OPTIMIZATION OF COLOR-BASED FOREGROUND / BACKGROUND SEGMENTATION FOR OUTDOOR SCENES

Louis St-Laurent[a], Donald Prévost[a], Xavier Maldague[b]
[a]National Optics Institute, Quebec City, Quebec, Canada
[b]Department of Electrical and Computer Engineering, Laval University, Quebec City, Quebec, Canada
louis.st-laurent@ino.ca

ABSTRACT
Performing foreground / background segmentation in outdoor scene is very challenging. Starting from a state-of-the-art color-based approach [8], we propose three improvements: the use of the *YCoCg* color space, a spherical association volume, and a cast shadows management approach. Using image sequences with ground truth at pixel-level, we quantitatively measured the performances of the proposed algorithm and demonstrated that it leads to a reduced processing time with improved detection accuracy. We also introduce a new public dataset of outdoor videos with ground truth.

KEY WORDS
Motion detection, background subtraction, *YCoCg* color space, shadow removal.

1. Introduction

Performing accurate and real-time automatic detection and tracking of human and vehicles in outdoor environment is very challenging, especially because of dynamic background, lighting changes and climate factors. For applications where the acquisition unit is fixed, background subtraction methods are almost always used as a first stage of foreground / background classification. In its simplest form, the background is modeled by a single Gaussian distribution, which may be satisfying for simple background. But when the scene contains dynamics areas, like oscillating trees, it is more suitable to use multiple Gaussians [18] or non parametric distributions [2][3][4][5][8][9][12].

As a starting point for background modelling, we selected the state-of-the-art background subtraction technique based on non-parametric codebook model proposed by Kim *et al.*[8]. According to the authors, this method is about three times faster than the Gaussian mixture model of Stauffer and Grimson [18] and the non parametric kernel-based model of Elgammal *et al.*[4], with similar segmentation accuracy. Since 2005, the codebook model has been the choice of numerous researchers [3][5][9].

The main idea of the non-parametric codebook model is that a quantization/clustering technique is used to generate a compressed form of background model. Every pixel is represented by a variable number of codewords (*CW*) depending on the background variability. At every new frame, every pixel is associated to the first sufficiently similar *CW*. If no codeword can be matched, a new one is created and added in a cache codebook. Codewords from the cache are promoted to permanent background codebook when they are repetitively matched, and codewords from permanent background codebook not matched since a long period of time are deleted.

In this paper, we propose three improvements to the codebook model: the use of the *YCoCg* color space, a spherical association volume, and a cast shadows management approach. The benefits of these improvements are a reduced processing time and an increased accuracy. It is worth to emphasis the fact that algorithms developed in this work are context-independent, which means that the methods can be used in numerous video surveillance applications.

2. Improved color-based segmentation

The use of the background subtraction technique proposed by Kim *et al.* [8] in our outdoor video monitoring applications allowed us to identify a few limitations. We propose in this section three generic improvements that can be applied to most background subtraction algorithms.

2.1 YCoCg color space

In outdoor environment, it has been demonstrated [6][12][19] that the difference between spectral power distributions of the sun and the sky often leads to a color shift toward blue when the sun is occluded. The direction of the color distortion may thus be used to discriminate cast shadow pixels.

The original codebook model [8] works in the *RGB* color space. In addition of being computationally expensive, its color distortion metric (eq. 1) can only measure the amplitude of the distortion, not its direction:

$$\delta = \sqrt{\left(R^2 + G^2 + B^2\right) + \frac{\bar{R}_k R + \bar{G}_k G + \bar{B}_k B}{\bar{R}_k{}^2 + \bar{G}_k{}^2 + \bar{B}_k{}^2}} \qquad (1)$$

We thus decided to look for a most suitable color space. Numerous studies have been published on the topic but their conclusions are quite disparate, mainly because

criteria and experimental conditions are variable. An observation that seems to stand out is that mixed color space (separated luminance and chrominance components) are the most effective for shadow / foreground segmentation. Based on some of the most rigorous papers ([1][7]), the *CIELAB* color space would be the most accurate, but it entails two disadvantages: it requires a color calibration for every illumination condition and its conversion from the *RGB* format is computationally expensive. Among the common mixed color spaces, the *YCbCr* is the one offering the fastest transformation.

Relatively recently, a new mixed color space named *YCoCg* has been presented by Malvar and Sullivan [13]. Currently used in *Dirac* and *H.264FRExt* video codecs, this color space distinguishes itself by the simplicity of its transformation equations:

$$\begin{bmatrix} Y \\ Co \\ Cg \end{bmatrix} = \begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 0 & -1/2 \\ -1/4 & 1/2 & -1/4 \end{bmatrix} \cdot \begin{bmatrix} R \\ G \\ B \end{bmatrix} \qquad \begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \end{bmatrix} \cdot \begin{bmatrix} Y \\ Co \\ Cg \end{bmatrix} \quad (2)$$

Contrarily to the *YCbCr* color space, aiming perceptual uniformity, the transformation equation of the *YCoCg* color space is closer to a principal component decomposition, which is more suitable for automated video analytic processes.

Another interesting characteristic of the *YCoCg* format is that the color distortion caused by cast shadows is mainly isolated into a single chrominance channel (*Co*), thus significantly simplifying the computation of the direction of the color distortion (eq. 8, section 2.3). As illustrated on Figure 1, the *Cg* component is only affected by a slight positive distortion, while the *Co* component suffers a strong negative distortion (even almost 1.5 times larger in amplitude than the distortion measured on *Cb* and *Cr* components). Every point of Figure 1 represents the mean distortion measured on a square of a Macbeth ColorChecker board when submitted to a cast shadow. We measured the same relation for different outdoor illumination conditions (clear sky, light cloud cover, at dawn, dusk, noon). It should be noted that automatic white balance must be disabled to preserve the color distortion.

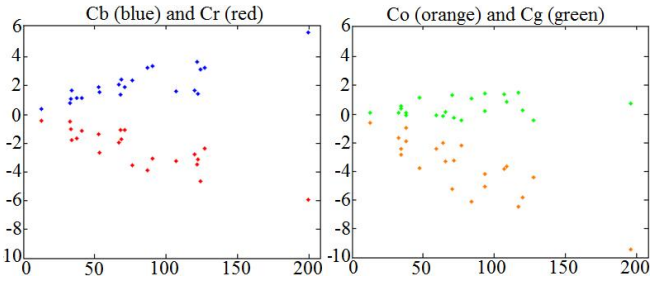

Figure 1: Color distortion of *Cb*, *Cr* (left) and *Co*, *Cg* (right) components as a function of luma (*Y*) of the background.

Surprisingly, we did not identify any studies related to video analysis using the *YCoCg* color space. However, we think that this unfamiliar format owns many interesting characteristics for efficient shadow segmentation. To sum up, we selected the YCoCg color space for the following arguments:

1. Mixed color space (as *CIELAB* and *YCbCr*);
2. Coefficients closer to a principal component decomposition rather than perceptual uniformity;
3. Color distortion caused by cast shadows isolated in a single chrominance channel (*Co*);
4. No color calibration required.

2.2 Spherical association volume

The association volume determines, for every pixel, the set of observed values that can be matched to a codeword, or more generally, to a distribution. The shape of this volume depends on the color space and the background model used. In [8], thresholds on color and intensity are defined independently. A cylindrical volume whose axes pass through the origin is thus obtained. The maximum allowable color shift determines the radius of the cylinder while the limits on brightness variation define its length. As mentioned previously, their color distortion metric is computationally expensive and only measures the amplitude of the distortion. Moreover, the technique used to determine the low intensity limit in [8] (eq. 3) does not allow the *CW* to adapt to long-term illumination changes:

$$I_{k,low} = \alpha I_k^{\max} \quad (3)$$

Indeed, when scene illumination decreases, the lower limit of the cylinder will not adapt accordingly since it is determined by the maximum intensity value observed in the past.

Based on the assumption that the noise of the sensor is multiplicative, it has been proposed to use a truncated cone [5] or a hybrid cone-cylinder [3] instead of a cylindrical volume to improve the sensitivity for low-intensity pixels. A first drawback of such volumes is that more parameters need to be set. Secondly, the assumption of multiplicative noise is not completely valid when analyzing a compressed video (MPEG-4). This is illustrated on Figure 2, where we notice that noise amplitude is relatively uniform all along the range of intensity values for the sequence with compression. A colorful scene was used to cover the entire range of intensity of each color channel. Every point on the graph represents the median of standard deviations (on 50 successive frames) of all pixels with a given mean intensity.

Another phenomenon that must be considered in outdoor environment is that high intensity areas are further affected by illumination variations. As shown on Figure 3, we see that the noise amplitude increases quasi linearly with pixel intensity. For this graph, the same colorful scene than for Figure 2 was used, but the sequence (MPEG4 compressed) was grabbed in presence of moving clouds, causing illumination variations. Note

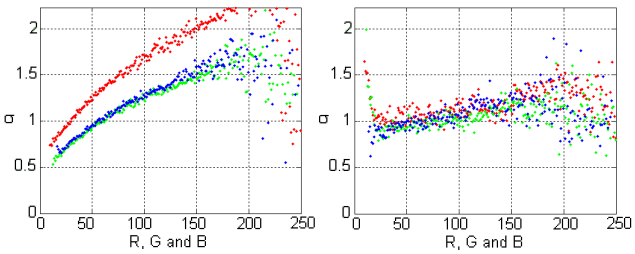

Figure 2: Standard deviation as a function of mean intensity without compression (left) and with compression (right).

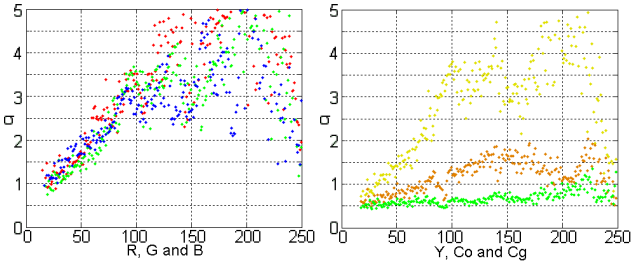

Figure 3: Standard deviation as a function of mean intensity of *RGB* (left) and *YCoCg* (right) components for a compressed video in which moving clouds create illumination variations.

that automatic adjustment of exposure time and iris aperture was disabled for the purpose of this analysis.

From these observations, we can state that it is justifiable and desirable to apply more permissive detection limits for high intensity pixels. The use of a conical volume of association is a possible approach to achieve this, but like Martel and Zaccarin [12] and *GMM* we chose to use a spherical volume with variable radius:

$$\left(\bar{Y}_k - Y_t\right)^2 + \left(\bar{C}o_k - Co_t\right)^2 + \left(\bar{C}g_k - Cg_t\right)^2 \le \left(\beta + \alpha \cdot \bar{Y}_k\right)^2 \quad (4)$$

The main interest in using a spherical association volume is its simplicity: only the radius needs to be determined. The coefficient β is the minimal radius allowed and must be adjusted in regards of video noise. The factor α is used to increase the radius for codewords with high luminance. $\bar{Y}_k$, $\bar{C}o_k$ and $\bar{C}g_k$ are luma, chroma orange and chroma green components of the k^{th} *CW*.

Using a spherical volume of association means that the three components of the color space are assumed to have a similar variance. Figure 3 shows that this assumption is valid for *RGB* components, but not for *YCoCg*. Indeed, we see that the amplitude of the noise affecting the chrominance channels (orange and green dots) is significantly lower than the noise amplitude of the luma component (in yellow). An association volume more suited to this color space would be an ellipsoid because luma and chroma semi-axis lengths (via coefficients β_Y and β_C respectively) could be adjusted independently:

$$\left(\frac{\bar{Y}_k - Y_t}{\beta_Y + \alpha \cdot \bar{Y}_k}\right)^2 + \left(\frac{\bar{C}o_k - Co_t}{\beta_C + \alpha \cdot \bar{Y}_k}\right)^2 + \left(\frac{\bar{C}g_k - Cg_t}{\beta_C + \alpha \cdot \bar{Y}_k}\right)^2 \le 1 \quad (5)$$

However, our experiments with an ellipsoidal association volume revealed that the application of more restrictive limits on chrominance channels leads to an increase of false detection rate. This behavior is mainly caused by compression artifacts around moving objects or close to edges. This noise increase goes unnoticed on the graph of Figure 3 since each point represents the median standard deviation of all pixels at a given intensity, thus filtering this localized behavior. For this reason, we believe that it is justifiable to use a spherical volume of association even with the *YCoCg* color space. To support this assertion, a comparison of performances obtained with spherical and ellipsoidal volumes is presented in section 3.

As a consequence of the choice of color space and association volume, we represent every pixel by L codewords (*CW*) of the form:

$$CW_{k=1\ldots L} = \left\{\bar{Y}_k, \bar{C}o_k, \bar{C}g_k, f_k, p_k, q_k, MNRL_k\right\} \quad (6)$$

Parameters f, p and q are respectively the number of matches, the time stamp of the first match, and the time stamp of the last match. The *MNRL* (*Maximum Negative Run-Length*) stores the length of the period (in number of frames) during which a codeword has not been matched. Note that parameters $I_{min,k}$ and $I_{max,k}$ of the original model [8], which are minimum and maximum observed brightness of codeword k, are removed since they are useless with our spherical association volume.

2.3 Cast shadows management

False detections caused by cast shadows have a considerable impact on detection accuracy, especially for outdoor video monitoring systems. Although a huge amount of publications address this problem, identifying shadow pixels still remains a challenge. In most efficient solutions, spectral properties are used in combination with spatial, geometrical or texture-based constraints. We propose here an approach based only on spectral properties. In addition of requiring a low computational cost, our technique may be used as a first level of foreground / background / shadow classification that could be post processed with additional constraints.

In [8], a low α value allows classifying most shadow pixels as background, but matched codewords are then erroneously updated with shadowed observations, leading to false detections behind slowly moving cast shadows. In the proposed approach, pixels affected by cast shadows are identified and not considered in the updating process. To avoid consecutively wrongly classifying some background pixels as cast shadows, new background codewords are created for cast shadows pixels belonging to a motionless track stationary for too long.

As with background codewords, an association volume for "shaded codewords" must be defined. In its simplest form, and as proposed by Cucchiara *et al.*[2], this volume can take the form of a rectangular prism (or a parallelepiped depending on the color space) by specifying independent thresholds for each component.

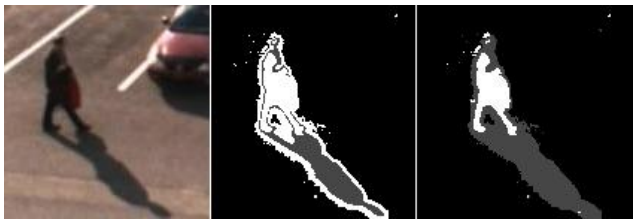

Figure 4: Discrimination of shadow pixels (in gray) with an association volume symmetrically centered on a distribution (center) and with the proposed volume ($\tau_{Y,up}$=1)(right).

The main limitation of such volume is that they are based on the assumption that occlusion of illumination source(s) do not cause color distortion. As mentioned in section 2.1, this assumption is generally invalid in outdoor environment due to the non-uniformity of the spectral power distribution of the sun.

To date, few authors have proposed solutions to take into account the color distortion [6][12][14][19] and all of them use the *RGB* color space. The solution of Martel and Zaccarin [12] seems one of the most accurate since they statistically model the expected shadowed appearance ("ambient illuminated appearance") of each pixel with kernel density estimators [4]. However, since the parameters of all existing distributions should be updated for every pixel, and at every frame, running such an algorithm in real time would be a challenge. In addition, we observed that an association volume symmetrically centered on a distribution or kernel is sometimes too discriminating. As illustrated by center images of Figure 4, it creates false detections along the boundary of dark cast shadows and on pixels slightly shaded by partial occlusion of the sky (occlusion of ambient illumination).

This behavior is mainly caused by the fact that a centered distribution does not model the progressive reduction of intensity observed at shadow borders or in areas where ambient illumination is occluded. This assertion is demonstrated by the right image of Figure 4 obtained by raising the $\tau_{Y,up}$ threshold to 1 (eq. 7). From these observations, and for real-time considerations, we chose to use the shadow association volume represented by the gray cylinder on Figure 5. Mathematically, an observation at time t would lie in the proposed cylinder if the two following conditions are satisfied:

$$\tau_{Y,low} \le \frac{Y_t}{\overline{Y}_k} \le \tau_{Y,up} = 1 \qquad (7)$$

$$\left| Co_t - \left(\overline{Co}_k \cdot \frac{Y_t}{\overline{Y}_k} - \omega \cdot \left(\overline{Y}_k - Y_t \right) \right) \right|^2 + \left| Cg_t - \overline{Cg}_k \cdot \frac{Y_t}{\overline{Y}_k} \right|^2 \le \beta^2 \qquad (8)$$

Our experiments and analysis showed that the Cg component provides very little additional information useful to identify shadow pixels. We can therefore simplify eq. 8 by removing the *Cg* term without affecting performances:

$$\left| Co_t - \left(\overline{Co}_k \cdot \frac{Y_t}{\overline{Y}_k} - \omega \cdot \left(\overline{Y}_k - Y_t \right) \right) \right|^2 \le \beta^2 \qquad (9)$$

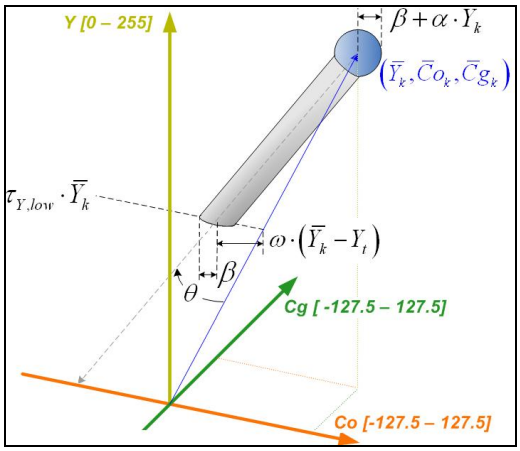

Figure 5: Proposed background *CW* association volume (blue sphere, eq. 2) and shadow association volume (gray cylinder).

The length of the cylinder is determined by the $\tau_{Y,low}$ parameter, which should be adjusted for illumination conditions. When the sky is completely clear and shadows are very dark, $\tau_{Y,low}$ must approach zero to form a longer cylinder. Conversely, when the sky is overcast and shadows are barely perceptible, $\tau_{Y,low}$ must tend towards one to form a short and discriminating cylinder.

The radius of the cylinder is fixed by parameter β (same as eq. 4). Angle θ showed in Figure 5 corresponds to the negative color distortion of component *Co*. The greater the magnitude of the distortion, the larger will be θ. It is important to understand that it is the parameter ω that determines the value of angle θ since a single ω value involves different θ angles depending on the position of the sphere (*CW*) in the *YCoCg* space. The expression $\omega \cdot (\overline{Y}_k - Y_t)$ was chosen to model the quasi-linear relationship between the distortion of the *Co* component and the absolute decrease in intensity. As for threshold $\tau_{Y,low}$, ω must be adjusted based on illumination conditions.

In summary, the proposed shadow association volume has the following interesting characteristics:

1. Only two additional parameters ($\tau_{Y,low}$ and ω) that can be adjusted automatically;
2. Takes into account color distortion;
3. Suitable for real-time application because $\tau_{Y,low}$ and ω are scene-based (same value for all pixels).

3. Quantitative performance analysis

Measuring the performance of a background / foreground segmentation algorithm is not a trivial task. Several factors may limit the validity of the results: limited quantity and quality of benchmark sequences, inappropriate performance metrics, type of post-processing applied on the detection mask, and non-optimal adjustment of parameters of compared algorithms. In this quantitative performance analysis, a special attention has been addressed to each of these factors.

We present and discuss the results obtained on eight outdoor sequences (listed in Table 1). Sequence *H1* is from the ViSOR repository (www.openvisor.org), *H3* from Université Laval (vision.gel.ulaval.ca/~CastShadows/), and *CAM* and *WS* are from University of Singapore (perception.i2r.a-star.edu.sg/). Sequences *CR*, *CP*, *MD* and *GF* were grabbed at our lab and were made publicly available (with many others coregistered thermal-color sequences) at www.ino.ca/Video-Analytics-Dataset.

Table 1: Selected public videos with ground truth at pixel-level.

Sequence	Frames	Resolution [pixels]	Frames with GT	Compression
HighwayI (H1)	440	320x240	10	Cinepak
HighwayIII (H3)	2237	320x240	7	None
Campus (CAM)	2438	160x128	20	None
WaterSurface (WS)	1632	160x128	20	None
Crossroads (CR)	760	320x240	12	MPEG4
ClosePerson (CP)	240	512x184	20	MPEG4
MultipleDeposit (MD)	2400	448x324	15	MPEG4
GroupFight (GF)	1482	452x332	22	MPEG4

Thanks to the ground truth at pixel-level, the number of true positives (*TP*), false positives (*FP*) and false negatives (*FN*) may be determined for every algorithm and every video. Among existing metrics, we chose the Jaccard coefficient (*J*) used by Rosin and Ioannidis [16]:

$$J = \frac{TP}{(TP + FP + FN)} \tag{10}$$

To present a more complete comparative study, we also report the detection rate and the false alarm rate:

$$DR = \frac{TP}{TP + FN} \qquad FAR = \frac{FP}{TP + FP} \tag{11}$$

To quantify the accuracy of identification of shadow pixels, we computed the shadow detection rate (η) and shadow discrimination rate (ξ) as defined by Prati *et al.*[15]:

$$\eta = \frac{TP_S}{TP_S + FN_S} \qquad \xi = \frac{\overline{TP_F}}{TP_F + FN_F} \tag{12}$$

where subscripts *S* and *F* refer to shadow and foreground, and $\overline{TP_F}$ represents the number of foreground pixels minus the number of foreground pixels wrongly classified as shadow. We also report the more intuitive weighted metric γ proposed by Soh *et al.*[17]:

$$\gamma = \frac{GT_S}{GT} \cdot \eta + \frac{GT_F}{GT} \cdot \xi \tag{13}$$

where $GT = GT_F + GT_S$, $GT_S = TP_S + FN_S$ and $GT_F = TP_F + FN_F$.

Detection thresholds of all compared algorithms have been adjusted to maximize the Jaccard coefficient. With the proposed association volumes, a total of four parameters need to be set. They are listed and described in Table 2.

The optimization procedure used was an exhaustive search, and we performed it independently for every sequence. We choose to use a set of optimized parameter for every video instead of a unique set of parameters for all sequences because, in a system operating continuously, the detection thresholds may be automatically adjusted based on illumination conditions. As mentioned in the center column of Table 2, every parameter of the proposed algorithm is related to a physical phenomenon.

Table 2: Detection thresholds to set with the proposed algorithm.

Param.	Related physical phenomenon	Typical range
β	Amplitude of sensor noise	[5 – 15]
α	Speed of illumination variations	[0.01 – 0.2]
$\tau_{Y,low}$	Maximum relative brightness attenuation caused by cast shadows	[0.15 – 0.75]
ω	Color distortion caused by cast shadows	[0 – 0.15]

3.1 Identification of cast shadow pixels

Table 3 presents the shadow related metrics for six test sequences. Best results for every video are printed in bold. Some detection masks obtained with the proposed spherical association volume are presented in Figure 6.

To be fair in our comparison, we adapted and integrated our cast shadow management algorithm (section 2.3) into the original method of Kim *et al.*[8]. A pixel is classified as shadow if no color distortion is measured with their metric (eq. 1), and if our condition on luminance reduction (eq. 7) is satisfied. This shadow enhanced version of the original method also allows empirically measuring the impact of the choice of the *YCoCg* color space (with spherical or ellipsoidal association volume) instead of the *RGB*-based color distortion metric of Kim *et al.*[8]. In fact, these results combine two contributions (color space + association volume), but they are hardly separable since an association volume designed for the *RGB* color space cannot be applied on a mixed color space like *YCoCg*.

Table 3: Identification of cast shadow pixels, quantitative results.

Seq.	Metric	Kim *et al.*[8] + shadow	Proposed spherical	Proposed ellipsoidal
H1	η	0.764	0.768	0.708
	ξ	0.695	0.692	0.743
	γ	0.723	0.722	**0.729**
H3	η	0.074	0.688	0.066
	ξ	0.926	0.674	0.957
	γ	0.727	0.677	**0.748**
CR	η	0.182	0.637	0.618
	ξ	0.886	0.848	0.846
	γ	0.755	**0.809**	0.804
CP	η	0.806	0.853	0.833
	ξ	0.716	0.853	0.854
	γ	0.763	**0.853**	0.843
MD	η	0.711	0.671	0.664
	ξ	0.770	0.792	0.806
	γ	0.761	0.774	**0.782**
GF	η	0.922	0.881	0.840
	ξ	0.705	0.732	0.759
	γ	0.778	0.782	**0.786**

At first glance, the *YCoCg* ellipsoidal association volume seems the most accurate (best γ for most videos). However, the difference with the spherical volume is non

significant for all sequences, except *H3* which illustrated a bias of the weighted metric γ when few cast shadow pixels are present. Indeed, we can see that the ellipsoidal volume got the best γ even though it detects (η) only 6.6% of shadow pixels. The same bias affects the results obtained with the original method of Kim *et al.*[8] (η = 7.4%).

The reason why η is so low with the ellipsoidal volume is that cast shadows of video *H3* present a strong distortion on *Cg* channel. This is somewhat abnormal since we didn't observe such behavior on any other video analyzed. A poor sensor quality, a dubious white-balance and the smoggy illumination conditions (giving a yellowish color to the sky) may cause the phenomenon. Since no distortion on *Cg* channel should be observed along our model (Figure 5), most cast shadows are thus wrongly classified as foreground. However, they are correctly identified as shadow with the spherical volume because it doesn't consider the *Cg* channel for cast shadow detection (eq. 9).

As predicted, the accuracy of cast shadow identification with the original method is inferior for sequences *H3*, *CR* and *CP* because of the significant color distortion. For such illumination conditions, the proposed association volume, which models the color distortion along the *Co* channel (eq. 9 and Figure 5), leads to better results.

Based on performances reported in previous works, we compared our shadow detection and discrimination rates with 8 and 2 others algorithms on sequences *HI* and *H3* respectively. Note that results for the *GMSM* method [11] comes from [10] for *H1* and from [12] for *H3*.

Table 4: Comparison with state-of-the-art methods.

Méthode	H1		H3	
	η	ξ	η	ξ
Proposed spherical	0.768	0.692	0.688	0.674
Proposed ellipsoidal	0.708	0.743	0.066	0.957
Kim *et al.*[8] + shadow	0.764	0.695	0.074	0.926
« Kernel-based » physical model [12]	0.705	0.844	0.684	0.712
GMM LGf (Local + Global features) [10]	0.721	0.797	-	-
GMSM [11]	0.633	0.713	0.585	0.444
SNP (Statistical non parametric) [15]	0.816	0.638	-	-
SP (Statistical parametric) [15]	0.596	0.847	-	-
DNM1 (Determ. non model-based) [15]	0.697	0.769	-	-
DNM2 (Determ. non model-based) [15]	0.755	0.624	-	-
GMM + texture + geometry [20]	0.672	0.902	-	-

Only based on results reported in Table 4, the compared methods cannot be sorted from worst to best because different operation points (η vs ξ) may have been chosen. As a proof, most of these algorithms have been implemented and compared by Zhang *et al.* [20] and distinct results from those reported in original papers have been obtained.

We can still draw some conclusions from Table 4. Our *YCoCg*-based method performs better than the *DNM2* approach because we obtain a higher shadow discrimination rate ξ (69.2% vs 62.4%) with a similar shadow detection rate η (76.8% vs 75.5%).

On video *H1*, ξ reported for the physical model [12] and the GMM LGf [10] are higher than what we measured with our ellipsoidal volume (84.4% and 79.7% vs 74.3%) while η are comparable. Similarly, on video *H3*, the physical model outperforms our spherical volume since we got a lower ξ (67.4% vs 71.2%) for a comparable η.

Globally, the physical model [12], the *GMM LGf* [10] and the combined method of [20] appear to be the most accurate. This observation was predictable because these approaches are significantly more computationally expensive. Indeed, in addition of pixel-based spectral information (including color distortion caused by cast shadows like our method), they also consider the texture in the local neighbourhood of every pixel. Comparatively to other methods, our *YCoCg*-based approach seems to perform slightly better, and it is faster than *GMM* and *GMSM*-based approaches.

3.2 Detection accuracy

Conditions expressed by equations 4, 7 and 9 give a preliminary detection mask in which every pixel is classified as background, foreground or shadow. Typically, some filtering is performed on this mask to remove noise prior to the blob labeling process. The three last columns of Figure 6 illustrate the enhanced detection masks obtained at the output of these filtering and blob labeling processes (to enhance details, we only illustrated a sub-area of the whole frames). Exactly the same cascades of operations are applied for every algorithm compared in Table 5: spatial filtering of candidate shadow pixels, morphological closure and blob labeling. We also integrated elimination of too small blobs and filling of holes (pixels printed in light gray in the three last columns of Figure 6) in the blob labeling process.

Table 5: Foreground detection accuracy, quantitative results.

Seq.	Metric	Kim *et al.* [8]	Kim *et al.*[8] + shadow	Proposed spherical	Proposed ellipsoidal
H1	*DR*	0.955	0.755	0.804	0.870
	FAR	0.472	0.166	0.189	0.222
	J	0.5150	0.6566	0.6775	**0.6972**
H3	*DR*	0.932	0.893	0.839	0.929
	FAR	0.266	0.250	0.151	0.255
	J	0.6966	0.6881	**0.7300**	0.7050
CAM	*DR*	0.931	0.915	0.735	0.853
	FAR	0.327	0.112	0.080	0.088
	J	0.6416	**0.8204**	0.6909	0.7881
WS	*DR*	0.970	0.950	0.945	0.954
	FAR	0.078	0.038	0.034	0.040
	J	0.8960	0.9155	0.9146	**0.9177**
CR	*DR*	0.906	0.824	0.883	0.861
	FAR	0.216	0.168	0.113	0.096
	J	0.7253	0.7065	**0.7934**	0.7890
CP	*DR*	0.855	0.659	0.779	0.745
	FAR	0.540	0.122	0.185	0.147
	J	0.4271	0.6035	**0.6624**	0.6562

Globally, the proposed algorithm performs better than the original approach, except for the *CAM* sequence. The counter-performance of the spherical volume on this video

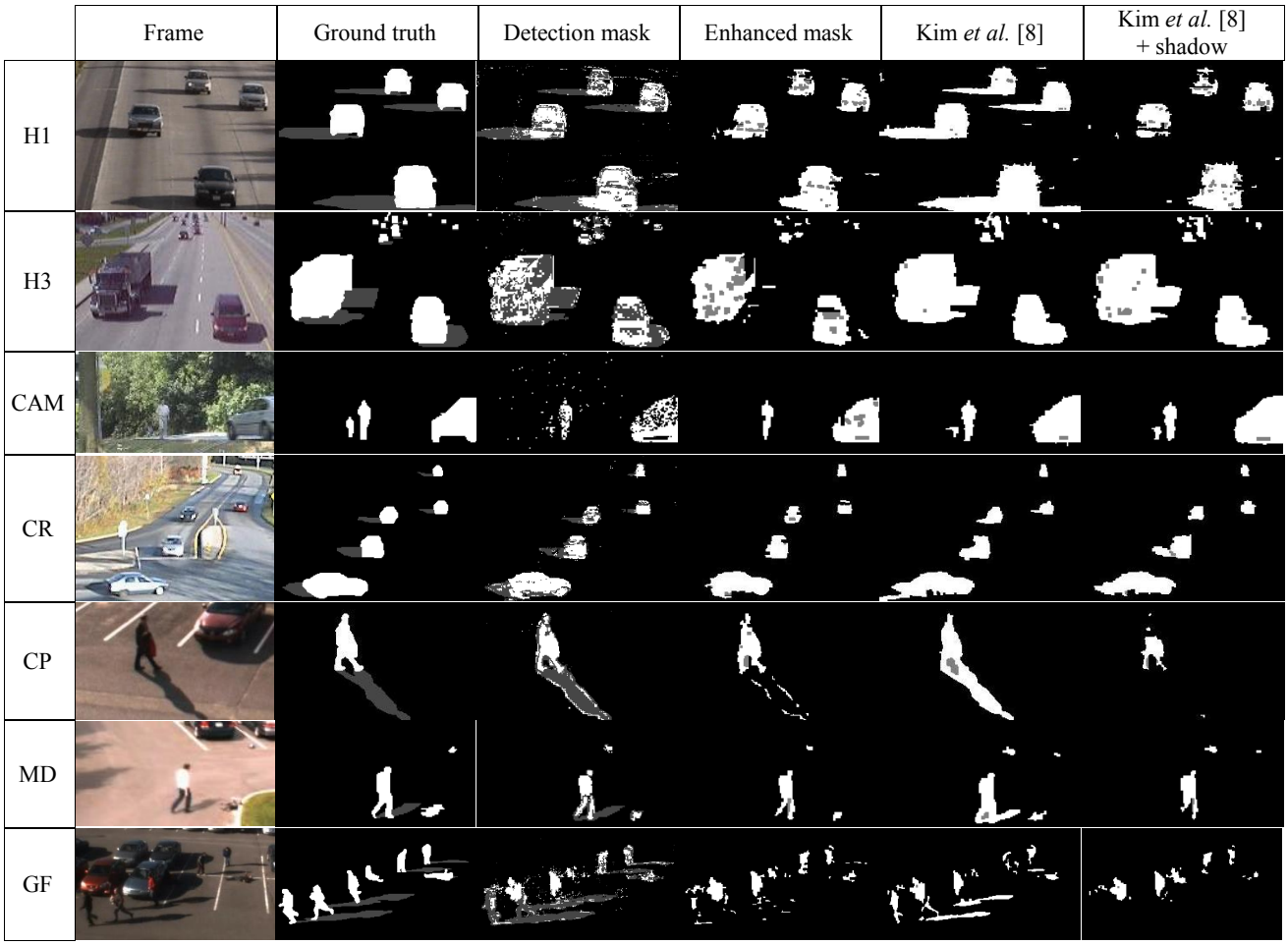

Figure 6: Examples of detection masks (column 3) and enhanced detection masks (columns 4, 5 and 6) obtained with the proposed algorithm (columns 3 and 4) and with the original codebook method [8] (columns 5 and 6).

is a consequence of using a unique detection threshold β on luminance and chrominance components (eq. 4). Indeed, some strong illumination variations are observed in this sequence (especially on the trunk of the tree), leading the spherical volume to use a more permissive β threshold. Consequently, a lower detection rate is obtained (73.5%).

It is interesting to note that the addition of the shadow module to the original method of Kim *et al.*[8] leads to a significant reduction of the false alarm rate on the sequence *CAM* even though there is very few shadow pixels in this video. This performance improvement is caused by the fact that many false detections are classified as shadow candidates and are subsequently labeled as background. We can also note that the addition of the shadow module to the original method contributes to significantly improve the accuracy for sequences *H1* and *CP*, but not for *H3* and *CR* where cast shadows are affected by color distortion.

Another observation is that the ellipsoidal volume performed slightly better than the spherical volume for sequences *H1* and *WS*, which are not compressed, but it is the spherical volume that performs slightly better for sequences *CR* and *CP*, which are MPEG-4 compressed. This confirms that the ellipsoidal volume is slightly more sensitive to color distortion caused by MPEG compression artefacts located along the edges of moving objects.

3.2 Memory and processing time

For every video, we measured the mean processing time per frame (including filtering and blob labeling processes) in detection mode. In initialization mode, a shorter processing time is required. As reported in Table 6, the spherical association volume is faster for all sequences, especially for those with high resolution and many background pixels. A 3 GHz Intel Core2 Duo CPU with 3.25 GB of RAM was used.

Table 6: Mean processing time per frame in ms.

Seq.	Kim *et al.* [8]	Kim *et al.*[8] + shadow	Proposed spherical	Proposed ellipsoidal
H1	11.58	13.07	**8.96**	11.03
H3	10.06	10.97	**7.25**	8.60
CAM	4.27	4.89	**3.87**	4.27
CR	10.40	10.84	**7.20**	8.33
CP	12.42	12.93	**8.22**	9.72

The computational load mainly increases with resolution, but it is also affected by the ratio of foreground /

background pixels. Indeed, classifying a pixel as background is faster than as foreground or shadow because conditions related to the shadow association volume (eq. 7 and eq. 9) do not have to be tested. It is why the processing rate on sequence *H1*, which contains large foreground objects, is about 22% slower than on *H3*.

In terms of memory requirement, the proposed algorithm requires 19.4% less memory for a VGA sequence (71.3 MB against 88.4 MB). The economy is due to the elimination of variables $I_{min,k}$ and $I_{max,k}$ in every *CW*. For this analysis, we limited the permanent background codebook to a maximum of 5 codewords per pixel.

4. Conclusion

Three generic improvements that could be applied to most background subtraction algorithms in outdoor environments have been proposed: the use of the *YCoCg* color space, a spherical association volume and a cast shadows management approach. We quantitatively demonstrated that the three elements lead to a reduced processing time with improved accuracy.

Optimal tuning of detection thresholds is often a major determinant of performances, and too little attention is generally addressed to this aspect. To operate continuously in outdoor environment, it is essential to adjust detection parameters to observation conditions. The proposed algorithm being used in the field on a 24 hours system since more than two years by a partner integrator, we have already worked on the design and implementation of such automatic mechanism. Our intention in a close future is to quantitatively evaluate its efficiency and demonstrate that the relatively high number of parameters of the proposed association volumes (Figure 5) could be easily set. It is worthwhile to mention that this performance analysis consists in a significant effort because it requires the acquisition and processing of very long benchmark sequences grabbed in various outdoor scenes and weather conditions.

As further development, it would also be interesting to integrate additional texture-based features into the proposed background subtraction algorithm, and to evaluate their impact on performances.

References

[1] C. Benedek & T. Sziranyi, Study on color space selection for detecting cast shadows in video surveillance, *Journal of Imaging Systems and Technology*, 17(3), 2007, 190-201.

[2] R. Cucchiara, C. Grana, M. Piccardi & A. Prati, Detecting moving objects, ghosts and shadows in video streams, *PAMI*, 25(10), 2003, 1337-1342.

[3] A. Doshi & M.M. Trivedi, Satellite imagery based adaptive background models and shadow suppression, *Signal, Image and Video Processing*, 1(2), 2007, 119-132.

[4] A. Elgammal, R. Duraiswami, D. Harwood, & L.S. Davis, Background and foreground modeling using nonparametric kernel density estimation for visual surveillance, *Proc. of the IEEE*, 90(7), 2002, 1151-1163.

[5] P. Fihl, R. Corlin, S. Park, T.B. Moeslund & M.M. Trivedi, Tracking of individuals in very long video sequences, *International Symposium on Visual Computing*, 2006.

[6] I. Huerta, M. Holte, T. Moeslund & J. Gonzàlez, Detection and removal of chromatic moving shadows in surveillance scenarios, *ICCV*, 2009.

[7] E.A. Khan & E. Reinhard, "Evaluation of color spaces for edge classification in outdoor scenes", *IEEE Int. Conf. on Image Processing*, vol.3, Genoa, Italy, 2005, 952-955.

[8] K. Kim, T.H. Chalidabhongse, D. Harwood & L. Davis, Real-time foreground - background segmentation using codebook model. *Real-time Imaging*, 11(3), 2005, 172-185.

[9] A. Leykin, R. Ran & R. Hammoud, Thermal-visible video fusion for moving target tracking and pedestrian classification, *IEEE Workshop on Object Tracking and Classification in and Beyond the Visible Spectrum*, 2007.

[10] Z. Liu, K. Huang, T. Tan & I. Wang, "Cast shadow removal combining local and global features", *IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, MN, 2007.

[11] N. Martel-Brisson & A. Zaccarin, "Learning and removing cast shadows through a multidistribution approach", *PAMI*, 29(7), 2007, 1133-1146.

[12] N. Martel-Brisson & A. Zaccarin, Kernel-based learning of cast shadows from a physical model of light sources and surfaces for low-level segmentation, *CVPR*, 2008.

[13] H. Malvar & G. Sullivan. YCoCg-R: A color space with RGB reversibility and low dynamic range, *Joint Video Team of ISO/IEC MPEG & ITU-T VCEG Meeting*, 2003.

[14] S. Nadimi, & B. Bhanu, Physical models for moving shadow and object detection in video, *PAMI*, 26(8), 2004, 1079-1087.

[15] A. Prati, I. Mikic, M.M. Trivedi & R. Cucchiara, Detecting moving shadows: algorithms and evaluation, *PAMI*, 25(7), 2003, 918-923.

[16] P.L. Rosin & E. Ioannidis, Evaluation of global image thresholding for change detection, *Pattern Recognition Letters*, 24(14), 2003, 2345-2356.

[17] Y.S. Soh, H. Lee & Y. Wang, Invariant color model-based shadow removal in traffic image and a new metric for evaluating the performance of shadow removal methods, *Lecture Notes in Computer Science*, 4099, 2006, 544-552.

[18] C. Stauffer & E. Grimson, Adaptive background mixture models for real-time tracking, *CVPR*, 1999.

[19] P.J. Withagen, F.C.A. Groen & K. Schutte, Shadow detection using a physical basis, *IEEE Instrumentation and Measurement Technical Conf.*, 2008.

[20] W. Zhang, X.Z. Fang, X.K. Yang & Q.M.J. Wu, "Moving cast shadows detection using ratio edge", *IEEE Trans. on Multimedia*, 9(6), 2007, 1202-1214.