A Dual Estimation Approach for Removing the Show-Through Effect in the Scanned Documents

Sabita Langkam and Alok Kanti Deb

Department of Electrical Engineering, Indian Institute of Technology Kharagpur, India

Introduction

The digital scans of double sided documents often suffer from distortions because the contents on the back side (verso) of the document often shows up on the front side (recto) in the scans and vice-versa either due to transparency of the paper or due to ink-bleeding. Such contamination in the scans of duplex printed documents called show-through effect is not trivial and this unwanted artifact needs to be eliminated. If the texts and images on the documents are not pure black and white but have varied gray levels, the problem calls for a rigorous show-through cancellation methods.

Approach

- Degraded scans are assumed to be linear and instantaneous mixing of the contents on the two sides of the original document.
- The show-through effect is given a state-space representation.
- A Dual estimation approach is proposed to cancel the show-through effect.

Mathematical Section

The scans of the recto and verso pages are assumed to be linear combination of clean images representing the sides of the duplex printed document. The show-through effect is thus modeled as

$$\begin{aligned} \mathbf{z}_k &= \mathbf{A}\mathbf{x}_k \\ \begin{bmatrix} z_k^1 \\ z_k^2 \end{bmatrix} &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_k^1 \\ x_k^2 \end{bmatrix} \end{aligned} \quad (1)$$

z_k^1: scanned image of one side of the double sided document,
z_k^2: scanned image of other side of the same document,
$\mathbf{A} \in \Re^{2\times 2}$: mixing matrix,
x_k^1: recto of the document,
x_k^2: verso of the document

The coefficients of matrix $\mathbf{A}$ are unknown. z_k^1 and z_k^2 are data, and x_k^1 and x_k^2 are sources to be estimated from the data when the mixing coefficients are unknown.
The state-space formulation of (1) can be written as

$$\begin{aligned} \mathbf{x}_k &= \mathbf{F}\mathbf{x}_{k-1} + \mathbf{p}_{k-1} \\ \mathbf{z}_k &= \mathbf{H}\mathbf{x}_k + \mathbf{m}_k \end{aligned} \quad (2)$$

The image pixels have been assumed to have temporal correlations which can be modeled as first-order autoregression.
State-space matrices $\mathbf{F}$ and $\mathbf{H}$ are unknown.

$$\mathbf{F} = \mathrm{diag}\,(\mathbf{f}_{11}, \mathbf{f}_{22}), \quad \mathbf{H} \in \Re^{2\times 2}$$

The process noise $\mathbf{p}_{k-1}$ and observation noise $\mathbf{m}_k$ are assumed additive, white and Gaussian.

$$\mathbf{p}_{k-1} \sim N(\mathbf{0}, \mathbf{Q}), \mathbf{m}_k \sim N(\mathbf{0}, \mathbf{R})$$

Dual Estimation

Both parameters and hidden states must be simultaneously estimated from only the observed data.
This work applies dual Kalman approach in which two Kalman filters run concurrently and generate state and parameter estimates.

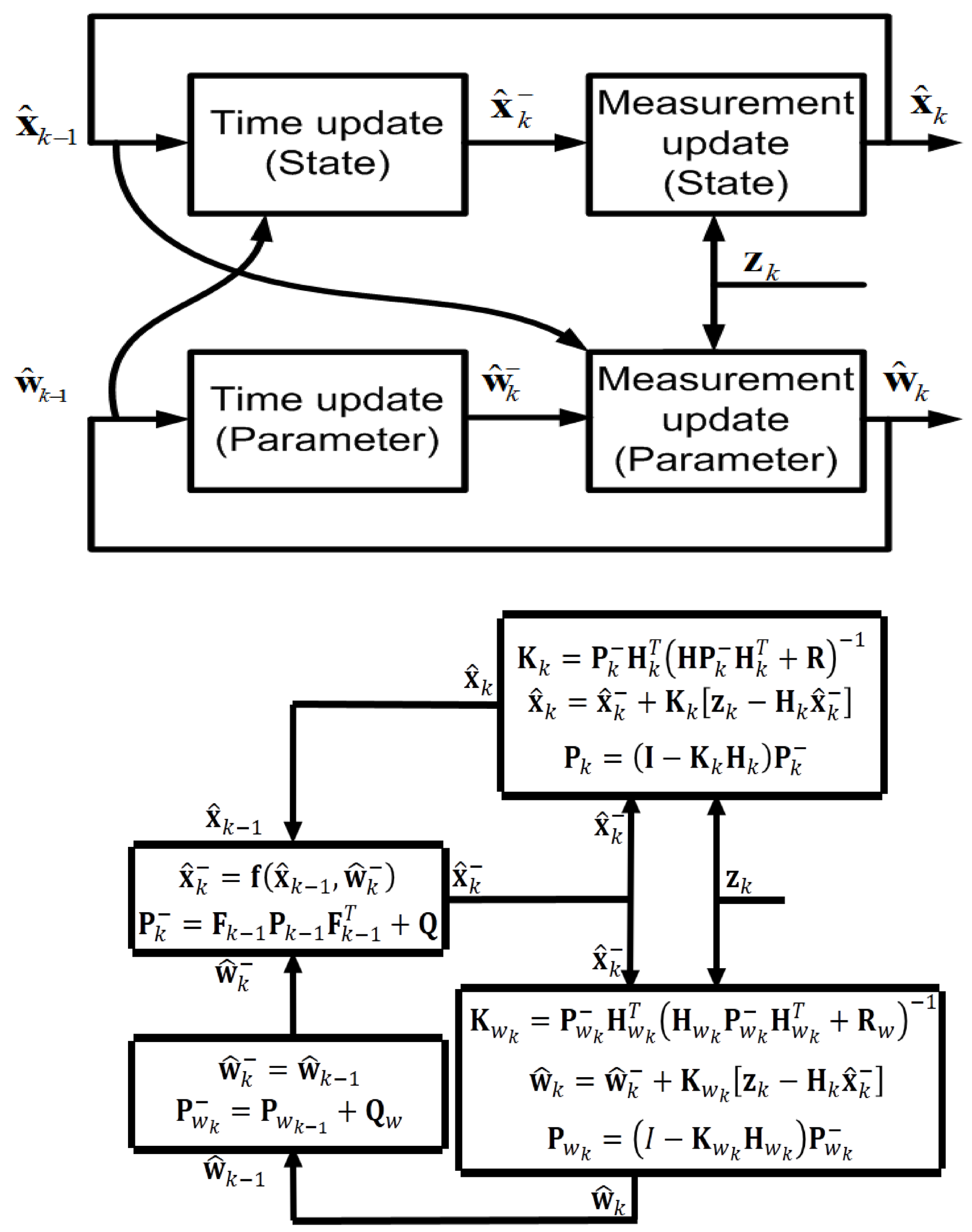

Figure 1: Dual Kalman

$\hat{\mathbf{x}}_k^-$: a priori state estimate;
$\hat{\mathbf{x}}_k$: a posteriori state estimate;
$\hat{\mathbf{w}}_k^-$: a priori parameter estimate;
$\hat{\mathbf{w}}_k$: a posteriori parameter estimate;
$\mathbf{P}_k^-$: a priori state error covariance;
$\mathbf{P}_k$: a posteriori state error covariance;
$\mathbf{K}$: Kalman Gain;
$\mathbf{Q}$: Covariance matrix of the process noise;
$\mathbf{R}$: Covariance matrix of the measurement noise;
$\mathbf{P}_{w_k}^-$: a priori state error covariance;
$\mathbf{P}_{w_k}$: a posteriori state error covariance;
$\mathbf{K}_{w_k}$: Kalman Gain;
$\mathbf{Q}_w$: Covariance matrix of the process noise;
$\mathbf{R}_w$: Covariance matrix of the measurement noise;

Results

Figure 2: First quadrant: Scanned images ([4]) (top) and clean images (bottom); Second quadrant: Scanned images ([3]) (top) and clean images (bottom); Third quadrant: Mixture images (top) and clean images (bottom); Fourth quadrant: Mixture images (top) and clean images (bottom)

Performance Analysis

The performance comparison in terms of mutual information (though values do not change significantly, it does indicate towards achieving visually improved images):

Table 1: Similarity measure: Mutual Information

Mutual Info.	I	II	III
Literature	0.2889 ([3])	0.2340 ([4])	0.3574 (FastICA)
Proposed	**0.2151**	**0.0615**	**0.3440**

Conclusions

- Linear state-space systems effectively represent show-through.
- Dual estimation approach successfully extracts clean images out of contaminated scans of double sided documents.
- Two Kalman filters running concurrently estimate clear images and thus removes show-through.
- An intelligent choice of noise parameters must be made and proper initialization is required.

References

[1] Sabita Langkam and Alok Kanti Deb. A Dual Estimation Approach to Blind Source Separation. *IET Signal Processing* , January, 2017. DOI: 10.1049/iet-spr.2016.0357

[2] Sabita Langkam and Alok Kanti Deb. Linear blind source separation: A dual state-parameter estimation approach. *39th National Systems Conference (NSC)* , 1–5, 2015. DOI: 10.1109/NATSYS.2015.7489092

[3] Qingju Liu and Wenwu Wang. Show-through removal for scanned images using non-linear NMF with adaptive smoothing. *IEEE China Summit & Int. Conf. Signal and Information Process. (ChinaSIP)*, 650–654, 2013. DOI: 10.1109/ChinaSIP.2013.6625422

[4] Boaz Ophir and David Malah. Show-through cancellation in scanned images using blind source separation techniques. *IEEE Int. Conf. Image Process. (ICIP)*, III-233–III-236, 2007. DOI: 10.1109/ICIP.2007.4379289

ICASSP 2017
42nd IEEE International Conference on Acoustics, Speech and Signal Processing
MARCH 5-9, 2017, NEW ORLEANS, USA
PAPER ID # 2199

Thanks to