

Classification and Annotation of Digital Photos Using Optical Context Data

Pinaki Sinha and Ramesh Jain
Department of Computer Science, University of California
Irvine, CA, USA
psinha@ics.uci.edu, jain@ics.uci.edu

ABSTRACT

Other than the pixel information, a digital photo of today has a host of other information regarding the photo shooting event. These information are captured by different sensors present on the camera and are stored as metadata. In this paper we exploit this meta information and derive useful semantics about the digital photo. We also compare our results with classical relevance models used for automatic photo annotation. We create a dataset of digital photos containing all information and report results on it. We also make the dataset available to the community for further experiments.

Categories and Subject Descriptors

H.3.3 [Information Storage Retrieval]: Information Search and Retrieval; I.4.9 [Image Processing and Computer Vision]: Applications

General Terms

Algorithms, Performance, Design

Keywords

Automatic Annotation, Digital Photo, Search, Context, Camera Metadata

1. INTRODUCTION

The traditional camera obscura is a device meant for capturing information present in the light as intensity values. However, the digital camera with multiple sensors present on it can capture much more information about the photo shooting event. Hence, the digital photograph of today is no longer just a collection of pixels. It has a host of sensory information stored as metadata. These information can be summarized if we model the digital photograph as a multilayered structure as shown in Figure 1. The layers are: i. Pixel/ Spectral layer and ii. Meta Layer. The pixel layer stores information recorded by the CCD (as pixel values). The content of the image is present in this layer. The Meta Layer stores the contextual information about a photo shoot. The Meta Layer can

further be divided into the following sublayers: a. Optical Meta Layer. It contains the metadata related to the optics of the camera; e.g., the focal length, aperture, exposure time etc. These metadata store important cues about the context in which the image was shot (like the lighting condition, depth of field and distance of subjects in the image). b. Temporal Meta Layer. It contains the time stamp of the instant in which the photo was shot. The time stamp of a single image in a standalone environment is not informative enough. But in a collection of images (e.g., photo albums) the time difference can shed valuable light on the content of the images [10]. c. Spatial Meta Layer. It contains the spatial coordinates of the places where pictures were shot. These coordinates are generated by the GPS systems attached to the camera. In case of camera phones, cell-tower ids help generate this data. d. Human Induced Meta Layer. This layer contains the tags/ comments/ ratings posted by people. Community tagging (in online photo albums) or voice tagging in mobile phones help generate data for this layer. e. Derived Meta Layer. This metadata is inferred from other information by various statistical modeling approaches. The taxonomy defined here helps us to explore the information sources present in a digital camera image. Presently, the spectral, optical and temporal layers are present in almost all digital photographs, while the spatial, human induced and derived meta layers might not be present.

It is estimated that 375 Petabytes of data is generated in form of digital photographs annually [22]. Hence it is absolutely necessary to organize and index this information for easy retrieval. Most researchers have used pixel measures like global features (color, texture and shapes) [18] [9] [13] or local features (edges, salient points or objects) [17] to characterize an image. However, these systems work efficiently in query by example or low level query situations. But image search using an example input image or query using low level features (e.g., color, texture values) might be difficult and non-intuitive to most people. Rather image search using keywords might be easier. To address this issue, correlations among image features and human assigned tags or labels have been studied. Image annotation has been modeled as machine translation [11], relevance models [14], or as a hierarchical topic model [6]. However, the results reported are only on the Corel Image Dataset. This dataset has been created by professionals, is restricted in its scope, and does not model the various real world photos shot by amateurs. Also, as suggested in [21] and [5] the semantic gap in image retrieval cannot be overcome using pixel features only. Further, computing the high dimensional pixels features is cost intensive and performs poorly while retrieving in heterogeneous image databases like the web. These are the primary reasons why the search engines rely only on text/ tags/ annotations to retrieve images. Recent research has used the Optical Meta Layer to classify and annotate images [7] [8] [22]. Boutell and Luo [7] use pixel values and optical

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIVR '08, July 7–9, 2008, Niagara Falls, Ontario, Canada.
Copyright 2008 ACM 978-1-60558-070-8/08/07 ...\$5.00.

metadata for sunset scene and indoor outdoor classification. They choose the most significant cue using K-L divergence analysis. Liu et al [16] use color, texture and camera metadata in a hierarchical way to classify indoor and outdoor images. However most of the researchers use the optical metadata without any strong reference to physics of vision (of why the images were being classified using the chosen cues).

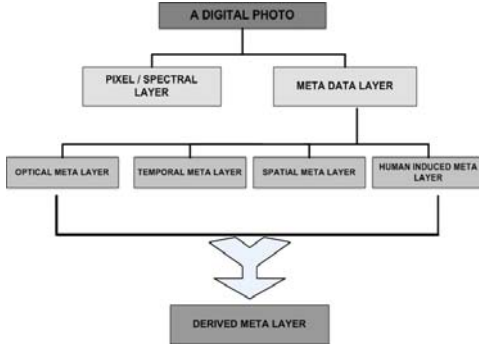


Figure 1: The digital photo: as a layered structure of information

1.1 Our Contribution

In this research, we use the Optical meta data to infer semantics about the digital photo. More precisely, we address the problems of classifying photos into mutually exclusive classes, automatically tagging new photos and retrieving photos corresponding to a textual query. In section 3 we discuss the Optical Context Data present in digital photos. Section 4 shows how we can find clusters in the Optical Context space and why these clusters make semantic sense. In section 5 we show how the useful the context data is in classifying photos into mutually exclusive classes. We build on the proposed relevance models for image annotation [15] and provide a new model using Optical Context in section 7. Finally, in section 8 we show the results for automatic annotation and ranked retrieval of digital photos. The results for annotation and retrieval have been reported using the context and content data independently and also when they are used together.

2. CREATING THE PHOTO DATASET

As discussed in [23] Corel dataset is a much easy dataset to annotate and retrieve. It has been shot by professionals in constrained environment. It does not model the challenges annotating and retrieving real world diverse photo collections. Further, unlike the ubiquitous digital photograph the Corel photos do not contain any contextual data. Hence we created a new dataset with photos from amateurs shot by regular point and shoot digital cameras. We create a dataset of 30K such photos. These photos have been shot by thirty different users, across five different continents over a period of six years. Hence this set models personal photo albums quite well. We tag 1500 photos from this set with some common nouns which will be helpful for annotation and retrieval. We also collected photos from Flickr [2]. Online photo albums like Flickr have photos with lots of unreliable and incorrect tags. If we use these tags, our inference will often be erroneous. Hence we filtered 3500 digital photos from this set and chose the correct tags. Flickr publishes all-time popular tags [1]. We chose the common nouns from this set. These common nouns were used to tag the photos. To facilitate future research we make these datasets available to the community [3].

3. OPTICAL CONTEXT DATA

The Exchangeable Image File (EXIF) Standard [4] specifies the camera parameters recorded during a photo shoot. The actual parameters recorded depend on the particular camera manufacturer. But there are certain fundamental parameters which are recorded by all popular camera models. These are: Exposure Time, Focal Length, F-number, Flash, Metering Mode and ISO. We examined several camera models from seven different manufacturers: Canon, Casio, Sony, Nikon, Fuji, Kodak, and Konica. All of them have these parameters in the EXIF header. Subject distance is an important parameter which can be used to make important inferences about the image content. However most camera models do not store this parameter. A detailed analysis of these parameters is provided in [20].

We extracted the optical context data from 30K digital photo corpus. Figure 2 shows the histogram of these data in the corpus. Distribution of exposure time is highly skewed (2(a)). Less than one percent of the images have exposure time more than 0.5 second. We show the log-exposure time distribution in Figure 2(b). Distribution of focal length in millimeters is shown in Figure 2(c). Since most of the images were shot by regular point and shoot cameras (which typically are wide angle with smaller relative aperture), majority of images have focal length in the range 0-100 mm. About 1-2% of the images have a focal length more than 100 mm. The distribution of F-Numbers is also skewed towards the right end of the spectrum (Figure 2(d)). Flash is a hex byte and its distribution (Figure 2(e)) shows various states of the flash; e.g., detection of reflected light, red eye detection mechanism and so on. ISO speed ratings are discrete values (Figure 2(f)) measuring the sensitivity of the image sensor. Higher the ISO more sensitive is the image sensor.

4. CLUSTERS IN OPTICAL SPACE

The skewed nature of the individual camera parameters indicate that none of them have enough discriminative power for meaningful classification when considered independently. Hence we decided to look at latent values which might be hidden in the optical metadata. The most important cue which is hidden in the Optical Meta Layer is the amount of ambient light during the photo shooting event. To estimate the ambient lighting condition we defined a metric based on the premise that amount of light entering a camera is directly proportional to: i. Exposure Time, ii. Aperture Area and iii. ISO speed rating of the CCD (which quantifies the camera’s sensitivity to incoming light). Further, the intensity of light incident on the image plane is inversely proportional to the squared distance between the lens and the image sensor (Inverse Square Law). The focal length provides an approximate measure of this distance. These laws from the physics of vision and camera, helped us to come up with a metric which quantifies the ambient light in an image. We call this the LogLight metric and is defined as follows:

Definition 1.

$$\text{LogLightMetric} = \lg \left(K \times ET \times AA \times ISO \div FL^2 \right)$$

where K is the proportionality constant, ET is the exposure time, AA is the aperture area, ISO is ISO speed rating and FL is the focal length. LogLight Metric will have a small value when the ambient light is high (the camera will have a low exposure time, small aperture and low ISO). Similarly it will have a large value if the outdoor light is small. Figure 3(a) and Figure 3(b) show the histograms of the LogLightMetric on photos shot without and with flash respectively from our 30K corpus.

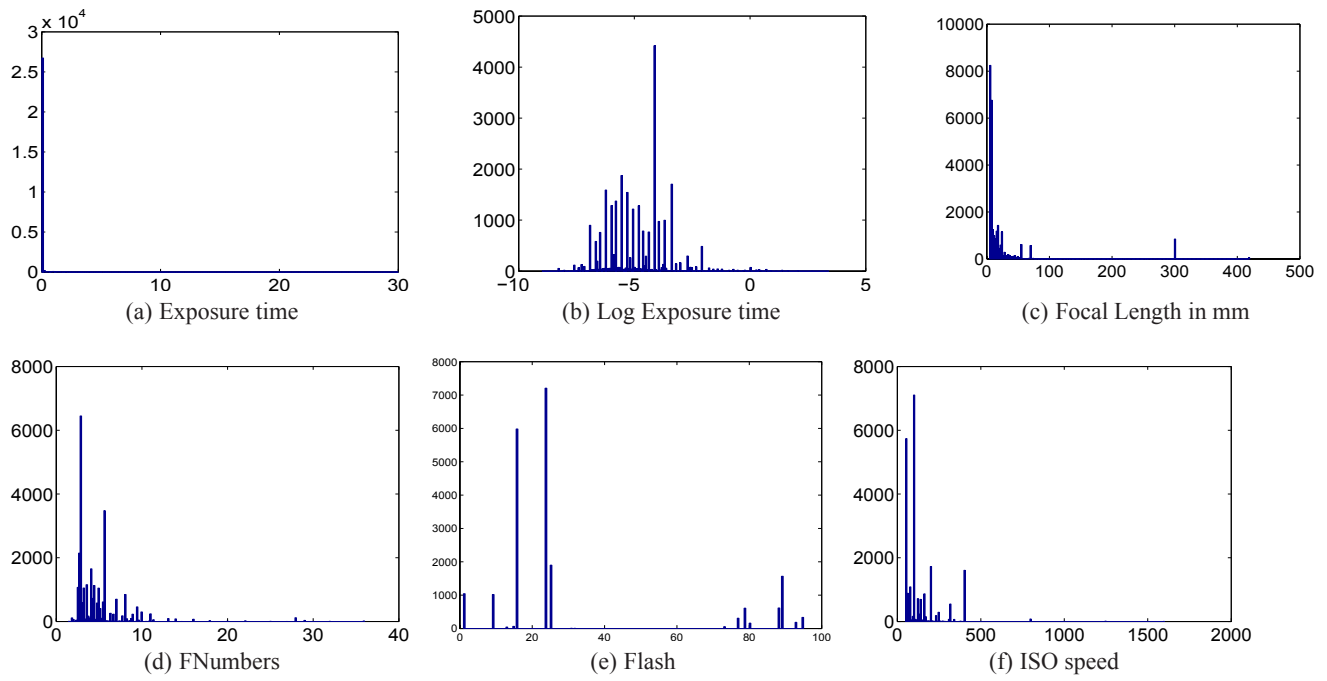


Figure 2: Distribution of Optical MetaData in the 30K digital photo corpus

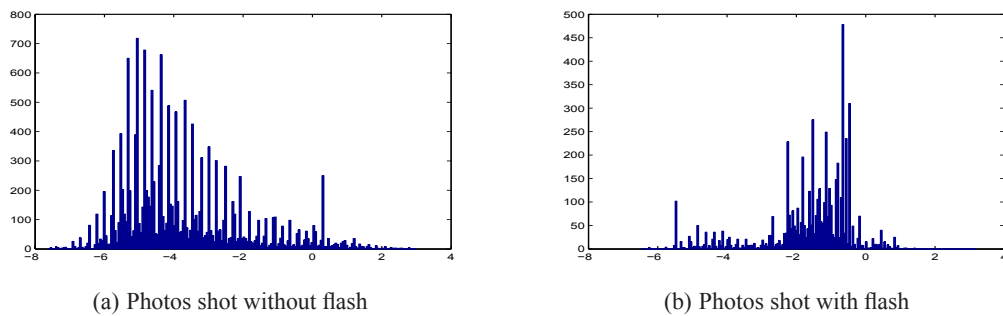


Figure 3: Distribution of LogLightMetric on Digital Photo set.

5. CLASSIFICATION OF PHOTOS

Prior to photo-shoot, a photographer adjusts the camera parameters to get a good representation of her subjects. Hence the intent of the photographer is somehow hidden in the optical data. In the automatic mode or different preset modes (e.g., landscape, night, portrait) the on-camera sensors compute these parameters for the users/ amateurs. One of the goals of this research is to investigate whether we can invert this process and generate some belief about the photographer’s or sensor’s intent. Hence we defined three mutually exclusive classes to determine the environment of the photo shooting event. These classes are: outdoor day, outdoor night and indoors. Each photo (from personal photo albums) must be from exactly one of these classes. The reason for choosing these classes, is that they tend to have very different lighting condition which should be represented in the LogLight metric. It must be noted here, that there can be photos where photographer is indoors and subject is outdoors or vice versa; however both in our 30K database and the Flickr dataset such photos are few and far between. Hence for this problem we shall ignore such photos. We tried to solve this classification problem using Optical Context only and also using Optical Context and Thumbnail pixel features. The pixel features are represented by color moment, color histogram, edge histogram and Gabor texture features. We divided our dataset into training and test sets. We used various classification algorithms. Decision trees seem to give the best results.

In table 1 we show the precision for the three classes and in table 2 we show the recall. The mean accuracies for classification using only Optical Context is 87.5% and using both Optical Context and Thumbnail Features is 89.2%. So we get a 2% improvement but we had to do significant amount of computation more to achieve this. However, when using Optical Context alone, outdoor night photos tend to confuse with indoor photos. Hence we see significant improvement of precision (16%) and recall(20%) of outdoor night photos when we use thumbnail content features. The apparent discrepancy between changes in accuracy and precision-recall is explained by the fact the percentage of outdoor day and indoor photos grossly outnumbers outdoor night photos. We should note here that Optical Context is represented in few bytes of data compared to kilobytes or megabytes of pixel data. Any type of statistical computation using these optical data are very efficient compared to that with high dimensional pixel features. Thus we believe the high degree of accuracy obtained by leveraging such small contextual data for our classification problem is worth reporting.

Table 1: Precision of the three mutually exclusive classes

Type of Data	Outdoor Day	Indoors	Outdoor Night
Only Optical Context	0.95	0.73	0.58
Optical Context+ Thumbnail Features	0.94	0.79	0.74

Table 2: Recall of the three mutually exclusive classes

Type of Data	Outdoor Day	Indoors	Outdoor Night
Only Optical Context	0.94	0.81	0.50
Optical Context+ Thumbnail Features	0.94	0.79	0.72

6. SEMANTICS HIDDEN IN THE OPTICAL LAYER

In this section we show how optical parameter space is related to human notion of semantics. This correspondence can be claimed if we show that certain subspaces of the optical parameter space are associated with specific human tags or semantics. We find clusters in the optical parameter space in an unsupervised manner. We model the LogLight distribution of photos shot with flash and without flash [Figure 3] as a mixture of Gaussians. Since we have no apriori knowledge of the structure of the optical parameter space, we use the Bayesian Model Selection [12] to find the optimal model and use Expectation Maximization algorithm to fit the model parameters. When EM is used to find the maximum likelihood, the Bayesian model selection can be approximated by the Bayesian Information Criterion [19]. Using this unsupervised clustering approach on the optical parameters from our 30K dataset, we find eight clusters that best explains the data. Next, we choose 3463 tagged photos from our Flickr dataset. For each photo, we find the probability of it being generated from each of the eight clusters. We assign the photo to the cluster having maximum probability. We assign all tags of the photo to that particular cluster. Figure 4 shows the prominence of different words in the various clusters. Figure 4(a) is a cluster containing photos shot in high exposures. The tags which are prominent in this cluster e.g., night, fireworks are typically shot in low light condition. Figure 4(b) are photos shot without flash. These photos have large field of view due to small focal length. We find these shots are mostly of nature. Figure 4(c) are photos shot with flash. They are of indoor events like parties. Figure 4(d) is a cluster mostly containing outdoor photos. However these were shot in lower light condition (high LogLight value). Hence flash was fired for these photos.

7. MODEL FOR AUTOMATIC ANNOTATION

The goal for automatic annotation is to predict words for tagging untagged photos. These words should also help in text based image retrieval. Lately, relevance model approach has become quite popular for automatic annotation and retrieval of images [14] [15]. In this approach automatic annotation is modeled as a language translation problem (translating from the language of humans to that of the image pixels and vice versa). Thus given a set of pixels (or pixel features) the task is to find how likely each word in the vocabulary is to tag the photo. In this research we use the Continuous Relevance Model [15] as a baseline. We then combine content and context data. We also provide an extension of the model, which takes into account the spatial positions of pixels within the image. We call this extended model as Spatial Continuous Relevance Model.

We define a Bayesian Network model to show the interaction between content and contextual information in a photo. A digital photo has pixels and words associated with it. Words are discrete elements generated from a finite vocabulary. In Figure 5, I is a discrete random variable over all photos in our training corpus T . W is a discrete random variable over the finite tag vocabulary. Pixels are represented by pixel features. We compute the pixel features as follows. We divide the whole image into rectangular blocks. For each block, we compute color, texture and shape features. Each feature vector has 42 dimensions (9 color moment, 16 color histogram, 12 Gabor texture features and 5 edge histogram features). Thus each rectangle in the photo is represented as a d dimensional feature vector. An image is composed on a fixed number of such vectors. In Figure 5, B is a random variable over the values in the pixel feature space.

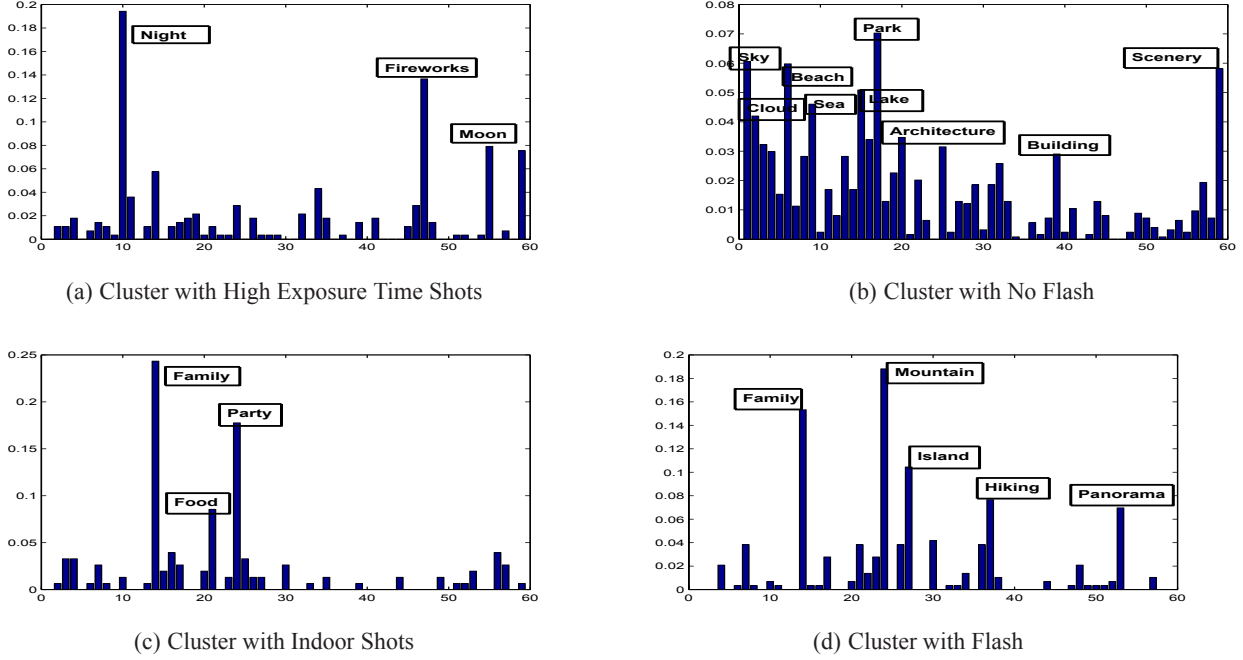


Figure 4: Distribution of Human Assigned Tags on Clusters in Optical Space

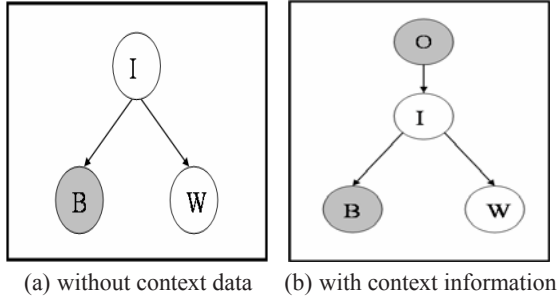


Figure 5: Bayes Net model of images

The goal of automatic annotation is to predict the W s associated with an untagged image based on B . This idea is presented in the Bayes Net model of Figure 5(a). B is the observed variable. The conditional probability of a word given a set of blocks is computed in 1.

$$\begin{aligned}
 P(w|B) &= \sum_I P(w, I|B) \\
 &= \sum_I P(w|I, B) P(I|B) \\
 &\propto \sum_I P(w|I) P(B|I) P(I)
 \end{aligned} \tag{1}$$

We model the distribution of word given an image $P(w|I)$ as a multinomial distribution with proper smoothing [14]. In our baseline model, we estimate the probability of feature vector given an image, $P(B|I)$, using nonparametric kernel density as in [15]. Thus, if we assume $C = \{c_1 \dots c_n\}$ be the set of pixel feature vectors

for an untagged test image J , we have to compute how probable is each of the feature vectors given a training image I . Now I has its own set of feature vectors $B = \{b_1 \dots b_n\}$. We compute the above probability as shown in equation 2.

$$P(c|B) = \frac{1}{n} \sum_{i=1}^n \frac{\exp \left\{ (c - b_i)^T \Sigma^{-1} (c - b_i) \right\}}{\sqrt{2^d \pi^d \|\Sigma\|}} \tag{2}$$

Equation 2 arises if we consider a Gaussian kernel at each of the rectangular blocks in the training image. We find the probability of a test block being generated by all the blocks in the training image and then average them out.

In the above approach, we are ignoring the spatial position of the blocks in the images. Consider, the following scenario. A photo has bluish rectangle generated by the sky on its top left corner. There is a training image with a bluish rectangle at the bottom part generated by water or someone wearing a blue shirt. If we are using the above technique, the photo of water will have a high probability of generating a sky block. This might lead to incorrect annotations. This problem arises as we are ignoring the block order in the images. We extended this model, by including a normalized smooth decay function centered around the position of the test block. Hence, the probability of a lower left block in a training image generating an upper right block in the test image will be less than a block at the upper right position. Thus equation 2 can be reformulated as 3.

$$\begin{aligned}
 &P(c|B) \\
 &= \text{SmoothDecay}(c) \left[\sum_{i=1}^n \frac{\exp \left\{ (c - b_i)^T \Sigma^{-1} (c - b_i) \right\}}{\sqrt{2^d \pi^d \|\Sigma\|}} \right]
 \end{aligned} \tag{3}$$

We experimented with some decay functions. Two dimensional logistic decay generated better results. The basic assumption behind

this approach is that in personal photo albums similar objects tend to appear in similar locations.

However this representation does not consider the contextual information which are present along with the digital photo. Hence we propose a new model integrating both content and context as shown in Figure 5(b). The context information is modeled as follows. As described in 6 we learn the optical clusters (O) using an untagged image database. Whenever a new image X comes, we assign it to the cluster O_j having maximum value for $P(X|O_j)$. Here O is a random variable over all clusters and it is observed. Thus the probability of a word given the pixel feature blocks and the optical context information can be computed as in 4.

$$\begin{aligned}
 P(w|B, O) &= \sum_I P(w, I|B, O) \\
 &= \sum_I P(w|I, B, O) P(I|B, O) \\
 &\propto \sum_I P(w|I) P(B, O|I) P(I) \\
 &= \sum_I P(w|I) P(B|I) P(O|I) P(I)
 \end{aligned} \tag{4}$$

Other than the pixel features computed on the entire image, we also compute features on the thumbnails of the image. All digital photos have a small resolution (typically 160x120) thumbnail version stored in their header. We extract this image and compute the same features, but now considering each thumbnail is made up of one rectangular block only.

8. RESULTS

All experiments for this research were done on two tagged digital photo datasets. One of them was created by tagging photos from our 30K dataset. We will refer to this as the Home Dataset. There are 1700 photos in this set. It has a tag vocabulary size 30. The other was created by crawling photos from Flickr and choosing the correct tags. There are 3500 photos in this set, with a tag vocabulary size of 59. We will refer to this as the Flickr Dataset.

8.1 Automatic Annotation

We divide our datasets into train, evaluation and test sets. We use train and evaluation sets to tune our parameters (the multinomial smoothing parameters and Gaussian covariance. Please refer to [14] and [15] for a detailed analysis). The annotation performance is measured by precision and recall values of the tags on the test set. We ran our experiments using the following algorithms and information sources. First we use only the Optical Context data for annotation. In this case, we will only have the variable O in Figure 5(b). Next we use only the image features and use the CRM model to solve the problem (Figure 5(a)). Then we find the results using Spatial CRM (SCRM). We then use only the Thumbnail Image features along with Context data for annotation. Finally, we use Optical Context with both CRM and Spatial CRM.

For each photo in the test set, our algorithms predict tags with probabilities assigned to them. We heuristically choose top five tags and assign them to the photo. Precision for a particular tag is the ratio between the number of photos correctly annotated to the number of all photos automatically annotated with the tag. Recall is the number of correctly tagged photos divided by the number of photos annotated with that tag in the ground truth data.

Tables 3 and 4 show the results of automatic annotation on the Home Dataset and Flickr Dataset respectively. We see in both the

Table 3: Annotation Performance on the Home DataSet

Type of Model	Mean Precision	Mean Recall
Image Features (CRM)	0.18	0.30
Image Features (SCRM)	0.20	0.32
Thumbnail Features +Optical Context	0.22	0.33
Image Features(CRM) +Optical Context	0.22	0.35
Image Features(SCRM) + Optical Context	0.26	0.35

datasets mean precision and recall significantly improves by using context information. In fact, using context and low dimensional thumbnail features we get much better than using image features alone. Further computing the lightweight thumbnail features is much more efficient than computing features on the entire image. Also in our Home DataSet (Table 3) SCRM performs better than CRM. In the Flickr DataSet (Table 4) including the entire image features with context data does not make a difference compared to using only thumbnail features and context. We should note here that these datasets have been created in completely unconstrained environments. The tag vocabulary has been created by the words entered by the user. Hence the tags are not really independent (as assumed in our models). Further, many tags are semantically similar and have been interchangeably used. For instance, the Flickr vocabulary has tags like scenery, landscape, panorama which have been interchangeably used. It also has tags like family, trip, roadtrip which are ambiguous and have been applied to all sorts of photos. We feel these reasons account for the low precision-recall values in the annotation task. The number of query words which retrieve at least one relevant image (tagged with the word in ground truth) varies across models. In the Home Dataset, Optical Context retrieves only 8 words, while other models retrieve 24 tags. In the Flickr Dataset, Optical Context retrieves only 10 words, while other models retrieve 44 tags. Hence we do not directly compare the mean precision and recall of the context only model with the other models. However in Table 5 we compare the precision recall of Optical Context, Image Feature and Thumbnail-Context models for the eight tags. We will make a few observations about the tags and their precision recall values. The tag wildlife has precision values of 0.71, 0.16 and 0.44 for the Optical Context, Image Feature and Thumbnail-Context models respectively. This is intuitive, as wildlife shots tend to have very distinct camera parameter values which distinguishes themselves from other photos. However using pixels only, we cannot achieve this distinction. The tag illumination (used for night time illuminations) has recall values of 0.24, 0.60 and 0.60 for the Optical Context, Image Feature and Thumbnail-Context models respectively. It turns out that Optical Parameters for night illuminations have similar values with those of indoor shots. Thus illuminations gets confused with group photo indoors or indoor party. But, night illuminations have pretty distinct pixel features. Hence, the image features can tag them properly. These two examples show that with optical context along with thumbnail features (less computation compared to image features) we can achieve the usefulness of both worlds of content and context.

Table 4: Annotation Performance on the Flickr Data Set

Type of Model	Mean Precision	Mean Recall
Image Features (CRM)	0.09	0.20
Image Features (SCRM)	0.11	0.22
Thumbnail Features +Optical Context	0.14	0.27
Image Features(CRM) +Optical Context	0.13	0.23
Image Features(SCRM) +Optical Context	0.13	0.24

Table 5: Annotation Performance on the Home Data Set for 8 tags

Type of Model	Mean Precision	Mean Recall
Optical Context	0.28	0.81
Image Features (CRM)	0.18	0.68
Thumbnail Features +Optical Context	0.29	0.76

8.2 Retrieval with Single Query Tag

In automatic annotation we heuristically assigned top K tags to a photo. We ignore the probability assigned to a tag for a particular photo. However in large photo databases, it is imperative to show photos relevant to the query word in some ranked order. Thus for a query word fireworks, if the system shows 10 images, then the photos should be sorted in the order of probability of having the tag fireworks. We use the standard metric Average Precision to measure the retrieval performance of our algorithms. Average precision corresponding to a query word (tag) is defined as the mean of the precision scores after each relevant document is retrieved. Mean Average Precision is the mean across all query tags in the vocabulary. Table 6 and 7 show the retrieval performance after top 10 photos are retrieved and also after all photos are retrieved.

8.3 Examples of Annotation and Retrieval

Figure 6 and Figure 7 show the automatically predicted annotations on test photos. We show the predictions by using three different models. Figure 8 and Figure 9 show the top 6 retrieved photos for the query word Wildlife using two different models.

Table 6: Mean Average Precision (MAP) for Retrieval on the Home Data Set

Type of Model	Top 10 Retrievals	All Queries
Optical Features	0.18	0.15
Image Features (CRM)	0.20	0.11
Image Features (SCRM)	0.22	0.14
Thumbnail Features + Optical Context	0.24	0.17
Image Features(CRM) + Optical Context	0.20	0.14
Image Features(SCRM) + Optical Context	0.24	0.16

Table 7: Mean Average Precision (MAP) for Retrieval on the Flickr Data Set

Type of Model	Top 10 Retrievals	All Queries
Optical Features	0.07	0.06
Image Features (CRM)	0.11	0.07
Image Features (SCRM)	0.13	0.08
Thumbnail Features + Optical Context	0.18	0.11
Image Features(CRM) + Optical Context	0.16	0.10
Image Features(SCRM) + Optical Context	0.16	0.10



Figure 8: Top 6 Images Retrieved for the query Wildlife. Using Optical Context and Thumbnail Features.



Figure 9: Top 6 Images Retrieved for the query Wildlife. Using Image Features Only.


		
Auto Annotation1: Park, Family, Scenery, Portrait. Auto Annotation2: Leaves, Forest, Park, Family. Auto Annotation3: Kids, Zoo, Park, Scenery.	Auto Annotation1: Park, Family, Scenery, Portrait. Auto Annotation2: Fireworks, Night, Concert, Family. Auto Annotation3: Moon, Park, Family, Portrait.	Auto Annotation1: Family, Portrait, Food. Auto Annotation2: House, Park, Family, Portrait. Auto Annotation3: Park, Family, Beach, Portrait.

Figure 6: Examples of Automatic Annotation on Untagged Images. AutoAnnotation1: Using Optical Context Only; AutoAnnotation2: Using Optical Context and Thumbnail Features; AutoAnnotation3: Using Image features Only.




		
Auto Annotation1: Park, Scenery, Family, Portrait. Auto Annotation2: Sea, Clouds, Scenery, Family. Auto Annotation3: Clouds, Park, Family, Sky, Scenery.	Auto Annotation1: Scenery, Park, Portrait, Flower. Auto Annotation2: Scenery, Park, Clouds, Family. Auto Annotation3: Bird, Park, Family, Scenery.	Auto Annotation1: Park, Family, Scenery, Portrait. Auto Annotation2: Hiking, Forest, Park, Family. Auto Annotation3: Bird, Park, Scenery, Family.

Figure 7: Examples of Automatic Annotation on Untagged Images. AutoAnnotation1: Using Optical Context Only; AutoAnnotation2: Using Optical Context and Thumbnail Features; AutoAnnotation3: Using Image features Only.

9. CONCLUSION

Optical context data is only a small fraction (less than one percent) of the information encoded in a high resolution digital image. As shown in the research, this small fraction has invaluable information about the photo shooting environment which can help mapping the pixels to human semantics. Further, if we combine this with thumbnail image features, the annotation and retrieval performances significantly improves. Also as we show in the results, combining the pixel features from the entire photo improves the results only slightly. However we achieve this with very large computation cost on high dimensional pixel feature space. The next step would be to investigate the fusion of other meta layers for finding image semantics. As for instance, including temporal and spatial information with optical context can help come up with better tags.

10. REFERENCES

- [1] <http://http://www.flickr.com/photos/tags/>.
- [2] <http://www.flickr.com/>.
- [3] <http://www.ics.uci.edu/~psinha/dataset.html>.
- [4] EXIF: Exchangeable image file format for digital cameras: version 2.2. Technical report, JEITI Association, 2002.
- [5] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In Proc of IEEE Intl Conf Computer Vision, volume 2, pages 408–415, 2000.
- [6] D. Blei and M. I. Jordan. Modeling annotated data. In Proc. ACM SIGIR, 2003.
- [7] M. Boutell and J. Luo. Bayesian fusion of camera metadata cues in semantic scene classification. In Proc. IEEE CVPR, 2004.
- [8] M. Boutell and J. Luo. Photo classification by integrating image content and camera metadata. In Proc. of ICPR, 2004.
- [9] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation maximization and application to image querying. IEEE Transactions in PAMI, 24(8):1026–1040, 2002.
- [10] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox. Temporal event clustering for digital photo collections. In Proc. ACM SIGMM, 2003.
- [11] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In Proc. ECCV, 2002.
- [12] C. Fraley and A. Raftery. How many clusters? which clustering method? answers via model based cluster analysis. The Computer Journal, 1998.
- [13] Y. Gong. Advancing content-based image retrieval by exploiting image color and region features. Multimedia Systems, 7(6):449–457, 1999.
- [14] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In Proc. ACM SIGIR, 2003.
- [15] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In Proc. NIPS, 2003.
- [16] X. Liu, L. Zhang, M. Li, H. Zhang, and D. Wang. Boosting image classification with lda based feature combination for digital photograph management. Pattern Recognition, 38:887–901, 2005.
- [17] D. G. Lowe. Distinctive image features from scaleinvariant key points. International Journal of Computer Vision, 60:91–110, 2004.
- [18] B. Manjunath and W. Ma. Texture features for browsing and retrieval of image data. IEEE Transactions in PAMI, 18:837–842, 1996.
- [19] G. Schwarz. Estimating dimensions of a model. The Annals of Statistics, 6:461–464, 1978.
- [20] P. Sinha and R. Jain. Concept annotation and search space decrement of digital photos using optical context information. In Proc of Multimedia Content Access: Algorithms and Systems, SPIE Electronic Imaging, 2008.
- [21] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content based image retrieval at the end of the early years. IEEE Transactions in PAMI, 22:1349–1380, 2000.
- [22] M. Tuffield, S. Harris, D. Dupplaw, A. Chakravarthy, B. Brewster, N. Gibbins, K. O’Hara, F. Ciravegna, D. Sleeman, N. Shadbolt, and Y. Wilks. Image annotation with photocopain. In International Workshop on Semantic Web Annotations for Multimedia, 2006.
- [23] T. Westerveld and A. de Vries. Experimental evaluation of a generative probabilistic image retrieval model on ‘easy’ data. In Proc of ACM SIGIR Multimedia Information Retrieval Workshop, 2003.