Analysis of Cross-Media Web Information Fusion for Text and Image Association- A Survey Paper

By M. Priyanka, B. Sunita Devi, S. M. Riyazoddin & M. Janga Reddy
CMR Institute of Technology, Hyderabad, Andhra Pradesh

Abstract - The image comprises of the text- and content-based features. Images can be represented using both text-and content-based features. Web information fusion can be defined as the problem of collating and tracking information related to specific topics on the World Wide Web. But the main concern is image and text association, a cornerstone of cross-media web information fusion. Two methods have been described. The first method based on vague transformation measures the information similarity between the visual features and the textual features through a set of predefined domain-specific information categories. Another method uses a neural network to learn direct mapping between the visual and textual features by automatically and incrementally summarizing the associated features into a set of information templates. Despite their distinct approaches, our experimental results on a terrorist domain document set show that both methods are capable of learning associations between images and texts from a small training data set. Another method encompasses a variety of techniques relating to document summarization and text- and content-based image retrieval.

Keywords : text, image, image processing, image and text association, fusion, context based image representation, search engine.

GJCST-F Classification: I.5.0

Strictly as per the compliance and regulations of:

© 2013, M. Priyanka, B. Sunita Devi, S. M. Riyazoddin & M. Janga Reddy. This is a research/review paper, distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Abstract - The image comprises of the text- and content-based features. Images can be represented using both text-and content-based features. Web information fusion can be defined as the problem of collating and tracking information related to specific topics on the World Wide Web. But the main concern is image and text association, a cornerstone of cross-media web information fusion. Two methods have been described . The first method based on vague transformation measures the information similarity between the visual features and the textual features through a set of predefined domain-specific information categories. Another method uses a neural network to learn direct mapping between the visual and textual features by automatically and incrementally summarizing the associated features into a set of information templates. Despite their distinct approaches, our experimental results on a terrorist domain document set show that both methods are capable of learning associations between images and texts from a small training data set. Another method encompasses a variety of techniques relating to document summarization and text- and content-based image retrieval. The text-based approaches described utilize the Unified Medical Language System (UMLS) synonymy to identify concepts in information requests and image-related text in order to retrieve semantically relevant images. Our image content-based approaches utilize similarity metrics based on computed "visual concepts" and low-level image features to identify visually similar images. The edge model for image and text association detects sharp changes in image brightness and captures important events and changes in properties of the world. An image-text association refers to a pair of image and text segment that is semantically related to each other in a web page. A sample image-text association is shown in Fig. 1. Identifying such associations enables one to provide a coherent multimedia summarization of the web documents.

Keywords : text, image, image processing, image and text association, fusion, context based image representation, search engine.

1. Introduction

The diverse and distributed nature of the information published on the World Wide Web has made it difficult to collate and track information related to specific topics. Although web search engines have reduced information overloading to a certain extent, the information in the retrieved documents still contains a lot of redundancy. Techniques are needed in web information fusion, involving filtering of irrelevant and redundant information, collating of information according to themes, and generation of coherent presentation. Document summarization is a commonly based technique used for information fusion and has been a topic of discussion among the scientists. However, most document summarization techniques focus only on text [11], [12], [13]. documents whereas in today’s world the documents on the net now consist of non text content such as images, video, and sound etc. thus collating multimedia information poses a great challenge in the web information fusion. The image comprises of the text- and content-based features. Images can be represented using both text-and content-based features. The purpose of detecting sharp changes in image brightness is to capture important events and changes in properties of the world. An image-text association refers to a pair of image and text segment that is semantically related to each other in a web page. A sample image-text association is shown in Fig. 1. Identifying such associations enables one to provide a coherent multimedia summarization of the web documents.

Figure 1 : An associated image text pair

The image-text learning association is similar to the task of automatic annotation but has important differences [15]. Whereas image annotation concerns annotating images using a set of predefined keywords, image-text association links images to free text segments in natural language. The methods for image annotation are thus not directly applicable to the problem of identifying image-text associations.

A key issue of using a machine learning approach to image-text associations is the lack of large

© 2013 Global Journals Inc. (US)
training data sets. Referring to Fig. 2, two associated image-text pairs (I-T pairs) not only share partial visual (smokes and fires) and textual features ("attack") but also have different visual and textual contents. As the two I-T pairs are actually on similar topics (describing scenes of terror attacks), the distinct parts, such as the visual content ("black smoke" which can be represented using low-level color and texture features) of the image in I-T pair 2 and the term "Blazing" (underlined) in I-T pair 1, could be potentially associated. Such useful associations, which convey the information patterns in the domain but are not represented by the training data set, implicit associations.

II. LEARNING ASSOCIATION RULES FOR IMAGE CONTENTS AND SEMANTIC CONCEPTS

Association rule mining (ARM) [14] is originally used for discovering association patterns in transaction databases. An association rule is an implication of the form $X \Rightarrow Y$, where $X, Y$ is a subset of $I$ (called item sets or patterns) and $X$ intersection $Y$ is empty. In the domain of market-basket analysis, such an association rule indicates that the customers who buy the set of items $X$ are also likely to buy the set of items $Y$. Mining association rule from multimedia data is usually a straightforward extension of ARM in transaction databases.

A pixel-based approach presented by Ding et al. [16] is based to deduce associations between pixels’ spectral features and semantic concepts. In the model each pixel is treated as a transaction, while the set ranges of the pixel’s spectral bands and auxiliary concept labels (e.g., crop yields) are considered as items. However, it was pointed out that using individual pixel as transaction may lose the context information of surrounding locations which are usually very useful for determine the image semantics. This motivated them to use image and rectangular image regions as transactions and items.

III. IMAGE ANNOTATION USING STATISTICAL MODEL

A major deficiency of the existing machine learning and statistic-based automatic multimedia annotation methods is that they usually assign a fixed number of keywords or concepts to a media object. Therefore, there will inevitably be some media objects assigned with unnecessary annotations and some others assigned with insufficient annotations. In addition, image annotation typically uses a relatively small set of domain-specific key terms (class labels, concepts, or categories) for labeling the images. Our task of discovering semantic image-text associations from Web documents however do not assume such a set of selected domain specific key terms. In fact, although our research focuses on image and text contents from the terrorist domain, both domain-specific and general-domain information (e.g., general-domain terms “people” or “bus”) are incorporated in our learning paradigm. With the above considerations, it is clear that the existing image annotation methods are not directly applicable to the task of image-text association. Furthermore, model evolution is not well supported by the above methods. Specifically, after the machine learning and statistic models are trained, they are difficult to update. Moreover, the above methods usually treat the semantic concepts in media objects as separate individuals without considering relationships between the concepts for multimedia content representation and annotation.

IV. IMAGE REPRESENTATION

The image comprises of the text- and content-based features. Images can be represented using both text-and content-based features. Text-based features include text that pertains to an image, such as in captions and \(\text{\textbackslash\textquotesingle mentions\textbackslash\textquotesingle} (\text{snippets of text within the body of an article that discuss an image}), \text{and content-based features include information derived from the image itself, such as shapes, colors and textures. The text- and content-based image representation is described below.\)

a) Text-Based Features

Each image in the ImageCLEFmed'10 collection [7] is represented as a structured document of image-related text. The representation includes the title, abstract, and MeSH terms1 of the article in which the image appears as well as the image's caption and mention. The content of an image's caption is organized into the well-formed clinical question framework following the method described by Demner - Fushman and Lin [2]. Extractors identify UMLS concepts related to problems, interventions, age, anatomy, drugs, and image modality. Textual Regions of Interest (ROIs) is extracted from image captions. A textual ROI is a noun phrase describing the content of an interesting region of an image which is identified within a caption by a pointer. For example, in the caption "MR image reveals hypo intense indeterminate nodule (arrow)," the word arrow points to the ROI containing a hypo intense indeterminate nodule. The above structured documents can be indexed and searched with a traditional search
engine or the extracted concepts may be combined with additional features (discussed below) for use in a multimodal representation. "Keywords" in a structured document Dj can be represented as an N-dimensional feature vector also.

\[ f_j^{\text{keyword}} = [w_{j1}, w_{j2}, \ldots, w_{jN}]^T \]

Where \( w_{j_k} \) denotes the weight (typically tf-idf) of keyword \( k \) in document \( j \).

b) Image Content-Based Features

In addition to the above textual features, the visual content of images is represented using various low-level global image features and several derived features intended to capture high-level semantic content. Low-level Global Features We represents the spatial structure and global shape/edge features of images with the Color Layout Descriptor (CLD) and Edge Histogram Descriptor (EHD) of MPEG-7 [1]. CLD is extracted to form the feature vector \( f_{\text{ald}} \) and EHD is extracted to form \( f_{\text{ehd}} \). Additionally, we extract the Color and Edge Directivity Descriptor (CEDD) and Fuzzy Color and Texture Histogram (FCTH) using the Lucene image retrieval (LIRE) library. CEDD incorporates color and texture information into \( f_{\text{cond}} \) and FCTH uses the high frequency bands of the Haar wavelet transform to form \( f_{\text{fcith}} \).

c) "Bag of Concepts" Feature

In a heterogeneous medical image collection, it is possible to identify specific local patches in images that are perceptually and/or semantically distinguishable, such as homogeneous texture patterns in gray-level radiological images, differential color and texture structures in microscopic pathology and dermoscopic images. The variation in the local patches can be effectively modeled as \( \text{visual concepts} \) [8] by using supervised learning- based classification techniques, such as Support Vector Machines (SVMs). For concept model generation, we utilize a multi-class SVM composed of binary SVM classifiers combined using the one-against-one strategy [3]. To train the SVM, a set of L labels are assigned as \( C = \{C_1, \ldots, C_L\} \) where each \( C_i \in C \) characterizes a visual concept. The training set consists of local patches generated by a fixed-partition and represented by a combination of color and texture moment-based features. The input to the system is the feature vectors for patches along with their manually assigned concept labels. Concept labels are assigned by fixed partitioning each image \( I_j \) into \( l \) regions as \( \{X_{j1}, \ldots, X_{j_k}, \ldots, X_{j_l}\} \) where each \( X_{jk} \in \mathbb{R}^d \) is a combined color and texture feature vector. For each \( x_{jk} \), its category \( cm \) is determined by the prediction of the multi-class SVM. Hence, instead of the low-level feature-based representation, an entire image is represented as a two-dimensional index linked to visual concepts. Based on this encoding scheme, an image \( I_j \) is represented as a vector of concepts \( f_{\text{con}} \).

\[ j = [w_{j1}, w_{j2}, \ldots, w_{jL}]^T \]

Where each \( w_i \) denotes the "tf-idf" weight of a concept \( c_i \); \( 1 \leq i \leq L \) in image \( I_j \), depending on its information content.

d) "Bag of Keypoints" Feature

Robust and invariant image features are also extracted that are commonly termed affine region detectors [6]. These regions simply refer to a set of pixels or interest points, which are invariant to affine transformations as well as occlusion, lighting, and intra-class variations. We use the Harris-affine detector to locate interest points [5] as a large number of overlapping regions. We then associate with each interest point a vector descriptor invariant to viewpoint changes and, to some extent, illumination changes computed from the intensity pattern within the point. We use a local descriptor developed by Lowe [4] based on the Scale-Invariant Feature Transform (SIFT), to describe the information in a set of scale-invariant coordinates. The SIFT descriptor is chosen to be invariant to viewpoint changes and, to some extent, illumination changes, and to discriminate between the regions. The above features are vector quantized by a self-organizing map (SOM)-based clustering. Finally, images are represented by a bag of these quantized features (i.e., a bag of keypoints). Hence, the model is applied to images by using a visual analogue of the bag of words model used in text retrieval [1].
There is a need to define a similarity measure \( \text{sim}_d(\langle v^1, t^1 \rangle, t^2, \ldots, t^n) \) for individual image regions, which can hardly convey any meaningful semantics without considering their contexts. For example, a yellow region can be a petal of a flower or can be a part of a flame. On the contrary, words in natural languages usually have a more precise meaning. Some of the models used for Vague-transformation of Image-Text associations are:

1) Statistical Vague Transformation in Multilingual Retrieval.
3) Dual-Direction Vague Transformation.
4) Vague Transformation with Visual Space Projection.

For the vague transformation technique to work on small data sets, we employ an intermediate layer of information categories to map the visual and textual information in an indirect manner. A deficiency of this method is that for training the transformation matrix, additional work on manual categorization of images and text segments is required. In addition, matching visual and textual information based on a small number of information categories may cause a loss of detailed information. Therefore, it would be appealing if we can find a method that learns direct associations between the visual and textual information. Some of the methods used for Fusion-based ART techniques are as follows:

1) A Similarity Measure Based on Adaptive Resonance Theory
2) Image Annotation Using Fusion ART
3) Handling Noisy Text

### Table: Vague Transformation vs. fusion ART

<table>
<thead>
<tr>
<th>Approach</th>
<th>Vague Transformation</th>
<th>fusion ART</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information is translated from one space to another space, so that information from different spaces can be compared.</td>
<td>Information is compared in their respective spaces and the results consolidated based on a multimedia object resonance function.</td>
<td></td>
</tr>
<tr>
<td>Learning methodology</td>
<td>Statistic based batch learning.</td>
<td>Incremental competitive learning.</td>
</tr>
<tr>
<td>Information encoding</td>
<td>Using transformation matrices to summarize the information based on predefined information categories.</td>
<td>Using self-organizing networks to learn typical categories of multimedia information objects.</td>
</tr>
<tr>
<td>Network size</td>
<td>Number of predefined information categories is small as information summarized in the transformation matrices is relatively more compact.</td>
<td>More category nodes are created.</td>
</tr>
<tr>
<td>Speed</td>
<td>Around 20 seconds for training; 20 seconds for testing, the training and testing time increases linearly with the number of data samples.</td>
<td>Around 120 seconds for training; 30 seconds for testing; the training and testing time increases exponentially with the number of learnt object templates.</td>
</tr>
<tr>
<td>Performance Stability</td>
<td>Unstable</td>
<td>Stable</td>
</tr>
</tbody>
</table>

#### Dual-Direction Vague Transformation for Cross-Media Information Retrieval

For the dual-direction vague transformation technique to work on small data sets, we employ an intermediate layer of information categories to map the visual and textual information in an indirect manner. A deficiency of this method is that for training the transformation matrix, additional work on manual categorization of images and text segments is required. In addition, matching visual and textual information based on a small number of information categories may cause a loss of detailed information. Therefore, it would be appealing if we can find a method that learns direct associations between the visual and textual information. Some of the methods used for Fusion-based ART techniques are as follows:

1) A Similarity Measure Based on Adaptive Resonance Theory
2) Image Annotation Using Fusion ART
3) Handling Noisy Text

```c
if (blue_diff < toleration) { blue_diff = 0; }
if (red_diff < toleration) { red_diff = 0; }
if (green_diff < toleration) { green_diff = 0; }
```

### Image and Text Association Presentation

The following description illustrates a process for image and text association:

```c
for x to width {
  for y to height {
    red_diff = red_of_pixel_in_first_image(x, y) - red_of_pixel_in_second_image(x, y);
    green_diff = green_of_pixel_in_first_image(x, y) - green_of_pixel_in_second_image(x, y);
    if (blue_diff < toleration) { blue_diff = 0; }
    if (red_diff < toleration) { red_diff = 0; }
    if (green_diff < toleration) { green_diff = 0; }
  }
}
```
diff_total = diff_total + ((blue_diff + red_diff + green_diff) / 3)}
}
Char *SrcImage tmp = (char*) Image Src;
Char *Src2Image tmp = (char*) Image Src2;
Const char *Image end = (char*) Image Src + width*height*3;
While (SrcImage tmp < Image end)
{blue_diff = *SrcImage tmp++ - *Src2Image tmp++;
red_diff = *SrcImage tmp++ - *Src2Image tmp++;
green_diff = *SrcImage tmp++ - *Src2Image tmp++;
If (blue_diff < toleration) {blue_diff = 0 ;}
If (red_diff < toleration) {red_diff = 0 ;}
If (green_diff < toleration) {green_diff = 0 ;}
diff_total = diff_total + ((blue_diff + red_diff + green_diff) / 3)}

VI. Edge Model For Image and Text Association

The purpose of detecting sharp changes in image brightness is to capture important events and changes in properties of the world. It can be shown that under rather general assumptions for an image formation model, discontinuities in image brightness are likely to correspond to:

- Discontinuities in depth,
- Discontinuities in surface orientation,
- Changes in material properties and
- Variations in scene illumination.

In the ideal case, the result of applying an edge detector to an image may lead to a set of connected curves that indicate the boundaries of objects, the boundaries of surface markings as well as curves that correspond to discontinuities in surface orientation. Thus, applying an edge detection algorithm to an image may significantly reduce the amount of data to be processed and may therefore filter out information that may be regarded as less relevant, while preserving the important structural properties of an image. If the edge detection step is successful, the subsequent task of interpreting the information contents in the original image may therefore be substantially simplified.

However, it is not always possible to obtain such ideal edges from real life images of moderate complexity. Edges extracted from non-trivial images are often hampered by \textit{fragmentation}, meaning that the edge curves are not connected, missing edge segments as well as \textit{false edges} not corresponding to interesting phenomena in the image – thus complicating the subsequent task of interpreting the image data.

Edge detection is one of the fundamental steps in image processing, image analysis, image pattern recognition, and computer vision techniques. During recent years, however, substantial (and successful) research has also been made on computer vision methods that do not explicitly rely on edge detection as a pre-processing step.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{image.png}
\caption{An image depicting Edge properties}
\end{figure}

\textbf{a) Edge Properties}

The edges extracted from a two-dimensional image of a three-dimensional scene can be classified as either viewpoint dependent or viewpoint independent. A \textit{viewpoint independent edge} typically reflects inherent properties of the three-dimensional objects, such as surface markings and surface shape. A \textit{viewpoint dependent edge} may change as the viewpoint changes, and typically reflects the geometry of the scene, such as objects occluding one another.

A typical edge might for instance be the border between a block of red color and a block of yellow. In contrast a line (as can be extracted by a ridge detector) can be a small number of pixels of a different color on an otherwise unchanging background. For a line, there may therefore usually be one edge on each side of the line.

\textbf{b) A simple edge model}

Although certain literature has considered the detection of ideal step edges, the edges obtained from natural images are usually not at all ideal step edges. Instead they are normally affected by one or several of the following effects:

- Focal blur caused by a finite depth-of-field and finite point spread function.
- Penumbral blur caused by shadows created by light sources of non-zero radius.
- Shading at a smooth object.

A number of researchers have used a Gaussian smoothed step edge (an error function) as the simplest extension of the ideal step edge model for modeling the effects of edge blur in practical applications. Thus, a one-dimensional image $f$ which has exactly one edge placed at $x = 0$ may be modeled as:

$$f(x) = \frac{I_r - I_l}{2} \left( \text{erf} \left( \frac{x}{\sqrt{2} \sigma} \right) + 1 \right) + I_l.$$  

At the left side of the edge, the intensity is

$$I_l = \lim_{x \to -\infty} f(x'),$$  

and right of the edge it is

$$I_r = \lim_{x \to \infty} f(x').$$  

The scale parameter $\sigma$ is called the blur scale of the edge.
VII.  Text Processing and Retrieval

To search a body of information (such as a collection of images or citations to the biomedical literature) for objects (scientific articles, images, case descriptions, etc.) relevant to a search query, take the following steps: (1) automatically represent the body of information in the structured form required by our search engine (ITSE) (2) formally represent the information need submitted by a user or inferred from a patient’s case using the Evidence Based Medicine framework of a well-formed question; and (3) translate the formal representation of the information need to a search query using the search engine query language. The body of information for our first research initiative consists of full-text scientific publications, whereas in our second initiative we process any type of free text associated with images stored in databases, or find image-related text in an article or case description that contains images.

Figure 5: Text Processing and Retrieval

a) Extracting image-related text: caption segmentation and mention extraction
   The first step in generation of an enriched citation is finding image captions and mentions. In most cases, finding image captions amounts to parsing the corresponding XML tags in such documents as PubMed Central articles or using regular expressions.

b) Understanding image description: pointers and ROI
   The extracted image-related text is further processed to identify Image Regions of Interest (ROI). ROIs are commonly described in the image caption and indicated by an overlay that facilitates locating the ROI. This is especially true for hard to interpret scientific images such as radiology images. ROIs are also described in terms of location within the image, or by the presence of a particular color.

c) Image representation (conceptual indexing)
   To provide a structured summary of the salient image content akin to that of the MeSH indexing of the biomedical articles, we explored MetaMap-based extraction of the salient indexing terms from the image-related text[9]. Studies show that image captions provide up to 80% of indexing terms (as judged by physicians trained in medical informatics), but additional filtering is needed for acceptable precision.

d) Generating structured documents (enriched citations) for retrieval
   Working with a collection of structured data in XML format provides ready access to document fields, such as title, abstract, conditions, treatment, keywords, etc. Access to document structure supports differential weighting of the occurrences of search terms in different document fields. Structured documents also support faceted search and document and query frame unification (if queries are represented using the same structure). Although the Essie search engine presently supports only the differential weighting of the search terms, creating the structured documents is worthwhile, as we can adjust the field weights for different tasks.

e) Automatic query generation
   The extraction and grouping approaches are based on the four components of a well-formed clinical question: Patient/Problem, Intervention, Comparison, and Outcome (PICO). To construct the PICO frames, Essie, MetaMap or eXtractor (CTX) for clinical narrative to map the text to the UMLS Metathesaurus and extract concepts relating to problems, interventions (drugs, therapeutic and diagnostic procedures), and anatomy are used.

f) Essie
   The Essie search engine, developed and used at NLM, features a number of strategies aimed at alleviating the need for sophisticated user queries[10]. These strategies include a fine-grained tokenization algorithm that preserves punctuation, and phrase searching based on the user’s query. Essie is particularly well-suited for information retrieval tasks in the medical domain since it performs concept-based indexing, automatically expands query terms using synonymy relationships in the UMLS Metathesaurus, and weights term occurrences according to their document location when computing document scores.

VIII. Conclusion

Enriching citations with relevant images can significantly improve literature retrieval for scientific research and clinical decision making. Images are retrieved using image features and visual keywords developed to describe their content. The visual keywords are used to find similar images, followed by IR techniques to improve the relevance of the visually similar images retrieved. Results show no benefit in combining textual and visual features for the retrieval tasks; modality detection is improved when utilizing both text- and content-based approaches. There are two distinct methods for learning and extracting associations between images and texts from multimedia web documents. The vague-transformation based method
utilizes an intermediate layer of information categories for capturing indirect image-text associations. The experimental results suggest that vague method is able to efficiently learn image-text associations from a small training data set. Also, it performs significantly better than the baseline performance provided by a typical image annotation model.

REFERENCES Références Referencias

1. Tao Jiang and Ah-Hwee Tan, Senior Member, IEEE “Learning Image-Text Associations” Manuscript received May 24, 2007; revised January 17, 2008; accepted June 25, 2008.


20. Learning Image-Text Associations Tao Jiang and Ah-Hwee Tan, Senior Member, IEEEIEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 2, FEBRUARY 2009.