2
Background and Related Work

In order to publish Deep Web multimedia data and, as consequence, make this content discoverable and available for search engines, annotations need to be generated in a format suitable for the Web. In this chapter, we review some approaches used to extract features from the multimedia data, as well as recent research aimed at publishing multimedia data annotation.

2.1
A Classification for Multimedia Metadata

To address the problem of multimedia data annotation, many solutions have been proposed in the literature. In the next sections, we present related work that provided some ideas to this dissertation.

To perform the annotation process, different types of information can be associated with images or videos [6]:

- *Content-independent metadata*: is related to the image or video content, but does not describe it directly. Examples are: author's name, date, location, cost of filming, etc.

- *Content-dependent metadata*: refers to low/intermediate-level features (color, texture, shape, motion, etc.).

- *Content-descriptive metadata*: refers to content semantics. It is concerned with relationships of image entities with real-world entities or temporal events, emotions and meaning associated with visual signs and scenes.

2.2 Video Annotation

A large number of approaches have been proposed to perform automatic annotation of video. After analyzing the literature, we use the following classification: *text-based approaches*, *audio-based approaches*, *visual-based approaches* [2, 7, 8].

2.2.1 Text-Based Approaches

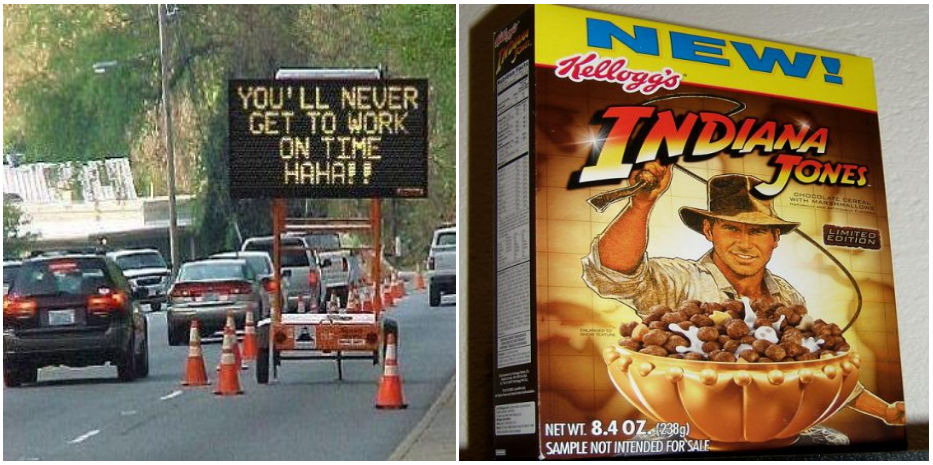
Text data present in video contain useful information for automatic annotation and indexing [9]. Consequently, most commercial video search engines such as YouTube[1], Google Videos[2], Baidu[3], and Truveo[4], use this information to index and retrieve video content. Basically users retrieve videos based on textual information such as description, filename, surrounding text, social tags, closed captions, or a speech transcript [7].

Text-based approaches try to leverage the capabilities of text in the video. This in turn is composed of two subclasses. The first captures the texts that appear immersed in the video, for example, advertisements, names on shirts of the actors, script that is part of the same screen (open caption) [2].

A description of this process is presented by [9], where the authors extract text data present in video. This process is commonly called "*text information extraction"* (usually known by the acronym TIE).

A TIE system receives as input a still image or a sequence of images (see the examples in Figure 1, Figure 2, Figure 3). The images can be in gray scale or color, compressed or uncompressed, and the text in the images may or may not move.

[1] http://www.youtube.com/
[2] http://video.google.com/
[3] http://www.baidu.com/
[4] http://www.truveo.com/

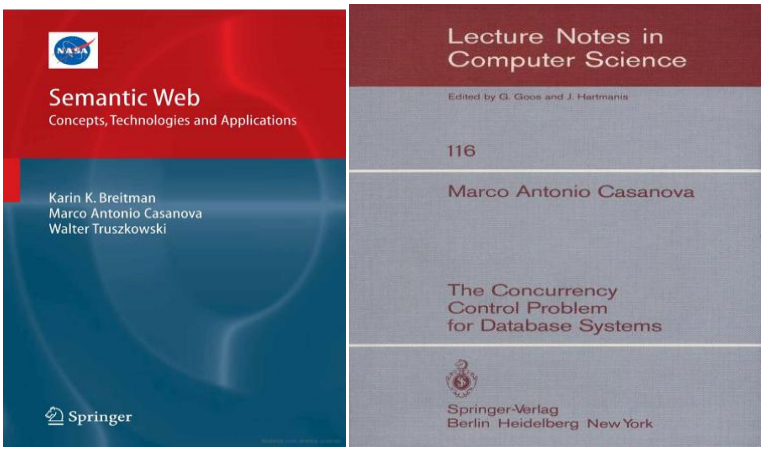


(a) (b)

Figure 1 (a) (b) Scenes text images.

(c) (d)

Figure 2 Shows captions directly on the background.



(e) (f)

Figure 3 (e) (f) Multi-color document images.

The TIE problem can be divided into the following sub-problems (see Figure 4):

- Text detection: refers to the determination of the presence of text in a given frame.

- Text localization: is the process of determining the location of text in the image and generating bounding boxes around the text.

- Text tracking: is performed to reduce the processing time for text localization and to maintain the integrity of position across adjacent frames. Although bounding boxes can indicate the precise location of text in an image, the text still needs to be separated from the background to facilitate its recognition. This means that the extracted text image has to be converted to a binary image and enhanced before it is fed into an OCR (Optical Character Recognition) engine.

- Text extraction: is the stage where the text components are segmented from the background.

- Enhancement of the extracted text components: is required because the text region usually has low-resolution and is prone to noise.

- Text transformation: is the process by which the extracted text images can be transformed into plain text using OCR technology.

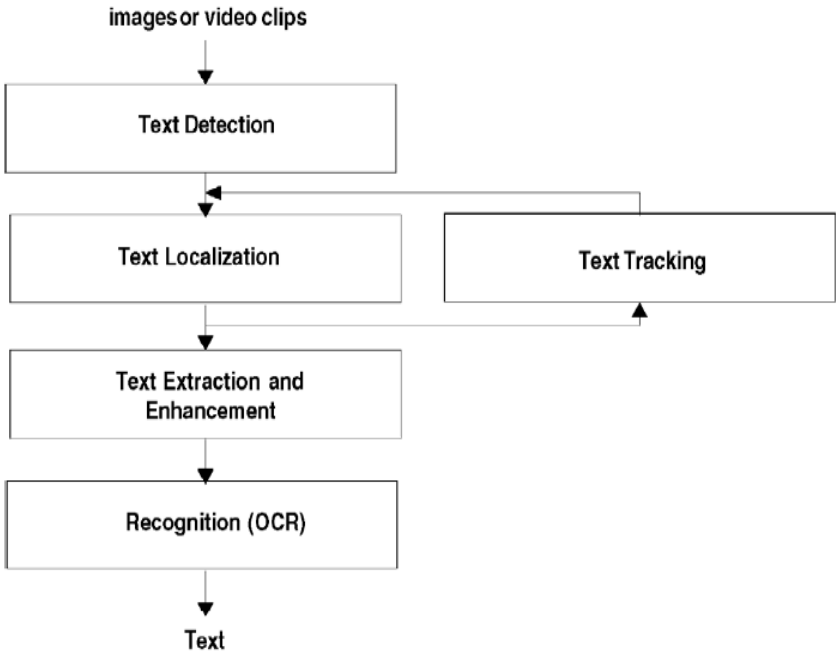


Figure 4 Architecture of a TIE system [9].

The second text-based approach uses the closed caption or legend attached to the video. *Closed captioning* is a term describing the process of displaying text on a television, video screen or other visual display to provide additional or interpretive information to individuals who wish to access it. Also, closed captioning is used as a method for letting hearing-impaired people know what is being said in a video by displaying text of the speech on the screen [10, 2]. The term "closed" in closed captioning indicates that not all viewers see the captions; only those who choose to activate it.

Open captions, sometimes called "burned-in" or "hardcoded", differs from closed captions in sense that open captions are visible to all viewers [10].

Usually, captions are not distinguished from subtitles. However, in the United States and Canada, these terms have different meanings: *subtitles* assume the viewer can hear but not understand the language or accent, or the speech is not entirely clear [10].

When there is no legend, it could be automatically generated using speech recognition tools (ASR) [2]. Speech recognition is used to convert speech waveform into a sequence of words. Hence, this text representation can be extracted from the spoken content.

It is essential to clarify that most ASR today have a very high word error rate (WER) – approximately 35-50% – using Start-of-the-art ASR systems [34, 2], which is normally caused by music and speech overlap. In addition, significant acoustic differences between speakers arise due to anatomical differences. Furthermore, an individual speaker's acoustics may be dependent on factors such as their state of health at the time the recording was made. Finally, a speaker's choice of words and speaking style may exhibit variations that relate to the social context [36]. All these complicating factors make transcribed text suffer from misspellings and omitted words. Nevertheless, since human speech is to some extent redundant, regardless of the aforesaid word error rate, retrieval effectiveness has proven to be fairly robust in the presence of recognition errors. Finally, we note that our aim in this dissertation is not to present an ASR technique, but a publishing process of spoken content that uses ASR services.

Current approaches to speech recognition are statistical. A *statistical speech recognition system* is composed of two subcomponents: the *language model* and the *acoustic model*. The language model governs the generation of word sequences (by estimating the probability of producing any given word sequence). The acoustic model, in turn, describes the generation of the audio signal from a word string. In the practice, these models are inverted to perform speech recognition: given an acoustic signal, find the word that most likely generated it. More precisely, given an acoustic model A, find a word W which maximizes the probability P(W|A) of A being an acoustic model of W [34, 36].

The *speech recognition lattice* is an alternative approach to generate text representation. It receives a set of acoustic models and also a set of acoustic observation, and then the lattice ASR system must apply the model constrains to the set of acoustic observation. This search may generate a very large lattice of possible word sequences. Since the size of the search space is huge, it is generally pruned on-the-fly to optimize the search. The lattice thus generated contains information about model scores and word timing information used by the recognizer.

Figure 5 illustrates a speech recognition lattice. In this lattice, each arc includes a word label and the probability of that arc being followed from the previous state. The best sequence of words that can be found in the lattice is

usually called the *1-best result*. If the result is not accepted, an alternative solution can be chosen from the lattice; this new search will produce the N-best sentence.

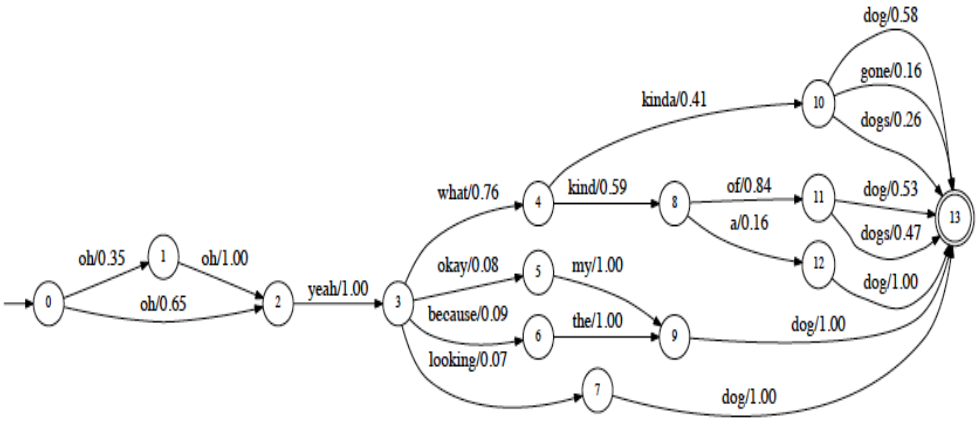


Figure 5 Example speech recognition lattice.

2.2.2
Audio-Based Approaches

The second approach uses audio features to indexing and retrieving videos. Given any audio piece, we can instantly tell the type of audio (e.g., human voice, music or noise), speed (fast or slow), the mood (happy, sad, relaxing etc.) and determine its similarity to another piece of audio. However, a computer sees a piece of audio as a sequence of sample values [11].

The first step to produce audio features from the video is extracting the audio signal. The signal is sampled at a certain rate (e.g., 22050 Hz) [2], then features are extracted in two levels: *short-term frame level* and *long-term clip level*.

A frame is defined as a group of neighboring samples which last about 10 to 40 ms. For a feature to reveal the semantic meaning of an audio signal, analysis over a much longer period is necessary, usually from one second to several tenths of seconds. Here we call such an interval an *audio clip*, which consists of a sequence of frames. Clip-level features usually characterize how

frame-level features change over a clip [12]. An example of frames and clips is shown in Figure 6.

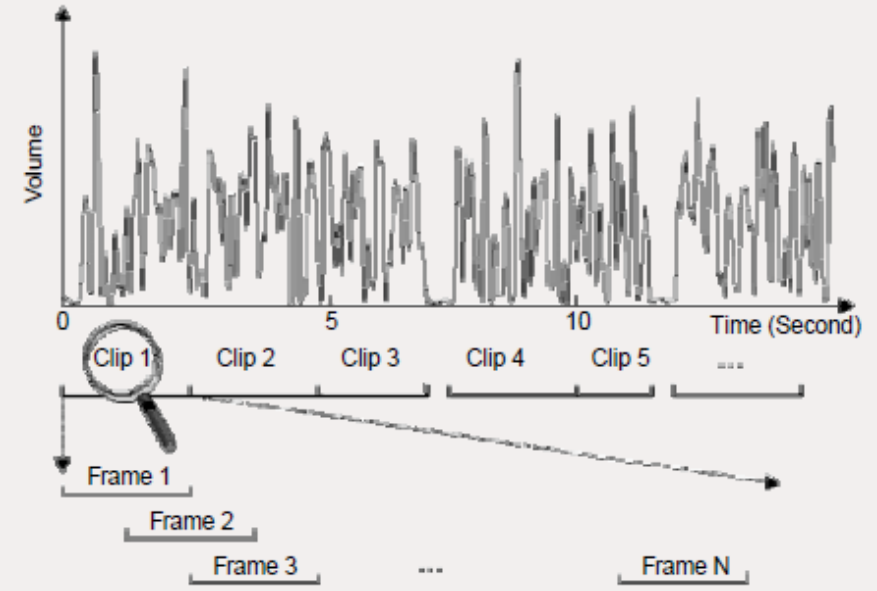


Figure 6 Decomposition of an audio signal into frames and clips [12].

Audio signals can be represented in the time domain (time-amplitude representation) or using the Fourier transform. A signal in the time domain can be transformed to the frequency domain (frequency-magnitude representation) [11, 2].

We review some of the features extracted from the time domain and the frequency domain.

Time Domain

The following features are calculated from the audio waveforms directly without any transformation:

- Volume
  - The most widely used and easy-to-compute frame feature is volume. (Volume is also referred as loudness, although strictly speaking, loudness is a subjective measure that

depends on the frequency response of the human listener.) Volume is a reliable indicator for silence detection, which may help to segment an audio sequence and to determine clip boundaries.

- Zero cross rate
    - To compute Zero crossing rate (ZCR), we count the number of signal amplitude sign changes in the current frame. Higher frequencies result in higher zero crossing rates. Speech normally has a higher variability of the ZCR than music. If the loudness and ZCR are both below thresholds, then this frame may represent silence.

- Silence ratio
    - The silence ratio indicates the proportion of the sound piece that is silent. Silence is defined as a period within which the absolute amplitude values of a certain number of samples are below a certain threshold. Note that there are two thresholds in the definition. The first is used to determine if an audio sample is silent. But an individual silent sample will not be considered as a silent period. Only when the number of consecutive quiet samples is above a certain time threshold are these samples considered to make up a silent period. The silence ratio is calculated as the ratio between the sum of silent periods and the total length of the audio piece. Speech normally has a higher silence ratio than music. News has a higher silence ratio than commercials.

Frequency Domain

These features are the result of the Fourier transform:

- Energy distribution.
    - From the signal spectrum, it is very easy to see the signal distribution across the frequency components. For example, we can see if the signal has significant high frequency components. This information is useful for audio

classification because music normally has more high frequency components than speech. So it is important to calculate low and high frequency band energy. The actual definitions of “low” and “high” are application dependent. For example, we know that the frequencies of a speech signal seldom go over 7 kHz. Thus we can divide the entire spectrum along the 7 kHz line: frequency components below 7 kHz belong to the low band and others belong to the high band. The total energy for each band can be calculated as the sum of power of each sample within the band.

- Bandwidth
  - The bandwidth indicates the frequency range of a sound. Music normally has a higher bandwidth than speech signals. The simplest way of calculating bandwidth is by taking the frequency difference between the highest frequency and lowest frequency of the non-zero spectrum components. In some cases “non-zero” is defined as at least 3 dB above the silence level.

- Pitch
  - Pitch is the fundamental frequency of an audio waveform and is an important parameter in the analysis and synthesis of speech and music. Sounds can be ordered according to the levels of pitch. Most percussion instruments, as well as irregular noise, do not give rise to a sensation by which they could be ordered. But we can still use pitch as a low-level feature to characterize the fundamental frequency of any audio waveforms. The typical pitch frequency for a human being is between 50-450 Hz, whereas the pitch range for music is much wider.

2.2.3 Visual-Based Approaches

Most of the video annotation approaches found in the literature are based on visual elements. This corresponds to the fact that humans receive much of

their information about the world through their sense of vision [2]. Visual features are extracted in three levels: frame, shot and scene. Very few approaches extract features from scenes because the complexity increases from frame to scene.

A *frame* is a still image composed of a set of pixels, where each pixel has a value (RGB); a *shot* can be represented by a single frame, known as the *keyframe*. Typically the *keyframe* is the first frame of a shot, although some authors use the term to refer to any single frame that represents a shot [2]. Also, a shot can be a set of consecutive frames. A *scene* is defined as a set shots, but it has a semantic meaning.

- Color-Based Features
    - Color histogram represents the color distribution of an image and is one of the most widely used color features. Many color spaces exist for representing the colors in a frame. Two of the most popular are the Red-Green-Blue (RGB) and Hue Saturation Value (HSV) color spaces [12]. The main drawback of histograms for classification is that the representation is dependent on the color of the object being studied, ignoring its shape, texture and spatial configuration. Color histograms can potentially be identical for two images with different object content which happens to share color information. Conversely, without spatial or shape information, similar objects of different color may be indistinguishable based solely on color histogram comparisons. There is no way to distinguish a red and white cup from a red and white plate [2].
- Texture
    - Whereas color is a property that can be measured at every pixel, texture is an important feature of a visible surface where repetition or quasi-repetition of a fundamental pattern occurs. Texture is of importance in classifying different materials, such as the line-like pattern in a brick wall, or the dot-like pattern of sand. In general, there will be different colors or textures in a key frame. This means that there will be many frame locations where there is a significant change

in visual data, in particular, a change in color or texture. [7, 12].

- Shape
    - Shape features can be represented using traditional shape analysis such as moment invariants, Fourier descriptors, autoregressive models, and geometry attributes. They can be classified into global and local features. Global features are the properties derived from the entire shape. Examples of global features are roundness or circularity, central moments, eccentricity, and major axis orientation. Local features are those derived by partial processing of a shape and do not depend on the entire shape. Examples of local features are size and orientation of consecutive boundary segments, points of curvature, corners, and turning angle [12].

2.3 Image Annotation

A large number of approaches have been attempted for image retrieval. In the literature, we found two basic approaches: *annotation-based image retrieval* (ABIR), also known as *text-based image retrieval*, and *content based image retrieval* (CBIR). ABIR is based on image textual descriptions, whereas CBIR performs the same process based on visual features, like color, texture, shape, etc. [13, 14, 15].

2.3.1 Text-Base Image Retrieval

The text-based approach can be tracked back to 1970s [15]. In such systems, some form of textual description of the image content is assumed to be stored with the image itself. Such descriptions are usually based on annotations made by human beings. To query the database, the user provides a keyword description of his information need. This description is then compared with the descriptions of the stored images, using text retrieval techniques [14].

Almeida et. al. [14] evaluated which parts of a Web document can be used to enhance an effective description of the images in the document. They first proposed an image retrieval model, based on Bayesian belief networks. Then, they considered several sources of text-based evidences within a Web document (URL, ALT, Page title, Image title, Surrounding text, etc.) to be used and combined through a belief network. Finally, they carried out four experiments using the aforesaid sources to determine which of those sources help to improve the ranking of the images in the web. They concluded that the combination of description tags with 40-term segment information provides the best description of an image, and this information can be used by an ABIR system.

Vani et. al. [6, 16] identified three types of image annotations: free text annotation, keyword annotation, annotations based on ontologies.

- Free text descriptions: in this approach the annotator can use a combination of word or sentences to describe the images. It seems easy to annotate, but in some cases it is difficult to describe the image content and the description can be subjective.

- Keywords: this process can be performed in two ways. The annotator chooses arbitrary keywords or he can restrict the keyword scope using a controlled vocabulary. Further, the keyword can be related with the image in two levels: the whole image or a single image region.

  Some problems arise using this approach. One is that it does not exist an unique vocabulary or an annotation standard to generate keyword; hence different images collections are annotated using different vocabularies. Thus, it is not easy, for a common user, to query an image collection, if he does not know the vocabulary used to annotate the images.

- Annotation based on ontologies: this process can solve the problem of keyword ambiguity. Indeed, ontologies contain concepts (entities), relationships and rules. For example, a puma could be an animal, a brand or a singer, etc.

In [17], the authors stated that using annotation alone does not establish the semantics of what is being annotated. To solve this problem, the authors explain the advantages of using Semantic Web languages and tools for the creation, storage, manipulation, interchange and processing of image metadata. The main idea is to associate resources with ontologies. They described the annotation process using images from several domains (cultural heritage, personal digital photo collections, television news archive, image collections at NASA, biomedical images). They concluded that Semantic Web technologies are sufficiently generic to support annotation of a wide variety of images in different domains. There is, however, a lack of commonly accepted, widely used vocabularies for image annotation. Also, a standard to address sub regions within an image is still missing.

This annotation process can be performed in an automatic or in a semi-automatically manner. Automatic annotation (also know as *automatic image tagging*) is performed using machine-learning techniques that learn the correlation between images features and textual values from examples of annotated images [18]. Semi-automatic annotation uses the automatic annotation process to generate annotations that are refined by encouraging the user to provide feedback while examining the retrieval results.

There are two disadvantages with manual text-based annotation. The first is that it is extremely labour-intensive, expensive, tiresome and time-consuming task. The second is that the annotation process can be subjective, since annotations depend on human perception. Nevertheless, text-based annotations are very effective and accurate.

2.3.2 Content-Base Image Retrieval

To solve the drawbacks in text-based retrieval system, *content-based image retrieval* (CBIR) was introduced in the early 1980s. In CBIR, images are indexed by their visual content, such as color, texture, shapes [18].

Low-level image feature extraction is the basis of CBIR systems. To performance CBIR, image features can be either extracted from the entire image or from regions. These features were reviewed in section 2.2.3.

Bosch A. et al. [19] presented a survey of the existing approaches to scene classification. They identified that low level methods can be classified in those that model the image as a single object and those that model a partition of the image in sub-blocks (see Figure 7).

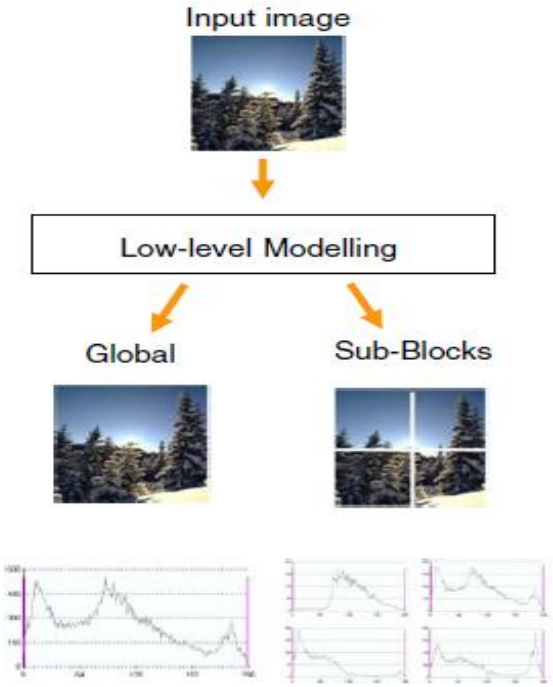


Figure 7 Scene classification. Low level.

In low level methods that model the image as a single object, the scene is described by low-level features from the entire image. Binary Bayesian classifiers attempt to capture high level concepts from the low-level features. For example, images can be classified as indoor or outdoor; outdoor images are further classified as city or landscape; finally, a subset of landscape images is classified into sunset, forest class. Depending on the classification problem, different measures can be extracted from the whole image (using spatial color moments, edge direction coherence vectors, etc.).

In low level methods that partition of the image in sub-blocks, the image is partitioned into several blocks, and the features are extracted from block level. Classifiers are employed to classify each sub-block based on the above features and finally the whole image is classified using a majority voting scheme from the

sub-block classification results. They used color, shape and texture features in combination with supervised learning methods to classify images into semantic classes (city, landscape, sunset, etc.).

Liu Y. et. al. [15] identified the state of the art techniques that try to narrow the semantic gap between the low level and high level features.

In techniques based on object ontology, the semantic is extracted in the follow way. First, intervals are defined for the low level features, and each interval is represented as an intermediated-level descriptor of images. The set of descriptors form a simple vocabulary, the so-called *object-ontology*, that supplies a definition of high-level query concepts. For example, 'sea' can be defined as a region of 'light blue' (color), 'multiform' (texture), and 'low' (spatial location).

Techniques based on relevance feedback (RF) try to learn the user's intention on the fly. With the user intention, it is possible to reduce the semantic gap between what queries represent (low-level features) and what the user thinks (high-level features). The process of a RF in CBIR is as below:

(a) The system provides initial retrieval results through query-by-example, sketch, etc.
(b) The user judges the above results as to whether, and to what degree they are relevant (positive examples) or irrelevant (negative examples) to the query.
(c) A machine learning algorithm is applied to learn the user's feedback. Then go back to Step (b).
(d) Steps (a)(b)(c) are repeated until the user is satisfied with the results.

2.4 RDF-a

The Resource Description Framework – in attributes (RDFa) is a technique that provides a set of markup attributes to augment the visual information on the Web with machine-readable hints [33].

These attributes are:

- "typeof": this specifies the RDF type(s) of the subject.
- "content": optional attribute that overrides the content of the element when using the property.

- “datatype”: this specifies the datatype of text specified.
- “property”: specifying a property for the content of an element.
- “about” and “src”: a URI that defines the resource the metadata is about.
- “rel” and “rev”: specifying a relation with another resource.
- “href” and “resource”: specifying the partner resource.

The following example, shows how to use the attribute “content”, here, although the string "2012-08-22" unambiguously identifies a date for a machine, it does not looks very natural for a human reader. Surely an English reader would prefer something like "20th of October, 2012".

```
<p>Date: <span property="http://purl.org/dc/terms/created"
content="2012-10-20">20th of October, 2012</span></p>
```

In this example is presented the attribute “datatype”. It is defined a date using "xsd:gYear" stands for http://www.w3.org/2001/XMLSchema#gYear, this is a standard datatypes defined by W3C's Datatype specification [33].

```
<span property="dc:date" datatype="xsd:gYear">2012</span>
```

RDFa enables Web publishers to do just that. Using a few simple HTML attributes, Web publishers can mark up human-readable data with machine-readable indicators for browsers and other programs to interpret. We assume that if a XHTML document is enriched with machine-readable data it will be significantly more discoverable, and therefore more usable.

2.5 Related Work

In this section, we discuss related research that addresses the problem of multimedia retrieval and publishing. We begin by observing that, in the last decade, content-independent metadata [20,21] has been used as an alternative to describe multimedia content [2]. Audio, video and image repositories [22,23,24] on the Web typically follow a similar approach. Larson et al. [25] explore the automatic generation of tagging and geotagging to improve video retrieval.

Alberti et. al. [26] address the spoken content retrieval problem in the context of last US presidential campaign. They generate the transcription of each asset using an automatic speech recognition service (ASR). The ASR service segments the audio, discards music and noise and then performs the transcription. Then, this content is used to “HTML-ize” the spoken document by filling in an HTML template with low level metadata, such as title and description. Although their approach uses content-descriptive metadata and content-independent metadata to describe spoken content, the content is not prepared to be machine-readable.

Repp et. al. [3] assert that the number of digital video recording has increased dramatically since recording technology became easier to use. They also comment that manual annotating is time-consuming and useless. To solve this problem, they present two algorithms to generate automatic semantic annotations for university lecturers, using features obtained from speech transcription, power point slides and closed caption. The complete approach uses a specific ontology to annotate content and makes it available using OWL-DL for semantic search engines. Although they provide semantic information to the search engines, the same semantic information is not available to assist humans on the Web.

Van et. al [29] argue that the Web is being transformed from a text-only medium into a more multimedia-rich medium. Thus, it is necessary to perform searches based on multimedia content. To address this issue, they automatically generate transcriptions of thousands of talks and news radio shows, using a speech recognition system. The transcription generated is used by an audio indexing and retrieval system for the Web.

Glass et. al. [27] discuss the components involved in a spoken content retrieval system to help finding spoken lectures. However, in their approach, content is still hidden to search engines, since the transcription files can only be accessed through a search form.

Finally, we observe that the work reported here is aligned with the activities of the W3C Media Annotations Working Group [28].