# Content Without Context is Meaningless

Ramesh Jain
Dept of Computer Science
University of California, Irvine
jain@ics.uci.edu

Pinaki Sinha
Dept of Computer Science
University of California, Irvine
psinha@ics.uci.edu

## ABSTRACT

We revisit one of the most fundamental problems in multimedia that is receiving enormous attention from researchers without making much progress in solving it: the problem of bridging the semantic gap. Research in this area has focused on developing increasingly rigorous techniques using the content. Researchers consider that *Content is King* and ignore everything else. In this paper, first we will discuss how this infatuation with content continues to be the biggest hurdle in the success of, ironically, content based approaches for multimedia search. Lately, many commercial systems have ignored content in favor of context and demonstrated better success. Given that the mobile phones are the major platform for the next generation of computing, context becomes easily available and more relevant. We show that it is not Content Versus Context; rather it is Content and Context that is required to bridge the semantic gap. In this paper, first we will discuss reasons for our approach and then present approaches that appropriately combine context with content to help bridge the semantic gap and solve important problems in multimedia computing.

## Categories and Subject Descriptors

I.4.m [**Image Processing and Computer Vision**]: Miscellaneous; H.4.m [**Information Systems Applications**]: Miscellaneous

## General Terms

Algorithms, Theory

## Keywords

context, content, image, perception, search, mobile, exif

## 1. INTRODUCTION

Multimedia content research community is facing the risk of becoming irrelevant. We need to critically examine our approaches and study why we are in the vicious circle of solving problems that nobody outside our own community cares. In this paper, we will examine the state of art in academic multimedia content research community and suggest that we approach important problems from a different perspective. We need to liberate ourselves from our current *koopmanduk* (Frog-in-the-Well) mentality, otherwise all our research will only result in making our approaches irrelevant to the mainstream computing community. In this paper, we propose to address one of the most fundamental problems: bridging the semantic gap. Based on research and emerging technology from multiple related areas, we adopt a new out-of-the-box perspective to bring revolutionary changes in the current research paradigm. At the first sight, it may appear to be something that is known, but we will show that despite a lot of lip-service to the use of context, it is mostly ignored. The current situation is exactly like that in the famous story: *The Emperor's New Clothes* [4].

In the last two decades, multimedia computing has evolved to become the dominant main stream, first in computing and now in mobile computing. The Web has clearly become more multimedia oriented. The popularity of mobile computing is because of numerous sensors in mobile phones which makes it a better audio-visual-interactive client than a personal computer. A simple analysis of the Web shows that in the last few years some remarkable multimedia technology has changed how people communicate. From multimedia centric perspective, the first major thing to arrive on the scene was the IPod which brought with it audio technology that was a major revolution. Next came the photo sharing service: Flickr. That has now become a major source of research for the image retrieval community. YouTube brought video to the mainstream and is now becoming a dominant source of data for video retrieval research. Facebook demonstrated that medium is no longer the message in mass communications. People use appropriate medium to communicate their message. The latest to become popular is Twitter which started with messages of 140 characters to encourage real time communication, but soon introduced links to text, pictures, videos, and audio to make real time mass messaging using appropriate medium.

Paradoxically, though multimedia researchers failed to contribute to emergence of multimedia in computing systems, they are becoming increasingly dependent on using these systems for their research. A simple look at any Multimedia Conference Proceedings clearly demonstrates that a significant fraction of the papers are related to the data from sites mentioned above [1]. We all use image, video, and audio search from popular search engines for finding content that we need. Interestingly, we write thousands of research papers on multimedia information retrieval, but it does not bother us that techniques developed by our community are rarely, if at all, used by the very search engines that we use.

[1] In ACM Multimedia Conferences 2008 and 2009 there were eight papers with the words Flickr or YouTube in the title. Many more papers must have used Flickr or YouTube datasets.


*MM'10,* October 25–29, 2010, Firenze, Italy.

We are at the beginning of a major new revolution that will make computing primarily multimedia. This revolution is being pioneered by mobile phones that are primarily multimedia capture and display devices that go beyond just audio-visual media to include tactile, GPS, accelerometers, gyroscopes, and several other interesting sensory sources. This makes it the richest multimedia device ever developed in large scale. Consider a very obvious situation, shown in Figure 1. There are several sources that feed different kinds of contextual information to a phone, ranging from location information to calendar, contacts, and information in clouds. When a photo is taken using this camera (and in near future most photos will be taken using such devices) one can effortlessly add information like EXIF++ as shown in the figure.

If multimedia content is exploding and is likely to continue, and all this research by very bright researchers from all over the world in content analysis is not being used in managing the content, then clearly there is something fundamentally wrong with our current thinking. Every community faces such moments in their history. These are the moments when a paradigm shift is required to adopt a new perspective to shape the future. Multimedia content community is facing this problem and must act if it wants to contribute to solving real problems and contribute to the progress of multimedia content management and access.

## 1.1 Machine Learning Hammer

Mark Twain once said: "To a man with a hammer, everything looks like a nail." His observation is definitely very relevant to current trends in content analysis. We have a Machine Learning Hammer (ML Hammer) that we want to use for solving any problem that needs to be solved. The problem is neither with learning nor with the hammer; the problem is with people who fail to learn that not every problem is a new learning problem [1]. Clearly, content analysis uses decisions at every level, starting from the lowest level of feature detection. In fact, every decision assumes existence of a model. For example even edge detection assumes a step discontinuity in intensity values or some other characteristics. The famous object recognition problem fundamentally tries to see whether a given feature pattern satisfies a model representing an object. The complexity in object recognition increases as the number of objects increase. The most difficult part in object recognition is defining models of objects clear enough so that each object occupies a distinct area in the feature space. This problem also requires identifying measurable features which will result in providing distinct areas in the feature space for each of the given set of objects. If we can identify such a feature set, then we can easily model each object by its appropriate feature values. The challenges are

- to identify a right set of features
- to identify feature values for representing each object

In reality, both problems are related. There is a right set of features for recognizing a given set of objects. Most of the content analysis focuses on the second problem and assumes that they have to live with a given set of features (such as color, texture, and shape) and try to use machine learning techniques for solving the second problem. This is because content analysis people discovered machine learning (because supervised and unsupervised learning approaches for classification have been around for more than at least 40 years [10] ) as a convenient hammer. Progress in storage and processing technology has facilitated application of solving the model building process. Unfortunately, we ignore the first problem and use our ML Hammer on whatever problem we are given. Surprisingly, we are happy even when we get (in most cases) 20% -30% accuracy in the results (the average precision of object detection in Pascal Challenge 2009 is in this range [11]).

## 1.2 Solving The Right Problems

Let us paraphrase a famous story in the context of this paper. A drunk multimedia researcher loses the keys to his house and is looking for them under a lamppost. Another researcher comes over and asks what he's doing. "I'm looking for my keys" he says. "Let me help you", says the new researcher and joins the effort. Soon there are many researchers looking for the keys. One of them got frustrated and asks: "where did you lose your keys". The original researcher replies, "I lost them over there", and points to a dark corner in the street. The new scientist looks puzzled. "Then why are you looking for them all the way over here? ", he asks. "Because the light is so much better here. We can formulate and solve the search problem much better here. Over there it is not easy to formulate because you can not see well", replies the original researcher. Finding the explanation reasonable, all researchers keep looking for the keys under the lamppost. After long rigorous and exhaustive efforts they conclude that the problem of finding lost keys is an unsolvable problem.

A famous real story is related to the milkshake by McDonald's [26]. McDonald's wanted to make their milkshake as a more effective product. A team of marketing researcher started analyzing standard statistical techniques, to find the taste, thickness, temperature and other basic features of milkshake to find what most people like. One researcher decided to ignore the features and study why people buy milkshake. The findings were startling. People bought milkshake not for taste but for giving them company over long drives without being messy to consume and being a good companion for long periods. In content analysis also, one needs to really understand why a particular media source is used and what need does this really address.

# 2. REVISITING THE MULTIMEDIA CONTENT PROBLEM

Multimedia content analysis is fundamentally the perception problem. In any perception problem, there are three components that must be considered:

- The data related to an environment
- The medium used to transmit physical attribute to the perceiver
- The perceiver

Human perceptual system has been explored from early days by philosophers from various perspectives. It has been very well realized, and rigorously articulated and represented, that we understand the world based on the sensory data that we receive using our sensors, and the knowledge about the world that we have accumulated since our birth. Both the data and the knowledge are integral component of the understanding. Before proceeding further we would like to point out one of the most important concepts from Perception that is very relevant for multimedia computing. Almost two century ago George Berkeley [6] asked: *If a tree falls in a forest and no one is around to hear it, does it make a sound?* Usually, sound is defined as the sensation excited in the ear when the air or other medium is set in motion. Thus, if there is no *receiving ear* then there is no sound. In other words, perception is not only data; it is a close interaction between the data and the perceiver.

Let us revisit the multimedia computing from fundamentals, because sometimes we keep doing something so long that we forget

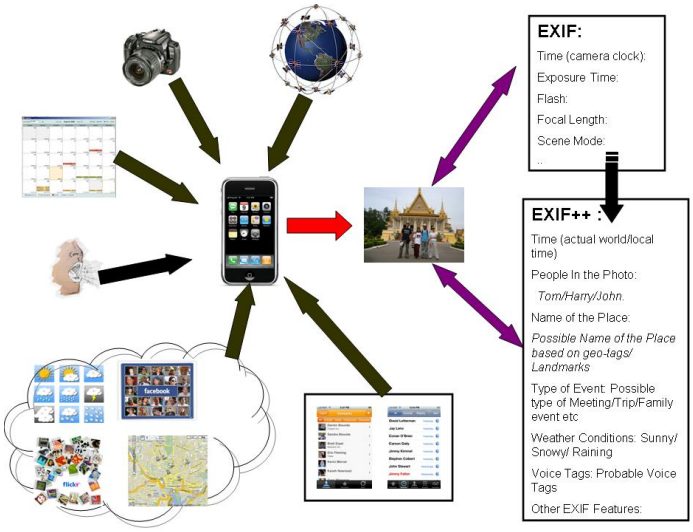


**Figure 1: Showing how each photo by modern cameras will be augmented with very rich metadata, that we call EXIF++**

why we started doing it in the first place. What is the multimedia content problem? In fact, where does the multimedia data come from? Why do we even need multimedia? Multimedia data, such as visual (photos and videos), aural, and other sensory data are captured for an event that unfolds over time. Each medium represents a particular physical spatio-temporal attribute of the event. Of course, an event represents changing relationships among objects and these are captured by different media. The data captured by any kind of sensor really represents these spatio temporal physical attributes of the environment. Objects are part of the environment and their physical attributes are also captured by the media. It is often forgotten that each sensor only captures one type of physical attribute from its *perspective* afforded from its physical location, including its orientation. Multiple sensors could be combined to create a synchronized signal representing the composite data obtained from these sensors. Thus, one uses appropriate number and types of sensors to capture all attributes of the event that may be of interest in a particular application. Multimedia is the right approach to capture event information and experiences. This is because each medium captures only one physical attribute and taken as a whole, the multimedia stream is capable of combining the correlated and complimentary information from individual streams to provide more holistic information and experience than possible using any one medium. None of the individual medium, including the most sophisticated human senses (the vision), can capture holistic experience in most applications. This is no accident that humans have five senses and combine them to experience events in the real world.

Equally important is the fact that each sensor captures data about the environment from its position and perspective. If its position or perspective is changed, then the data and experience also change. For interpretation of the data, one must know the position and perspective. Moreover, many sensors, like cameras, have several other parameters (e.g., focal length, aperture diameter, flash, etc.) that determine the capture of the data and hence they are very important in understanding and analyzing the experience represented by the data.

The most important component in multimedia computing systems also happens to be the most ignored component: *the user*. Each user is unique and while interacting with a system, the context may be different. Interpretation of the data is not only user dependent but also dependent on the context of the user. It is a common knowledge that if you give the same photo to different people and ask them to assign tags to represent the photo, there may be as many different tags as the number of people assigning tags. Moreover, many studies have demonstrated if you give the same photo to a person at two different times in different contexts, then the tags assigned are different. The concept of Rorschach [12] tests is based on the theory that an interpretation of data is as much, or more, dependent on the person than the data.

## 3. CONNECTING DATA AND USERS

Multimedia computing addresses a problem that many other fields like computer vision, databases, and information retrieval face: connecting data and users. As shown in the Figures 2 and 3, data exists in many forms: ranging from bits to alphanumeric documents to photos and videos. On the other hand, users of the data in a modern computing environment may come from many different education backgrounds, of different cultures, and of different socio-economic status. The challenge is how to connect a user with a data source so the user can use the data he needs to solve his application. A key point to remember is that a user is never interested in what and where the data is; she is only interested in solving the problem at hand.

The major hurdle in connecting users to the data is often referred to as the semantic gap. The term was first used in connection with going beyond query by example [25], but was better defined later in [29], where it was stated: *We opine that most of the disappointments with early retrieval systems come from the lack of recognizing the existence of the semantic gap and its consequences for system setup. The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation. A linguistic description is almost always contextual, whereas an image may live by itself.*

To understand semantic gap, let us consider Figure 2. This figure shows that the data operations in a computer start at the bit level and can be structured to represent various data concepts such as documents, photos and videos. A user on the other hand always thinks in terms of objects and events and builds other concepts based on

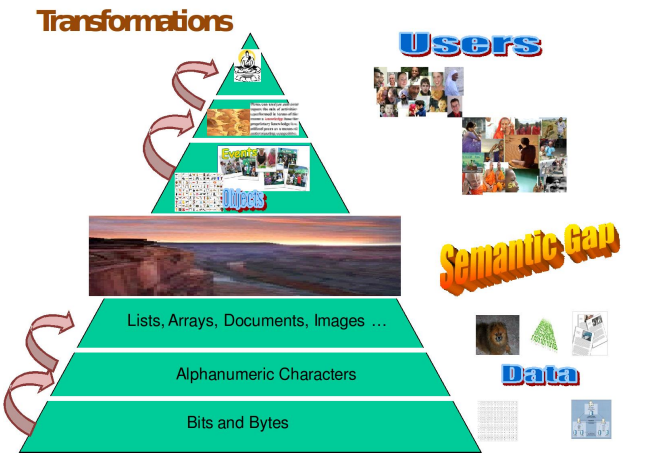


**Figure 2: Semantic Gap Between Users and Data**

the basic notion of objects and events. The transformation of data level concepts such as photos to user level concepts such as objects and events is the challenge that must be solved by content analysis, and used in content organization and retrieval.

## 4. CONTENT AND CONTEXT

Content has received significant attention by multimedia research community. Content analysis leading to content-based retrieval represents majority of research done in multimedia and related fields like computer vision. By content, it is commonly understood that we refer to substantive information perceived by a user in the data that is represented by a particular file. Thus a photo may contain a person standing near a car next to a house. The challenge faced by content analysis is the famous problem of pattern recognition. One faces this problem in all sensory data. The problem is to segment the data into meaningful parts and to use known models of objects of interest to label all segments of an image. And this is where one runs into a tricky problem: We need segments to recognize objects but we also need objects to segment the data. It is possible to formulate this problem such that one can use models of potential objects to segment and then see how best the segments fit object recognition. One may potentially use an optimization framework to accomplish this. The problem gets complicated and almost intractable because in some cases, like in images, a higher dimensional space (e.g., 3D) is mapped into 2-dimensional space resulting in loss of information, making the problem impossible to solve unless some strong assumptions are made. Moreover, in many sensors the signals from multiple objects get added making it almost impossible to solve the problem. The only way to simplify the problem appears to be to reduce the number of potential objects that could be in the data and other information that can help in using appropriate parameters to filter noise from the data. No wonder, people have been trying to solve the mystery of human perception for several centuries and are still without a clue. Closer to multimedia, people have been working on image recognition and speech recognition (note only speech recognition, not audio recognition) and are still far from being close to solving these problems even with the powerful computing infrastructure that we have today. The successful solutions usually are for limited domains, meaning the number of objects is limited in those applications, making the problem more tractable. Let us look at a related concept: context. Context is defined in standard dictionaries and reference sources as:

- The set of circumstances or facts that surrounds a particular event, situation, etc.
- The interrelated conditions in which something exists or occurs: environment or setting.
- Determinant of meaning.

In technical areas, context started receiving attention in the last decade and has been receiving increasing attention. A review of context is provided in [19]. Surprisingly, people in multimedia (and computer vision) try to analyze content with minimal use of context. It appears that content analysts assume that content must be analyzed independent of the context. This is intriguing, considering that many researchers try to derive inspiration and ideas from cognitive psychology. Irwin Rock [24] and Richards [15] have strongly championed the role of knowledge in many different forms in visual perception. They believed that context plays at least as important, in most cases more significant role, as content.

Most people referring to semantic gap ignore the primary reason for the gap. Human sensory processing uses context extensively. Many philosophers and cognitive scientists, including one of the most noted in the 20th century, Karl Popper [22] and Ulric Neisser [21] have created models of all human actions that include context and prior knowledge about an application as an integral component of understanding data. Media processing research, however, has focused on content assuming that interpretation can be done based only on the data values. Researchers trying to bridge the semantic gap often forget that *the linguistic description is always contextual, whereas an image may live by itself.* Looking at the research in media processing, it appears that researchers want to avoid context and want to use only content. Consider a simple case to understand how context can significantly simplify analysis: A photo is taken and needs to be interpreted. If one knows when the photo was taken and at that time what was the illumination level in the scene, one could use appropriate parameters for segmentation and interpretation of images. Moreover, if the photo was taken in Iowa, one should not expect beaches or mountains.

Modern digital cameras have become very sophisticated event capture devices. Unlike their predecessors, these cameras not only capture intensity values but also many scene parameters in EXIF [18]. All these parameters have rich information about the scene which is captured in the projection as the photo. In many cases, some of the camera parameters like depth of field or field of view represent the intent of the photographer as to what she wants to capture. Thus, digital cameras are getting closer to the active human eye that projects an image but also knows the conditions under which the image was acquired.

Our belief to use context was inspired by several researchers in psychology of perception, neurophysiology, and cognitive sciences. Many notable researchers like David Waltz [31], Irwin Rock [24], Richard Gregory[15], and V. S. Ramachandran [23] have emphasized the importance of using knowledge and context in perception. A perceiver is at least as important as the data. We believe that the context or knowledge that could be used in analyzing photos comes in five different classes:

1. Context in Content: Relationship among different objects and even in their subparts in real world can be utilized in analysis of data. This has been studied in early days in computer vision and has started receiving attention again [8].

2. Device Parameters: Environmental parameters of the digital devices at the time of photo taking play an important role in

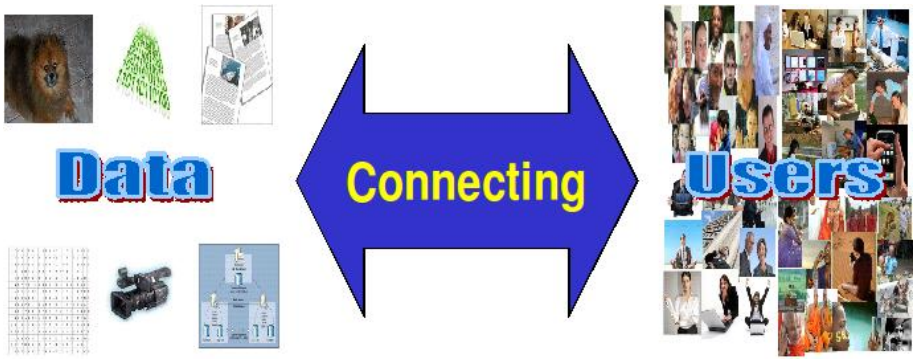


**Figure 3: The Connection Between Data and User**

the analysis of data. Unlike their predecessors, modern digital cameras are more *event capture* devices rather than photographic devices. All EXIF parameters represent information about the event for which a photo or a video represents experiential data. This information is essential for interpretation and organization of data. Search engines know the importance of context data such as creation date of the document, author of the page, domain of the site, and other such information and use this in their ranking algorithms effectively to find relevant pages. This sensor based information can further be fused with data from the web to generate a really rich multimedia platform (Figure 1).

3. Data Acquisition Context: Knowledge about the person taking photos, location, and environmental conditions at the time of photo acquisition (e.g., sun angle, cloudy, rainy, night, indoor, etc.) affect the content of the image. These could be easily used in analysis.

4. Perceiver: Cognitive scientists know the importance of the perceiver. Rorschach tests are a clear demonstration of the knowledge and personality of the perceiver in interpretation of visual data. Surprisingly, computer vision and multimedia community has never realized the importance of this. All research effort by search companies to prospect click-stream data for building personal profiles to help in presenting relevant results to users shows that at least in the context of text-search they understand the importance of this knowledge source.

5. Interpretation Context: Real world situation in which the data is interpreted results in focus on different aspects of the data. A botanist looks at a garden with different goals and interprets it differently than a person interested in enjoying the beauty of flowers.

## 4.1 Why is Context Useful?

Most of the challenging problems in multimedia analysis fall in the genre of recognition and classification. We have to define features for the multimodal data, compute similarity measures between them, and then build models for various tasks. All these tasks entail some variation of the comparison operation. This task becomes exceedingly difficult given the amount of multimedia data (on the web or otherwise) we have in this age of real time web search, because we have a large and heterogeneous search space. With the availability of abundant computation power, the comparison operation among large number of data points may not be the primary problem. The noise that creeps in while comparing these huge data and the inability of content to have enough distinguishing power, makes the problem intractable. Context reduces the search space drastically with insignificant cost. Thus it is likely to remove a lot of noise while operating in the reduced space. Freuder proposed the idea of Verification in Vision [13] to make the computer vision problem pragmatic. The idea behind his hypothesis is, vision systems usually work well in a specific application. Hence, one can compare a generic recognition problem to a verification problem. Thus, given a test object the task is to find NOT 1 of $N$, but rather 1 of $M$, where $M \ll N$. This makes the problem computationally tractable and noise free. Use of context helps in mapping a heterogeneous recognition problem to a more refined verification task, which ideally should generate better results.

Success of search engines with text documents, and even images and videos, has clearly shown the importance of the eco-system that goes from the time a document is created to the time it is searched and used. The success of search engines even in searching for images is a good demonstration, as discussed in the next section, of their use of context. We believe that explicit consideration of the above classes of context will help in the analysis and management of multimedia data. In the next section, we show some examples of success achieved by systems in using some of these classes of context. By considering all these sources, one can improve the performance of these systems significantly.

# 5. EXAMPLES OF USE OF CONTEXT

To show the efficacy of context in processing content, in this section we consider some systems that show efficacy of using context. We will use several examples to show efficacy of some sources of context listed above. These examples are drawn from different sources. In most cases they only use one type of context. We should note here that, this Brave New Topic paper deals with the idea of bridging the semantic gap by use of context. This paper does NOT talk about one particular application or its efficiency. Hence, rigorous statistical results on the use of context in a particular application domain is out of scope in this paper.

## 5.1 Context in Content

The role of context in computer vision was emphasized even in very early days by Barrow and Tenenbaum in systems like MYSYS [5]. In this work it was vocally stated that: "In scene analysis, it is frequently impossible to interpret parts of an image taken out of context. Different objects may have similar appearances, while objects belonging to the same functional class can have strikingly different appearances (e.g., chairs). Ambiguous local interpretations must be ruled out by using contextual constraints to achieve a

meaningful, globally consistent interpretation of the whole scene." Relaxation labeling [20] was very popular in computer vision to use context about thirty years ago. The basic idea in this approach is to utilize knowledge of local relationships among objects that may appear in a scene and use these local relationships to propagate local interpretations repeatedly to refine overall interpretations of the image. Thus one may use simple facts like *"a computer monitor should be on a desk"*, *"floor is likely to be at the bottom in an image"*, and *"desk is on the floor in an office scene"*. A set of such constraints among all objects can then be used iteratively in the interpretation process. The process starts with recognition of all possible regions and assigning them all plausible levels. The relaxation process then iteratively eliminates all implausible levels. When this process terminates, each region is assigned the best possible interpretation based on the constraints, or the knowledge, available.

Recently there has been an increase in interest in the use of context in computer vision, most notably in the fields of object detection [14, 8, 30, 9, 17]. In most of these works the context is the relationship among potential objects, represented as spatial relationships among corresponding regions, and used for either filtering or constraint propagation like approaches. Interestingly, the earliest recorded research in computer vision, the Blocks World research at MIT [31] was really a systematic study of using constraints of the domain for interpretation of images. Since, use of context has been a topic of discussion in such cases, we will not discuss this in more details here.

## 5.2 Context Only Image Search: Commercial Systems

The most commonly used example of use of context in search are the commercial applications of image search from any major search engine, like Google, Bing, or Yahoo. Suppose that you search for images with keyword Obama, rose, Tendulkar, or cars. If one takes a look at the top ranked retrieved results, most of them are correct. Surprisingly, as is well known, most of these results are obtained without even processing the image itself, that is without even looking at the content. These search systems only use the context provided to them from sources such as the name of the file containing the picture, surrounding text on the page where the picture file appears, and the topic of the page. These search engines perform much better than any content based retrieval system that we have seen, including the ones that one of the authors was involved in developing.

## 5.3 Device Parameters: EXIF

EXIF data is attached to all digital pictures and contains very valuable information about camera parameters used in taking photos. Some of these parameters affect the part of the scene imaged and the intensity values of pixels, while others give very valuable contextual information about the data acquisition context. We discuss several experiments related to the role of device parameters and data acquisition context in the following paragraphs.

### *Experiment 1.*

**EXIF Parameters and Human Semantics** : In this experiment we investigate how the EXIF camera parameters are correlated without human assigned tags. We crawl Flickr to download 2000 photos with predefined set of tags like scenery, sunset, family etc. We remove noises by manual inspection. All of these photos have been shot by digital cameras with EXIF data in the header. We then cluster the photos based on Exposure Time, Focal Length, F-Number, Flash and ISO (optical metadata in EXIF). Each of the Figures 4, 5 and 6 show the probability distribution of human induced tags in different EXIF based clusters. We find that photos shot with high exposure and no flash are more likely to have tags like night, moon, etc (Figure 4). Photos with low exposure and shorter focal length and no flash have tags predominantly associated with outdoor events (Figure 5). Another cluster with slightly longer exposure with flash are generally associated with indoor events (Figure 6). Thus even without looking at the pixels we get a good prior on the possible tags which can be assigned to a photo.

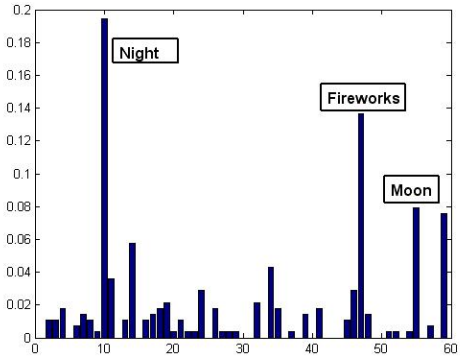


**Figure 4: Tag distribution in a cluster with No Flash and Large Exposure time**

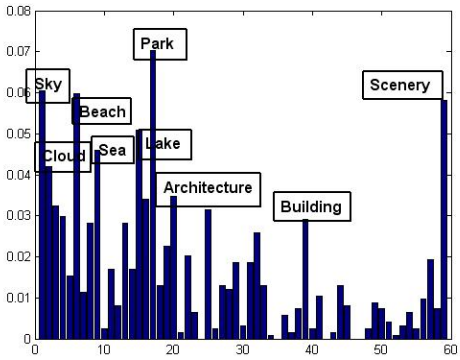


**Figure 5: Tag distribution in a cluster with NO Flash, Low Exposure and Low Focal Length (large Field of View)**

### *Experiment 2.*

**Disambiguation of photo capture conditions using EXIF Metadata** : In this experiment we test automatic photo annotation system with and without using context. We set up an automatic image annotation engine, built on a ground truth set created from the Flickr data. We used a probabilistic multinomial model to predict the annotation as discussed in [27] and [28] which can fuse content and context data. The task is to predict the annotations on the test photo in Figure 7. If we feed pixel features to our model, the annotations predicted in order of decreasing score are: scenery, city streets, illuminations and wildlife. If we use both EXIF (camera parameters) and pixel features, the annotations predicted are indoor, party, portrait, indoor group photo. There is a clear discrepancy between tags predicted by the content and the context channel. Why is it so? The answer lies in the way the photo was shot. This is

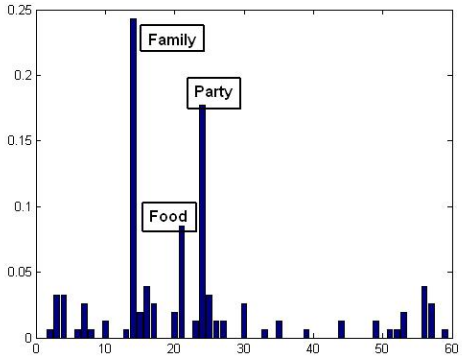


**Figure 6: Tag distribution in a cluster with Flash, Larger Exposure**

**Figure 7: Example Test Photo**

actually a photo of a photo. The image originally appeared in the cover page of a magazine and we shot a photo of it using a standard digital camera in an indoor environment. Since the content and context channels capture two entirely different semantics about an image, the tags are so different. The context channel in the test photo guides the system to the possible subset of tags which apply to indoor photos.

The next task is to predict the tags of the photo in Figure 8. The tags predicted based on only pixel features are: scenery, city streets, group photo outdoors, wildlife. The tags based on EXIF and pixel features are group photo indoors, indoor party, indoors artifact, illuminations. Why is there a discrepancy? The background is confusing. It has a lot of green component usually seen in outdoor photos. Hence the pixel feature based tagging algorithm gets confused and predicts noisy tags. However the event capture conditions are represented well in the EXIF parameter space. Hence the tags based on EXIF and pixel features better explain the semantics of the photo.

*Experiment 3.*

**Classification using Pixel Features and EXIF**: This experiment is about classification of test photos. We define three mutually exclusive classes: indoor, outdoor day and outdoor night. We build a logistic regression model to predict the class names on test photos based on a training set. Tables 1 and 2 show the precision and recall of the classification task while using pixel features and EXIF data (optical context), separately and together. We find that the EXIF data is by itself efficient enough for the classification task. In case of outdoor day photos, there is not much improvement if we include the pixel features. Only in case of outdoor night photos, precision and recall improves if we use pixels. This is probably because sometimes the camera parameters for indoor and night shots are very similar.


**Figure 8: Example Test Photo**

**Table 1: Precision of the Classification Task**

| Type of Data | Outdoor Day | Indoors | Outdoor Night |
|---|---|---|---|
| EXIF(Context) | 0.95 | 0.73 | 0.58 |
| EXIF and Pixels | 0.94 | 0.75 | 0.74 |

## 5.4 Data Acquisition Context: Location Recognition with and without GPS

If we try to predict the location of the photo in Figure 9, based on pixel feature similarity with geo tagged images crawled from the Web (as in [16]), we should get a high probability for China/Asia and related geographic region. However, if we look into the gps tag of this image, it will be clear that this photo was shot in a theme park in Orlando, Florida, USA. A careful look at the people in the picture will show you that most of the people in it appear to be from the USA (not apparent in the low resolution version). Thus pixels by themselves can be very misleading.

## 5.5 Rorschach Tests

Many psychological tests and even many photos that commonly appear in psychology literature are pictures shown to people, who are asked to specify what they see in those pictures. In these experiments, the goal is to know about the personal characteristics of the perceiver. These tests demonstrate that the interpretation of data depends on the perceiver [12]. Common phrases used by people such as "*Do you see what I see in this picture?*" are clear indication of the well recognized role of the perceiver.

## 5.6 Interpretation Context: Domain knowledge

Let us look at Figure 10. Let us try to guess what this picture is. Now suppose you are told that this is an Atomic Force Microscope image; you will either think about cellular images or atomic level images of materials. In fact what you are seeing is a recent break-

**Table 2: Recall of the Classification Task**

| Type of Data | Outdoor Day | Indoors | Outdoor Night |
|---|---|---|---|
| EXIF (Context) | 0.94 | 0.81 | 0.50 |
| EXIF and Pixels | 0.94 | 0.79 | 0.72 |

**Figure 9: Where was this photo shot?**

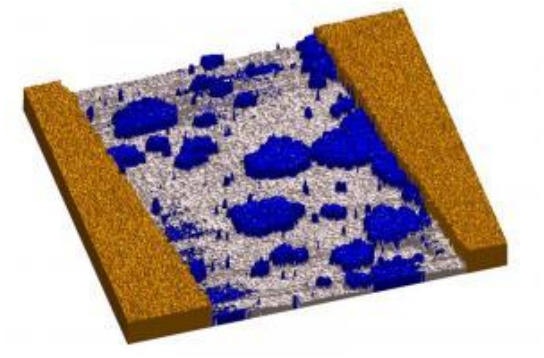

**Figure 10: What does this image show ?**

through in developing chemical sensors that could be developed cost-effectively [2]. Most possibly, you and I can not understand what this is because we do not know how to interpret these images because we do not have the context and associated knowledge. But this was an important image in the announcement of this breakthrough. What this shows is that without domain knowledge, it is almost impossible for people to analyze and understand content.

## 5.7 Context for Photo Management in Smartphones

Next generation digital photos will be captured by smartphones like the IPhone and the Android sets. As shown in Figure 1, smartphones are privy to a lot of heterogeneous information sources which will help augment photos with a lot of contextual knowledge. In this section we will highlight some scenarios where context can play an useful role for managing photos on smartphones.

*Identifying People.* Most consumer cameras can detect frontal faces while shooting photos. However face recognition or automatically assigning name tags to faces is an open problem. In case of smart phones this problem can be reduced to a much easier face identification problem (comparison in a constrained set). Our personal calendars provide us with the name of people we are meeting at any particular point of time. If a portrait / group photo was shot using the smartphone, we just need to compare the faces appearing in it with the faces of people we were supposed to meet at that time. The latter can come from any social network or photo sharing site. If the system does not get a good match, it can go through the list of recent callers / callee and try matching with their faces. Still if it does not get a good matching face we can go through our contact list to find a good matching face. Thus we iteratively increase our search space for finding good retrieval results based on context, starting from a really small high valued subspace. This makes the problem much easier to solve.

Let us elucidate this point with a concrete example. We present a small experiment of using social network knowledge for automatic face tagging in personal photos. Consider the photo shown in Figure 11. This was shot at a gathering of friends. All the participants in the meeting were on Facebook. The meeting was set up as an event on Facebook, with links to all the participants. We set up the experiment as follows:

- We gather the names and profile pictures of each participant in the meeting from Facebook.
- Additionally, we gather five tagged faces for each participant from Facebook.
- We use the OpenCV [7] face detection module to detect faces in the photo shown in Figure 11.
- We represent each face using Local Binary Pattern (LBP) features [3], which has been shown to be pretty robust to varying illumination conditions.
- We try to identify the faces, using K-Nearest Neighbors in the feature space of the faces.

The face detector detected 11 faces from the photo. Two were not detected (probably due to occlusion). We compute the LBP features for all faces in the test photo as well as from the faces retrieved from Facebook. We use K-Nearest neighbor (with K = 5) to find matches between each test face and the faces extracted from Facebook. We assign each test face to a particular person if there exists a majority vote in the five nearest neighbors. In this particular test experiment, 7 out of 11 people were correctly tagged. Face recognition in consumer photos are not known to perform very well due to with variation in lighting and geometry in an unconstrained setting. But with contextual knowledge (like the event and social data here), we can get a reduced search space to look for a good match and hence get a considerably satisfactory result. Note that we reduced the problem of recognition to a matching problem with no supervised training or face modeling phase.

*Identifying objects.*

Object localization and identification is another challenging problem. Usually people shoot landmarks or objects of special interest using their cameras. Using the EXIF data the system can infer if the photo was shot indoors or outdoors ( i.e., day or night). It can also estimate a possible size of the object based on the focal length, field of view and subject distance. Geo location will help us to narrow to down to a small set of important objects to match to (e.g., landmarks, flowers or food) which are commonly shot in that area. We can get this popular object data by crawling the web which has images (with tagged objects) shot in the geographical vicinity. Comparison to this much refined set is likely to generate better results.

*Event tagging based on public / private calendars.*

People shoot a lot of photos in their life events, e.g., parties, trips, meeting et al. It is very relevant and useful to tag photos based on the events. It may be very difficult to automatically tag a photo with an event name ( e.g., John's birthday) or even a generic class name (e.g., indoor party) based on pixels features and EXIF

**Figure 11: Identification of People in an Event using Social Network Knowledge**

data. However, personal calendars can help on such cases to properly tag photos with event names. Further, people also participate in public events like concerts, baseball games, parades etc. There are abundant sources of event repositories on the web. Based on a user's location, and the events taking place in the vicinity, it may be possible to predict the proper event name (e.g., Giants vs Red Sox Game at the AT&T center in San Francisco) with reasonably good accuracy.

# 6. CONCLUSION

Our goal in this paper is to critically examine current perspectives and approaches in multimedia content analysis (and related fields such as computer vision) towards one of the most difficult challenge in this field: bridging the semantic gap. Despite a very large number of research papers in this area, semantic gap remains as challenging as ever. This is well acknowledged in many research papers and clearly demonstrated by the lack of progress. We believe that to make progress in bridging the semantic gap for content analysis leading to interpretation, organization, and access to increasing volumes of multimedia data, we must look at the problem from a new perspective. In this paper, we revisit the perception problem and adopt a perspective inspired by philosophy, cognitive science, and modern search engines. This perspective suggests that discovery and utilization of all sources of knowledge in the ecosystem of a multimedia content capture and analysis may provide an easier approach to bridge the semantic gap. Unlike current content analysis approaches that try to extract as much as possible from content data, this approach suggests collection and application of metadata from all potentially useful sources. Similar approaches have been successful in search engines where one uses several sources of information to interpret data to get answers. Given the urgency to solve multimedia information management task and the fact that current content based approaches have failed to deliver intended results, multimedia research community has no other option but to explore application of metadata and all other context information.

# 7. REFERENCES


[1] *Google Research Blog:* `http://googleresearch.blogspot.com/2010/04/lessons-learned-developing-practical.html`.

[2] *Science Daily Article:* `http://bit.ly/QHiJh`.

[3] T. Ahonen, A. Hadid, et al. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2037–2041, 2006.

[4] H. Andersen and V. Burton. *The emperor's new clothes*. Sandpiper, 2004.

[5] H. Barrow and J. Tenenbaum. MSYS: A system for reasoning about scenes. *Technical Note, AI Center SRI*, 121, 1976.

[6] G. Berkeley and C. Krauth. *A treatise concerning the principles of human knowledge*. JB Lippincott & co., 1878.

[7] G. Bradski, A. Kaehler, and V. Pisarevsky. Learning-based computer vision with Intel's open source computer vision library. *Intel Technology Journal*, 9(2):119–130, 2005.

[8] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *Proc. ICCV*, 2009.

[9] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert. An empirical study of context in object detection. 2009.

[10] R. Duda, P. Hart, and D. Stork. *Pattern classification*. Citeseer, 1971.

[11] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results. 2010.

[12] J. Exner. *The Rorschach: A comprehensive system*. John Wiley & Sons.

[13] E. Freuder. Recognition of Real Objects. *Vision Flash, No. 33*, 1972.

[14] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. 2009.

[15] R. Gregory. The intelligent eye. 1970.

[16] J. Hays and A. Efros. IM2GPS: estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, pages 1–8, 2008.

[17] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15, 2008.

[18] Japan Electronics and Information Technology Industries Assoc. 2.2 Digital Still Camera Image File Format Standard (Exif) Version 2.2 .

[19] Y. Joung, M. El Zarki, and R. Jain. A user model for personalization services. In *Fourth International Conference on Digital Information Management, 2009. ICDIM 2009*, pages 1–6, 2009.

[20] J. Kittler and J. Illingworth. Relaxation labelling algorithms–a review. *Image and Vision Computing*, 3(4):206–216, 1985.

[21] U. Neisser. *Cognitive psychology*. Appleton-Century-Crofts New York, 1967.

[22] K. Popper and S. Popper. *Objective knowledge: An evolutionary approach*. Clarendon Press, 1979.

[23] V. Ramachandran and S. Anstis. The perception of apparent motion. *Scientific American*, 254(6):102–109, 1986.

[24] I. Rock. The logic of perception. 1983.

[25] S. Santini and R. Jain. Beyond query by example. In *Proceedings of the sixth ACM international conference on Multimedia*, page 350. ACM, 1998.

[26] C. Shirky. *Cognitive Surplus: Creativity and Generosity in a Connected Age*. Penguin Press, 2010.

[27] P. Sinha and R. Jain. Classification and annotation of digital photos using optical context data. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 309–318. ACM, 2008.

[28] P. Sinha and R. Jain. Semantics In Digital Photos: A Contenxtual Analysis. In *Proceedings of the 2008 IEEE International Conference on Semantic Computing*, pages 58–65. IEEE Computer Society, 2008.

[29] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1349–1380, 2000.

[30] A. Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191, 2003.

[31] D. Waltz and P. Winston. The psychology of computer vision, 1975.