Introduction

- Videos of presentations are employed in a large variety of systems for different purposes
  - Distance or E-learning
  - Automatic generation of conference proceedings
  - Student presentations
- Challenges
  - Many videos are already archived
    - Lack of additional sources of information (e.g. electronic copies of slides)
  - Low quality
    - Unconstrained camera movements
    - Slides Clipped
    - Compression
  - Lack of Structure
    - Not recorded by professional cameramen
    - Shots cannot be used as clue
    - Not edited
- Result: Fully automatic method for summarizing and indexing unstructured presentation videos based on text extracted from the projected slides
- Integration into summarization and presentation tools such as the VAST MultiMedia Browser (see image above)

Proposed Approach

1. Segment video into semantically distinct shots based on slides
   - Studies have been conducted in order to assess the reliability of slides as a summarization tool [He et al. 2000]
   - No electronic copies of the slides
   - Changes in text used to assess slide changes

2. Slides Text Detection
   - Geometric Constraints
     - $R_{min} \leq R_{edge} \leq R_{max} \leq 10000$
     - $2 \leq R_{valid} \leq R_{valid} \leq 3$
     - $6 \leq R_{edge} \leq R_{edge} \leq 5$
   - Edges Connected Components
   - Alignment Regions Merge Constraints
     - $merge(R_i, R_j) \iff \left\{ \begin{array}{l} R_i \cap R_j > 0 \\ |R_i - R_j| < 10 \end{array} \right.$
   - Edge Density Constraint
     - $E_{density} > 0.2$

3. Slides Text Recognition
   - Double Text Regions Size with Bilinear Interpolation
   - Tesseract OCR Engine
   - Training with 15 character sets
   - Height 30pt
   - Most popular fonts used in presentation slides

Local Adaptive Otsu (LAO) Binarization

Optimize for threshold $T$ maximizing between-class variance in sliding window

$$T(x, y) = \mu(x, y, W) \left[ 1 + k \left( \frac{\sigma(x, y, W)}{R} - 1 \right) \right]$$

[Optimal version of Sauvola’s algorithm]

$$T(x, y) = n_b(T) n_f(T) (\mu_b(T) - \mu_f(T))^2$$

[Localized version of Otsu’s algorithm (Otsu 79)]

- Assumption: bimodal distribution (foreground/background)
- Fast implementation with Integral Histogram [Porikli et al. 05]

Results

8 presentation videos, 1hr 45 mins, ~13 Slides each

- Slides Text Detection - 500 Frames
  - 400 with text
  - 100 no text
  - $\text{Precision}_{\text{SIMPLE}} = \frac{TA_{GT} \cap TA_{E}}{TA_{E}}$
  - $\text{Recall}_{\text{SIMPLE}} = \frac{TA_{GT} \cap TA_{E}}{TA_{GT}}$

- Binarization - 54 Detected Regions, 2177154 annotated pixels

- Slides Text Recognition - 2276 words, 13804 characters

Example of indexing function:
- The word Energy is recognized in slides across 4 different presentations

1. www.aquaphoenix.com/reasearch/vastmm