There are currently three dominant approaches to clustering and example-based web page retrieval. We propose a 4th approach.

1. **Link structure**
   - *User's perspective:* For a web user, the visual appearance is more discriminating than the document structure, e.g., TABLE vs DIV tables: [http://gatterbauer.name/tables/DIV_table.html](http://gatterbauer.name/tables/DIV_table.html)
   - *Similar appearance hints at similar content* (language independent?)
   - *Preprocessing:* for automated information extraction approaches targeting visual structures
   - *Exploration:* Looking for visually similar web pages is a fun and new way to explore the Web.

2. **Content**
   - Remove pictures, replace text, use the visual box model:
     - Consider an asymmetric version of a two player verification game (similar to ESP):  
       - Player 1 sees one input image (the visual box model) and chooses one of four other images she considers most similar to the input
       - Player 2 sees the chosen image and chooses among the other four images the most similar
     - *Goals:* (i) to learn human judgment of visual similarity, (ii) training and evaluation set for our visual edit distance
     - For a simple demo, visit: [http://pbit.at/vsc/](http://pbit.at/vsc/)

3. **Document structure**
   - Calculate the visual edit distance as the weighted sum of all differences ($j$):
     - adjacency and alignment violations ($V$) [e.g. 4 and 8 adjacent in P1, not P2]
     - transformations of box groupings ($T$) [e.g. dimension and color of group 1-4]
     - missed matches ($M$) [e.g. boxes without matches and numbers]

     $\sum_{j \in V \cup T \cup M} w_t(j)$

     $w_t(j)$ ... type of difference

     $w(j)$ ... weight of this difference type

4. **Visual structure**
   - Find the bipartite matching with the minimal visual edit distance

   $BM = \min_{j \in V \cup T \cup M} \sum_{j} w_t(j)$

   Two caveats: (i) hard metrics, (ii) learning of weights required

---

**Key points**

1. We propose a 4th and yet unexplored approach to clustering and navigating the "Visual Web" (contrast with the Linked, Syntactic or Semantic Web): Our focus is on the visual appearance of web pages.
2. We propose to capture the visual appearance of a web page by removing images, replacing text with repeated letters and focusing on the visual box model, and define a general visual edit distance between two visual box models.
3. We propose a two player verification game whose purpose is to learn the feature weights of our similarity measure.

**Open point:** can human perception of visual similarity be captured simply in terms of low level visual features? Or do we need to build a measure that reasons in terms of more abstract intermediate concepts?


**Links** (also in the paper):
- Visual similarity comparison demo: [http://pbit.at/vsc/](http://pbit.at/vsc/)
- Example DIV table: [http://gatterbauer.name/tables/DIV_table.html](http://gatterbauer.name/tables/DIV_table.html)