Clustering SVG Shapes

Integrating SVG with Data Mining and Content-Based Image Retrieval

Michel Kuntz
Fachhochschule Kaiserslautern
Zweibrücken, Germany

SVG Open 2010
Presentation Overview

• Context, Problem, Prototype Solution
• Minimal (non-SVG) Background
• Top-level Software Architecture
• End-user View of Main Functionalities
• Experiments in 3 Application Areas
• Looking Back and Taking Stock
Raster vs. vector

• Clustering bitmap images is active research
  – Multimedia Data Mining
  – Content-Based Image Retrieval

• Clustering visual data expressed by means of the SVG language till now unexplored
Evolving Context for SVG

• SVG documents:
  – output from authoring tools
  – input to rendering engines
  – information resources -- to be analysed, organised, and retrieved

• clustering SVG shapes contributes to filling gap in analytical tools for SVG data

• adapt best-of-breed results from shape modelling and clustering to fit SVG
Why would SVG users want to ... cluster SVG shapes?

• get assistance in exploring SVG data
• improve reusability of existing SVG code
• avoid reinventing what already exists (both for creativity and efficiency reasons)
Prototype Solution

• we designed and implemented software that
  – collects shapes from SVG doc rendering tree
  – represents shapes as polygon vertices
  – computes Fourier Descriptors for polygons
  – clusters FD sets into perceptually similar groups
  – enables display and manipulation of results
      ( cluster tree and individual cluster contents )
(non-SVG) Background

• modelling shape
  – wanted: good match between perceptual similarity and values of model's parameters
  – using raw data directly is no solution

• Discrete Fourier Transform
  – raw data => coefficients => descriptors

• clustering (unsupervised classification)
  – finding natural groups (i.e. clusters) in data
  – (partitional vs.) hierarchical: nested clusters
shape contour (solid) and polygon approximation (dashed)
better approximations (red) with more coefficients
tree resulting from hierarchical clustering
(dendrogram cut to form 4 clusters)
Top-level architecture

- shape acquisition
  - polygon vertices
  - description generation
    - Fourier descriptors
      - clustering algorithm
        - cluster hierarchy
      - manipulation of results
Defining “shape“

• General geometry
  – external form, contour, boundary curve

• SVG-syntax
  – rect, circle, ellipse, polygon, path

• Perceptual
  – intuitively: what meets the eye

• “Semantic“
  – (representation of) everyday objects
End-user View

• Select SVG documents to cluster
• ( Check to see what shapes they contain )
• Launch clustering run
• Browse complete tree of nested clusters
• Cut tree and see summary results
• Investigate cluster memberships
kuntz

clustering SVG shapes
# Select SVG documents

<table>
<thead>
<tr>
<th>include</th>
<th>uri</th>
<th>% clusterable</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc02.svg</td>
<td>100.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc01.svg</td>
<td>000.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc04.svg</td>
<td>100.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc03.svg</td>
<td>100.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc06.svg</td>
<td>100.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc05.svg</td>
<td>100.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc07.svg</td>
<td>100.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc08.svg</td>
<td>100.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc10.svg</td>
<td>000.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc11.svg</td>
<td>050.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc09.svg</td>
<td>050.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc12.svg</td>
<td>000.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc13.svg</td>
<td>000.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc15.svg</td>
<td>100.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc14.svg</td>
<td>100.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc16.svg</td>
<td>100.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc17.svg</td>
<td>100.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc18.svg</td>
<td>100.00</td>
</tr>
<tr>
<td>✔️</td>
<td>file:/home/michel/svg2010/clipart/shapes/flowchart/fc20.svg</td>
<td>100.00</td>
</tr>
</tbody>
</table>
Browse complete tree
Cut tree and see summary
Investigate cluster memberships

clustering SVG shapes
Experiments in 3 Application Areas

• Clip art
  – Relatively good results
  – Problem: perceptual vs. semantic shapes

• Technical drawings
  – CAD: SVG "line" elements – unclusterable
  – Map data: uneven mix of un- / clusterable

• Glyphs of Chinese characters
  – Relatively good results
  – Test bed for N syntactic => 1 perceptual
Looking back and taking stock

• What we did right
  – actually doing it -- at all

• What we did wrong
  – naive implementation of some clustering
  – FD dissimilarity function too simple

• What needs to be done
  – better shape modelling and clustering
  – move on from simple to composite shapes
  – more and better analytical tools for SVG