Dynamic Spatial Predicted Background for Video Surveillance

Yaniv Tocker[1], Rami R. Hagege[2], Joseph M. Francos[1]
[1]Department of Electrical & Computer Engineering, Ben-Gurion University, Be'er-Sheva, Israel [2]Lifetbot: http://lifetobot.com/

Abstract

We propose a novel method for video foreground-background separation that models the scene as a superposition of illumination effects. The model predicts each pixel's value using a linear estimator comprised by a few other pixels of the scene. Our method achieves real-time performance using minimal hardware, which is a crucial consideration for embedding such a system on surveillance cameras.

Introduction

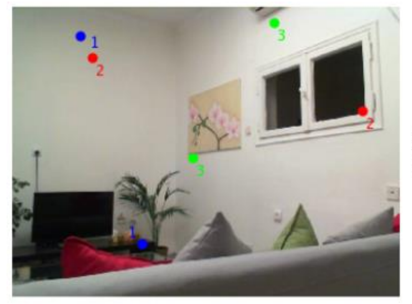
- Goal: Given a video clip taken by a static camera, to separate the objects of interest ("Foreground") and irrelevant information ("Background")
- Basic step in video analysis: recognition & tracking
- E.g.: Intelligent visual surveillance, & Human-Machine interaction (Microsoft's Kinect)
- Main challenges: noisy images, shadows, illumination changes (gradual\sudden), dynamic background and computational load

Video #1 – Foreground/Background separation example

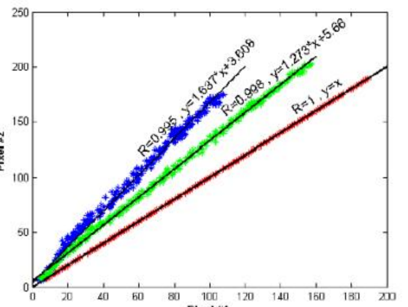
Yaniv Tocker - ytocker@gmail.com

DSPB – Dynamic Spatial Predicted Background

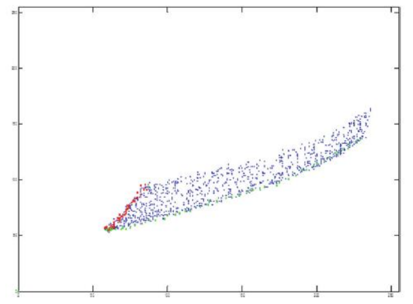
- Physically inspired to handle illumination changes
- Pixel correlations depend on number of light sources:



(a) 3 random pixel pairs



(b) 1 varying light source & Linear fit



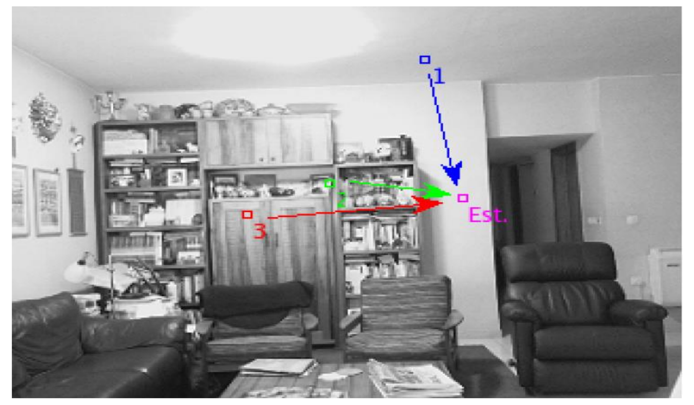
(c) 2 varying lights source

Fig 1. Pixel correlations & light sources

- Pixels can be estimated by knowing their and others history:
- $\{p_k\}_{k=1}^{m}$ - set of randomly chosen pixels, referred as "control pixels"
- $I(p) = M^p \cdot A$ (1) , $I(p) \in \mathbb{R}^1$ - Brightness level of pixel p
  $M^p \in \mathbb{R}^{1\times N}$ - weights of light sources to outcome
  $A \in \mathbb{R}^{N\times 1}$ - Light source powers
- $I_{\{p_k\}_{k=1}^{m}} = M \cdot A,\ I_{\{p_k\}_{k=1}^{N}} \Rightarrow A = M^{-1} I_{\{p_k\}_{k=1}^{m}}$ (2)
  (2)->(1): $I(p) = M^p \cdot A = M^p M^{-1} I_{\{p_k\}_{k=1}^{N}} \Rightarrow I(p) = T \cdot I_{\{p_k\}_{k=1}^{m}}$ (3)
- $\Rightarrow I_t = T^* \cdot I_{\{p_k\}_{k=1}^{m}}$ (4)
- Problem: inefficient to estimate T^*, so force linear solution
- Solution: Optimal linear estimator:

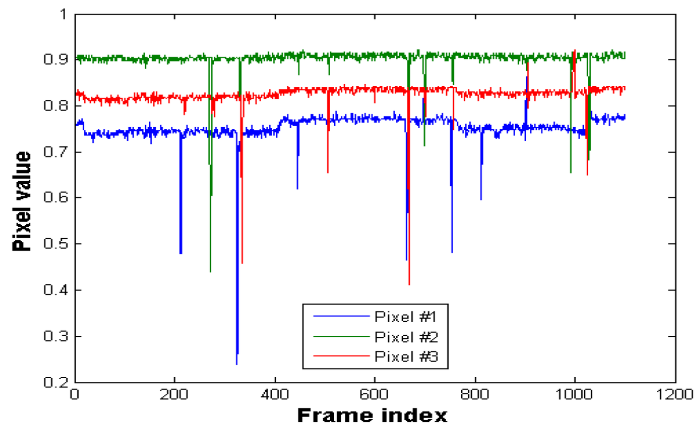
$$I_t = \mathbb{E}_{I_t} + Cov_{I_t P} \cdot (Cov_P)^{-1} \cdot (P_t - \mathbb{E}_P) \qquad P = I_{\{p_k\}_{k=1}^{m}}$$

- Means & correlations calculated empirically (using past frames)
- Each pixel is estimated by a set of **5 pixels**
- **Background Initialization:** From the 5th frame!
- **Background Maintenance:** updating the model parameter as the video continues: $\mathbb{E}_x^{i+1} = \alpha \mathbb{E}_x^i + (1-\alpha)\mathrm{x}^i$
- **Foreground Detection:** $FG(x,y,t) = |I(x,y,t) - BG(x,y,t)| > 3 \cdot \sigma(x,y,t)$
- **Problem #1:** Control pixels can be occluded or noised
- Solution: 3 estimators instead of 1 -> 3 candidates for each pixel -> take median
- Problem #2: Correlations are more dominant to a small surrounding area
- Solution: - k-means on median image using first 5 frames
  - Estimation done separately for each cluster
  - BG image obtain by a mosaic of the sub-areas

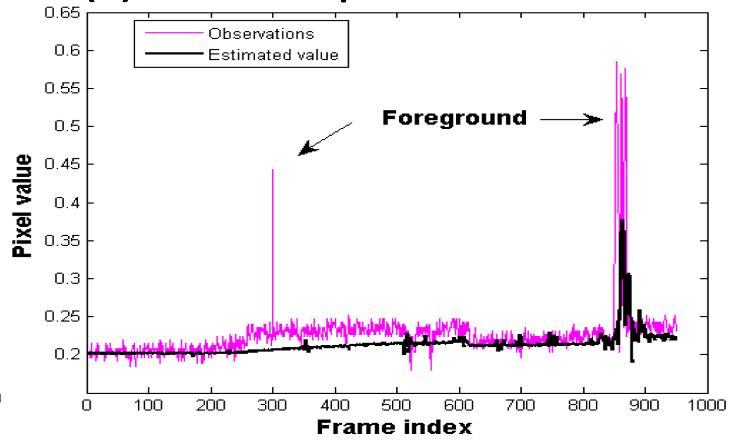


(a) Control pixels estimates another

(b) Estimated pixel is occluded



(c) Control pixels values



(d) Estimated pixel vs. observed

Fig 2. Estimator comprised of 3 pixels

Fig 3. "Light switch" video from LIMU dataset, Top – Frames, Middle Background estimation, Bottom – Foreground mask

Fig 4. "Wallflower" Ground truth frames with foreground masks of tested methods

Evaluation & Results

- **Evaluation metrics: Precision:** $\mathrm{Pr} = \frac{TP}{TP+FP}$, **Recall:** $Re = \frac{TP}{TP+FN}$, **Specificity:** $Sp = \frac{TN}{TN+FP}$, **F-measure** $= \frac{2PrRe}{Pr+Re}$, **Frames per second:** fps
- **Tested Methods**: 1) GMM zivkovic et al. [1] 2) ViBe [2], 3) RPCA - ALM [3]
- **LIMU Dataset:** 5 video clips.
  5000 frames with 240×320 resolution
  Ground truth each 15th frame from frame 500
- **Wallflower Dataset:** 7 short videos, unique challenges

Method	zGMM	ViBe	RPCA	DSPB
Pr	0.5466	0.5838	0.7663	0.8263
Re	0.5534	0.3555	0.6836	0.4368
Sp	0.9262	0.9426	0.9959	0.9984
F	**0.4380**	**0.2819**	**0.7012**	**0.5335**
fps	137.29	209.76	0.55	278.25

Table 1. LIMU results

Method	zGMM	ViBe	RPCA	DSPB
Pr	0.6230	0.7722	0.7115	0.6981
Re	0.5661	0.4051	0.5537	0.4293
Sp	0.7827	0.9680	0.9271	0.9122
F	**0.4891**	**0.4813**	**0.5903**	**0.5035**
fps	352.06	466.74	3 .64	209.96

Table 2. Wallflower results

Conclusions

- A novel Hybrid FG-BG separation system
- Involves spatial information (correlations) combined with pixel temporal statistics
- Physically inspired to deal with illumination changes – Gradual or Sudden
- Simple method with low computational requirements – performs in real-time

References

1. Zoran Zivkovic, "Vision modules for a multi sensory bridge monitoring approach," in Proceedings of the 17th International Conference on Pattern Recognition, ICPR, pp. 1051–1054. 2004
2. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," IEEE Transactions on Image Processing, vol. 20, no. 6, pp. 1709–1724, 2011.
3. D Goldfarb, S Ma, and K Scheinberg, "Fast alternating linearization methods for minimizing the sum of two convex function," Math. Program. Ser. A, 2010.