Mobile Video Capture of Multi-Page Documents

Jayant Kumar (UMD College Park)
Raja Bala, Hengzhou Ding, Phillip Emmett (Xerox Research Center Webster)

International Workshop on Mobile Vision
June 23, 2013
Motivation

- Traditional document imaging involves “one page at a time” capture
- User has to carefully position and capture each frame
- Time consuming and laborious for multi-page documents
- document scan Apps: Camscanner, MDScan, JotNot, apps from banks

booklets  menus  magazines  ID cards
Proposed Approach

- Single video capture session captures all pages/artifacts
- User concentrates only on framing and alignment
- Requires that whole artifact/document is visible and motionless for short duration
- Method automatically selects high-quality frames
  - HD capture in modern smartphones sufficient for high OCR accuracy

*Video capture using Samsung Galaxy S3, 8 MB picture resolution*
The Algorithm...

1. Mobile video capture
2. Temporal and spatial subsampling
3. Page turn detection (frame-difference, morphological operations)
4. Hand movement detection (local luminance profile)
5. Camera motion detection (using inertial sensors)
6. Image quality score (SVM trained using OCR data)
7. Page selection for output PDF
Page Turn Detection

- **Adjacent frame differencing**
- **page turn**
- **quiet period**

- **Frame difference followed by Gaussian smoothing**
  - yields large blobs corresponding large-scale motion during page turn event
- **Morphological erosion using a circular structuring element**
  - to remove fine-scale motion arising from shake and/or jitter
- **Integration over a few frames (K)**
  - to accumulate enough evidence for continuous motion
Detection of Hand Presence & Movement

- search for frames occurring just prior to page-turn that exhibit significant luminance change with respect to a reference frame
Camera Motion Detection

- iPhone 4S emits acceleration data calibrated for gravity (G)
- Record magnitude of 3D acceleration vector for each video frame
- Retain frames with magnitude $\leq 0.02G$

Example:

2nd frame (high motion) 11th frame (low motion)
Image Quality Prediction

- Machine learning technique trained to predict OCR accuracy of document images

Unsupervised feature learning + SVM classifier

Quality Score Prediction

- Low
- High

OCR 15%
OCR 5%
OCR 75%
OCR 95%
Unsupervised Feature Learning Framework

- **CORNIA Algorithm (Ye et al., CVPR 2012)** comprising 3 steps:
  - Codebook construction
  - Feature Encoding & Pooling
  - Regression


Codebook Construction

\[ C = \{C_1, C_2, ..., C_N\} \]

- Patch Selector
- Appearance descriptor (raw image patch)
- A set of local features
  
  Feature extraction

- Feature Extraction
- Clustering
- Codebook
Encoding

- Soft-assignment encoding

\[ C = \{C_1, C_2, ..., C_N\} \]

\[ f_{d \times 1} \rightarrow \text{Find K-nearest neighbors} \]

\[ s_i = \text{similarity}(f, C_j) \quad C_j: j\text{-th nearest codeword of } f \]

\[ \begin{bmatrix} 0 \\ s_1 \\ 0 \\ s_2 \\ 0 \\ \vdots \\ s_K \end{bmatrix} \]

K non-zero elements
Pooling

M local feature vectors

Average Pooling

Max Pooling

\[ X = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{M1} \\ x_{12} & x_{22} & \cdots & x_{M2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1N} & x_{2N} & \cdots & x_{MN} \end{bmatrix} \]

\[ \tilde{X} = \begin{bmatrix} \text{mean}(x_{11} \ldots x_{M1}) \\ \text{mean}(x_{21} \ldots x_{M2}) \\ \vdots \\ \text{mean}(x_{1N} \ldots x_{MN}) \end{bmatrix}_{M \times 1} \]

\[ \tilde{X} = \begin{bmatrix} \max(x_{11} \ldots x_{M1}) \\ \max(x_{21} \ldots x_{M2}) \\ \vdots \\ \max(x_{1N} \ldots x_{MN}) \end{bmatrix}_{M \times 1} \]
Linear Support Vector Machine

Classification into high-quality and low-quality
- does not provide a way to select image based on score

Alternative: Linear Regression

\[
\min \left( C \sum_{n=1}^{N} E_{\varepsilon}(y(x_n) - t_n) + \frac{1}{2} \|w\|^2 \right)
\]

Bimodal distribution of OCR accuracy
(obtained using ABBYY Fine reader)
Extensions

SVM trained with high quality (>= 90%) & low quality (<= 10%)

(SVM margin)

\[ q_i = W^T X_i + b \]

\[ q_j = W^T X_j + b \]

\[ q_k = W^T X_k + b \]

\[ \ldots \]

\[ \ldots \]

\[ \ldots \]

Final quality score Q: \( \text{mean}(q_i) \)
Image Quality Prediction

True score

Predicted score

$f(t) = \frac{100}{1 + \exp\{-at - b\}}$

a and b are obtained via least-squares
Experiments

- **Training data**
  - smartphone video footage of multi-page text-intensive documents with different font sizes and styles
  - a total of 4800 video frames, 1443 randomly selected for training
  - variety of representative lighting conditions and distortions such as shake, translation, motion, glare, shadow

- **Parameters**
  - Patch-size $M = 11$, Codebook size $K = 50$,
    No. of patch descriptors $= 1000$

- **Test Data**
  - 5 printed artifacts presented to 10 subjects in random order:

<table>
<thead>
<tr>
<th>Artifact</th>
</tr>
</thead>
<tbody>
<tr>
<td>a duplex stapled 6-page conference paper</td>
</tr>
<tr>
<td>a simplex unstapled 4-page press release</td>
</tr>
<tr>
<td>a duplex 5-page brochure</td>
</tr>
<tr>
<td>an A3-sized bank application form (front/back)</td>
</tr>
<tr>
<td>a bank check (front/back)</td>
</tr>
</tbody>
</table>
Still Vs. Video Output

- **User1**
  - still
  - video

- **User2**
  - still
  - video

**Note:** auto-crop and auto-enhancement can be applied to all selected frames (not shown)
User Studies

Average Capture Time

- Average Capture Time
- Preference for standard vs. proposed capture experience

- Duplex stapled
- Simplex unstapled
- Bank Check
- Bank form
- Brochure

still
video
User Studies

- Preference for Still vs. Video capture

![Bar Chart]

- Duplex stapled
- Simplex unstapled
- Bank Check
- Bank form
- Brochure

still
video
Quantitative Evaluation

- Ten subjects each captured a 4 page text document using still vs. video mode.
- ABBYY FineReader was used to perform OCR.

\[
\text{Accuracy} = \frac{\text{# correction}}{\text{# total characters}} \cdot s
\]

<table>
<thead>
<tr>
<th>OCR Accuracy</th>
<th>75th Percentile</th>
<th>Median</th>
<th>25th Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Still Capture</td>
<td>99.9</td>
<td>99.6</td>
<td>98.8</td>
</tr>
<tr>
<td>Video Capture</td>
<td>99.8</td>
<td>98.8</td>
<td>87.1</td>
</tr>
</tbody>
</table>

*Time performance:* approximately equal to the duration of video.
Video Demo
Conclusion

- Video HD resolution in today’s smartphones can provide adequate text quality
- We presented a mobile app for capturing and extracting high quality page images from video capture
- Experiments show that video capture offers a superior experience while posing minimal sacrifice to image quality
- Method is applicable for any printed material such as checks, business cards, credit cards, packages, etc.
- Future work includes
  - Algorithm speedup, including GPU implementation
  - Utilizing spatial IQ measure to fuse multiple video frames
  - Incorporation of real-time feedback to guide video capture