Supplementary Material:
Minimum Delay Moving Object Detection

Dong Lao and Ganesh Sundaramoorthi
King Abdullah University of Science & Technology (KAUST), Saudi Arabia
{dong.lao, ganesh.sundaramoorthi}@kaust.edu.sa

1. Description of Videos

Videos are available on the following site: https://sites.google.com/kaust.edu.sa/mindelaydetection/home.

1. FAR3.mp4, FAR5.mp4, FAR7.mp4 - Sample results of all methods tested on our new dataset. Purple masks indicate the algorithm's current best guess of the detection (but this is not necessarily the final output of the algorithm, which is at detection time). When the video stops (possibly before the end of the video), the masked region (purple or green) is the algorithm's detection. The mask in green indicates a correct detection by the algorithm (measured against ground truth). Although our dataset consists of 78 videos, only a subset are shown (due to limited space).

   - FAR3.mp4 - The threshold of ours and other methods are chosen to operate at a false alarm rate of 0.3.
   - FAR5.mp4 - The threshold of ours and other methods are chosen to operate at a false alarm rate of 0.5.
   - FAR7.mp4 - The threshold of ours and other methods are chosen to operate at a false alarm rate of 0.7.

   Higher false alarm rate leads to less delay in our method. On average, the theory indicates that our method should have less delay than other approaches operating at a fixed false alarm rate. The videos indicate that that is true on most sequences.

2. LRT.mp4 - Samples of our results shown together with a plot of the likelihood ratio Λ_t. The four windows show the detections for different thresholds b indicated. Larger b means longer delay but at fewer false alarms; small b means smaller delay but more false alarms.

   As in the other sequences above, when the video stops, purple or green masks indicate the algorithm's detection. Purple masks before the video stops are the algorithm's best guess of the detection at the current time, but not the final output of the algorithm. Green indicates a correct detection (verified against ground truth), see the paper for the criterion for correct detection.

2. Details of Derivations

Below, "Eqn X" refers to equation X in the main paper. (X) indicates a reference to equation X in this supplementary material.

- **Derivation of Eqn 5 and 6**

  Using Eqn 3 and Eqn 4, we see that $\eta_t(x) = I_{t+1}(w_{t,t+1}(x)) - I_t(x) \sim \mathcal{N}(0, \sigma_{\eta,i})$. We see that for points $x_1, x_2, \dots$ that are distinct and using the independence of the noise process in space,

$$p(\mathbf{I}_{t:t+1}, x_1, \dots, x_l | v_t^i, O_t^i, R_t^i) = p[\eta, x_1, \dots, x_l | v_t^i, O_t^i, R_t^i] = \prod_l p[\eta_t(x_l) | v_t^i, O_t^i, R_t^i] \tag{1}$$

$$= \prod_l p[I_{t+1}(x_l + v_t^i(x_l)) - I_t(x_l) | O_t^i, R_t^i]. \tag{2}$$

By the statistics of the noise process, we have that

$$p[I_{t+1}(x+v_t^i(x)) - I_t(x)|O_t^i, , R_t^i] \propto \begin{cases} \exp\left(-\frac{1}{2\sigma_{\eta,i}^2}|I_{t+1}(x+v_t^i(x)) - I_t(x_l)|^2\right) & x \in R_t^i \backslash O_t^i \\ \exp\left(-\frac{\beta}{2\sigma_{\eta,i}^2}\right) & x \in O_t^i \end{cases}.$$

Since we choose the occlusion to be regions of high residual, i.e.,

$$O_t^i = \left\{ x \in R_t^i \, : \, \frac{1}{2\sigma_{\eta,i}^2}|I_{t+1}(x+v_t^i(x)) - I_t(x_l)|^2 > \beta \right\}, \tag{3}$$

we have that

$$p[I_{t+1}(x+v_t^i(x)) - I_t(x)|O_t^i, R_t^i] \propto \exp\left\{ \frac{1}{2\sigma_{\eta,i}^2}\rho\left[I_{t+1}(x+v_t^i(x)) - I_t(x)\right] \right\}. \tag{4}$$

where $\rho(y) = \min\{|y|^2, \beta\}$.

Therefore, substituting (4) into (2), we find that

$$p(\mathbf{I}_{t:t+1}, x_1, x_2, \ldots | v_t^i, R_t^i) \propto \exp\left(-\sum_l \frac{1}{2\sigma_{\eta,i}^2}\rho\left[I_{t+1}(x_l + v_t^i(x_l)) - I_t(x_l)\right] \right). \tag{5}$$

Passing to the continuum limit and accounting for the fact that a point x may belong to one of the n regions R_t^i, we arrive at

$$p(\mathbf{I}_{t:t+1}|v_t^i, R_t^i) \propto \exp\left(-\sum_{i=0}^{n-1} \int_{R_t^i} \frac{1}{2\sigma_{\eta,i}^2}\rho\left[I_{t+1}(x+v_t^i(x)) - I_t(x)\right] \, \mathrm{d}x \right). \tag{6}$$

- **Derivation of Eqn 7**

  Using the conditional independence of pairs of adjacent image frames, given the regions and frame-by-frame displacements, we have that

$$p_1(\mathbf{I}_{t_1:t_2}|\mathbf{v}_{t_1:t_2}^i, \mathbf{R}_{t_1:t_2}^i) = \prod_{t=t_1}^{t_2-1} p_1(\mathbf{I}_{t:t+1}|v_t^i, R_t^i). \tag{7}$$

  Substituting, (6) into the above, we have that

$$p_1(\mathbf{I}_{t_1:t_2}|\mathbf{v}_{t_1:t_2}^i, \mathbf{R}_{t_1:t_2}^i) \propto \prod_{t=t_1}^{t_2-1} \exp\left(-\sum_{i=0}^{n-1} \int_{R_t^i} \rho_i \left[I_{t+1}(x+v_t^i(x)) - I_t(x)\right] \, \mathrm{d}x \right) \tag{8}$$

$$= \exp\left(-\sum_{t=t_1}^{t_2-1} \sum_{i=0}^{n-1} \int_{R_t^i} \rho_i \left[I_{t+1}(x+v_t^i(x)) - I_t(x)\right] \, \mathrm{d}x \right). \tag{9}$$

  Now to maximize the quantity above, we may minimize $-\log p_1(\mathbf{I}_{t_1:t_2}|\mathbf{v}_{t_1:t_2}^i, \mathbf{R}_{t_1:t_2}^i)$ since $-\log$ is a monotone decreasing function, as is done in Eqn 7 in the paper.

- **Derivation of Eqn 10**

  Define the quantity in Eqn 7 as Q, we can switch the order of summation

$$Q = \sum_{t=t_1}^{t_2} \sum_{i=0}^{n-1} \int_{R_t^i} \rho_i[I_{t+1}(w_{t,t+1}^i(x)) - I_t(x)] \, \mathrm{d}x = \sum_{i=0}^{n-1} \sum_{t=t_1}^{t_2} \int_{R_t^i} \rho_i[I_{t+1}(w_{t,t+1}^i(x)) - I_t(x)] \, \mathrm{d}x. \tag{10}$$

Given an $s \in \{t_1, \dots, t_2\}$, we can perform a change of variables as $x = w^i_{s,t}(y)$ for the inner integral. We see that $\mathrm{d}x = \det \nabla w^i_{s,t}(y)\,\mathrm{d}y$. Then

$$Q = \sum_{i=0}^{n-1} \sum_{t=t_1}^{t_2} \int_{w_{s,t}(R^i_t)} \rho_i[I_{t+1}(w^i_{t,t+1}(w^i_{s,t}(y))) - I_t(w^i_{s,t}(y))] \det \nabla w^i_{s,t}(y)\,\mathrm{d}y \tag{11}$$

$$= \sum_{i=0}^{n-1} \sum_{t=t_1}^{t_2} \int_{R^i_s} \rho_i[I_{t+1}(w^i_{s,t+1}(y)) - I_t(w^i_{s,t}(y))] \det \nabla w^i_{s,t}(y)\,\mathrm{d}y \tag{12}$$

$$= \sum_{i=0}^{n-1} \int_{R^i_s} f^i(x)\,\mathrm{d}x, \tag{13}$$

where

$$f^i(x) = \sum_{t=t_1}^{t_2} \rho_i[I_{t+1}(w^i_{s,t+1}(x)) - I_t(w^i_{s,t}(x))] \det \nabla w^i_{s,t}(x). \tag{14}$$

Note that we have used the fact that $w^i_{t,t+1} \circ w^i_{s,t} = w^i_{s,t+1}$. Note that these maps denote smooth diffeomorphisms so that the inverse of the map exists and the composition holds. This is because the maps have been extended smoothly onto the entire domain Ω from their initial definition within the regions.

The functional Q does not take into account the fact that the motion signal across multiple frames could be ambiguous (in textureless regions in R^i_s or when all motion residuals are large). This gives rise to the modification seen in Eqn (10), where the motion is unreliable to estimate. This incorporates color histograms and a prior for smoothness. Note that the prior for smoothness used is the standard length regularization of the region boundaries:

$$\sum_i \int_{\partial R^i_s} \mathrm{d}s, \tag{15}$$

where s is the arclength measure and ∂R^i_s is the boundary of the region.

- **Derivation from Eqn 15 to Eqn 16.**

The definition of $\Lambda_{t_c,t}$ is

$$\Lambda_{t_c,t} = \frac{\max_{\mathbf{R}^i_{t_c:t}, \mathbf{v}^i_{t_c:t}} p_1[\mathbf{I}_{t_c:t} | \mathbf{R}^i_{t_c:t}, \mathbf{v}^i_{t_c:t}, i = 0, \dots, n-1]}{\max_{\mathbf{v}^0_{t_c:t}} p_0[\mathbf{I}_{t_c:t} | \mathbf{v}^0_{t_c:t}]}. \tag{16}$$

This is because to approximate the unconditional probabilities, we approximate the marginalization by choosing it to be the maximum value of the probability over the conditioned variables (this is equivalent to choosing a delta function for the prior probabilities of the conditioned variables).

Since $-\log$ is a monotone decreasing function, the maximization problem can be changed to a minimization of the $-\log$ of the quantity being maximized. This leads to

$$-\log \Lambda_{t_c,t} = \min_{\mathbf{R}^i_{t_c:t}, \mathbf{v}^i_{t_c:t}} -\log p_1[\mathbf{I}_{t_c:t} | \mathbf{R}^i_{t_c:t}, \mathbf{v}^i_{t_c:t}, i = 0, \dots, n-1] - \min_{\mathbf{v}^0_{t_c:t}} -\log p_0[\mathbf{I}_{t_c:t} | \mathbf{v}^0_{t_c:t}] \tag{17}$$

Noting that

$$-\log p_0[\mathbf{I}_{t_c:t} | \mathbf{v}^0_{t_c:t}] = \int_\Omega \rho_i\left[I_{t+1}(w^0_{t,t+1}(x)) - I_t(x)\right]\,\mathrm{d}x \tag{18}$$

$$-\log p_1[\mathbf{I}_{t_c:t} | \mathbf{R}^i_{t_c:t}, \mathbf{v}^i_{t_c:t}, i = 0, \dots, n-1] = \sum_{i=0}^{n-1} \int_{R^i_t} \rho_i\left[I_{t+1}(w^i_{t,t+1}(x)) - I_t(x)\right]\,\mathrm{d}x, \tag{19}$$

using (6). Note we have removed the normalization term (to make proper probabilities) for convenience, and this does not affect the optimizer. Using that

$$\mathrm{Res}^i_t(x) = \rho_i[I_{t+1}(w^i_{t,t+1}(x)) - I_t(x)] = \frac{1}{2\sigma^2_{\eta,i}} \rho[I_{t+1}(w^i_{t,t+1}(x)) - I_t(x)], \tag{20}$$

we arrive at Eqn 16 in the paper.

- **Derivation of Eqn 18**

  We apply QD to the signal $r_t = \frac{1}{|\Omega|}\int_\Omega \text{Res}_t^{NL}(x)\,\mathrm{d}x$. We assume that before t_c the distribution is $\mathcal{N}(\mu_0, \sigma)$ and at and after it is $\mathcal{N}(\mu_1, \sigma)$. Since the means are unknown, we estimate them from the data, and by maximizing the likelihood, we arrive at

$$\mu_0 = \hat{\mu}_{1:t_c-1} = \frac{1}{t_c - 1}\sum_{s=1}^{t_c-1} r_s \tag{21}$$

$$\mu_1 = \hat{\mu}_{t_c:t} = \frac{1}{t - t_c - 1}\sum_{s=t_c}^{t} r_s. \tag{22}$$

  QD leads to the following likelihood ratio:

$$\log\frac{p[r_s \sim \mathcal{N}(\mu_1,\sigma), s = t_c, \dots, t]}{p[r_s \sim \mathcal{N}(\mu_0,\sigma), s = t_c, \dots, t]} = \log p[r_s \sim \mathcal{N}(\mu_1,\sigma), s = t_c, \dots, t] - \log p[r_s \sim \mathcal{N}(\mu_0,\sigma), s = t_c, \dots, t] \tag{23}$$

  (after taking $-\log$ of the ratio of the pre- and post-change distributions). The previous expression leads (using independence of the r_s) to

$$\begin{aligned} -\frac{1}{2\sigma^2}\sum_{s=t_c}^{t}(r_s - \mu_1)^2 + \frac{1}{2\sigma^2}\sum_{s=t_c}^{t}(r_s - \mu_0)^2 &= (\mu_1 - \mu_0)\frac{1}{2\sigma^2}\sum_{s=t_c}^{t}(2r_s - \mu_1 - \mu_0) && (24)\\ &= (\mu_1 - \mu_0)\frac{1}{2\sigma^2}(t - t_c - 1)[2\mu_1 - \mu_1 - \mu_0] && (25)\\ &= \frac{1}{2\sigma^2}(t - t_c - 1)(\mu_1 - \mu_0)^2 && (26)\\ &= \frac{1}{2\sigma^2}(t - t_c - 1)(\hat{\mu}_{t_c:t} - \hat{\mu}_{1:t_c-1})^2. && (27) \end{aligned}$$

  Since we only seek to find the maximum value of this with respect to t_c, we can ignore the constant factor at the start. This results in the F-statistic shown in Eqn 18.

3. Extended Discussion

- **Speed Analysis**: We provide some analysis of run-times of various components of our system. Note that these speeds are on our initial un-optimized C++ / MatLab implementation, and many speed-ups can be done, as discussed later.

  We provide a per-frame analysis of the major costs of our algorithm, Algorithm 3 in the paper on a 640×480 spatial resolution video:

  - Line 3: Classic-NL optical flow - 1 min
  - Lines 4-7 (F-statistic computation): less than 1 sec
  - Lines 8 (warp composition and first test): less than 1 sec
  - Line 9 (initialization to motion segmentation, i.e., only the initialization in Algorithm 1): 2 sec (only if condition in Line 8 passes, otherwise 0 sec)
  - Line 9 (full motion segmentation, i.e., full Algorithm 1): 1-2 min, using 12 core parallelization, assuming 50 frames are segmented together (only if condition in Line 8 passes, otherwise 0 sec)
  - Line 10 (likelihood computation): 5 sec

  We make comments on the main bottlenecks:

  - As can be seen above, the main bottleneck is Classic-NL optical flow. We haven't currently used GPUs, but there are GPU versions of similar optical flows for which there are real-time implementations on GPUs [2]. There is also a real-time CPU version with similar results as Classic-NL [1].

- The other bottleneck is the gradient descent for motion segmentation. It should be noted that if only the initialization to Algorithm 1 is used, and no gradient descent is used, then this portion is no longer a bottleneck. The main paper demonstrated that without applying the gradient descent (no refinement) that not much performance degradation in terms of delay and detection is incurred (see Figure 10).

  Thus, the motion segmentation gradient descent can be skipped if one is not interested in a highly precise segmentation, leading to a method that takes on the order of a few seconds per frame.

References

[1] T. Kroeger, R. Timofte, D. Dai, and L. Van Gool. Fast optical flow using dense inverse search. In *European Conference on Computer Vision*, pages 471–488. Springer, 2016. 4

[2] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint Pattern Recognition Symposium*, pages 214–223. Springer, 2007. 4