Motion based Segmentation in videos

Visual motion

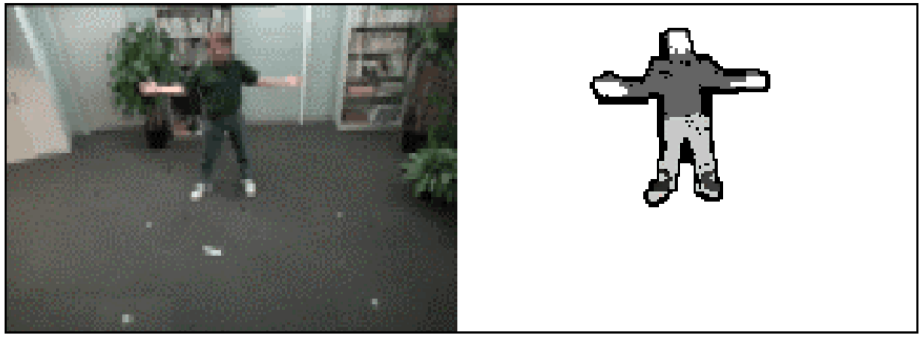
- Powerful cue used by humans to extract objects and regions.
- Motion arises from
  - Objects moving in the scene.
  - Relative displacement between the sensing system and the scene

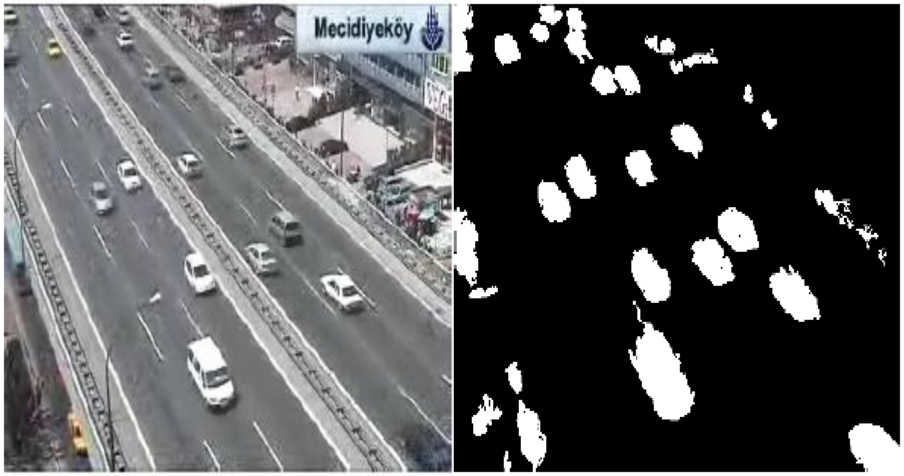
Moving object segmentation

- Visual world is dynamic and we constantly come across video scenes with many moving objects like vehicle, pedestrians etc.
- Automated analysis of such dynamic video scene activities/events require 1) detection 2) classification 3) tracking
- Applications - Video surveillance, Traffic monitoring, Human action recognition, Human-computer interaction etc.

Background Subtraction

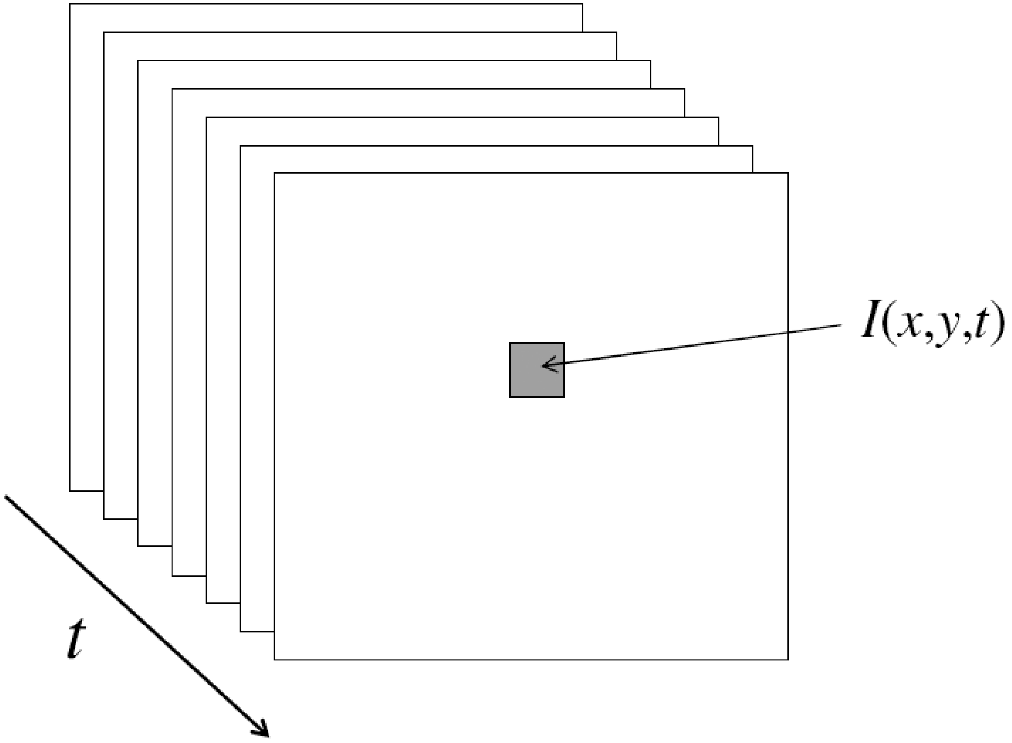
- Given a video frame from a static camera observing a scene, goal is to separate the moving foreground objects from the static *background (part of the scene that does not change position with time)*





Video as an “Image Stack”

- A video is a sequence of frames captured over time
- Now our image data is a function of space (x, y) and time (t)

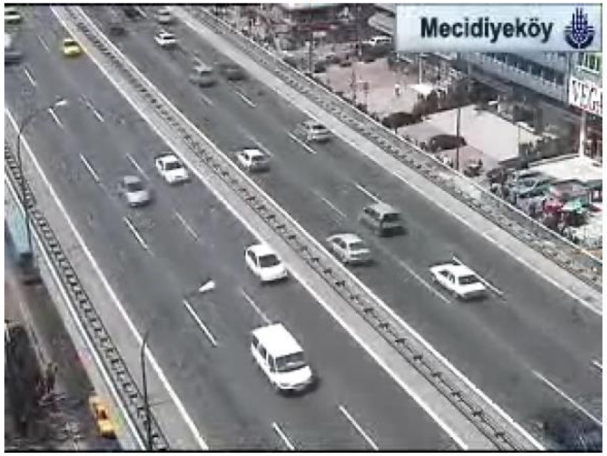


Simple Approach

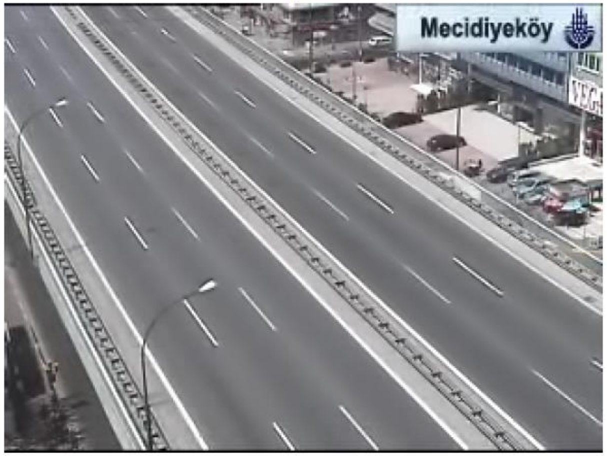
Image at time t:
$I(x, y, t)$
$\Downarrow$

Background at time t:
$B(x, y, t)$
$\Downarrow$

$|$



$-$



$| > Th$

1. Estimate the background for time t.
2. Subtract the estimated background from the input frame.
3. Apply a threshold, Th, to the absolute difference to get the **foreground mask**.

But, how can we estimate the background?

Why Background modeling?

- Several factor cause change in the background like
  - Illumination changes
  - Natural wind motions causing background parts to move
  - Shadows, flicker etc.
  - Moving object itself may halt/become stationary (eg. car gets parked)

Test Video

Frame Differencing

- Background is estimated to be the previous frame. Background subtraction equation then becomes:

$$B(x, y, t) = I(x, y, t-1)$$

$$\Downarrow$$

$$|I(x, y, t) - I(x, y, t-1)| > Th$$

- Depending on the object structure, speed, frame rate and global threshold, this approach may or may **not** be useful (usually **not**).

$|$

$-$

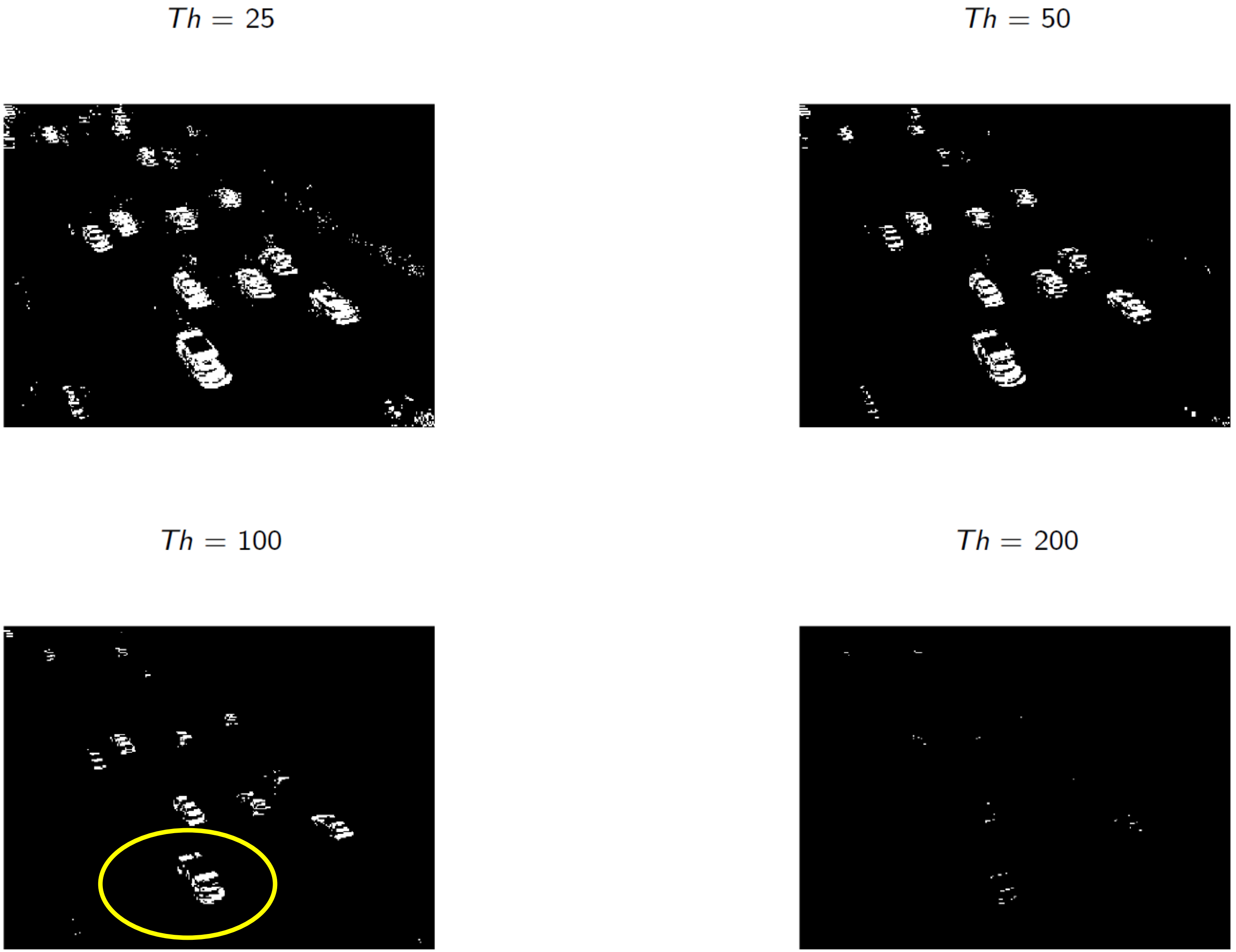
$| > Th$

Result

Frame Differencing

- Major flaws
  - For objects with uniformly distributed intensity values, the interior pixels are interpreted as part of the background.
  - Objects must be continuously moving otherwise it becomes part of the background.
- Two major advantages.
  - modest computational load.
  - background model is highly adaptive
- A challenge with this method is in determining the threshold value.

Frame Differencing



$Th = 25$

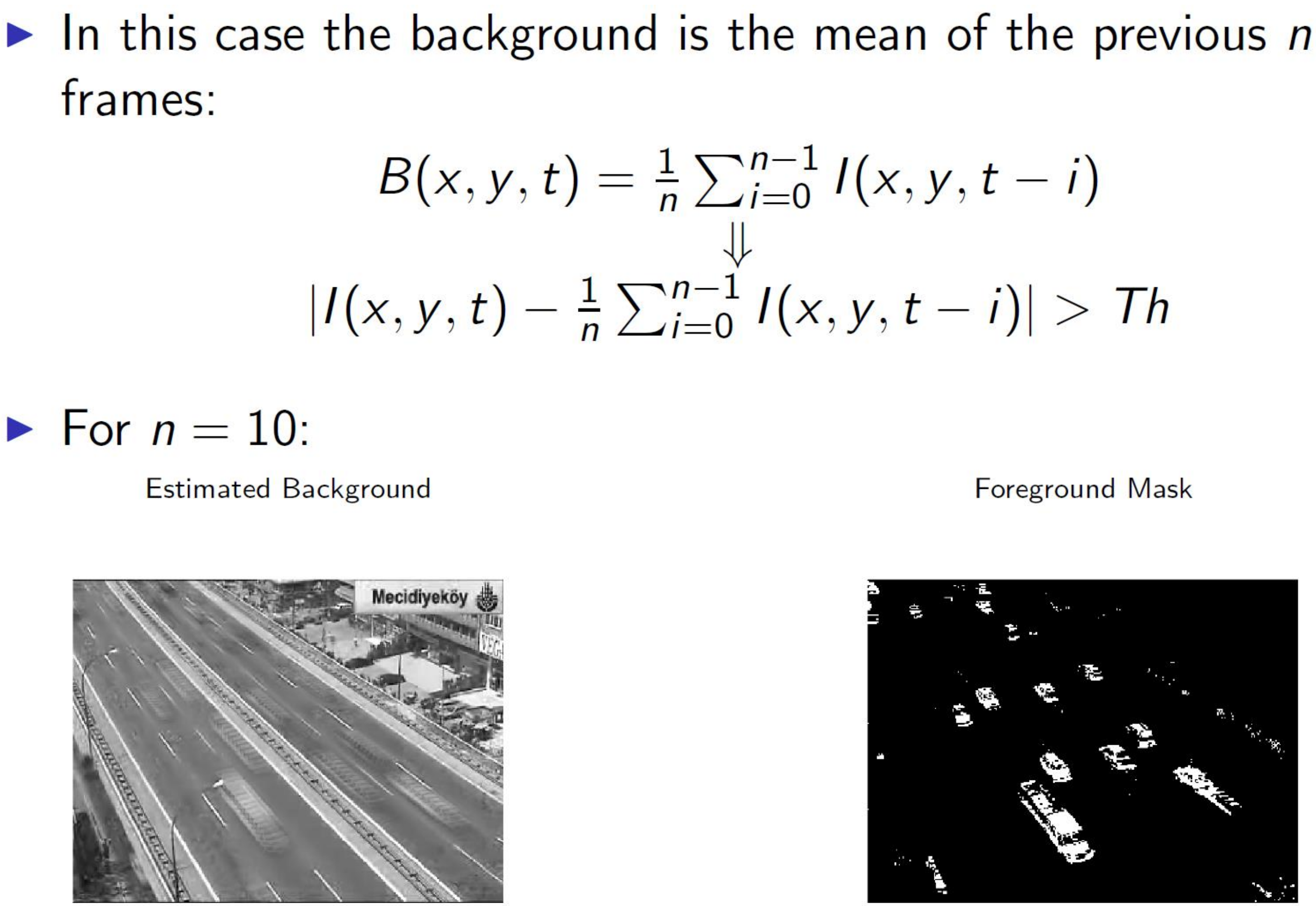
$Th = 50$

$Th = 100$

$Th = 200$

The accuracy of this approach is dependent on object speed, frame rate, threshold value.

Mean Filter



- In this case the background is the mean of the previous n frames:

$$B(x,y,t) = \frac{1}{n}\sum_{i=0}^{n-1} I(x,y,t-i)$$

$$\Downarrow$$

$$\left|I(x,y,t) - \frac{1}{n}\sum_{i=0}^{n-1} I(x,y,t-i)\right| > Th$$

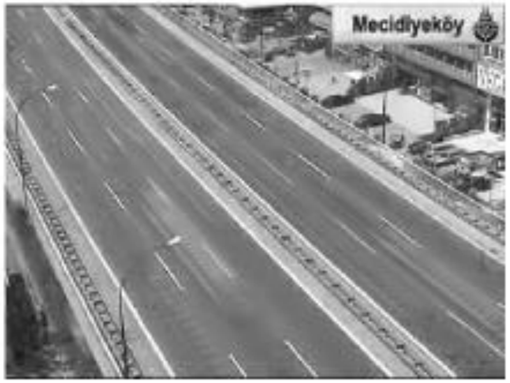
- For $n = 10$:

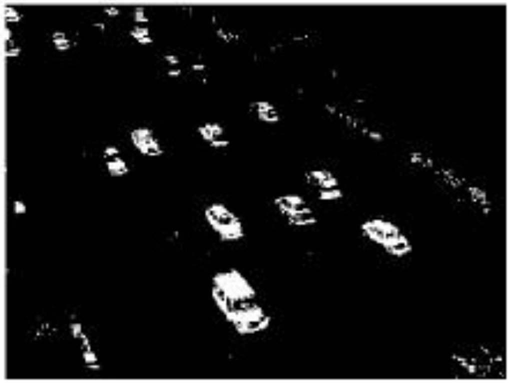
Estimated Background

Foreground Mask

where *N* is the number of preceding images taken for averaging. This averaging refers to averaging corresponding pixels in the given images. *N* would depend on the frame rate and the amount of movement in the video.

Mean Filter





- For $n = 50$:

Estimated Background

Foreground Mask

Median Filter

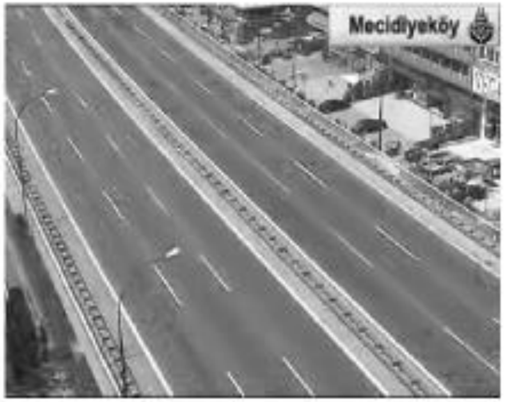
- Assuming that the background is more likely to appear in a scene, we can use the median of the previous n frames as the background model:

$$B(x,y,t) = median\{I(x,y,t-i)\}$$

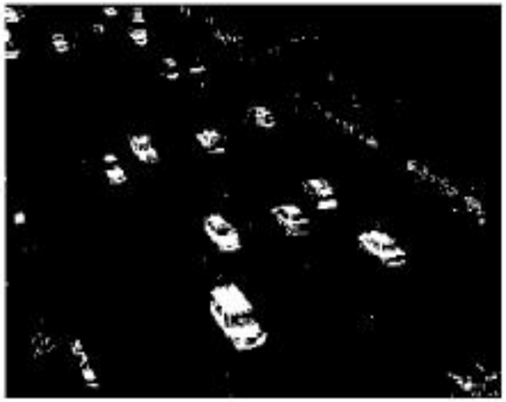
$$\Downarrow$$

$$|I(x,y,t) - median\{I(x,y,t-i)\}| > Th \text{ where}$$
$$i \in \{0,\ldots,n-1\}.$$

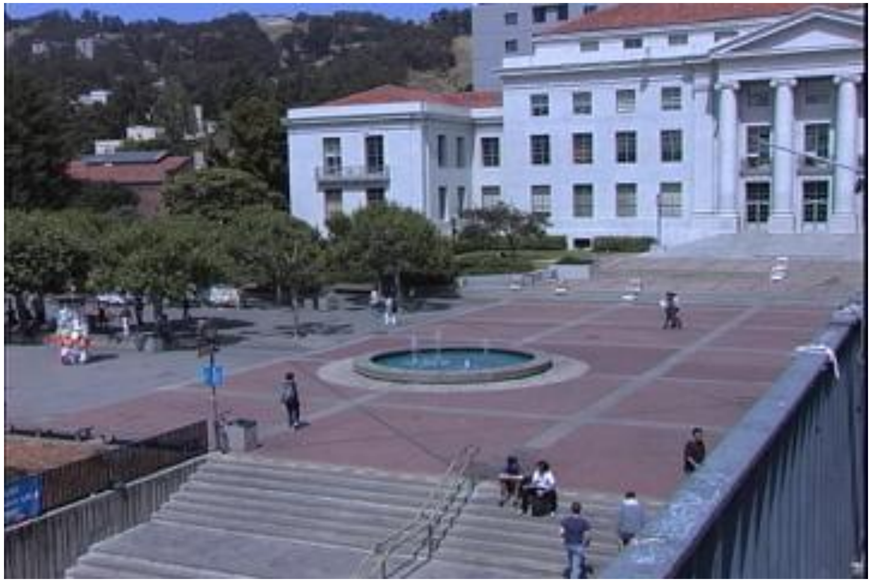
Estimated Background

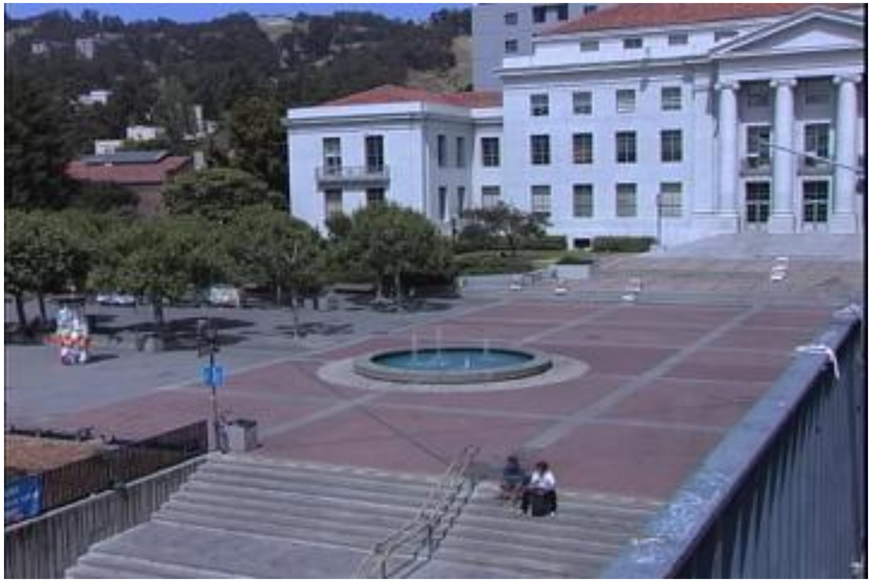


Foreground Mask

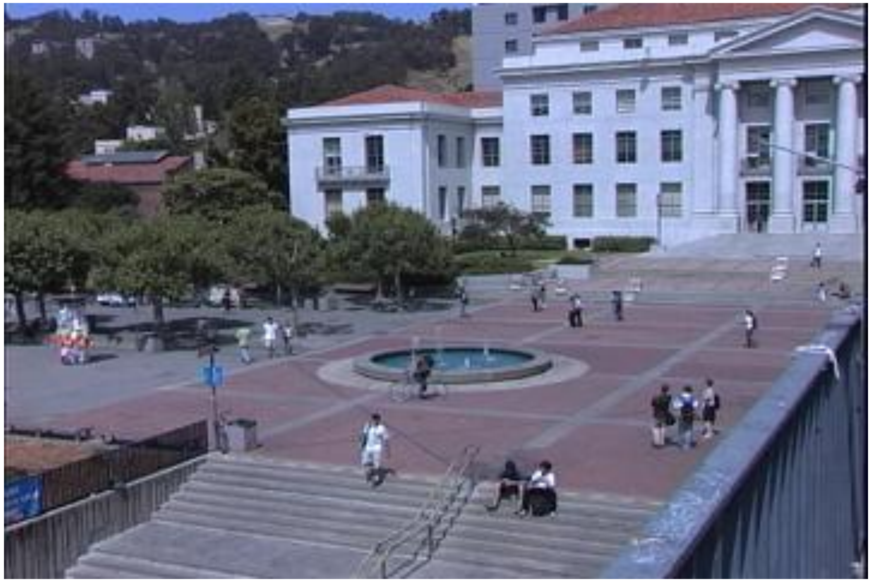


Average/Median Image

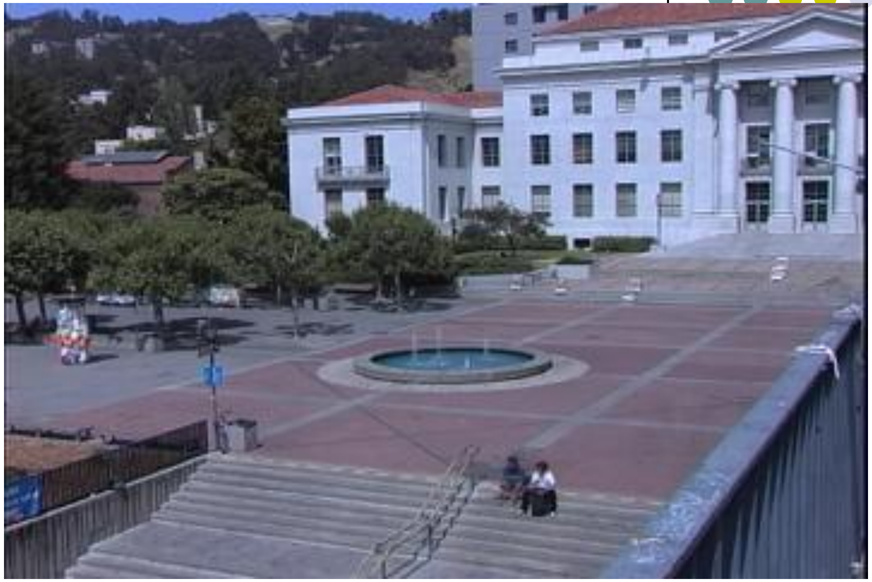




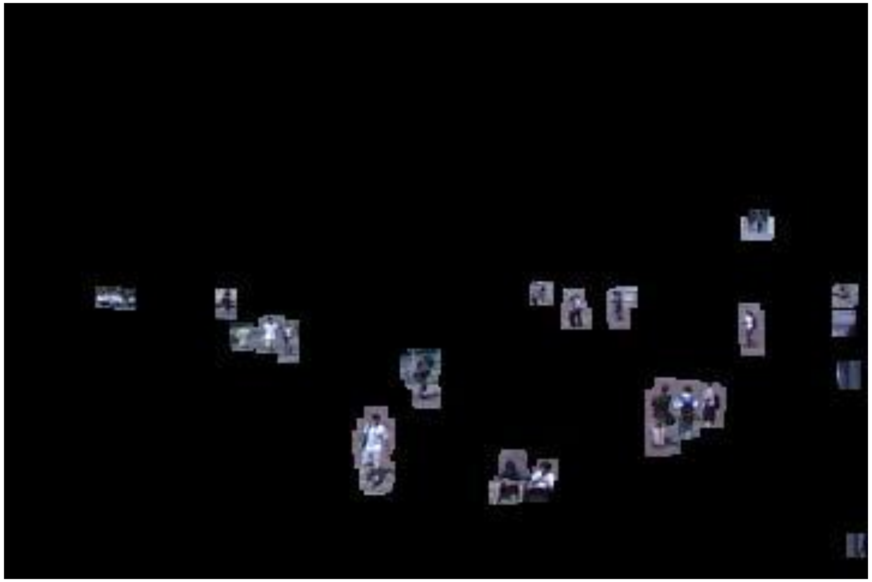
Background Subtraction



-



=



Approximate Median Filter

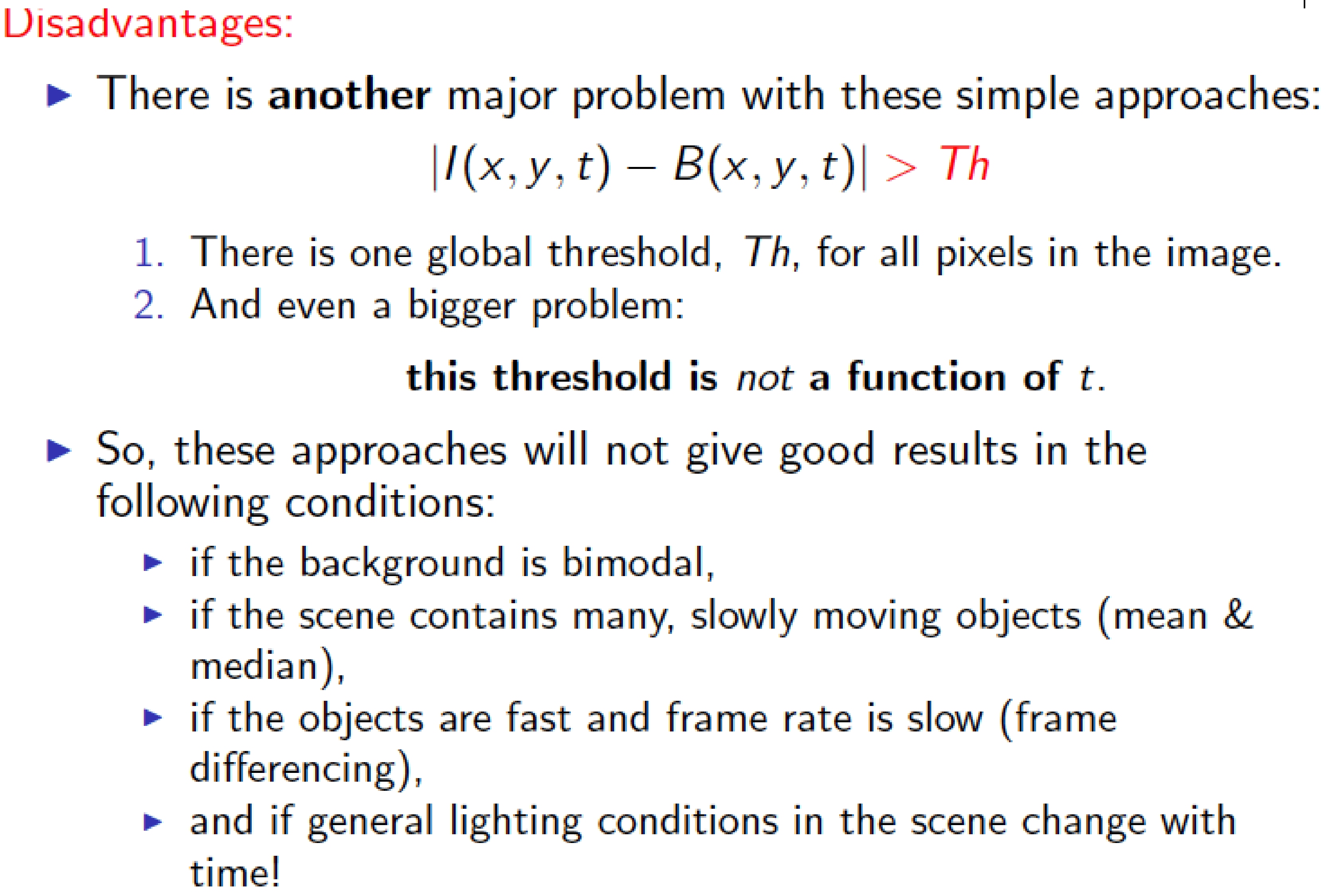
To overcome the large memory and computation requirement:

- If a pixel in the current frame has a value larger than the corresponding background pixel, the background pixel is incremented by 1.
- Likewise, if the current pixel is less than the background pixel, the background is decremented by one.
- The background eventually converges to an estimate where half the input pixels are greater than the background, and half are less than the background—approximately the median (convergence time will vary based on frame rate and amount movement in the scene.)

Background model

Results

Main Shortcoming



Disadvantages:

- There is **another** major problem with these simple approaches:

$$|I(x,y,t) - B(x,y,t)| > Th$$

1. There is one global threshold, Th, for all pixels in the image.
2. And even a bigger problem:

this threshold is *not* a function of t.

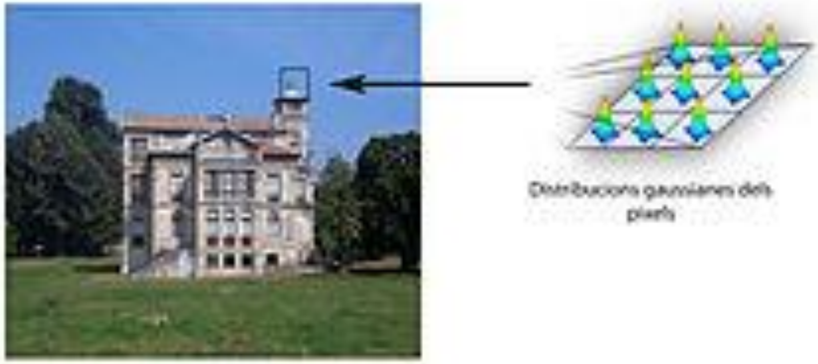
- So, these approaches will not give good results in the following conditions:
  - if the background is bimodal,
  - if the scene contains many, slowly moving objects (mean & median),
  - if the objects are fast and frame rate is slow (frame differencing),
  - and if general lighting conditions in the scene change with time!

Running Gaussian Average

- For every pixel, fit one Gaussian PDF distribution **(μ,σ)** on the most recent ***n*** frames (this gives the background PDF).
- To accommodate for change in background over time (e.g. due to illumination changes or non-static background objects), at every frame, every pixel's mean and variance must be updated.
- background PDF update (running average):

$$\mu_{t+1} = \alpha F_t + (1-\alpha)\mu_t$$

$$\sigma_{t+1}^2 = \alpha(F_t - \mu_t)^2 + (1-\alpha)\sigma_t^2$$



Running Gaussian Average

- Initialization: For variance, we can, for example, use the variance in x and y from a small window around each pixel and take $\mu_0 = I_0$
- We can now classify a pixel as background if its current intensity lies within some confidence interval of its distribution's mean.

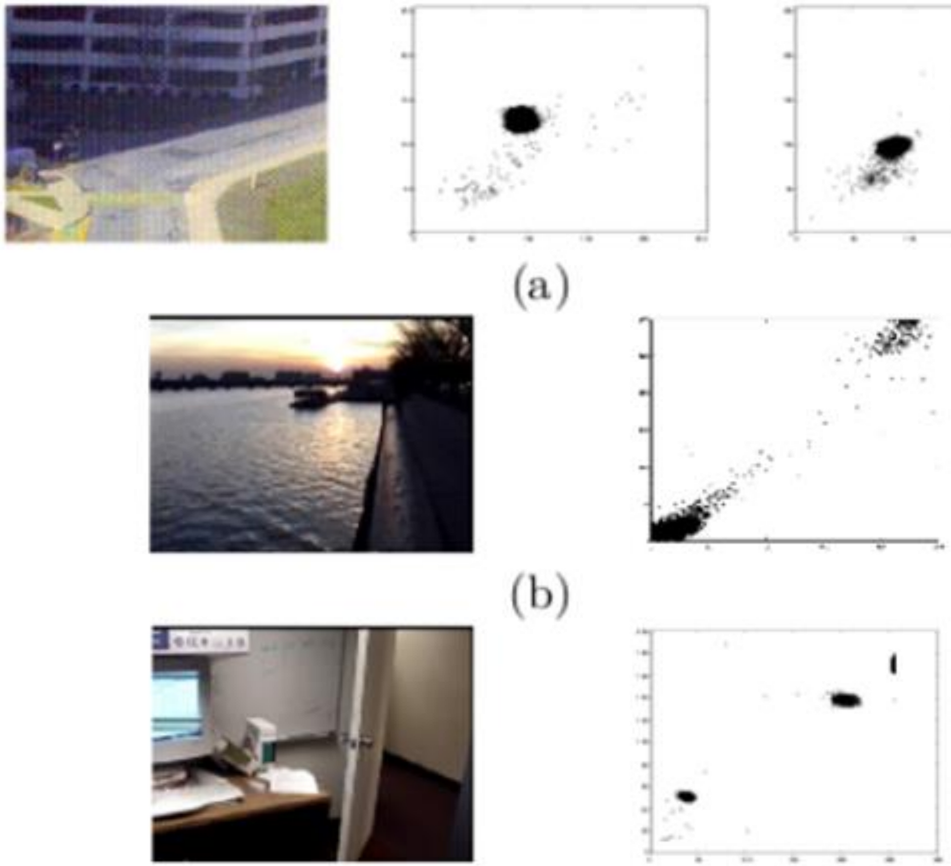
$$\frac{|(I_t - \mu_t)|}{\sigma_t} > k \longrightarrow Foreground$$

$$\frac{|(I_t - \mu_t)|}{\sigma_t} \leq k \longrightarrow Background$$

- It does not however cope with multimodal background.

Gaussian Mixture Model

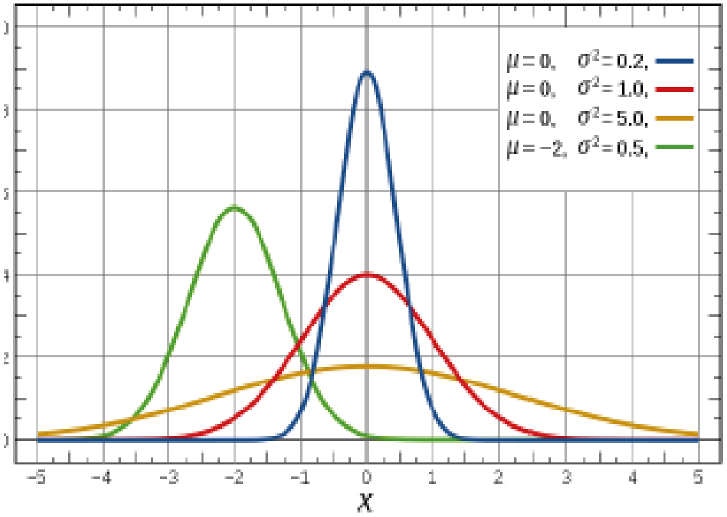
- A background pixel may belong to more than one pattern class each having a specific PDF



(a)

(b)

(c)



Mixture model

- Model the history of a pixel distribution by a weighted combination of K adaptive gaussians to cope with multimodal background distributions;

$$p(x) = \sum_{j=1}^{K} w_j \cdot N(x \mid \mu_j, \Sigma_j) \qquad \sum_{j=1}^{K} w_j = 1 \quad \text{and} \quad 0 \le w_j \le 1$$

- The parameters of Gaussians - weights, mean and variance. How to learn/update the background model?
- The mixture of Gaussians actually models both the foreground and the background: how to pick only the distributions modeling the background?

Model Adaptation

- An on-line K-means approximation is used to update the Gaussians.
- If a new pixel value, X_{t+1}, can be matched to one of the existing Gaussians (within 2.5σ), that Gaussian's $\mu_{i,t+1}$ and $\sigma^2_{i,t+1}$ are updated as follows:

$$\mu_{i,t+1} = (1-\rho)\mu_{i,t} + \rho X_{t+1}$$

and

$$\sigma^2_{i,t+1} = (1-\rho)\sigma^2_{i,t} + \rho(X_{t+1} - \mu_{i,t+1})^2$$

where $\rho = \alpha\mathcal{N}(X_{t+1}|\mu_{i,t},\sigma^2_{i,t})$ and α is a learning rate.

- Prior weights of all Gaussians are adjusted as follows:

$$\omega_{i,t+1} = (1-\alpha)\omega_{i,t} + \alpha(M_{i,t+1})$$

where $M_{i,t+1} = 1$ for the matching Gaussian and $M_{i,t+1} = 0$ for all the others.

Model Adaptation

- The mean and variance for unmatched distributions remain the same.
- If none of the K distributions match the current pixel value, the least probable distribution is discarded and
- A new distribution with the current value as its mean value, an initially high variance, and low prior weight, is added.

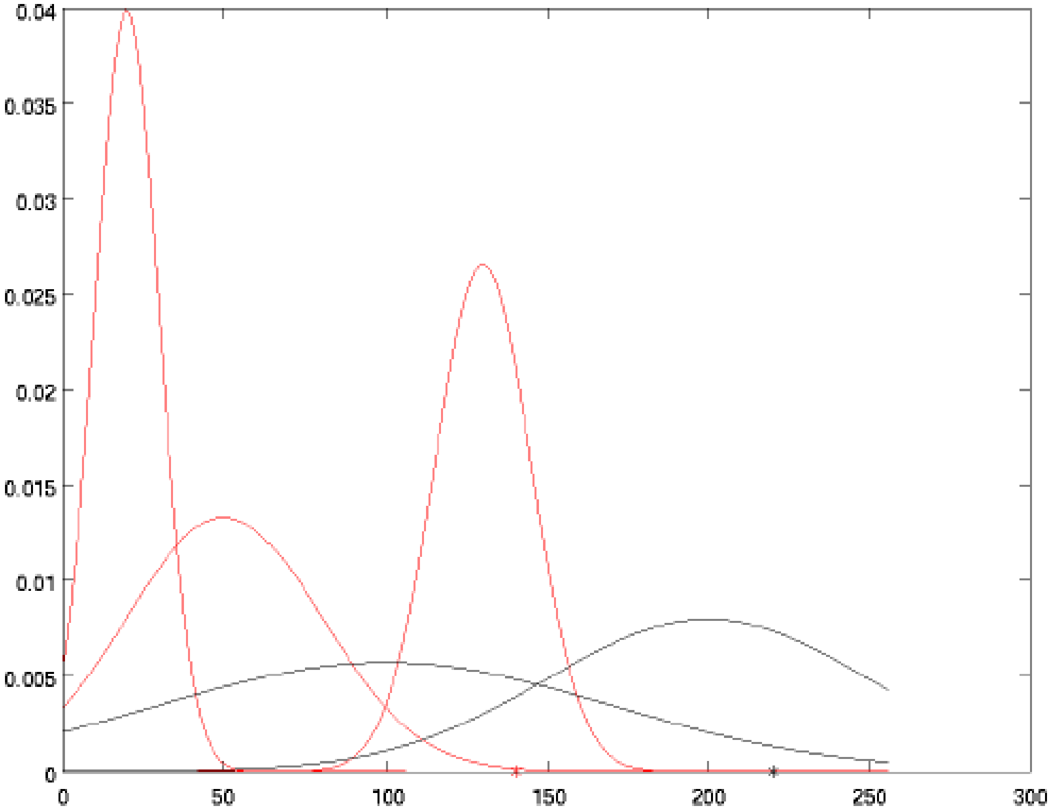
Background Model Estimation

- Heuristic: the Gaussians with the **most supporting evidence** and **least variance** should correspond to the background (Why?).
- The Gaussians are ordered by the value of ω/σ (high support & less variance will give a high value).
- Then simply the first B distributions are chosen as the background model:

$$B = argmin_b(\sum_{i=1}^{b} \omega_i > T)$$

where T is minimum portion of the image which is expected to be background.

Background Model Estimation

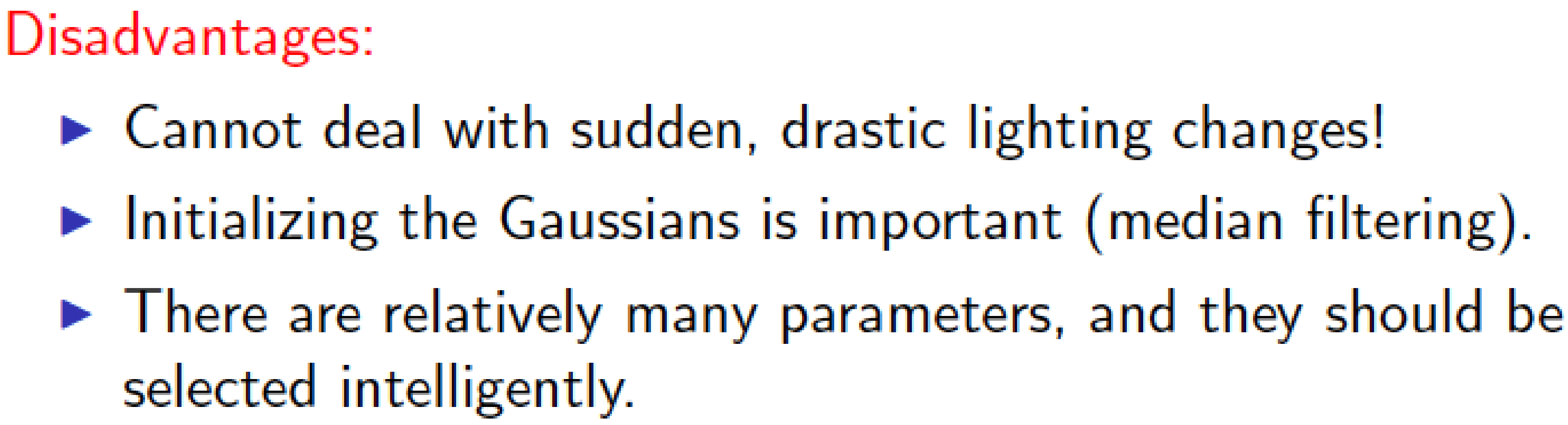


- After background model estimation **red** distributions become the background model and **black** distributions are considered to be foreground.

Advantages Vs Shortcomings

Advantages:

- Same as adaptive Gaussians but now can also cope with multimodal background distributions



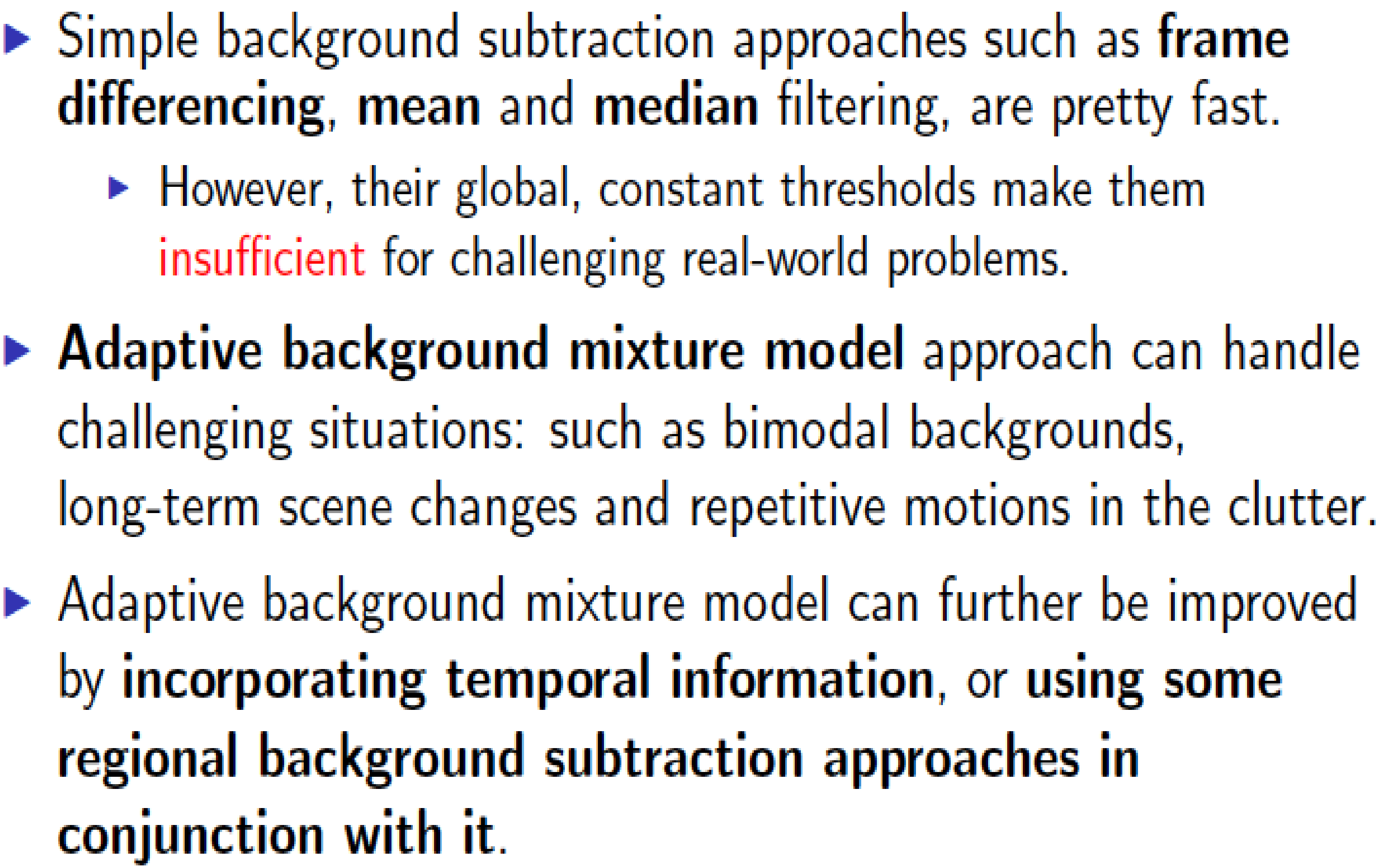
Disadvantages:

- Cannot deal with sudden, drastic lighting changes!
- Initializing the Gaussians is important (median filtering).
- There are relatively many parameters, and they should be selected intelligently.

Background model

Results

Summary



- Simple background subtraction approaches such as **frame differencing**, **mean** and **median** filtering, are pretty fast.
  - However, their global, constant thresholds make them insufficient for challenging real-world problems.
- **Adaptive background mixture model** approach can handle challenging situations: such as bimodal backgrounds, long-term scene changes and repetitive motions in the clutter.
- Adaptive background mixture model can further be improved by **incorporating temporal information**, or **using some regional background subtraction approaches in conjunction with it**.