A Revaluation of Frame Difference in Fast and Robust Motion Detection

Davide A. Migliore*
AIRLab
Politecnico di Milano, Italy
migliore@elet.polimi.it

Matteo Matteucci
AIRLab
Politecnico di Milano, Italy
matteucc@elet.polimi.it

Matteo Naccari
ISPG Lab
Politecnico di Milano, Italy
naccari@elet.polimi.it

ABSTRACT

In this paper we propose a robust approach to detect moving objects for video surveillance applications. We demonstrate that a jointly use of frame by frame difference with a background subtraction algorithm allows us to have a strong and fast pixel foreground classification without the need of previous background learning. The Joint Difference algorithm uses frame difference information to correct pixels classification made by a background subtraction algorithm while selectively updating the background model according to such classification. In this way we should perform motion segmentation also in presence of environmental changes such as illumination variations or "waking persons". The algorithm is capable of 15 fps tracking of moving people on 640×480 unsampled color images; results on both VSSN06 and Wallflower [8] benchmark videos are presented.

Categories and Subject Descriptors: I.4.6 Segmentation: Pixel classification, I.4.8 Scene Analysis: Motion

General Terms: Algorithms.

Keywords: Motion segmentation, Tracking, Background Modeling.

1. INTRODUCTION

One of the artificial vision goals is to emulate some features of the human visual system, such as the skill of recognizing object movements and tracking their behavior in a complex environment. The first step in tracking applications is to detect moving objects in the environment, classifying object pixels and gathering them in connected areas named as "blobs" characterized by features that allow their identification; this reduces the problem complexity giving a global perception of the scene. Doing this way, a connection between blobs and real moving objects in the environment is created and object behavior in the scene is replaced by blob behavior on the image plane. In this process, classical issues of tracking arise: *motion segmentation*, *tracking* and *data association* (i.e., blobs in different frames must be connected to the same moving object in order to retrace the real path of moving entities).

*research activity supported by the IIT Foundation (www.iit.it).

VSSN'06, October 27, 2006, Santa Barbara, California, USA.

In this paper we present a novel approach named *Joint Difference* to solve the first task, motion segmentation, obtaining a fast and robust foreground/background segmentation without any a priori knowledge of the environment. A brief review of classical motion detection algorithms is presented in Section 2. Section 3 describes the Joint Difference algorithm while Section 4 presents our experimental validation.

2. PREVIOUS WORK

As far as motion segmentation is concerned, two main approaches have been proposed in scientific literature: image difference and background subtraction. The former consists of a thresholded difference between frame at time t and frame at time $t-1$; this method is very performant in computational terms and grants a prompt object motion detection between two frames; however, it suffers two well-known drawbacks [1] caused by frame rate and object speed (see Figure 1): foreground aperture and ghosting. To solve these issues Kameda and Minoh [5] propose a variation on this method: "the double difference". This approach operates a thresholded difference between frames at time t and $t-1$ and between frames at time $t-1$ and $t-2$, combining them with a logical AND. However if the moving objects have not enough texture this procedure does not allow an accurate motion detection and the object position is not estimated in real time. Collins et al.describe the VSAM project [1], this algorithm exploits image difference between frames at time t and $t-1$ and the difference between t and $t-2$ to erase ghosting; it also keeps in memory a background model to solve the foreground aperture problem. This system is widely used in outdoor environments with a low depth of field images, but suffers a few drawbacks on variable depth shots. On the other side, background subtraction uses a background model to be compared with the actual frame. This algorithm solves issues of image difference and gives good results in motion segmentation. Luminance variations and objects that begin or end their motion in the scene – *waking person* – require a proper background update procedure. To solve these problems many techniques have been developed; the most simple ones update the background by a convex composition of background pixels at time $t-1$ and image pixels at time t; update weight is eventually variable with pixel classification (light weight for background pixels and heavy weight for foreground pixels). Wren et al. describe the PFinder system [9], with the assumption that the scene is less dynamic than the object to be tracked and that the background is distributed according to a single Gaussian distribution. Although Pfinder can deal with small or gradual changes in the background, it fails when the background scene involves large or sudden changes, or has multi-modal distributions (such as small repetitive movements). The W^4 system [4] modeled the background scene by maximum and minimum intensity

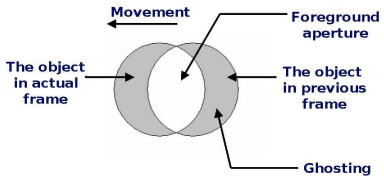


Figure 1: Drawbacks of Image Difference algorithm

values, and the maximum intensity difference between consecutive frames in training stage. However, the background model from W4 may be inaccurate when the background pixels are multi-modal distributed or widely dispersed in intensity. The pixel-level Mixture of Gaussians (MOG) [7] background model has become very popular because of its efficiency in modeling multi-modal distribution of backgrounds (such as waving trees, ocean waves, light reflection, etc) and its ability to adapt to a change of the background (such as gradual light change, etc.). Friedman and Russell [2] modeled the intensity values of a pixel by using a mixture of three Normal distributions and applied the proposed method to traffic surveillance applications. Stauffer and Grimson [7] presented a method that models the pixel intensity by a mixture of K Gaussian distributions. Toyama et. al. [8] implemented MoG and compared the result of MoG with that of "Wallflower" claiming superiority of the latter. In this paper, we show that the jointly use of simple methods, like frame difference and background subtraction, can be a faster alternative of MoG in dealing with shadow removal, in background update, and in background subtraction.

3. THE JOINT DIFFERENCE

To detect a moving object within the current frame we propose an hybrid technique that uses both frame by frame difference and background subtraction. To classify pixels in a robust way, at time t we combine the difference between the actual frame F_t and previous frame F_{t-1} with the difference between actual frame and the actual model of background B_t. To bootstrap this process, the first image of the sequence is used as initial background model and it is dynamically updated with the new frame according to motion segmentation. We distinguish five classes in this process because we want to detect: *foreground*(FG), *background*(BG), *ghosting*(GH), *foreground aperture*(FA), and *waking person*(WP). A foreground pixel belongs to the moving object. On the contrary, a background pixel belongs to the (fixed) background of the scene; the other three classes are more interesting. To better describe the ghosting problem let be p a foreground pixel at location (x, y) in F_{t-1}; in F_t it will be in a different location, so the difference between the two images in (x, y) will have a foreground pixel instead of a background one. In Figure 1 we can notice that the ball is moving from right to left and in the difference image ghosting is present on the right. On the other hand, foreground aperture is due to pixels with similar texture. In Figure 1 the white zone between the two black circles shows this problem. The *waking person* is a different problem due to background subtraction (see Figure 2); an object that belongs to the background model suddenly starts to move at time t. From this frame, motion detection identifies a second object in the position where previously was the moving object. This is due to the difference between the old background and the actual frame where the waking person is moving. To solve these problems we have exploited a joint background subtraction and frame by frame difference to properly classify pixels. A pixel belongs to foreground when frame by frame difference is above a given threshold τ_I and difference between background and actual

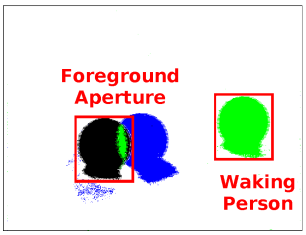


Figure 2: Waking person blob contains only pixels classified as WP (green), while Foreground blob contains mostly pixels classified as FG (black) and Ghosting contains mostly pixles classified as GH (blue)

frame is above threshold τ_B. Instead, a pixel belongs to background if frame by frame difference is below τ_I and difference between background and actual frame is below τ_B. Moreover, we can classify ghosting by exploiting the fact that frame by frame difference is above τ_I and difference between background and actual frame is below τ_B. With this simple reasoning, we can not distinguish between waking person and foreground aperture; we just know that these two phenomena appear when frame by frame difference is below than τ_I and background difference is above τ_B. We can thus summarize this simple classification procedure as:

$$\Delta_I > \tau_I \quad \wedge \quad \Delta_B > \tau_B \Rightarrow FG \tag{1}$$
$$\Delta_I < \tau_I \quad \wedge \quad \Delta_B > \tau_B \Rightarrow FA \vee WP \tag{2}$$
$$\Delta_I > \tau_I \quad \wedge \quad \Delta_B < \tau_B \Rightarrow GH \tag{3}$$
$$\Delta_I < \tau_I \quad \wedge \quad \Delta_B < \tau_B \Rightarrow BG \tag{4}$$

being Δ_F and Δ_B defined as $\Delta_F = |F_t(x, y) - F_{t-1}(x, y)|$ and $\Delta_B = |F_t(x, y) - B_{t-1}(x, y)|$. (Figure 2 shows the result at the end of this classification procedure).

An immediate distinction between foreground aperture and waking person is impossible. At this point, to face the ambiguity between FA and WP, pixels are collected into blobs without doing any distinction among foreground, foreground aperture and waking person. Once pixels are collected in blobs, we can exploit the fact that a waking person blob will contain only pixels classified as WP while a foreground blob contains mostly foreground pixels (Figure 2). This can be easily obtained by comparing the ratio of foreground pixels in the blob with the total number of pixels:

$$N_{FG} \geq \gamma N_{Tot} \quad \Rightarrow FG \tag{5}$$
$$N_{FG} < \gamma N_{Tot} \quad \Rightarrow WP. \tag{6}$$

To get a stronger detection of waking person a further constraint can be added for those objects that started their motion into the scene; if a pixel has been classified as $FA \vee WP$ for a given number of frames it can be safely classified as a waking person since (by definition) a foreground aperture is something that moves with the object. This classification process is based on simple fast rules that can be exploited both to get a more precise blob detection and an adaptive background update as described further on.

Sometimes there is some little imprecision due to camouflage; thresholds reduction can solve this problem, however it increases the image difference noise, so they have to be chosen looking for a compromise between good motion detection and noise filtering. In our system, the background model is initialized with the first image of the video sequence, and then updated, as in [9], using: $B_t = (1 - \alpha)B_{t-1} + \alpha F_t$.
We exploit pixel classification to tune background adaptation using different α values. When the pixel is classified as foreground then

(a) Video 2 image (b) Joint Difference (c) Ground-Truth

(d) Video 6 image (e) Joint Difference (f) Ground-Truth

Figure 4: Joint Difference result on Video 2 (first row) and Video 6 (second row)

α is set to 0 to avoid background corruption, if it is classified as background then α is set to a low level to slowly update the model. Finally, if a pixel is classified as a waking person α is set to the highest level, so that background is quickly restored. Image difference is computed through a pixel by pixel single channel difference in the RGB color space; after this, the image is thresholded to become binary. Before binarization it is useful to introduce a shadow filter to get a refined blob growing. The works by Geusebroek et al [3] and Prati et al. [6] show that, in the HSV space, the difference between a shaded zone and the same zone enlightened regards the luminance component, leaving the components of hue and saturation nearly unchanged. Exploiting this result, a simple filter is used in the classified blobs; we define the x and y distance between two points with constant S in the HSV cylinder as:

$$HSx_i = S_i \cos(H_i) \quad (7)$$

$$HSy_i = S_i \sin(H_i) \quad (8)$$

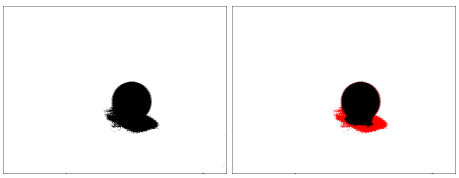
then a shadow is identified if the three following inequalities between foreground frame F_t and background model B_t are verified:

$$|HSx_F - HSx_B| < \tau_x \quad (9)$$

$$|HSy_F - HSy_B| < \tau_y \quad (10)$$

$$\alpha < \frac{V_F}{V_B} < \beta \quad (11)$$

An example of this shadow filter is shown in Figure 3, red pixels are those detected by inequalities (9), (10) and (11).



(a) No shadow detection (b) With shadow detection

Figure 3: An example of shadow detection.

4. EXPERIMENTAL VALIDATION

The Joint Difference algorithm has been tested using VSSN06 and Wallflower test videos, comparing the resulting binary image with a ground truth one, to show us that our system is better than other algorithms not only for speed, but also in classification quality. The VSSN06 benchmark gives us videos to test the average error of classification. In this case, we used a sequence without illumination changes (Video 2 in Figure 4) and a sequence with slow illumination changes (Video 6 in Figure 4). Data in Table 1 represent average errors calculated as:

$$\frac{FalsePositiveAverage + FalseNegativeAverage}{TotalPixel} \quad (12)$$

The Wallflower dataset (see Table 3) grant us a good benchmark to compare our algorithm with others proposed in literature. Table 3 shows the performance of each approach. We can notice that the Joint Difference performances, using jointly single difference with background subtraction, are better than other algorithms like Wallflower and Mixture of Gaussian, especially in video with camouflage (C) and slow illumination change (TOD). In presence of waving trees (WT) our algorithm detected more false positive than false negative because the single difference and the background subtraction failed to detect this drawback; however this behavior is coherent with the policy of video surveillance system: "A false allarm is better than a no detection". Moreover the algorithm has shown good performance, comparable with MoG, also with bootstrap (B), moved object (MO) and foreground aperture (FA). An interesting result is obtained in presence of a light switch (LS): we notice that with a background reset on sudden illumination changes we can obtain a good motion segmentation due to the use of single difference to improve the background model updating.

Algorithms	**Video 2**	**Video 6**
Joint Difference	0.39%	0.40%
Background Subtraction (1)	0.48%	0.61%
Background Subtraction (2)	0.52%	0.73%
Single Difference	1.42%	1.45%

Table 1: Background Subtraction was performed with a Gaussian Model: threshold 2*stdev (1) and 2.5*stdev (2).

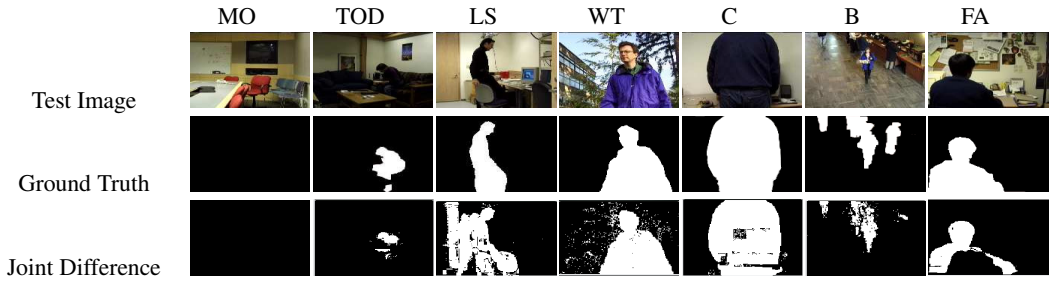


Table 2: Wallflower Dataset: in the first row original images, in the second hand segmented and in the last our results.

Algorithm	Error Type	moved object	time of day	light switch	waving trees	camouflage	bootstrap	foreground aperture	Total Errors	Total Errors*
Frame difference	false neg.	0	1165	2479	3509	9900	1881	3884	27311	17957
	false pos.	0	193	86	3280	170	294	470		
Mean + threshold	false neg.	0	873	1116	17	194	415	2210	29996	10479
	false pos.	0	1720	15116	3268	1638	2821	608		
Mean + covariance	false neg.	0	949	1857	3110	4101	2215	3464	35133	14686
	false pos.	0	535	15123	357	2040	92	1290		
Mixture of Gaussians	false neg.	0	1008	1633	1323	398	1874	2442	27053	9587
	false pos.	0	20	14169	341	3098	217	530		
Block correlation	false neg.	0	1030	883	3323	6103	2638	1172	21683	14160
	false pos.	1200	135	2919	448	567	35	1230		
Temporal derivative	false neg.	0	1151	752	2483	1965	2428	2049	46167	27342
	false pos.	1563	11842	15331	259	3266	217	2861		
Bayesian decision	false neg.	0	1018	2380	629	1538	2143	2511	31422	14640
	false pos.	0	562	13439	334	2130	2764	1974		
Eigen-background	false neg.	0	879	962	1027	350	304	2441	17677	13269
	false pos.	1065	16	362	2057	1548	6129	537		
Linear prediction	false neg.	0	961	1585	931	1119	2025	2419	27027	10002
	false pos.	0	25	13576	933	2439	365	649		
Wallflower	false neg.	0	961	1585	931	1119	2025	320	11478	7280
	false pos.	0	25	375	1999	2706	365	649		
Joint difference	false neg.	0	853	466	55	630	1160	2451	**10875**	**5916**
	false pos.	0	7	3543	895	135	166	514		

Table 3: Comparison with other algorithms in Wallflower paper. The * is total error without light switch and waving trees videos

5. ADDITIONAL AUTHORS

Andrea Bonarini, email: bonarini@elet.polimi.it, AIRLab Politecnico di Milano)

6. REFERENCES

[1] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, and L. Wixson. A system for video surveillance and monitoring. Technical report, 2000.

[2] N. Friedman and S. Russell. Image segmentation in vedio sequences: A probabilitic approach. In *Proc. Thirteenth Conf. on Uncertainty in Artificial Intelligence*, pages 175–181, 1997.

[3] J. M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts. Color invariance. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(12):1338–1350, 2001.

[4] I. Haritaoglu, D. Harwood, and L. Davis. Who, when, where, what: A real time system for detecting and tracking people. In *Proceedings of the Third Face and Gesture Recognition Conference*, pages 222–227, 1998.

[5] Y. Kameda and M. Minoh. A human motion estimation method using 3-successive viedo frames. In *ICVSM*, pages 135–140, 1996.

[6] A. Prati, I. Mikic, M. Trivedi, and R. Cucchiara. Detecting moving shadows: Algorithms and evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:918–923, 2003.

[7] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, pages 246–252, 1999.

[8] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *ICCV*, pages 255–261, 1999.

[9] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.