Impediments to General Purpose Content Based Image Search

Melanie A. Veltman
CIS Department, University of Guelph
Guelph, ON N1G 2W1, Canada
+1.519.824.4120
mveltman@uoguelph.ca

Michael Wirth
CIS Department, University of Guelph
Guelph, ON N1G 2W1, Canada
+1.519.824.4120
mwirth@cis.uoguelph.ca

JingBo Ni
CIS Department, University of Guelph
Guelph, ON N1G 2W1, Canada
+1.519.824.4120
jni@uoguelph.ca

ABSTRACT

Challenges faced by prevailing text metadata paradigms for online image search have inspired overwhelming research in Content Based Image Retrieval (CBIR). A multitude of approaches have been introduced within the literature, yet relatively few image search engines have been made publicly available on the web. Aside from challenges facing the user, such as describing a visual query using keywords, or finding an appropriate example image to initiate a visual search, all systems must inevitably grapple with the sensory and semantic gaps [Smeulders et al. 2000], which essentially represent a loss of information in the abstraction process. In this work, we challenge commonly suggested approaches to improving CBIR and illustrate drawbacks of relying on textual data, as well as visual data, in general CBIR search. We provide cogent examples using online visual search engines Behold™[1], Tiltomo Beta[2], Pixilimar[3], and Riya™ Beta[4]. These examples demonstrate the effect of semantic ambiguities in natural language, which extend to search terms and text tags.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis – *color, object recognition, feature evaluation and selection, classifier design and evaluation.*

General Terms

Algorithms.

Keywords

Image analysis; image processing; computer vision; information retrieval; visual similarity; semantic gap.

1. INTRODUCTION

Content Based Image Retrieval (CBIR) refers to a field of technologies that aim to use the *content* of digital images to index, search, and manage them, in contrast to the text-based approaches used by the widely popular Google™, Yahoo! ™, and Alta Vista™[5] [1] Image Searches. Considerable success, in terms of both popularity and functionality, has been had in the text-based image search realm, but this type of retrieval is ultimately limited by the availability of textual metadata associated with images. Although manually assigned tags, when available, may indeed afford precise and accurate information at the proper level of abstraction, these concepts themselves are subjective, so it naturally follows that the metadata will undoubtedly suffer the same subjectivity. The results of a Google™ Image Search for 'kitty', illustrated in Figure 1, demonstrate that in large scale image retrieval, the metadata provided by one user may be of little use for another user's search. This challenge makes itself especially apparent in cases where a language difference may render text data entirely irrelevant, as discussed in detail by Popescu and Kanellos[2008]. The popular proverb 'a picture is worth a thousand words', reveals the enormity of the task of CBIR. In a single glance, an image may have the power to convey more information than a thousand words - in any language - could ever possibly hope to.

The viability of CBIR has been widely acknowledged since Smeulders' thorough analysis of the field at the 'end of the early years'[Smeulders et al. 2000], and has been reiterated in surveys[Datta et al. 2005, Kherfi et al. 2004, Veltkamp and Tanasa 2000]. It is further validated by an explosion in research: Datta et al [2005, 2008] provide solid evidence of exponential growth within the field since Smeulders' publication in 2000. Due to the enormity of the field and the wide range of disparity found among approaches, applications of CBIR can be classified in numerous ways. Smeulders categorizes systems by the pattern of computation, and also by the type of search carried out by the user, discussing category search, target search, and search by association[Smeulders et al. 2000]. When segregated by task, categories may include object recognition, i.e., [Lowe 1999], image annotation, i.e., [Hervé and Boujemaa 2007], image classification, i.e., [Li et al 2000, Li and Wang 2006, Yavlinsky and Heesch 2007], and others. Obviously, for image matching and description tasks, a query image must be provided. IBM [Ashley et al. 1995] and others use the term Query By Image Content

C3S2E-09 2009, May 19-21, Montreal [QC, CANADA]

[1] http://www.behold.cc
[2] http://www.tiltomo.com
[3] http://ideeinc.com/products/piximilar
[4] http://www.riya.com
[5] Alta Vista™ Image Search allows the user to filter result images by 'Color', 'Black and White', or 'All Colors'.

(QBIC) to refer to systems which use an image or sketch as the query. Other query types include text or semantic search and combined searches. Another predominant distinction is made based on computation, whereby approaches are defined as either global or local.

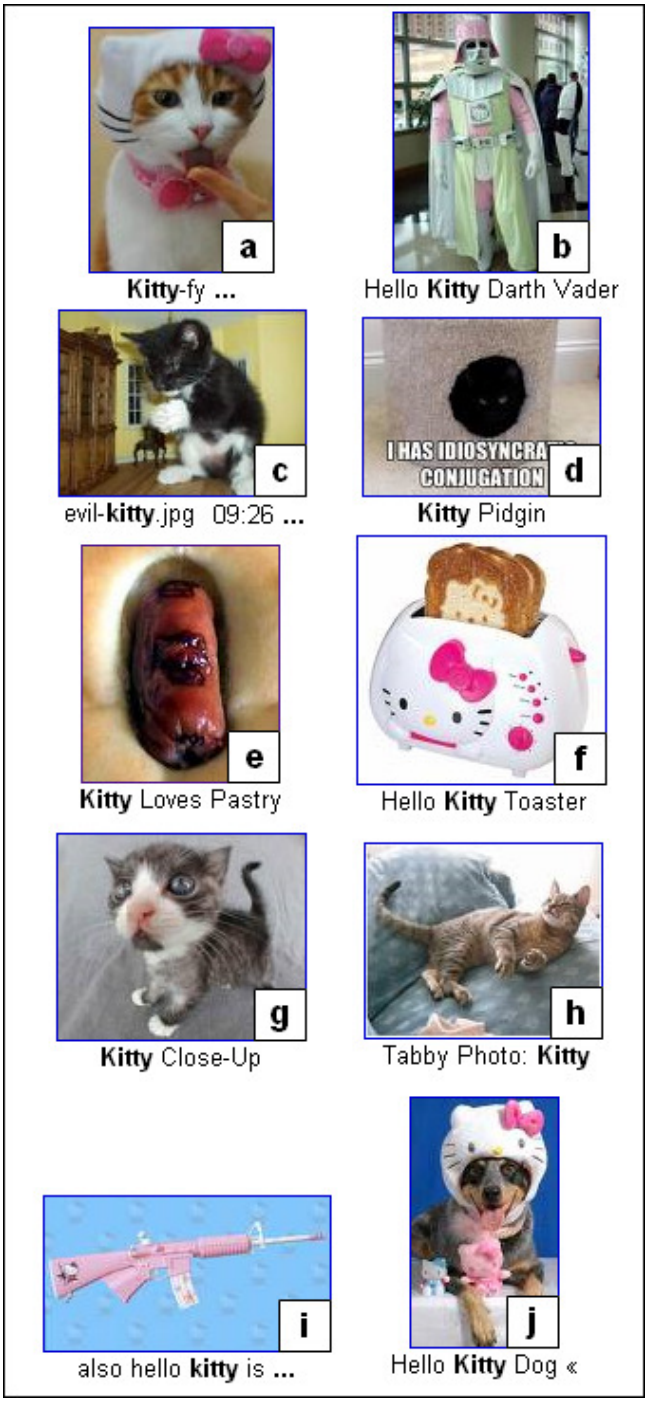

Figure 1. Highest ranked Google™ Image search results for 'kitty' include (b) a man in a pastel warrior costume, (e) a sausage in pastry, (f) a toaster, and (i) a pink rifle. Source: http://images.google.com.

1.1 Descriptor Assignment

Regardless of the query type, goal, or any other differentiating factor, CBIR implies at least two fundamental computational tasks, the first of which is describing the image data mathematically. Presumably quite contrary to the human visual system, image processing systems must, at some point, represent visual information as a set of numerical vectors (pixels). Methods for understanding the pixel information include global approaches, which take into account the image data in its entirety, and local approaches, which disregard some regions of an image - like, for example, sky, grass, or other backgrounds - in favour of focusing on pixels or regions of interest. In both cases, the description task entails some form of simplification, whereby the pixel information is reduced to facilitate efficient indexing and matching.

On the basis that they are indeed non-trivial abstractions, all approaches, both global and local, are leaky abstractions[Spolsky 2002], rendering the goal of finding the 'perfect' way to describe images unattainable. It may even be argued that the preliminary task of capturing a 3D scene as a 2D set of pixels is also a leaky abstraction: human scene understanding entails the use of stereo vision to assimilate depth and distance, whereas images capture only one viewpoint. Hence, information that can be inferred about a 3D object or scene based on the available 2D image data, i.e., the shape, depth, and spatial locations of objects, is essential to scene understanding and lost in the digital translation. Smeulders refers to this perceptual disparity as the 'sensory gap' [Smeulders et al. 2000]. Theoretically, all approaches will remain leaky abstractions until the human visual system (HSV) and scene understanding capabilities are fully understood and digitally replicated.

Regardless of these limitations, research in the field remains optimistic, and applications continue to build upon both paradigms, with varying levels of success. Although local approaches are understood by some to be somewhat similar to the HSV [Hervé and Boujemaa 2007, Li et al. 2000], finding the proper level of abstraction, i.e., determining which regions or properties are most important in determining the 'meaning' of the image data, remains a challenge. The information required to appropriately classify or match an image may indeed exist in the pixel data, but if it is not represented well in the descriptor, a match will not be made. Hervé and Boujemaa [2007] propose the term 'numerical gap' to refer to the disparity between the information provided by the pixel data in an image and its computed mathematical descriptor.

1.2 Descriptor Comparison

After mathematically describing an image, the second fundamental task lies in finding an appropriate method to perform comparisons between mathematical descriptors to determine similarities between images or appropriate annotations. The design of the similarity determination or matching process is heavily dependent on the describing scheme, and also on the task. In most cases, an exact image match is not the exclusive goal of the search. Even in the case of targeted searches, matching and returning 'replicas' of an image that may have undergone any number of transformations (i.e., rotation, blurring, scale changes) is desirable to returning no results at all. Most retrieval methods aim to match under various image conditions by incorporating

robustness into either the descriptor or the method used for matching. An example of one such robust implementation is TinEye[2], a visual search engine created by Idée labs, which allows a user to search the web for query images under circumstances of colour changes, text overlay, flipping, cropping, etc., for the purpose of enforcing copyright.

1.3 Semantic Comparison

In the case of text queries, where there is no query image for comparison, images are generally selected based on their mathematical similarity to any number of conceptual models the system has learned. The training concepts chosen by researchers and application designers range in both usefulness (see [Lampert 2006] for an entertaining discussion, which reveals that the 332 concepts used in Li and Wang's [2006] ALIPR™ [3] system include *thing* and *photo*) and semantic complexity. Consider, for example, an image which captures a goal being scored in a game of soccer. A low-level semantic definition of such an image may include terms and phrases like *grass*, *person*, *ball*, etc., whereas higher level descriptions, i.e., those which require more interpretation, may include semantic concepts like *sport*, *soccer*, or even *goal*.

Furthermore, different semantic meanings can be inferred from the same image by different users, and also by the same user when facing a different task. Smeulders' survey [Smeulders et al. 2000], and most of the literature that has succeeded it, refer to this concept as the 'semantic gap'. Pavlidis[2008] refers to the disparity between semantic meaning and pixel level statistics as a 'semantic abyss', and reveals that some CBIR matching schemes fail to find semantic similarity between a query image and replicas transformed by changes in only contrast and luminance.

In the absence of reliable metadata to help bridge the semantic gap, it is commonly suggested that the words surrounding or detected in an image may help a system understand the context of the image [Wang et al. 1997, Kherfi et al. 2004, Pavlidis 2008]. In this work, we present drawbacks to using textual information in visual search, and discuss the notion of semantic importance in this regard. We also demonstrate and discuss semantic ambiguities in search phrases, complexities in semantic hierarchies, and their effect on search results. We provide cogent, illustrative examples using publicly available online visual search engines Behold™, Tiltomo, Piximilar and Riya™ . Due to the fact that each of these engines searches a different dataset of images, a proper and thorough assessment of their efficacy cannot be achieved. We comment on observed challenges specific to each of the approaches, and provide precision results for a common object search, *television*, recognizing the bias in our results due to the highly subjective nature of semantic relevance.

2. LIMITATIONS

2.1 Behold™

Being aware of the semantic gap, it is a reasonable prediction that relying on visual data alone for classification and description tasks will provide, at times, erratic results. A simple example of this phenomenon can be seen in search results from the publicly available Behold™ prototype visual search engine, which claims to recognize visual concepts in high quality Flickr™ photos. Users can perform a simple text search; filter a text search using 28 categories including *animal*, *beach*, *car*, *face*, etc.; or browse images by category. Browsing images in the *city* category, highly ranked search results for the class, illustrated in Figure 2, include images of sunsets, clouds, and beaches --- each distinct categories in the Behold™ system. Algorithms to detect the presence of vertical edges - which, in images, are generally indicative of manmade structures - may improve classification of *city* images by excluding images of sunsets and clouds devoid of cityscapes, but will not eliminate images of sailboats, as are seen in Figure 3(a) and (b), the results of browsing the category *skyline*. Note also that *boat* is also a category. The numerical gap [Hervé and Boujemaa 2007] presents itself here as a failure to extract and make use of discriminatory information that does indeed exist in the images.

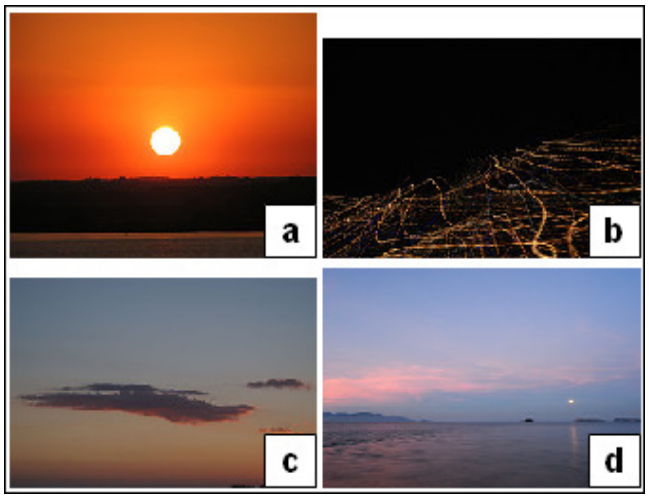

Figure 2. Highest ranked images returned in the category *city* include (a) a *sunset*, (c) a *cloud*, and (d) a *beach*, each of which are themselves categories. Source: http://www.behold.cc.

The visual similarity of the *skyline* categorized images in Figure 3(a), (b), and (c) lies mostly in the spatial layout of coloured regions. In a similar fashion, the classes *sunset* and *silhouette* are highly correlated, as illustrated in Figure 3(d), (e), and (f). These results are predictable, knowing that the system is based on the work of Yavlinsky and Heesch[2007]. In addition to using global colour histograms and a local colour structure descriptor to identify groups of near identical images, similarity judgements rely heavily on thumbnail similarity features - vectors of 1,888 values representing the grey levels of the original image, scaled down to 44x27 pixels [Yavlinsky and Heesch 2007].

The search results clearly demonstrate that visual similarity does not necessarily correspond with semantic similarity, and also reveal that many images may depict more than one category, like *sunset* and *silhouette* in Figure 3(e), and *skyline*, *city* and *sunset* (arguably) in Figure 3(f). Additionally, many of the categories overlap semantically at some level of interpretation. This concept of balance between the inclusiveness of a category definition and discriminatory power carries over into visual similarity, whereby learning categorized training images may skew conceptual models and cause significant overlapping. The ALIPR™ system [Li and Wang 2006] overcomes these challenges by associating numerous concepts with each image, an approach which seems to be more

[2] http://ideeinc.com/products/tineye/
[3] http://alipr.com/

cohesive with human natural language patterns. A recent report on keyword usage statistics for general online search queries [Unknown 2008] indicates that the average number of keywords in English language searches is 2-3, and in predominantly English speaking countries, single keyword searches represent fewer than 13% of all searches.

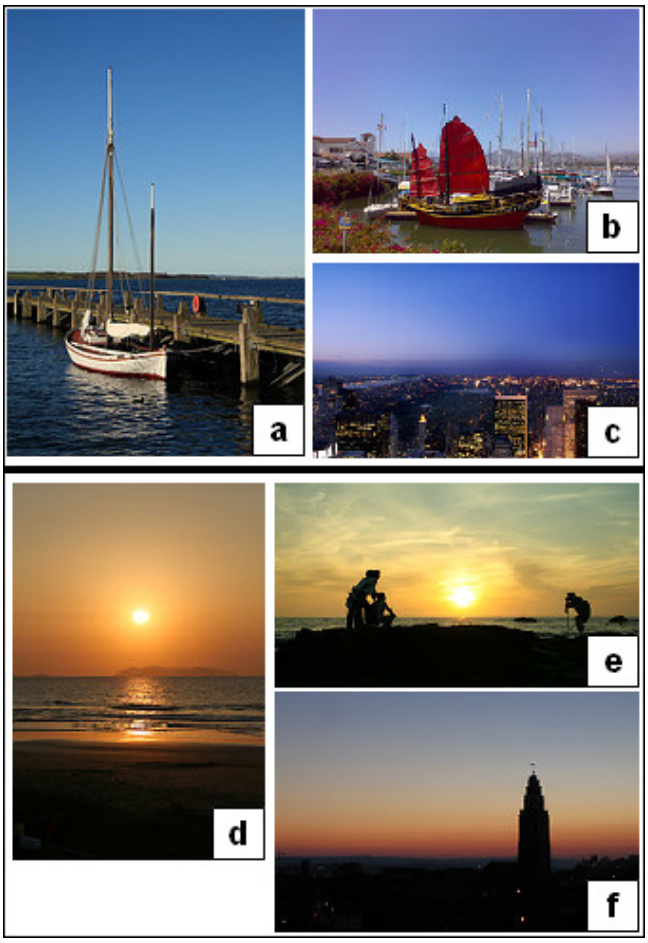

Figure 3. Behold™ categories overlap. (a), (b), and (c) are categorized as *skyline*, even though *boat* is a distinct category. (d), (e), and (f) are categorized as *sunset*, yet also depict *silhouette* and *skyline*. Source: http://www.behold.cc.

2.1.1 Television keyword search

In addition to category browsing, Behold™ provides keyword search functionality. Here, we search for the keyword *television*, and attempt to assess precision, as the ratio of the number of relevant results to the total number of images returned, *n*. The Behold™ engine searches more than 1,040,000 Flickr™ images, and returns a total of 148 for the *television* keyword search. There could presumably be many more relevant images in the search set, and our results here essentially represent a measure of the validity of the Flickr™ tags. Since we do not have access to the dataset in its entirety, a measure of completeness (i.e., recall) is not attainable.

The keyword search *television* is vague, and desired results of such a search are dependent on the user's search goals. Relevant images could range from those clearly depicting an entire television, to those loosely portraying a television-related theme. Being aware of these ambiguities, we define three search goals by which to measure precision: tight object search (*TOS*), with the goal of finding images of unobstructed televisions; loose object search (*LOS*); and a theme search (*THS*). The classification of the return images as relevant for each of these conditions is a highly subjective process, and to clarify the biases imposed, we further specify the categories as follows. For an image to be classified as relevant under *TOS*, each of the following statements must be true about the image, viewed as a thumbnail: the television is arguably the primary subject of the image; more than 60% of the television is visible and unobstructed; it would be difficult to mistake another object in the image as the focus of the image. To qualify as relevant for *LOS*: the image must contain a television, which may or may not be partially obstructed, and the television is clearly not the primary subject of the image. Images which meet *TOS* conditions are also considered relevant for *LOS*, and irrelevant images are those which do not contain a recognizable television. For *THS*, the least restrictive search, we consider relevant those images which meet *TOS* or *LOS* conditions, and any images depicting a television-related theme, i.e., a television tower, watching television, a remote control. In this case, irrelevant images are those for which it is difficult to determine, at a glance, why the image would fall under the category television.

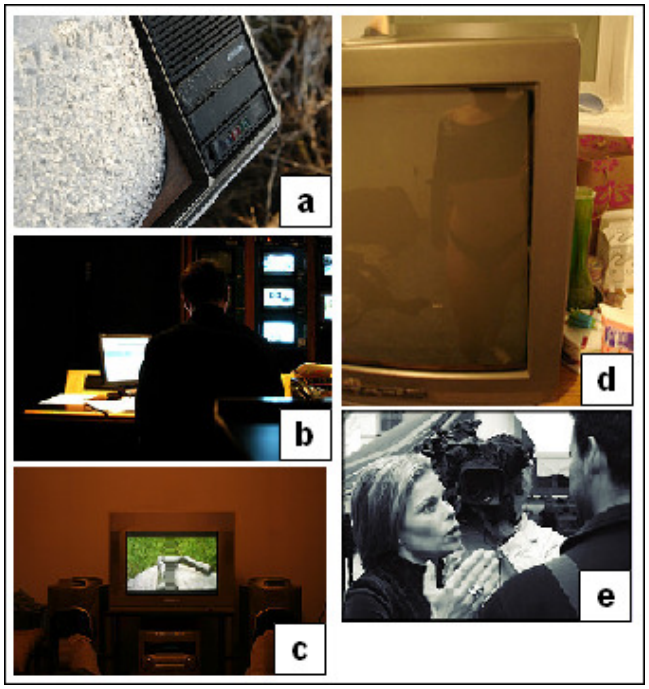

Figure 4. Sample of images returned for the Behold™ keyword search *television*. Source: http://www.behold.cc.

At a first glance, Behold™ returns seemingly relevant images for the *television* search. A sample of the highest ranked results are illustrated in Figure 4. Of these, (c) meets our criteria for *TOS*; (a) and (d) qualify for *LOS*; and (b) and (e) are considered relevant for *THS*. We consider the first *n=30* return images in our precision calculations, based on a limitation of the Tiltomo engine. Normalizing *n* across the four systems ensures that calculations are not skewed with lower ranked results for the engines which return more than 30 images. Of the first 30 images

returned by Behold™, 5 can be categorized as *TOS* relevant, 9 as *LOS*, and 10 as *THS*, leaving 6 irrelevant images. Our subjective assessment finds that Behold™ performs, on this particular search, with precision rates of 16.7%, 46.7%, and 80% for *TOS*, *LOS*, and *THS* searches respectively.

2.2 Tiltomo Beta

Tiltomo Beta is another online visual search engine for high-quality Flickr™ images. In addition to keyword search, Tiltomo allows users to browse 'similar' images on the basis of subject/theme or colour/texture similarity. Unlike Behold™, Tiltomo does not classify images, but rather, allows for the refinement of results based on visual similarity. Available documentation[4] defines "similar" as either Image Subject/Color/Texture or 100% Color/Texture similarity. Tiltomo searches either 133,417 general Flickr™ images, or 137,400 in the Catchy Colors group, and displays 30 return images at a time.

2.2.1 Television keyword search

To begin, we perform a keyword search for *television* on the Catchy Colours (default) database. Samples of the results are shown in Figure 5(a)-(e). Out of the 30 image matches returned, only five contain a television, and only two can be categorized as *TOS*. Subjective precision rates for *TOS*, *LOS*, and *THS* searches are 6.7%, 23.3%, and 46.7% respectively. These numbers are significantly lower than those achieved by Behold™ on a similar (i.e., Flickr™), but much larger image set. Assuming the assignment of Flickr™ tags is equally reliable and sporadic among both sets, we can infer that these results demonstrate the negative impact of the smaller size of the Tiltomo search set, which is roughly one tenth of the size of the one searched by Behold™.

2.2.2 Television similarity search

For the similarity search, we choose the first image from the keyword results that most closely adheres to *TOS* conditions. In this case, the query image for the thematic search is that in Figure 5(a). We assess the subject/theme search because its focus on the subject of the image closely relates to our search goals. Samples of the search results, illustrated in Figure 5(d) and (f)-(j), demonstrate that theme determination is indeed heavily dependent on colour information. Clearly, the algorithm has recognized that the query image is more mono-hue than multi-colour, and has returned results with similar colour patterns. The results reveal the power of semantic concepts which are drawn not only from the television, but also from its surroundings and environment. In particular, the image in Figure 5(d), which depicts a television in a car, may be a perfect match for the query image in the case of someone searching for images of *abandoned televisions*.

However, quite apart from their colour likeness, these images, along with most of the search results, at best questionably depict a television. Although the documentation refers to the search as Image Subject/Color/Texture, the recognition of a subject seems questionable. For this particular similarity search, subjective precision rates of 3.3%, 10%, and 20% are achieved for *TOS*, *LOS*, and *THS* searches, rendering the extracted visual information is less reliable than assigned tags. Readers are referred to Pavlidis'[2008] work for a thorough and critical discussion of the limitations of reliance on colour information in CBIR.

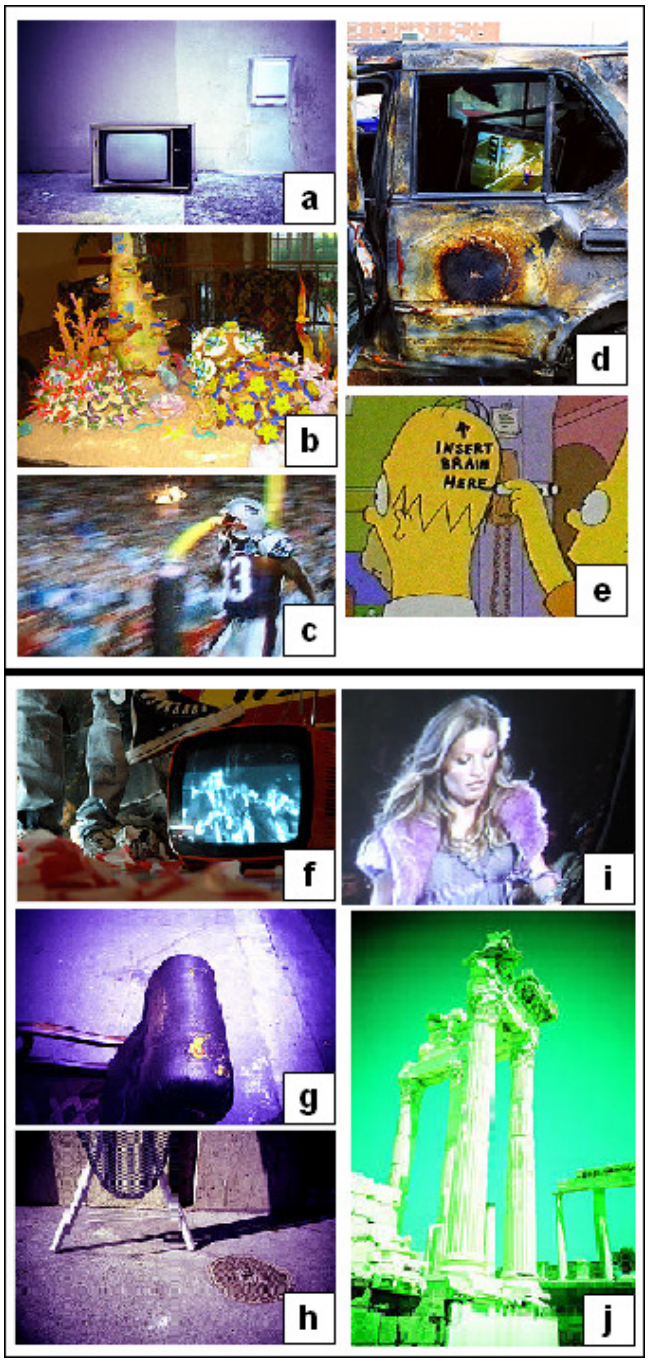

Figure 5. Top, (a) through (e): Tiltomo Beta search results for keyword *television*. Bottom, (f) through (j): search results for thematic similarity based on image (a); (d) also appears in thematic similarity results. Source: http://www.tiltomo.com.

2.3 Piximilar

Piximilar, a visual search engine provided by Idée, offers similar search capabilities with its Multicolour[5] and Visual Search[6],

[4] http://www.tiltomo.com/about.php

[5] http://labs.ideeinc.com/multicolour/

demos of which are available online. The website boasts that the digital signature computed for each image makes use of hundreds of attributes, including colour, texture, and luminosity, as well as object and region shapes and complexity. The Multicolour search allows a user to select up to 10 colours from a palette of 120, and returns images from Flickr™ or Alamy[7] stock photos comprised of similar colours. The Visual search begins by displaying a random set of images to the user, and alternatively, allows the user to enter a keyword. Searches can then be refined by selecting a desirable image from the random or returned set, or through the use of keywords. Although the Piximilar™ Visual search makes no claims about semantic or conceptual similarity, a discussion of these is provided here on the basis that the digital signature makes use of object and region shapes and complexity.

2.3.1 Television keyword search

Our Piximilar *television* keyword search returns 2,250 images from the Alamy™ stock set of 2,823,639. A sample of the search results are shown in the top half of Figure 6. As with the previous keyword searches, the images returned are visually diverse and depict a wide variety of semantic concepts. Based on the first 30 images returned, precision rates for *TOS*, *LOS*, and *THS* search are calculated to 16.7%, 26.7%, and 60%. These *LOS* and *THS* precision rates are lower than those achieved by Behold™, even though the Alamy™ set is more than two and a half times the size. Clearly, for this particular search, the Flickr™ tags are more reliable than those in the Alamy™ set.

Reasoning behind some of the seemingly inappropriate matches is revealed upon further inspection of the image tags. For example, Figure 6(b), the image of a man stepping out of a car, is annotated with the terms: *business*, *businessman*, *car*, *city*, *copyright*, *dmitri*, *entrepeneur*, *governor*, *interros*, *interview*, *jeremy*, *local*, *mall*, *moscow*, *new*, *nicholl*, *offical*, *olympic*, *political*, *politician*, *politics*, *prepares*, *rich*, *russia*, *shopping*, *station*, ***tver***, *wealth*, *zelenin*. Presumably, the match in this case was made between the abbreviated keyword *tv* and the annotation *tver*, a city in Russia – an excellent example of tags being rendered irrelevant by language differences.

Five of the keyword search results, in Figure 6(e)-(g) clearly, even at the thumbnail size, depict the concept of *watching television*. This is inferred by: the position of the person or people in the image; the pattern of light diffusion, as in (e), (f), and (g); and/or the existence of a remote control. Most notably, in only one of these *watching television* images, (c), is a television present. The addition of a second keyword, *watching*, results in far more appropriate results, with a subjective precision of 83%, i.e., out of the first 30 images returned, 25 clearly depict the notion of watching television, although only 5 of the images (16.7%) contain a television.

2.3.2 Television visual search

For the visual refinement search, again we choose the first *TOS* image; in this case, the image in Figure 6(a). The search returns a total of 39 images and we calculate *TOS*, *LOS*, and *THS* precision rates of 30%, 30%, and 53.3%. The samples provided in the lower half of Figure 6 suggest that the digital signatures used to index and match the images are highly dependent on colour information, as well as tag data. Further investigation finds that many of the monochromatic people images are tagged with words loosely related to television, like *actor*, *screen*, *personality*. Most likely, relationships between these terms in the hierarchy of learned concepts are being used by the system's visual search. For this particular search, theme results for *people in black and white* would be equally as successful, with 53.3% precision. Like Tiltomo, the Piximilar visual search appears to capture the essence, style, and theme of the images, more so than the objects.

Figure 6. Top: Piximlar Visual search results for keyword *television*. Bottom: search results for thematic similarity based on image (a). Source: http://www.tiltomo.com. Original images source Alamy™: http://www.alamy.com/.

[6] http://labs.ideeinc.com/visual
[7] http://www.alamy.com/

2.4 Riya™

Using proprietary image analysis and search techniques, the Riya™ Beta visual search engine provides the functionality to find similar faces and objects in 9,887,329 images on the web. Users can perform object, tag, people, and keyword searches, or browse through known categories like *guitar*, *high-chair*, and *sunglasses*. A personal search service allows users to train the system on faces or objects within their own image collection for autotagging, and users are able to tag public images with keywords, as well as rate existing tags.

2.4.1 Television search

We continue our *television* search, using here the tag, keyword, and object searches. The tag search returns a total of 83 images, and among the 30 images with the highest Riya™ Rank, yields precision rates of 13.3%, 93.3%, and 96.7% for *TOS*, *LOS*, and *THS* searches. The results are not as visually diverse as those returned by Behold™, Tiltomo, or Piximilar, with many of the returned images depicting scenes from a common event, with the same television in the background, like the bottom three images in Figure 7. For many of the return images, *television* is not listed as a user assigned tag, but rather has been assigned automatically through object detection. We infer that these are intrinsic causes for the high precision rates for *LOS*, and by extension, *THS*. Additionally, the image search set is the largest among the four search engines, more than three times the size of the Alamy™ set searched by Piximilar.

The difference between the keyword and object search functionality is unknown, but the results of these two searches differ dramatically. Keyword search returns 171 result images, with dismal precision rates of 3.3%, 3.3%, and 20% for the 30 highest ranked results for *TOS*, *LOS*, and *THS* searches. By contrast, the object search returns 44 images, and achieves 10%, 90%, and 96.7% precision for the searches, respectively. These values are noticeably similar to those achieved by the tag search, and many of the return images are common to both sets. These results indicate that the object detection, at least in the case of television objects, is highly successful. Highlights of the keyword search are provided in the top half of Figure 7, and the images in the bottom half of the figure are common to both the object and tag search result sets. Note that (b) is also included in the keyword search results.

None of the top 30 keyword search result images, with the exception of Figure 7(b), visually resemble a television, or even a television-related theme at first glance. Further inspection of the result images reveals some of the logic behind the unexpected similarity rankings: Riya™ uses detected text information to index images. The overwhelming majority of the top 30 keyword search results, 96.7%, have been matched based on detection of the text 'television' in the images. The word is detected in a newspaper clipping in Figure 8(a), and in Figure 8, a sign posting the cost of television service on the wall in a hospital room results in an image of a mother and newborn being indexed with *television*. Interestingly enough, the only image in the top 30 keyword results which has not been retrieved on the basis of text detection is the image of a television, Figure 7(b). On the basis that this image has no associated tags, we infer that it has been retrieved on account of its filename which contains the word television. These results provide clear evidence that the Riya™ system has not succeeded in determining the semantic importance of the object and text tags. The keyword search is very effective at returning images containing the keyword text.

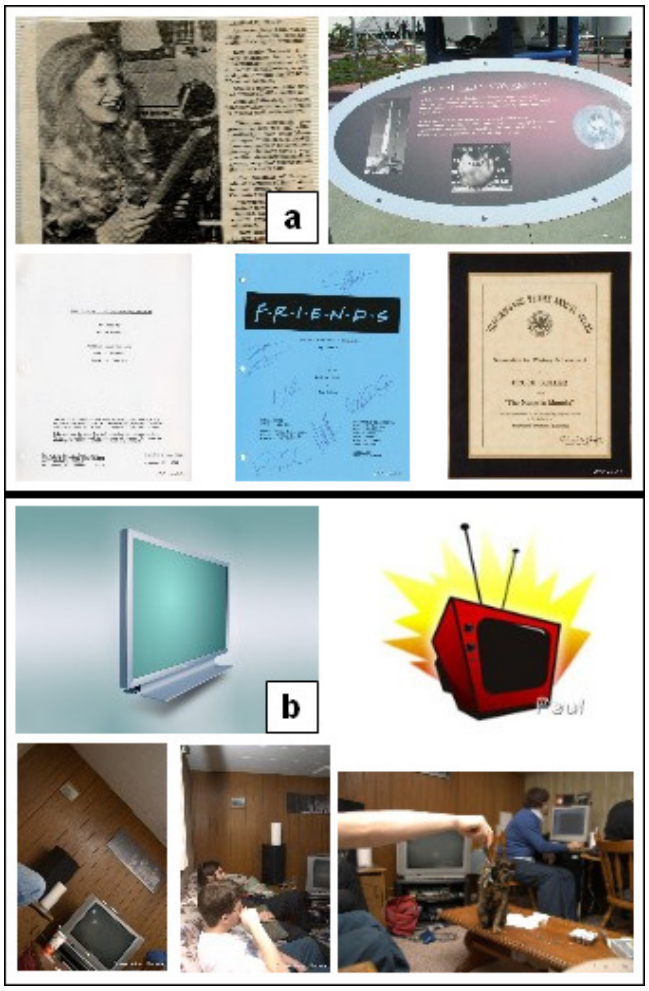

Figure 7. Top: Riya™ Beta search results for keyword *television*. Bottom: search results for *television* object and tag searches. Source: http://www.riya.com.

Slightly more accurate results occur while browsing one of Riya™'s known categories, *penguin*, although the consequence of the text retrieval is just as apparent. Out of 82 result images, 75 (91.4%) clearly, though subjectively, contain penguins, and only 7 images, or 8.6% of the result images, are questionably semantically similar to the concept *penguin*. These include: a picture of Paris Hilton with a Riya™ rank of 64, and a photo of a bride and groom ranked 66, both with filenames containing 'penguin'; and 5 of the 6 most similar images, which can be explained by text retrieval.

In browsing another known object category, *Nokia 7610*, all of the images returned contain a clear depiction[8] of a Nokia phone (although conformity to *7610* was not verified), resulting in a subjective accuracy rate of 100%. Likewise the category *Samsung phones* contains 175 images, each depicting a Samsung phone.

[8] Again, as above, this is a subjective measure.

Indeed, it seems evident that there are fewer false positives in the case of prelisted categories. However, results for the semantically similar search term 'mobile phone' yields disappointing results, including images of business cards, matched on the basis of detected text information, and a literally 'mobile' telephone booth illustrated in Figure 9. For this work, the latter can be considered more poignant than disappointing, as it exposes semantic nuances of this common natural language term. In total, there were 73 matches for 'mobile phone', far less than the number of *Samsung phones*, indicating that the system lacks knowledge of a logical hierarchy that exists in this case.

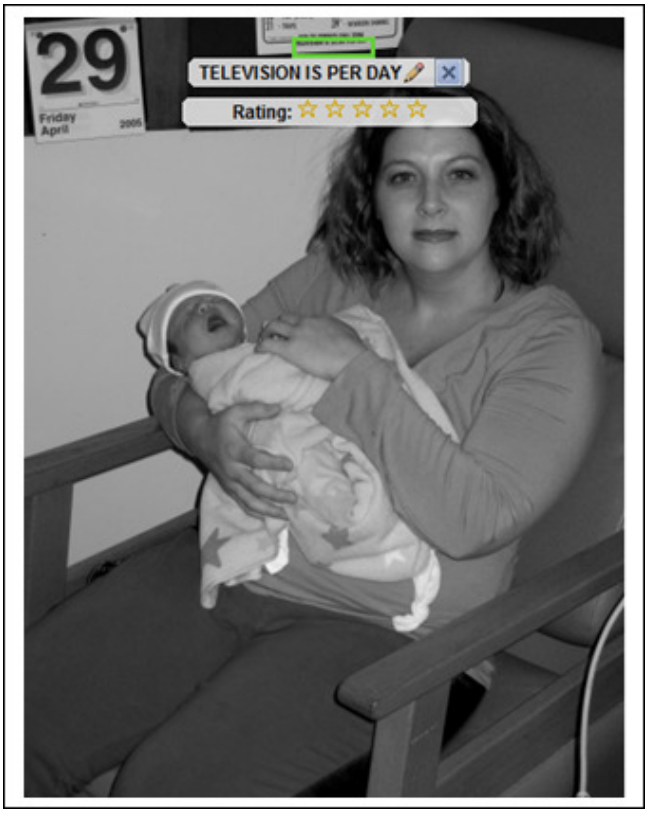

Figure 8. Sample image from Riya™ *television* search results. Source: http://www.riya.com/fileServer?p=04f77c53c5aaa244f872f5a210477b0b09d5b72e.jpg.

3. CONCLUSION

In this work, we have subjectively assessed the precision of online general purpose visual search engines Behold™, Tiltomo Beta, Piximlar, and Riya™ Beta. We have used our results to illustrate a few known limitations inherent to relying solely on visual data in visual search tasks. Additionally, we reveal negative consequences of overreliance on textual data in image search, providing solid evidence that the novelty of text recognition within images needs to be balanced with some knowledge of the semantic importance of the text detected. Clearly, as seen in the television examples, the existence of an exact word match is not sufficient to determine a semantic or object match. Furthermore, Google™ Image Search results make it apparent that the same caution should be heeded when assigning importance to user assigned tags. Amidst a number of initiatives to have users enter or rate image tags, like Google™ Image Labeller and ALIPR™, the tags will undoubtedly be more effective if their validity, usefulness, and semantic importance were scrutinized prior to their use in matching tasks. Kherfi et al.[5] suggest the use of a hierarchical taxonomy in classification tasks, and the 'mobile phone' Riya™ search illustrates that this may be of benefit.

The Behold™, Tiltomo, and Piximilar similarity searches perform well at the task of returning visually similar images, resulting in high levels of precision for theme searches. By contrast, the Riya™ object recognition results are precise for loose object searches, indicating success in object recognition, at least for our particular searches. Further research is needed to determine if Riya™'s superior performance on known categories is due to overfitting - if this is the case, results for unknown searches should decrease in quality, while results for known categories become more precise as the system learns from more images. The Riya™ keyword search is highly effective in recognizing text in images, although this information has shown to be of little value in object and theme searches.

This brief investigation, by providing simple examples of a few key challenges to CBIR for visual search, evidences that possibilities for future work in designing and evaluating CBIR systems are vast in number and dimension.

Figure 9. An image returned by RiyaTM keyword search mobile phone illustrates the term's semantic nuances. Source: http://www.riya.com/fileServer?p=f5aa11334da425175b5473fa2fa26dbd102a6824.jpg

4. ACKNOWLEDGMENTS

The authors wish to extend their gratitude to Prof. Judi McCuaig for meaningful insights into this work.

5. REFERENCES

[1] Ashley, J., Flickner, M., Hafner, J., Lee, D., Niblack, W., and Petkovic, D. 1995. The query by image content (QBIC) system. In Proceedings of the 1995 ACM SIGMOD international Conference on Management of Data (San Jose, California, United States, May 22 - 25, 1995). M. Carey and D. Schneider, Eds. SIGMOD '95. ACM, New York, NY, 475. DOI= http://doi.acm.org/10.1145/223784.223888

[2] Datta, R., Joshi, D., Li, J., and Wang, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40, 2 (Apr. 2008), 1-60. DOI= http://doi.acm.org/10.1145/1348246.1348248

[3] Datta, R., Li, J., and Wang, J. Z. 2005. Content-based image retrieval: approaches and trends of the new age. In Proceedings of the 7th ACM SIGMM international Workshop on Multimedia Information Retrieval (Hilton, Singapore, November 10 - 11, 2005). MIR '05. ACM, New York, NY, 253-262. DOI= http://doi.acm.org/10.1145/1101826.1101866

[4] Hervé, N. and Boujemaa, N. 2007. Image annotation: which approach for realistic databases? In Proceedings of the 6th ACM international Conference on Image and Video Retrieval (Amsterdam, The Netherlands, July 09 - 11, 2007). CIVR '07. ACM, New York, NY, 170-177. DOI= http://doi.acm.org/10.1145/1282280.1282310

[5] Kherfi, M. L., Ziou, D., and Bernardi, A. 2004. Image Retrieval from the World Wide Web: Issues, Techniques, and Systems. *ACM Comput. Surv.* 36, 1 (Mar. 2004), 35-67. DOI= http://doi.acm.org/10.1145/1013208.1013210

[6] Lampert, A. 2006. The broken promise of automatic image tagging, November 2006. Available online: http://www.sgi.nu/diary/2006/11/16/the-broken-promise-of-automatic-image-tagging/. Accessed January 08, 2009.

[7] Li, J., Gray, R.M., and Olshen, R.A. 2000. Multiresolution image classification by hierarchical modeling with two-dimensional hidden markov models. *IEEE Transactions on Information Theory*, 46, 5 (Aug. 2000), 1826 – 1841. DOI= 10.1109/18.857794

[8] Li, J. and Wang, J. Z. 2006. Real-time computerized annotation of pictures. In *Proceedings of the 14th Annual ACM International Conference on Multimedia* (Santa Barbara, CA, USA, October 23 - 27, 2006). MULTIMEDIA '06. ACM, New York, NY, 911-920. DOI= http://doi.acm.org/10.1145/1180639.1180841

[9] Lowe, D.G. 1999. Object recognition from local scale invariant features. In *Proceedings of the 7th IEEE International Conference on Computer Vision*, (Kerkyra, Greece, September 20 – 27, 1999). 2, 1150–1157. DOI=10.1109/ICCV.1999.790410

[10] Pavlidis, T. 2009. Limitations of content based image retrieval, June 2008. Available online: http://www.theopavlidis.com/technology/CBIR/PaperB/Apr08.htm. Accessed January 08, 2009.

[11] Popescu, A. and Kanellos, J. 2008. Multilingual and content based access to flickr images. In *Proceedings of the 3rd International Conference on Information and Communication Technologies: From Theory to Applications*. (April 2008). ICTTA'08. DOI= 10.1109/ICTTA.2008.4530012

[12] Smeulders, A. W., Worring, M., Santini, S., Gupta, A., and Jain, R. 2000. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 12 (Dec. 2000), 1349-1380. DOI= http://dx.doi.org/10.1109/34.895972

[13] Joel Spolsky. The law of leaky abstractions, November 2002. Available online: http://www.joelonsoftware.com/articles/LeakyAbstractions.html. Accessed January 08, 2009.

[14] Remco C. Veltkamp and Mirela Tanasa. Content-based image retrieval systems: A survey. Technical report, Dept. of Computing Science, Utrect University, 2000.

[15] Wang, J. Z., Wiederhold, G., and Firschein, O. 1997. System for Screening Objectionable Images Using Daubechies' Wavelets and Color Histograms. In *Proceedings of the 4th international Workshop on interactive Distributed Multimedia Systems and Telecommunication Services* (September 10 - 12, 1997). R. Steinmetz and L. C. Wolf, Eds. Lecture Notes In Computer Science, vol. 1309. Springer-Verlag, London, 20-30.

[16] Yavlinsky, A. and Heesch, D. 2007. An online system for gathering image similarity judgements. In *Proceedings of the 15th international Conference on Multimedia* (Augsburg, Germany, September 25 - 29, 2007). MULTIMEDIA '07. ACM, New York, NY, 565-568. DOI= http://doi.acm.org/10.1145/1291233.1291372