SUMMARY This paper describes a multiresolutional Gaussian mixture model (GMM) for precise and stable foreground segmentation. A multiple block sizes GMM and a computationally efficient fine-to-coarse strategy, which are carried out in the Walsh transform (WT) domain, are newly introduced to the GMM scheme. By using a set of variable size block-based GMMs, a precise and stable processing is realized. Our fine-to-coarse strategy comes from the WT spectral nature, which drastically reduces the computational steps. In addition, the total computation amount of the proposed approach requires only less than 10% of the original pixel-based GMM approach. Experimental results show that our approach gives stable performance in many conditions, including dark foreground objects against light, global lighting changes, and scenery in heavy snow.

key words: foreground segmentation, Gaussian mixture model, fine-to-coarse strategy, Walsh transform, variable block size

1. Introduction

Foreground segmentation is a fundamental, but tough task in many computer vision applications such as video surveillance and auto-cruising. The goal of the foreground segmentation process is to detect moving objects from the current frame. Many background subtraction techniques are widely studied for this purpose in recent years [1].

One of the best formulations is to introduce a Gaussian mixture model (GMM) into every pixel in the background. Stauffer and Grimson have proposed the GMM scheme which consists of the RGB components of a pixel [2]. This scheme is suitable for the approximation of a multimodal distribution of the background, and can be used to cope with dynamic scenes. However, the pixel-level processing has crucial drawbacks, such as a large amount of computation and that of memory capacity.

In order to overcome these problems, the employment of spatial correlation is incorporated into the GMM scheme. For the real-time processing, a hierarchical data structure is employed for the GMM scheme in [3]. This approach decreases the processing amount, compared to the GMM without using the structure. A block-based simplified GMM scheme is employed for a DSP implementation in [4]. The pixel-level foreground processing is limited within the detected blocks, based on the result of the subsequent simplified GMM scheme as coarse processing. These hierarchical approaches achieve the reduction of the calculation amount rather than the improvement of segmentation performance.

Further reduction of the computation amount is achieved by the use of a Discrete Cosine Transform (DCT) based single Gaussian model [5]. In addition to the single Gaussian model, the use of only luminance component information contributes simplification of the computational complexity. This method employs two feature parameters, the DC parameter and an AC parameter which correspond to the lowest frequency coefficients and a set of low frequency coefficients, respectively. These parameters can be considered to represent intensity and texture information. The AC feature parameter is defined by a sum of low frequency coefficients. Therefore, its distribution is assumed to be a single Gaussian, because of the Central Limit Theorem. However, the employment of the single Gaussian has caused false decision under dynamic background motions, as well as generating separated segments from a single foreground object.

In this paper, a precise, stable and computationally efficient foreground segmentation method is proposed. Our framework has two novel components: a multiple block sizes GMM and a fine-to-coarse approach, which are carried out in the Walsh transform (WT) domain. Different spatio-temporal performance is achieved by the multiple block sizes GMM. The process in a small block size mode gives precise but unstable results. On the other hand, stable but rough results are produced in a large block size mode. By combining these results, precise and stable performance is achieved. The fine-to-coarse processing is efficiently performed by introducing Walsh spectrum feature parameters. According to the WT spectral nature, a set of fine processing can be realized by an interim result for a coarse processing. Another possible approach is to employ DCT, as motion pictures are normally compressed by MPEG or H.264. However, a direct application of compressed video in the DCT or another transform domain is not possible except Intra coded Pictures (I-pictures), which do not employ the frame difference and appear once in every group of several pictures [6]. In order to realize frame-by-frame foreground segmentation, a computationally simple transform is preferable.

The rest of the paper is organized as follows. Our background models in the WT domain are described in Sect.2. The fine-to-coarse processing and the use of multiple thresholds are explained in Sect.3. The stable performance of our algorithm and its evaluation are shown.
2. Gaussian Mixture Model in Walsh Transform Domain

2.1 Walsh Spectrum Feature Parameters

Walsh spectrum feature parameters to be applied to the GMM are determined by using coefficient sets of the vertical, horizontal and diagonal directions. Every coefficient has strong spatial correlation within the set. By using WT of the luminance component, feature parameters are defined as follows.

\[
\begin{align*}
f_v &= \sum_{i=0}^{3} j^2 \times |WAL(i, 0)| \\
f_h &= \sum_{j=0}^{3} i^2 \times |WAL(0, j)| \\
f_d &= \sum_{i=0}^{3} \sum_{j=0}^{4-i} (i^2 + j^2) \times |WAL(i, j)|, \\
\end{align*}
\]

where \(|WAL(i, j)|\) is an absolute value of the Walsh coefficient of coordinate \((i, j)\), \(f_v\) is the DC coefficient, and \(f_v\), \(f_h\) and \(f_d\) are weighted sums of AC coefficients in vertical, horizontal and diagonal directions, as shown in Fig. 1. The weighted sum of spectrum components is the purpose for the compensation of the decreasing nature in these coefficients, and the exclusion of the high frequency coefficients is for the elimination of noise in these coefficients. In our experiment, the segmentation performance is not affected by including higher frequency coefficients beyond coordinate 4 in \(f_v\), \(f_h\) and \(f_d\). In contrast, the use of low frequency coefficients in these parameters achieves higher performance. Therefore, our method employs the coefficients under the coordinate 4. The similar approach is employed in [5].

Figure 2(a) shows a part of an image sequence which was taken in rain at an outdoor scenery (pixels: \(1920 \times 1080\)). By using 900 frames from this sequence, the distribution of the \(f_v\), \(f_h\) and \(f_d\) are measured and summarized as histogram charts, respectively. The results are shown in Figs. 2(b)–(d). The variances of \(f_v\) and \(f_h\) in Fig. 2 are \(\sigma_v^2 = 6.72 \times 10^6\) and \(\sigma_h^2 = 3.01 \times 10^4\), respectively. The reason of a smaller value of \(\sigma_h^2\) comes from the raining condition. Rain drops constantly move in the vertical direction. Moreover, the shape of each distribution modeling parameters at time \(t\) are defined as follows,

\[
\begin{align*}
X_{DC,t} &= f_v \\
X_{AC,t} &= [f_v, f_h, f_d]^T. \\
\end{align*}
\]

2.2 Gaussian Mixture Model Formulation

In this section, the GMM formulation, originally proposed in [2], is briefly reviewed in order to introduce the GMM to the Walsh parameters. Let \(X_t\) be a modeling parameter at time \(t\), and the probability of the observing parameter is assumed by a mixture of \(K\) Gaussian distributions in the following way.

\[
P(X_t) = \sum_{i=1}^{K} \omega_i \times N(\mu_i, \Sigma_i),
\]

where \(\omega_i\) is the weight of the \(i\)-th Gaussian, \(N(\mu_i, \Sigma_i)\) indicates the Gaussian probability density function of \(i\)-th component, having the mean \(\mu_i\) and the covariance matrix \(\Sigma_i\). The number \(K\) is determined by the available memory capacity and computational amount. Typically, \(K\) is selected from 3 to 5 in many articles. In our scheme, \(K = 3\) is used in every block size mode. Therefore, our GMM employs much more Gaussians in comparison to other
systems. The GMM is composed of one or more Gaussians to the background model, and the foreground is modeled by remaining Gaussians. Therefore, $B$ out of $K$ distributions are considered as the background model, based on the weight of each Gaussian and its standard deviation. If $X_i$ satisfies

$$|X_i - \mu_{ij}| < T,
$$

and the Gaussian with $\mu_{ij}$ is included in the $B$ distributions, then $X_i$ is classified as the background. $T$ is a threshold to judge whether the current state of $X_i$ is matched in the distribution of the $i$-th Gaussian. When $X_i$ satisfies Eq. (4), the update scheme is applied to the $\mu_{ij}, \Sigma_{ij}$ and $\omega_{ij}$ of the distribution. In our method, this update scheme is based on the Ref. [5].

Note that the covariance matrix for $X_{AC_i}$ in our method is defined as $\Sigma_{AC_i} = \text{diag}[\sigma^2_{11}, \sigma^2_{22}, \sigma^2_{33}]$, where $\text{diag}[-]$ indicates the diagonal matrix. This assumes three feature parameters are independent and have different variances. For $X_{DC_i}$, the covariance matrix becomes a scalar value, $\Sigma_{DC_i} = \sigma^2_{DC_i}$. In our framework, WT-based GMM schemes are carried out on different block size modes. The tentative segmentation results on a block are obtained from OR operations of DC and AC decisions. The use of OR decision includes the changing of the intensity and the texture, in other words. Moreover, the threshold $T$ may have different values in each block size mode. A contribution of these schemes will be described later.

3. Fine-to-Coarse Approach

3.1 Advantage on Walsh Transform

Walsh transform (WT) is one of the orthogonal transforms, and its calculation requires only additions and subtractions. Moreover, WT has an efficient property with respect to the expansion of the block size. Let us consider $N = 2^n$, where $n$ is a positive integer. In one dimensional case, the Walsh spectrum $w_N$ of a $N$ element block can be calculated by the adjacent two $N/2$ element blocks of $w_{N/2}$ and $w'_{N/2}$ within the $N$ element block as follows.

$$\begin{align*}
\omega_N^{(2k)} &= w_{N/2}^{(k)} + w_{N/2}^{(k)}, \\
\omega_N^{(2k+1)} &= w_{N/2}^{(k)} - w'_{N/2}^{(k)},
\end{align*}
$$

(5)

where $k = 0, 1, \cdots, \frac{N}{2} - 1$. By calculating Eq. (5), the dyadic ordered Walsh transform is realized. A sequency ordering, which corresponds to the frequency, can also be computed in a similar approach [7]. As a result, a transform block size can be easily enlarged without using the inverse transform. In case of the FFT, this kind of a scheme is impossible, due to the existence of twiddle factors. Therefore, the use of WT-based feature parameter is feasible for our fine-to-coarse strategy which drives the multiple block sizes GMM scheme.

3.2 Fine-to-Coarse Strategy in Our Method

The fine-to-coarse approach in our framework is composed of two steps. At the first step, the block size is expanded from a small size to a large size. By using the previously described Walsh spectral nature, Walsh spectrum components in smaller block sizes are definitely calculated on the way to the final spectrum components as shown in Fig. 3. By employing this procedure, the GMM scheme is performed on every block size mode. Block sizes typically vary from 4 to 64. At the second step, the tentative segmentation results, obtained from different block sizes, are combined by AND operations.

The threshold $T$ in Eq. (4) can be modified in every block size mode for further improvement of performance. In our algorithm, $T_{DCN}$ and $T_{ACN}$ are defined for DC and AC decisions, respectively:

$$\begin{align*}
T_{DCN} &= \delta_{DCN} \times \mu_{DC} + 2.5 \times \sigma_{DC} \\
T_{ACN} &= \delta_{ACN} \times \mu_{AC} + 2.5 \times [\sigma_v, \sigma_h, \sigma_d]^T,
\end{align*}
$$

(6)

where $\delta_{DCN}$ and $\delta_{ACN}$ are parameters which may have different values on each block size. If an accurate standard deviation of a distribution can be estimated, $T$ should be defined by a value of the 2.5 times the standard deviation because of Gaussianity [1],[2]. However, our method initially gives a high variance without using an estimation of the parameters. Hence, in order to reduce the effect of outliers, $T$ is defined by linear combination of the mean and the standard deviation. Therefore, $T$ is also controlled by the mean which is related to the observing parameter.

Normally, $\delta_{DCN}$ is constant and $\delta_{ACN}$ has a linear relation from block size $N = 8$ to $N = 32$ as shown in Fig.4. However, the parameters of $\delta_{DCN}$ and $\delta_{ACN}$ at the finest and the largest blocks are modified in the following reasons. As the largest block size mode produces a rough detection of an object, the threshold value is set to a little bit higher than the normal one. In the same way, the threshold value of the finest block size mode is decreased to detect a precise contour of an object without any consideration of background noise.

In addition, morphological opening/closing operations are applied for each segmentation. In particular, by using this operation, the decision of large block size modes, which comes from the higher threshold, can reliably cover
Fig. 4 The parameters of $\delta_{DC}$ and $\delta_{AC}$ for the thresholds $T_{DC}$ and $T_{AC}$, shown in Eq. (6).

Fig. 5 The procedure of the introduced Selective Fast WT (SFWT). $W^2_{N/2}(i, j)$ indicates the Walsh coefficient of coordinate $(i, j)$ in the $N \times N$ block $b$.

Four adjacent $N/2 \times N/2$ blocks $N \times N$ block

(a) The relationship of block positions

(b) The calculation flow of SFWT

For further reduction of calculation amounts of Eq. (1), the selective Fast WT (SFWT) is introduced, because our parameters require only low frequency coefficients. The complete procedure of the SFWT from $W^0_N(b = 1, \ldots, 4)$ to $W^1_N$ is shown in Fig. 5. At the first step, $W^1_{N/2}(i, j), \ldots, W^4_{N/2}(i, j)$ are extracted, where $i, j \in [0, 1]$. The relationship of block positions is shown in Fig. 5(a). At the following step, shown in Fig. 5(b), these components are calculated into Walsh spectrum components in a wider block size, based on the relation of Eq. (5). The results are arranged in the order of sequency and allocated to $W^1_N$. Consequently, the calculation of the transform is fit to that of the feature parameter, defined in Eq. (1). By applying the similar manner to $W^1_N, \ldots, W^4_N$, Walsh spectrum components of $W^1_N$ can also be calculated.

The comparison of computational steps among conventional sequence ordered FWT [7], DCT [8] and SFWT for each $N \times N$ transform block is shown in Fig. 6. The total number of additions and that of multiplications are shown in Figs. 6(a) and (b), respectively. The solid line indicates the calculation amount of SFWT, and it always shows that the required additions in SFWT is less than 30% of that in conventional FWT. This is because the SFWT is a rearranged version of FWT which calculates only required low frequency coefficients. By employing SFWT, computational steps can be drastically reduced in our fine-to-coarse strategy.

4. Evaluation

4.1 Computational Complexity

Our method is compared with the algorithms [2], [5], which contains different formulations in terms of the background modeling. Among three algorithms, the background modeling process accounts for computational complexity in the individual algorithm.

For simplicity, let us assume that the frame size is $M \times M$. The pixel-wise GMM algorithm [2] deals with the RGB components in each pixel. Hence the computational complexity is proportional to $3M^2$. In contrast, the number of operations for the modeling scheme decreases in the
block-based algorithm. The DCT-based algorithm [5] employs two feature parameters and the block size is set to 8. Therefore, the required operation amount is proportional to \(\frac{M^2}{n^2}\). In addition, the use of a single Gaussian model leads to less computation. The computational steps will be nearly less than a percent. In our algorithm, a variable block size is set to \(2^n(n = 2, 3, \cdots, 6)\) and four feature parameters are used in every block size. Therefore, the required operation amount is expressed as the following geometric progression,

\[
4M^2 \sum_{j=2}^{6} \frac{1}{(2^n)^j} = \frac{M^2}{3}.
\]

As a result, computational complexity of our algorithm is proportional to \(\frac{M^2}{n^2}\), which is around 10% steps, compared to pixel-wise GMM. The summary is shown in Table 1.

Note that our experiment does not include the algorithm proposed in [3]. The main target of the method is the reduction of the required computational amount. However, the reduction differs from 1/4 to 1/8, depending on the frame size and picture contents. The performance of this method has been reported to be similar to the pixel-wise GMM or slightly less.

The computation times for each algorithm are measured on MATLAB 7.5 platform as examples, where each computational time is not optimized. All the experiments are implemented with Intel Core Duo CPU 2.0 GHz, 2 GB RAM and Mac OS X. The average time, which is required to process in each frame, is summarized in Table 2. In the background modeling, our method requires less than 10% of time, compared to the pixel-wise GMM. This result corresponds to the analysis, shown in Table 1. On the other hand, the DCT based method is significantly fast. This result comes from the algorithm simplification of the single Gaussian model approach. The single Gaussian scheme causes around 3 times faster than that of the mixture of 3 Gaussians in Table 2, due to the complexity difference of Gaussian processing. Additional speed-up factors will be presented in MATLAB control functions which are often used in the GMM scheme. However, the transform calculation cost occupies a large portion in the DCT based method. Due to Walsh transform and the introduced SFWT requires only less than 10%, compared to the DCT.

All the methods employ a post processing in order to improve the segmentation performance. Typically, a blob removal scheme is applied to eliminate isolated false decisions [2], [5]. Because of the pixel-wise processing, the computation of post processing in ref. [2] requires much time, compared to the other two block-wise processing, shown in Table 2. Our method employs the morphological operation in addition to the scheme. Therefore, in terms of the post processing, our method results in a slightly longer computation time than that of DCT based one.

### 4.2 Experimental Results

In this section, the performances of three algorithms are evaluated by using three test sequences. Test sequences were taken in outdoor with a CMOS sensor camera, recording to HDV format. The size of the sequences is stretched to 1920 \times 1080 pixels. The results of foreground segmentation are shown in Fig. 7. The first row shows four selected test frames, and the second row indicates ground truths for these test frames. Foreground pixels in the ground truth is manually labeled for the reason that it is important to distinguish the object of interest as a single segment. The third row shows the results of the pixel-wise GMM method [2], and the fourth row is that of the DCT-based single Gaussian model method [5]. In experiment, the parameters on these methods are set to the suggested values which are described in the references. The results of our method are shown in the fifth row.

Before going into the considerations, Table 3 summarizes the evaluation results of performance. As the metrics, the false positive (FP) and the false negative (FN) are employed. These parameters indicate the number of background/foreground pixels which are wrongly classified, respectively. By using the FP and the FN, false positive ratio (FPR) and false negative ratio (FNR) are defined as follows.

\[
\text{FPR} = \frac{\text{the number of background pixels wrongly classified (FP)}}{\text{the number of background pixels in the ground truth}}
\]

\[
\text{FNR} = \frac{\text{the number of foreground pixels wrongly classified (FN)}}{\text{the number of foreground pixels in the ground truth}}
\]

The test frame 1 features that two men are crossing each other and their brightness is dark. The segmentation result of the pixel-wise GMM method includes many false negatives within an object. This false decision comes

### Table 1 Evaluation of the computational complexity of background modeling methods. (For simplicity, the frame size is assumed by \(M \times M\).)

<table>
<thead>
<tr>
<th>Modeling</th>
<th>Transform</th>
<th>Feature parameter</th>
<th>process</th>
<th>Total amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref. [2]</td>
<td>GMM</td>
<td>–</td>
<td>(RGB)</td>
<td>(3M^2)</td>
</tr>
<tr>
<td>Ref. [5]</td>
<td>Single GM</td>
<td>DCT</td>
<td>2</td>
<td>((M/8)^2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(AC)</td>
<td>(/ 32)</td>
</tr>
</tbody>
</table>
| Ours     | GMM       | WT                | 4       | \(M^2 \sum_{n=2}^{3} \frac{1}{(2^n)^j}\) | \(M^2 / 3\)

### Table 2 Evaluation of the average computation time for the algorithms [sec./frame]. (Each time is measured on MATLAB without optimization.)

<table>
<thead>
<tr>
<th>Modeling</th>
<th>Transform</th>
<th>Post processing</th>
<th>Total amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref. [5]</td>
<td>0.47</td>
<td>5.70</td>
<td>1.08</td>
</tr>
<tr>
<td>Ours</td>
<td>122*</td>
<td>0.52</td>
<td>1.13</td>
</tr>
</tbody>
</table>

* If the smallest block size is set to 16, the time can be reduced to 8.20 [sec./frame].
from the dark objects and from the process on spatially independent pixels. On the DCT-based method, stable results are obtained in terms of false positives, due to spatial information. However, a single foreground object is separated into two objects, due to the single Gaussian model. Our method achieves robust performance with respect to the FNR in Table 3. In addition, a foreground object can be segmented as a single region.

The test frame 2 and 3 includes the complicated motions in the background, which are caused by leaves around the pond and waves at the surface of the pond. The test frame 2 is under global lighting changes, due to the covered/uncovered sun by a cloud. The pixel-wise GMM method presents a high FPR, due to the global lighting changes, because this method uses only the pixel level processing. However, thanks to the mixture model, a FNR becomes low. On the other hand, the transform domain processing approaches show stable performance under the global lighting changes.

The test frame 3 includes an uncovered region. When a still object in the background starts to move, an uncovered region is generated. Therefore, such a region should be classified as the background. The pixel-wise GMM method gives the false positives at the region. This is because the background model is slowly updated in this method. In the DCT-based method, a part of the foreground person’s back is lost. This false decision seems to be caused by the similarity of the intensity between the background and the foreground object. Our method retains better decision in this situation, because of the use of the multiple block size modes.

The test frame 4 was taken in the heavy snow situation. The snow falls toward a diagonal direction. The pixel-wise GMM method labels the snow as the foreground. The DCT-based method shows better performance in the FPR.

**Table 3** Performance of algorithms on test sequences. (FPR and FNR indicate False Positive Ratio and False Negative Ratio, respectively.)

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Test seq.1</th>
<th>Test seq.2</th>
<th>Test seq.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref. [2]</td>
<td>FPR 0.0228</td>
<td>0.1320</td>
<td>0.0738</td>
</tr>
<tr>
<td></td>
<td>FNR 0.5014</td>
<td>0.1066</td>
<td>0.1107</td>
</tr>
<tr>
<td>Ref. [5]</td>
<td>FPR 0.0043</td>
<td>0.0093</td>
<td>0.0024</td>
</tr>
<tr>
<td></td>
<td>FNR 0.1942</td>
<td>0.2144</td>
<td>0.4478</td>
</tr>
<tr>
<td>Ours</td>
<td>FPR 0.0115</td>
<td>0.0090</td>
<td>0.0182</td>
</tr>
<tr>
<td>(w/o post</td>
<td>FNR 0.0234</td>
<td>0.0481</td>
<td>0.2728</td>
</tr>
<tr>
<td>processing)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 7 Foreground segmentation results on test sequences.
of foreground objects under several conditions, such as dark object against light, global lighting changes, uncovered region, repetitive motions of background and scenery in heavy snow.

References