03-7

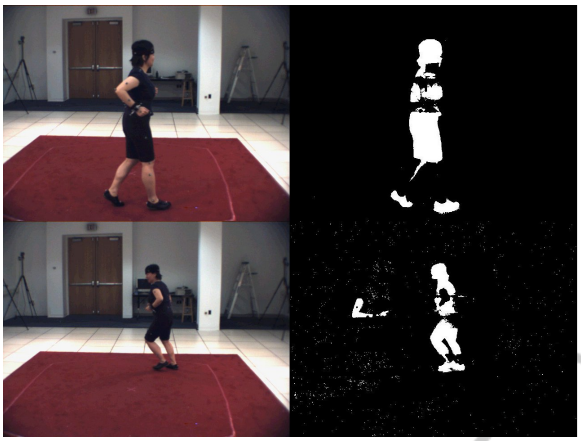

Figure 1: Output of the Basic $\Sigma - \Delta$ Algorithm for a sequence of the Human Eva Dataset.

pixel descriptors in a single frame, i.e. the regular $\Sigma - \Delta$ uses a single pixel intensity. This may be better approached by introducing local information. The present investigation proposes an extension of the $\Sigma - \Delta$ background subtraction algorithm, focusing on region features rather than on pixel intensity.

2.1 The $\Sigma - \Delta$ Operator

The non linear operator $\Sigma\Delta$ increases the correlation between adjacent frames by oversampling a signal at higher rates than the specified by the Nyquist theorem. This operator dynamically updates a background model $M_t(x)$, by comparing each image $I_t(x)$ with the current background model $M_t(x)$, using a simple updating rule: If $I_t(x)$ is greater (lower) than $M_t(x)$, then a positive increase (decrease) Δ is performed. The absolute difference $|I_t(x) - M_t(x)|$ is used to compute an estimate of the per pixel variance $V_t(x)$, based on this estimate pixels are classified as either foreground or background (Manzanera and Richefeu, 2007), a detailed description can be found on algorithm 1.

2.2 Region Features

As stated above, a main limitation of the $\Sigma\Delta$ background subtraction is that it operates exclusively over the intensity values of a pixel through an image sequence I_t, restricting thereby the accuracy and robustness of the background estimation process. We approached herein this problem by projecting each frame I_t into a multiresolution space. Unlike a classical multiresolution decomposition, the different image scales are not herein obtained by a simple downsampling of the original image, but rather by a local pixel neighbourhood smoothing upon which a block Haar wavelet analysis is carried out. In our scheme

Algorithm 1: Basic $\Sigma - \Delta$ Algorithm.

Initialization: $M_0(x) = I_0(x)$
for each Frame t **do**
  $M_t(x) = M_{t-1}(x) + sgn(I_t(x) - M_{t-1}(x))$
  $\Delta_t(x) = |M_t(x) - I_t(x)|$
end for
Initialize: $V_0(x) = \Delta_t(x)$
for each Frame t **do**
  for each pixel x such that $\Delta_t(x) \neq 0$ **do**
    $V_t(x) = V_{t-1}(x) + sgn(N \times \Delta_t(x) - V_{t-1}(x))$
    if $\Delta_t(x) < V_t(x)$ **then**
      $D_t(x) = 0$
    else
      $D_t(x) = 1$
    end if
  end for
end for

we use a set of features calculated from the Speeded Up Local Descriptor (SULD), proposed by Zhao et al. (Zhao et al., 2009) and now on used as a pixel descriptor. The low frequency is computed as the average of a neighbourhood centred at the pixel, while the high frequency is calculated by firstly averaging a spatial shifted version of the previously used neighbourhood, and then differences between the up-down (left-right) shifted neighbourhoods are stored. The calculated values are closely related to the gradients and therefore to the edges along the x and y axes, as seen in figure 2. Each image pixel is associated to a feature vector with three components containing an average of the different scale pixel descriptors, i.e., the neighbourhood sizes. This allows to systematically remove finer details or high-frequency information from an image, achieving a compact description of the most relevant information which is usually preserved through multiple scales. Therefore, the first step of our approach was to build, for each pixel, a multidimensional feature vector containing the local first order information.

These features are calculated for each of the n channels of the image and used as input of the $\Sigma - \Delta$ algorithm, after normalization, yielding a $3n$ dimensional descriptor for each pixel.

2.2.1 Efficient Feature Calculation

Two of the features are calculated as the difference of the sum of pixel intensities within two shifted boxes, either vertically or horizontally. This can be efficiently computed using the summed area table known as the integral image (Viola and Jones, 2001), case in which an image (ii) replaces a pixel value (i) with the sum of the intensity of every pixel located above and

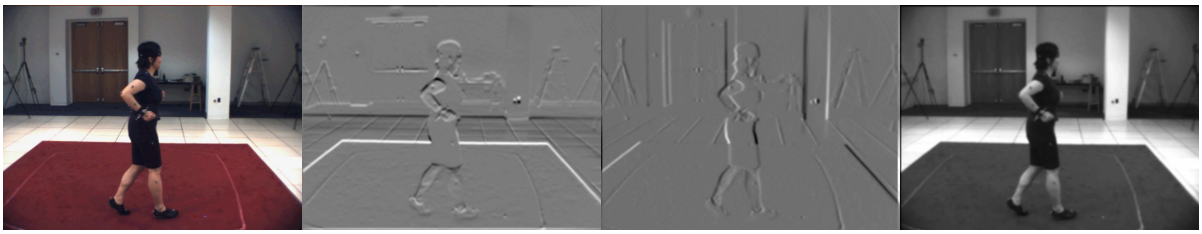

Figure 2: Descriptors, from left to right: original image, vertically and horizontally filter response maps, sum of values in region.

before it, formally:

$$ii(x,y) = \sum_{x' \leq x} \sum_{y' \leq y} i(x',y') \quad (1)$$

The use of the integral image optimizes calculation of the region intensity sum as:

$$\sum_{j<x' \leq k} \sum_{m<y' \leq n} i(x',y') = ii(j,m) + ii(k,n) - ii(k,m) - ii(j,n) \quad (2)$$

2.3 Foreground Classification Criteria

The basic $\Sigma - \Delta$ algorithm uses a simple classification criterion: the last pixel intensity variation ($\Delta_t(x)$) is compared with an estimation of the cumulated variance ($V_t(x)$), if the result is positive then the pixel is marked as foreground, otherwise it is considered as background (see algorithm 1). This metric does not fit our multidimensional representation: while the mentioned criterion may be applied to each feature, another metric must be built to produce a final decision from the obtained set of per-feature decisions. To overcome these limitations, we propose a multidimensional metric that associates the feature vector to a single scalar value, obtained by integrating on every feature component and shifting from the $[-1,1]$ to the $[0,2]$ interval. Each image pixel is assigned then to a particular (P_t) value, an estimate of the regional changes, the higher (lower) a P_t value is the more (less) likely the corresponding pixel in I_t is a foreground pixel. A change is then defined if the history of regional changes is smaller than the change reported by the current local analysis. For achieving so, we exploit the characteristics of the histogram's waveform of P_t, where background pixels are near 0 and their number is significantly larger than the foreground pixels. Hence we build two classes, one with a small (large) number of bins which contains most (few) scene pixels: the background (foreground). We are interested in a value that maximizes the intra-class variance by comparing the variances of the two previously defined classes. For doing so, let us suppose that we have k different bins, starting from an initial bin, a class is composed of a set of bins that are progressively increased by including new bins into the class. The algorithm includes new bins in each class by running forward (backward) over the histogram, starting from 0 and k for the background and foreground classes, respectively. The goal is to stop when the variance of the two classes is alike and its magnitude is maximum. We search then for a bin (γ) where the consecutive per group variances are close and large in magnitude for both classes as follows: for a histogram with k bins let

$$\alpha_i = var(bin_0 ....bin_{i-1}) - var(bin_0 ....bin_i) \quad (3)$$

the consecutive variance of a background estimation composed of bins 0 to i, likewise let

$$\beta_i = var(bin_k, ..., bin_{i+1}) - var(bin_k, ..., bin_i) \quad (4)$$

the difference of variances for the foreground group up to bin i. A set of candidate bins Γ_i is stablished with

$$i \in \Gamma \iff \frac{\alpha_i}{\beta_i} \approx -1 \quad (5)$$

Among all the candidates in Γ_i we choose γ as as the one with the larger magnitude in the variance differences i.e.

$$\gamma = \max_{\Gamma_i} |\alpha_i \beta_i| \quad (6)$$

2.4 Dataset Description

Validation was carried out with a subset of the Human Eva Dataset (Sigal et al., 2010), composed from 3 different subjects, each captured from 2 different cameras for a total of 6 sequences. Each sequence was manually labeled, as frame n has almost the same foreground and background of frame $n \pm 1$ labeling was done only once per 10 frames, additionally the labeling only started at the 40th frame this accounts for an initial estimation of the background (stabilization) of both algorithms.

3 EVALUATION AND RESULTS

There are well know metrics to evaluate the perfor-

mance of a binary classification, however most of these metrics assume that there is approximately a balanced quantity of elements in the classes. In this dataset the foreground usually amounts to less than the 10% of pixels in the image, hence we choose the true positive rate (TPR), and the Matthews Correlation Coefficient (MCC), the former is independent of the class distribution, while the later is designed to measure the quality of the classification even with unbalanced classes.

During the evaluation of the algorithm it was clear that scales (box sizes) larger than 11 were not appropriate for the segmentation of relative small objects in movement (like the hands and forearms), also the body boundaries are not properly located. Therefore we first seek for a combination of scales between 1 and 11 that provides the best results, for this particular dataset the selected scales were 1,3 and 5. The results are summarized in tables 1 & 2.

Table 1: True positive Rate.

Sequence	Regular $\Sigma-\Delta$ TPR	Proposed $\Sigma-\Delta$ TPR
1	37.94%	67.46%
2	55.73%	68.95%
3	31.61%	69.23%
4	57.63%	73.25%
5	48.62%	71.94 %
6	65.86%	78.05%

Table 2: Matthews Correlation Coefficient.

Sequence	Regular $\Sigma-\Delta$ MCC	Proposed $\Sigma-\Delta$ MCC
1	0.557	0.683
2	0.707	0.678
3	0.524	0.721
4	0.725	0.731
5	0.656	0.713
6	0.767	0.774

The TPR of the proposed method outperforms the regular ΣΔ in every test sequence, this can be attributed to the better detection of the limbs in motion, specially the shins and forearms (see figure 6).

An interesting feature of the proposed algorithm can be analysed with table 1, our method offers a large improvement for sequences 1, 3 and 5 (30.06% average) however the improvement for sequences 2, 4 and 6 is smaller (13.71% average). This is related to the background of the sequences, on the first group the background has several objects of different colors on it i.e. it contains borders, the background of the later group has a single color and is nearly flat (see figure 3). The absence of borders lowers the effectiveness of the proposed algorithm as the input information for the ΣΔ comes mainly from the intensities of neighbouring pixels.

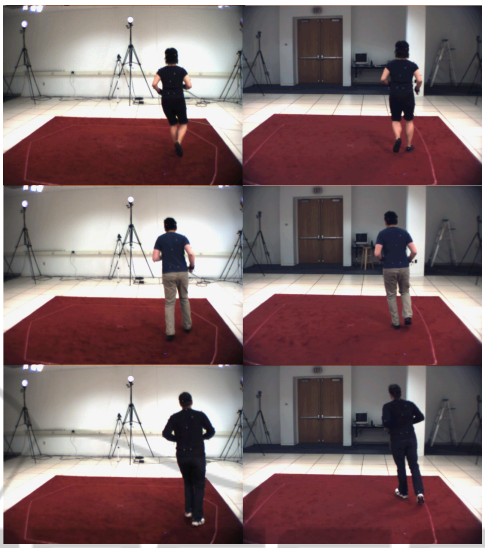

Figure 3: Sequences 2,4,6 on the left side, sequences 1, 3, 5 on the right side.

While the TPR shows a significant improvement of our algorithm over the regular ΣΔ, the MCC shows cases where there is not a significant improvement over the base algorithm. This can be attributed to the nature of the dataset, where the moving object (human body) is present and in motion on the first frames, this generates ghosts on every scene for both algorithms, these ghosts last longer in our algorithm thus increasing the amount of False Positives on the first frames, drawing down the average MCC for the sequence.

This can be seen in figures 4 and 7, while the first frames show an MCC for the proposed algorithm under the MCC of the regular sigma delta, on the next frames (when the ghost starts to fade) the MCC of our algorithm is better, even in the second sequence, where out algorithm had an average MCC under the regular ΣΔ (fig. 4).

Again the nature of the background seems to have influence on how long the ghosts last, scenes 1, 3 and 5. have ghosts that last shorter than the ghosts in scenes 2,4,6.

3.1 Performance

As stated on section 2, one of the main features of the ΣΔ Background subtraction is its computational efficiency, therefore we briefly analyze the performance penalty of the multiscale features and the new classification criterion.

A GNU Octave implementation of both algorithms was tested on an core i7 processor at 3.3 Ghz, on this set up the average the regular ΣΔ can process 6.72 million pixels per second. The speed of the proposed extension depends on the number of scales

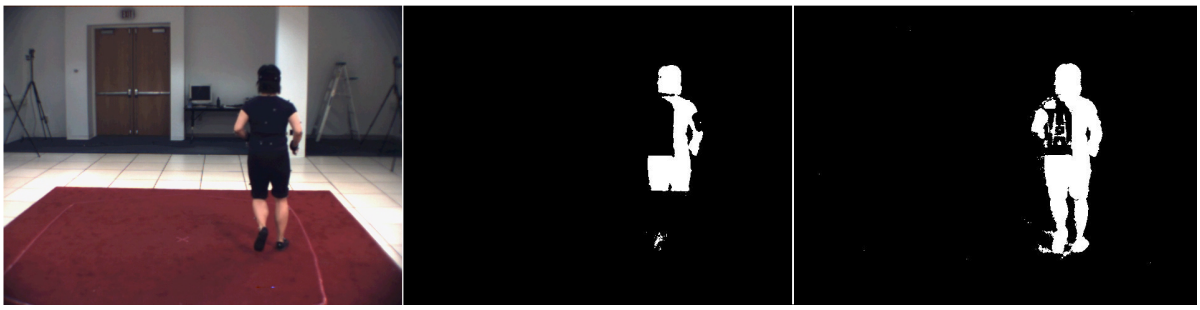

Figure 6: Results of the segmentation, from left to right, Original image, regular ΣΔ segmentation, proposed algorithm segmentation.

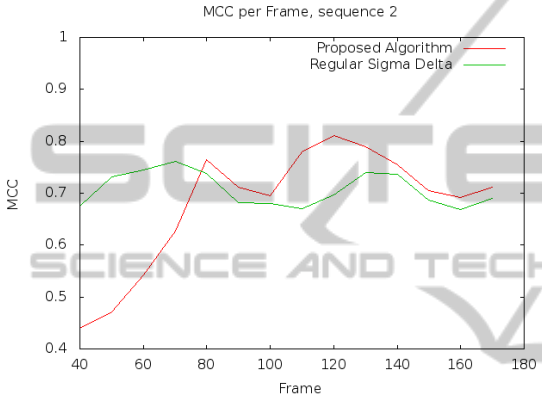

Figure 4: Comparison of the per frame MCC for sequence 2.

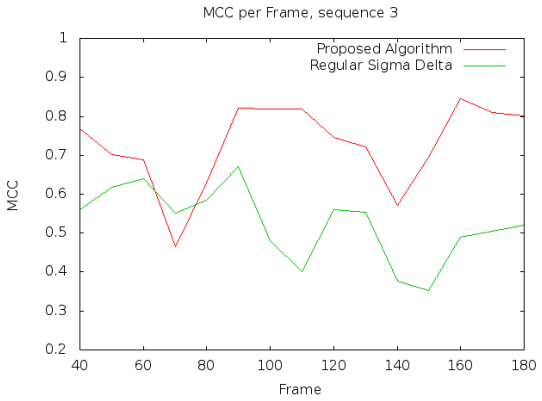

Figure 5: Comparison of the per frame MCC for sequence 3.

Table 3: Proposed algorithm processed pixels per second.

Number of scales	Average pixels per second (millions)
1	1.61
2	1.31
3	1.06
4	0.89
5	0.77

used for the analysis, we calculated the average speed of the proposed algorithm for a number of scales between 1 and 9, the results are summarized on table 3. Although the base ΣΔ is faster, when the proposed algorithm is compared with other variations of the ΣΔ operator for background subtraction proposed on the literature (Manzanera, 2007)(Lionel Lacassagne, 2009) (Richefeu and Manzanera, 2006) it shows an average performance (see figure 7).

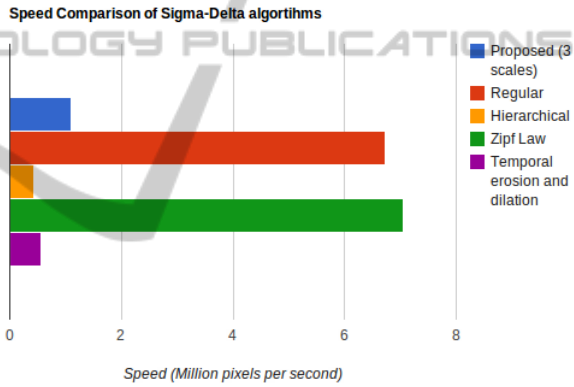

Figure 7: Speed of other ΣΔ algorithms (million pixels per second).

4 CONCLUSIONS

An novel method for segmenting the human silhouette in video sequences based on the ΣΔ background subtraction, was introduced on this paper, this method offers a significant improvement in the background segmentation over the base ΣΔ, at the expense of computational cost.

The proposed algorithm enhances the pixel description with local features, allowing a multiscale representation of each frame, which results in an improved detection of the human body silhouette, this methods improves the quality of the segmentation, specially at the arms and lower limbs, which is critical for tasks that require a proper description of the dynamics of the human body, as gait analysis and video surveillance.

REFERENCES

Kirtley, C. (2005). *Clinical Gait Analysis: Theory and Practice*. Churchill Livingstone.

Lionel Lacassagne, A. M. . A. D. (2009). Motion detection: Fast and robust algorithms for embedded systems. In *IEEE International Conference on Image Processing*.

Manzanera, A. (2007). Sigma-delta background subtraction and the zipf law. In *Progress in Pattern Recognition, Image Analysis and Applications*.

Manzanera, A. and Richefeu, J. (2007). A new motion detection algorithm based on [sigma]-[delta] background estimation. *Pattern Recognition Letters*, 28(3):320–328.

Richefeu, J. and Manzanera, A. (2006). A new hybrid differential filter for motion detection. In Wojciechowski, K., Smolka, B., Palus, H., Kozera, R., Skarbek, W., and Noakes, L., editors, *Computer Vision and Graphics*, volume 32 of *Computational Imaging and Vision*, pages 727–732. Springer Netherlands.

Sigal, L., Balan, A. O., and Black., M. J. (2010). Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV)*, 87.

Viola, P. and Jones, M. (2001). Robust real-time object detection. In *International Journal of Computer Vision*.

Zhao, G., Chen, L., and Chen, G. (2009). A speeded-up local descriptor for dense stereo matching. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 2101–2104.