Feature-Independent Context Estimation for Automatic Image Annotation*

Amara Tariq Hassan Foroosh
The Computational Imaging Lab., Computer Science,
University of Central Florida, Orlando, FL, USA

Abstract

Automatic image annotation is a highly valuable tool for image search, retrieval and archival systems. In the absence of an annotation tool, such systems have to rely on either users' input or large amount of text on the webpage of the image, to acquire its textual description. Users may provide insufficient/noisy tags and all the text on the webpage may not be a description or an explanation of the accompanying image. Therefore, it is of extreme importance to develop efficient tools for automatic annotation of images with correct and sufficient tags. The context of the image plays a significant role in this process, along with the content of the image. A suitable quantification of the context of the image may reduce the semantic gap between visual features and appropriate textual description of the image. In this paper, we present an unsupervised feature-independent quantification of the context of the image through tensor decomposition. We incorporate the estimated context as prior knowledge in the process of automatic image annotation. Evaluation of the predicted annotations provides evidence of the effectiveness of our feature-independent context estimation method.

1. Introduction

Image search and retrieval systems rely heavily on the availability of the textual descriptions of images to satisfy the textual queries of users. Such systems can benefit greatly from image annotation systems which aim at producing accurate and concise textual descriptions for images. Automatic image annotation is a very challenging problem as low-level visual features (HOG, mean and standard deviation of color channels, edge filters, etc.) used to describe the *contents* of an image do not represent any ready and comprehensible connection to the textual description of the image. This distance between low-level visual features and textual description of image is described as *semantic gap*. The *context* of the image may help reduce this gap. The *context* information may be used as prior knowledge in the process of automatic image annotation to bridge the gap between low-level representation of *content* of the image and its textual description.

Estimation of the *context* of image is essential to the task of incorporating this information in image annotation process. The meta-data or any additional data available with images may provide understandable *context* but the availability of such additional information is not a very practical assumption. In the ideal scenario, the *context* should be estimated from the image itself. Some forms of image representation may be suitable for *context* representation. For example, visual features representing scene of an image may be used as prior knowledge in the process of identification of details of the scene, i.e., *content* of the image. Still, the *context* estimation task is tied to some for of visual features with the inherent problem of *semantic gap*.

In this paper, we propose a feature-independent and unsupervised *context* estimation process. The proposed process does not depend upon availability of additional information with images or any form of visual features. It involves tucker decomposition of tensors, typically used for video processing. We devise a unique strategy to transform images into suitable tensors, which are capable of providing useful *context* information for individual images. We incorporate estimated *context* in the process of automatic image annotation as prior knowledge. The evaluation of this process over two popular image annotation datasets, i.e., IAPR[1] and ESP game[2], provides encouraging evidence of the effectiveness of our *context* estimation strategy.

The rest of this paper is arranged as follows. We present a survey of image annotation and tensor decomposition related literature in section 2. The problem is properly formulated in section 3. The proposed feature-independent *context* estimation strategy is presented in section 4. In section 5, we describe the annotation scheme incorporating *context* information. We explain the intuition behind our *context* estimation strategy in section 6. Sections 7 and 8 describe the results of our evaluation and conclusion, respectively.

*This work was supported in part by the National Science Foundation under grants IIS-1212948 and IIS-091686.

[1] http://www.imageclef.org/photodata

[2] http://hunch.net/ jl/

2. Related Work

Automatic image annotation is a well-studied problem and various strategies have been previously proposed to solve this problem. One popular class of frameworks is inspired by relevance models used for machine translation problem from the domain of natural language processing[12, 16, 6, 22]. These methods estimate joint probability of words with some visual representation of images by assuming a generative model solved by expectation over training data. These frameworks are computationally efficient and moderately accurate. Several methods have been proposed that achieve significantly better performance than the relevance model based frameworks, at the cost of increased computational complexity. These methods rely on some iterative optimization of the system, assuming that the nearest-neighbors of an image encode information of the appropriate tags for that image[9, 17, 3, 29]. Object identification tools from computer vision have been employed in the annotation process[21, 15]. Such methods usually work with very limited vocabulary sets.

Some systems have been proposed to use auxiliary information available with images in the process of image annotation. Usually these systems work with images from news datasets as these images have accompanying news articles[7, 8]. Auxiliary information provides *context* for each image that helps reduce the *semantic gap* between visual features and textual descriptions. Tariq et al. proposed to estimate *context* without auxiliary information, using scene analysis of images[27].

In this paper, we propose to remove the dependence of *context* estimation process on auxiliary information as well as any form of visual features.

Tensors have been used as a natural representation scheme for videos, text document collections and image ensembles [4, 14, 28, 1, 10]. Tensor analysis and decomposition algorithms have been applied to the tasks such as action recognition and motion detection, in the domain of video analysis[23, 19, 30, 26]. Kolda et al. presented a detailed study of tensor decomposition methods along with their applications[13].

In this paper, we present a novel strategy for forming useful tensors from independent images and using decomposition of tensors as a source of *context* information to be employed in image annotation process.

3. Problem Formulation and Notations

Our aim is to estimate useful *context* information to be incorporated in the process of automatic annotation of images with proper textual tags. We assume the availability of training data which is a reasonable assumption made by every annotation and prediction system. The training dataset, denoted by $\mathbb{I}$, consists of images and their textual descriptions. Let each training sample be denoted by I. There is a fixed set of vocabulary, say $\mathbf{V}$, made up of words used in the descriptions of all $I \in \mathbb{I}$. Let the size of $\mathbf{V}$ be N. There are test images available, each denoted by I_o. These are the images which need to be annotated with proper words from the vocabulary set $\mathbf{V}$. The goal of the annotation system is to come up with a subset $\{w_1, w_2, ..., w_B\}$ of vocabulary words for I_o such that each w_b of this subset is a suitable annotations for image I_o.

We assume that the*context* information is encoded in the group structure of training images where training images in one group have some 'relation' to other members of the group. This 'relation' shared by the members of one group is defined in section 4.1. Each test image has some association to these groups which encodes its *context* information.

4. Estimation of Context Information

Estimation of *context* is a three-step process in the proposed system. The first task is to form a limited number of groups of the training data such that all images in one group have some 'relation' with each other. The next step involves formation and then decomposition of tensors, made up of images of each group. In the last step, *context* of test images is estimated by modifying tensors and comparing new decomposition results against those generated in the second step.

Assume that each training image $I \in \mathbb{I}$ has its textual description encoded in a vector $\mathbf{v}$ of length N. Each entry of v_n of vector $\mathbf{v}$ indicates presence or absence of the corresponding word of the vocabulary in the textual description of image I.

4.1. Context groups

In the first step, we need to construct groups, termed *context groups*, of images such that all images in one group have some 'relation' to each other. Our system requires two things from this process; *1)* each image group should be a representation of some *context* that is capable of aiding the prediction of appropriate annotations for images, *2)* images of one group should have sufficient visual similarity to each other so that the group could be used as basis for formation of visual signature of the *context*.

The textual descriptions of the images in the training set are available. It is intuitive to assume that the textual description of an image predicts its visual representation. We calculate *tfidf* representation of the textual description of each image I, denoted by $\mathbf{v}'$ which is a vector of length N. If N_{nI} is the number of times n_{th} word appears in the description of image I and N_n is the cumulative frequency of n_{th} word in the dataset, n^{th} entry of $\mathbf{v}'$ is

$$v'_n = \frac{N_{nI}}{N_n} \tag{1}$$

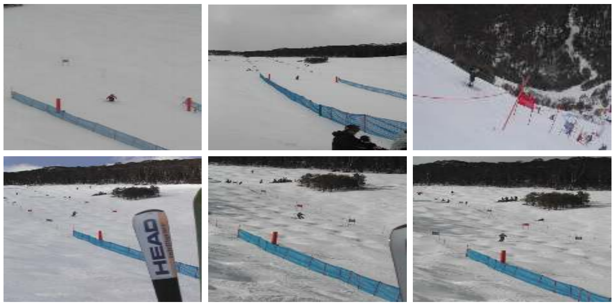

Figure 1. An example of *context group* formed on the basis of similarity in textual descriptions

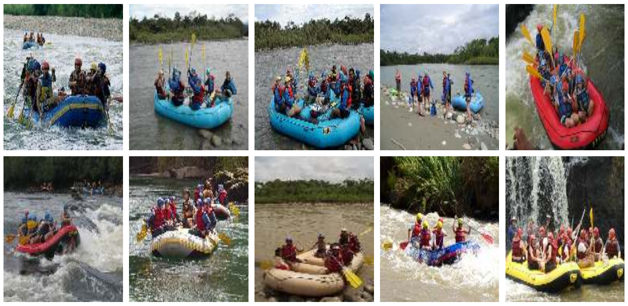

Figure 2. An example of *context group* formed on the basis of similarity in textual descriptions

The images in the training set are clustered based on the cosine similarity between their *tfidf* vectors. The properties of *tfidf* representation ensures that this process groups images with the same *distinctive* words in their textual descriptions, together. Each image group will be able to uniquely provide evidence for those words. For example, if the word 'sky' occurs commonly in descriptions across all the training data, it is part of 'general' vocabulary of the dataset and is not a *distinctive* feature of any of the image groups. On the other hand, if 'snow' is a tag for a few images, the group of those images can uniquely provide evidence for the tag 'snow'. Moreover, the images grouped together have high similarity among their textual representations. Therefore, we justifiably assume reasonable visual similarity between images of one group. We employed an iterative hierarchical clustering process with cut-off threshold to control the number of clusters at each iteration. Large clusters were further split in subsequent iterations to keep *context group* size distribution as uniform as possible.

4.2. Tensor formation and decomposition

One *context tensor* $\mathcal{T}_c \in \mathbb{R}^{X\times Y\times Z}$ is constructed for each of the image groups formed in the first step. The images in one group are resized to a fixed height Y and width X, converted to gray-scale, processed though a Gaussian blurring filter and concatenated together to form the tensor. Three dimension, i.e., x, y and z, of this tensor represent image width, image height and image indices, respectively.

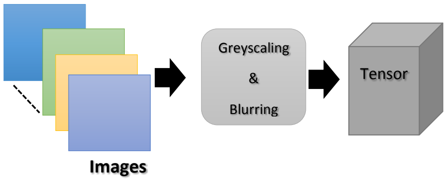

Figure 3. Tensor formation: images of one group are stacked together to form one tensor

Notice that the goal of this process is to estimate an overall signature of the *context* while *context* is encoded in the *distinctive* words of image descriptions in one group. This *context signature* should be made insensitive to fine visual details so that when association of a new image to any of the *context signatures* is assessed, it focuses on global similarity between the new image and member images of that *context group* and not on the local details of images. Therefore, images of one group are all processed by a blurring Gaussian filter to remove sharp distinctions because of edges.

The next step is the decomposition of the *context tensor* through Tucker decomposition to find a compact *signature* of the *context group*. Tucker decomposition is a popular technique to project tensor $\mathcal{T}_c \in \mathbb{R}^{X\times Y\times Z}$ onto a smaller core tensor S and three matrices P, Q, and R such that

$$\mathcal{T}_c \approx \mathcal{S}\times_1 P\times_2 Q\times_3 R = \sum_{x=1}^{X}\sum_{y=1}^{Y}\sum_{z=1}^{Z} g_{xyz}p_x\circ q_y\circ r_z, \quad (2)$$

where $P \in \mathbb{R}^{X\times K}$, $Q \in \mathbb{R}^{Y\times K}$, and $R \in \mathbb{R}^{Z\times K}$ are the orthogonal matrices, $\mathcal{S} \in \mathbb{R}^{K\times K\times K}$ is the core tensor and $K \leq min(X, Y, Z)$. The $\overline{\times}_i$ operator denotes the multiplication between a tensor and a vector in mode-i of that tensor, whose result is also a tensor, namely, $\mathcal{A} = \mathcal{B}\overline{\times}_i\alpha \Longleftrightarrow (\mathcal{A})_{jk} = \sum_{i=1}^{I}\mathcal{B}_{ijk}\alpha_i$.

We apply a rank-1 decomposition, i.e., K is set to 1. In this case, P, Q, R are vectors with lengths equal to the width of the image, the height of the image and the size of the *context group*, respectively. Vector $R \in \mathbb{R}^{Z\times 1}$ is the most important for our system. This vector represents the similarity/dissimilarity of one image to its neighboring images in the tensor $\mathcal{T}_c$. Since all images concatenated together belong to one *context groups*, i.e., they are all visually similar as they all have highly similar textual descriptions, there should be only small variations in the entries of this vector. Vector R is the compact *signature* for the *context group*.

4.3. Context estimation

The next step is to quantify *context* of test images in terms of their association with different *context signatures*. Let each test image be represented as I_o. There is no textual description available for I_o. As we explained in previous section, *context signature* is a vector of length R with little variation across its entries as it is the result of tucker decomposition of a tensor made up of R visually similar images belonging to one *context group*. If a foreign entity, e.g., a test image I_o, is inserted into this tensor at any location, say l, it will disturb entries at and around index l in the vector R. The amount of disturbance will be proportional to the dissimilarity between I_o and the members of that *context group*.

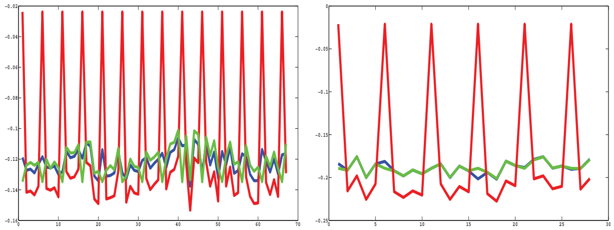

Figure 4. Comparison of rank-1 tucker decomposition with visually similar and dissimilar image inserted into a tensor; Blue curve: original tucker decomposition vector R, Green curve: New tucker decomposition vector R' with an image visually similar to images of the *context group* inserted into *context tensor*, Red curve: New decomposition vector R' with visually dissimilar image inserted into tensor.

To estimate association of a test image I_o with a *context group*, it is inserted at locations separated by a fixed interval, say L, in the corresponding *context tensor* $\mathcal{T}_c$ by swapping images at those locations for I_o. New vector R' is computed through tucker decomposition. The difference between R and R' is an inverse measure of the association of I_o with the *context* represented by the *context group* corresponding to $\mathcal{T}_c$. We estimate conditional probability distribution for I_o given every possible *context group* as

$$P(\mathcal{T}_c|I_o) = \frac{\exp(-(R'-R)^T \Gamma^{-1}(R'-R))}{\sqrt{2\pi|\Gamma|}} \quad (3)$$

Γ is covariance matrix, assumed to be of form $\gamma\mathbf{I}$ where $\mathbf{I}$ is identity matrix and γ can be selected empirically over some held-out portion of data. This probability distribution encodes the association of the test image I_o with available *context groups*, in turn encoding its association with sets of *distinctive* words of each group.

As mentioned earlier, words occurring too frequently are given less weight in the process of forming *context groups*. To service such words, we also form a 'general' *context group* consisting of all training images. Each test image I_o is assigned the same conditional probability, given 'general' *context group*, and $P(\mathcal{T}_c|I_o)$ is renormalized so that it sums to 1. Let α denote the renormalization weight which is empirically estimated by cross-validation over a held-out portion of training data.

Note that no visual features have been employed in the 3-step process. Instead, a comprehensive estimate of *context* in terms of a probability distribution is obtained using the textual labels of the training data and processing of raw images through tucker decomposition.

5. Context-sensitive Automatic Image Annotation

In this section, we will present our strategy for automatic image annotation that incorporates *context* information, estimated as distribution $P(\mathcal{T}_c|I_o)$ in section 4.

Our annotation model is inspired by relevance models, used in machine translation to estimate joint probability between words of two different languages. The proposed model induces *context*-dependence in a weighted expectation procedure to find joint probability of vocabulary words and visual features of images.

5.1. Mathematical Model

Each image is assumed to be made up of A number of visual units, i.e., $\mathbf{r} = \{r_1, r_2, ..., r_A\}$. These visual units are formed by dividing each image through a grid of fixed size. Color and texture qualities of each section of the grid form the representative vector for that section. The textual description of the image is represented by set $\mathbf{w} = \{w_1, w_2, .., w_B\}$ such that each $w_b \in \mathbf{V}$ where $\mathbf{V}$ is the vocabulary set. Size of set $\mathbf{w}$, say B, is assumed to be the same for all test images.

Context information is incorporated by assuming that there exists a set of all *context* categories, i.e., $\mathbb{T}$ such that each $\mathcal{T}_c \in \mathbb{T}$ corresponds to one *context group* and, in turn, one *context tensor* (defined in sections 4.1 and 4.2). Each training image I belongs to one of these $\mathcal{T}_c$. Every test image I_o has certain conditional probability distribution over all $\mathcal{T}_c$, denoted by $P(\mathcal{T}_c|I_o)$. The *context* is encoded in $P(\mathcal{T}_c|I_o)$. Tariq et al. presented a similar case for *context* estimation by scene analysis[27]. Therefore, we employ similar relevance model weighted by *context*, i.e., $P(\mathcal{T}_c|I_o)$, to estimate joint probability of words and visual units of I_o.

1. pick a *context category* $\mathcal{T}_c \in \mathbb{T}$ with probability conditioned over the test image I_o, i.e., $P(\mathcal{T}_c|I_o)$

2. pick image I from training set $\mathbb{I}$ with probability $P(I|\mathcal{T}_c)$

3. for $a = 1, 2, ..., A$

   (a) pick a visual unit r_a from conditional probability $P_{\mathbb{R}}(.|I)$

4. for $b = 1, 2, ..., B$

   (a) pick a word w_b from conditional probability $P_{V_{\mathcal{T}_c}}(.|I)$

The goal of the system is to maximize joint probability of $\mathbf{r}$ and $\mathbf{w}$ conditioned over I_o, given by following equation.

$$P(\mathbf{w}, \mathbf{r}|I_o) = \\ \sum_{\mathcal{T}_c \in \mathbb{T}} P(\mathcal{T}_c|I_o) \sum_{I \in \mathbb{I}} P(I|\mathcal{T}_c) \prod_{b \in B} P_{\mathbf{V}_{\mathcal{T}_\mathbf{c}}}(w_b|I) \prod_{a \in A} P_{\mathbb{R}}(r_a|I) \tag{4}$$

Image description of training image I is assumed to have multiple Bernoulli distribution over the vocabulary set of the *context group* it belongs to, i.e., $\mathbf{V}_{\mathcal{T}_\mathbf{c}}$. Thus, $P_{\mathbf{V}_{\mathcal{T}_\mathbf{c}}}(w_b|I)$ is w_b*-th* component of this distribution.

$$P_{\mathbf{V}_{\mathcal{T}_\mathbf{c}}}(w_b|I) = \frac{\mu \delta_{w_b} + N_{w_b c}}{\mu + N_{\mathcal{T}_c}} \tag{5}$$

$N_{w_b c}$ denotes the number of members of the *context group* $\mathcal{T}_c$ with word w_b in their descriptions. $N_{\mathcal{T}_c}$ is the total number of members of $\mathcal{T}_c$. δ_{w_b} is set to 1 if description of image I has word w_b in it. Otherwise, It is set to 0. μ is an empirically selected constant.

Section 4.3 explains the estimation of $P(\mathcal{T}_c|I_o)$ by equation 3 while $P(I|\mathcal{T}_c)$ is estimated as the following step function.

$$P(I|\mathcal{T}_c) = \begin{cases} 1/N_{\mathcal{T}_c}, & \text{if } I \in \mathcal{T}_c \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

$P_{\mathbb{R}}(r_a|I)$ is the density estimate for generating visual unit r_a given a training image I where r_a is a visual unit of interest, i.e., it belongs to the test image I_o. Gaussian kernel is employed for this density estimate. If training image I is assumed to be made up of set of visual units $\{i_1, i_2, ..., i_A\}$, then

$$P_{\mathbb{R}}(r_a|I) = \frac{\exp(-(r_a - i_a)^T \Sigma^{-1} (r_a - i_a))}{\sqrt{2\pi|\Sigma|}} \tag{7}$$

This equation uses Gaussian density kernel with covariance matrix Σ which can be taken as $\beta\mathbf{I}$ for convenience where $\mathbf{I}$ is the identity matrix. β determines smoothness around point i_a and can be empirically selected on held-out set of data. Note that this estimate signifies importance of spatial coherence between I and I_o as it compares the visual units at the same grid location, indicated by subscript a.

This model incorporates estimated *context* information for each test image I_o in the form of $P(\mathcal{T}_c|I_o)$ and our experiments show that this information improves the performance of the annotation system.

6. Implications of Tensor Decomposition

In this section, we present a brief analysis of tensor formation and decomposition along with the complexity of the decomposition process and its implications regarding *context* estimation for images.

The idea of tensor formation and decomposition has been widely explored in text mining and video analysis communities. Tensor provides a comprehensive representation for videos such that each frame of the video is a 'slice' in a tensor. Two out of three dimensions are representative of frame width and height while the third dimension represents time. Thus, tensors are highly suited for temporal analysis of videos. Our contribution in this work is to come up with a comprehensive tensor formation strategy for images which have no temporal connection to each other. In our case, the third dimension is used for image indices only.

Tucker decomposition of three-way tensors is a higher-order extension of Principal Component Analysis (PCA) of matrices[13]. It is a rank based estimation which results in the decomposition of the tensor in three matrices and one core tensor where size of the core tensor is pre-specified. Assume that the three dimensions represent words, authors and keywords for a tensor made up of documents with available authorship and keywords information. Three decomposed matrices U, V and W represent association of words with word-groups, authors with author-groups and keywords with keyword-groups, respectively. The number of word-groups, author-groups and keyword-groups are specified by the size of core tensor. Core tensor encodes how groups relate to each other.

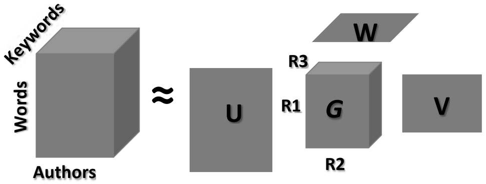

Figure 5. Tucker decomposition: $U = words \times word\text{-}groups, V = authors \times author\text{-}groups, W = keywords \times keyword\text{-}groups$, $R1, R2$ and $R3$ represent word, author and keyword groups

The proposed *context* estimation strategy employs rank-1 decomposition, i.e., core tensor is a scalar and matrices are now vectors. Idea is that the system already knows that all images in one tensor belong to one group, based on the

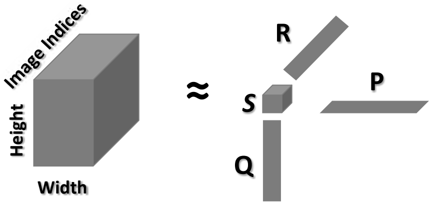

Figure 6. Rank-1 Tucker decomposition: $\mathcal{S}$ is a scalar, P, Q and R are vectors, $R = Image\text{-}indices \times 1$ where 1 represent the single *context group* represented by tensor.

similarity in their textual descriptions. This type of group information is potentially useful in the final task of the proposed system, i.e., image annotation. The purpose of tucker decomposition is to find out how individual elements of one group relate to the overall group so that the system may determine if some entity belongs to the group or not. Ideally, there should be little variation in the vector along the dimension of indices of images as all images are similar to each other. If a foreign entity is plugged in, this vector is perturbed. The amount of perturbation provides an estimate of how much similar/dissimilar the foreign entity is, to the group. If foreign entity is the test image, as in section 4.3, this process estimates how much the test image is similar/dissimilar to the *context group* at hand.

6.1. Computational Complexity

Computational complexity of the proposed *context* estimation scheme depends on the strategy used for Tucker decomposition. Popular existing algorithms for Tucker decomposition, such as *higher order orthogonal iterations* (HOOI)[5], are based on *alternating least square* (ALS). Phan et al. proposed a method, computationally less expensive than HOOI[24]. ALS method is not guaranteed to converge to a global optimum or a stationary point, but if it converges under certain conditions, then it has local linear convergence rate[2]. Alternatively, differential-geometric Newton method provides convergence guarantee with quadratic local convergence rate and per iteration cost of $\mathcal{O}(H^3D^3)$ for a tensor $\mathcal{T} \in \mathbb{R}^{H\times H\times H}$ and core tensor $\mathcal{S} \in \mathbb{R}^{D\times D\times D}$[11].

7. Evaluation

In this section, we will present the effects of incorporation of *context*, estimated by the proposed feature-independent strategy, in relevance model based automatic image annotation process.

7.1. Datasets

We used two popular image annotation datasets, i.e., IAPR-TC 12 and ESP game, to evaluate our system. IAPR dataset consists $19,846$ images taken by tourists and each image is described carefully in a few sentences. The descriptions of images are processed by TreeTagger[3] for tokenization, lemmatization and part-of-speech tagging of the tokens. Frequently occurring nouns are picked to form vocabulary set. ESP game dataset consists of images labeled by players of ESP game. A smaller subset of size $21,844$ has been popularly used to test different image annotation systems. We also experimented with the same subset. The description for each image is already in the form of tokens/ words which are used to form the vocabulary set.

We used the same split of data in training and test sets (90% for training, 10% for test) for both datasets as used by other image annotation systems. IAPR and ESP datasets have been generally tested over vocabulary sets of 291 and 269 most frequently occurring words, respectively, by various image annotation systems. In our system, training data is split into *context groups* and each group provides evidence for the *distinctive* words of descriptions of images in that group. Thus, the vocabulary varies from collection of samples of one *context group* to the other, instead of being fixed to a specific number for all data. But we made sure that approximately the same number of unique words (291 for IAPR and 269 for ESP) appear in the final output, i.e., annotations predicted for test images, by adjusting parameters of our system. In this paper, results have been reported over these unique words to keep them comparable to those of other systems. The overall vocabulary sets are of lengths 1002 and 2032 for IAPR and ESP game datasets, respectively.

Our system, in the first stage, forms groups of training images on the basis of similarity between their textual descriptions, i.e., 595 *context groups* for IAPR and 637 *context groups* for ESP game. Too small groups are dropped as they correspond to infrequently occurring words out of the overall vocabulary set. A training image is pruned out of its *context group* if the distance of its corresponding entry in *context signature* (R), from mean of R, is more than the standard deviation of R. Such images fail the condition of visual similarity with their *context group*. All test images which are sufficiently close to at least one *context group* are part of the test set.

7.2. Image Representation

We employed a grid-based representation of images in the automatic image annotation process. This representation requires division of each image through a grid of fixed size. A grid of 5×6 is used in our experiments. Each grid section is assigned a vector representing color and texture qualities of that specific portion of the image. In our experiments, this vector is of length 46 and contains 18 color features (mean and std. deviation of each channel of RGB, LUV and LAB color-spaces), 12 texture features (Gabor energy computed over 3 scales and 4 orientations), 4 bin HoG and discrete cosine transform coefficients. This set of image features has been commonly used by many previously proposed image annotation systems[16, 6]. We observed that increasing the grid size beyond 5×6 does not improve performance of the system.

Guillaumin et al. employed a combination of holistic and local visual features and reported improvement in the performance of their system[9]. More recently, Chen et al. and Verma et al. used the same features in their systems [3, 29]. We also ran additional experiments with this combination

[3] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

of feature set and observed performance improvement.

Note that these features are used in the relevance model based image annotation process and have nothing to do with the *context* estimation procedure. Initial stage of our system is feature-independent, estimating *context* information for test images by processing raw images only.

7.3. Results

In general, image annotation systems are used to produce as many annotations for each test image I_o as is the average number of words per image in training data. Commonly used evaluation parameters are mean values of precision and recall per word and number of words with positive recall (N^+). We used the same evaluation criterion. Tables 1 and 2 show performance comparison of our system against many previously proposed strategies over IAPR-TC 12 and ESP datasets, respectively. Two variations of our system have been thoroughly tested. Two notations, i.e., **context-RM** and **context-RM-b**, represent settings in which our system employs grid-based visual features and features presented by Guillaumin et al., respectively, for weighted expectation based annotation process. Table 3 presents samples of words with very high and very low recall for both datasets.

	%age mean Precision	%age mean Recall	N^+
CRM[16]	21	15	214
MBRM[6]	21	14	186
BS-CRM[22]	22	24	250
JEC[20]	25	16	196
Lasso[20]	26	16	199
HGDM [18]	29	18	–
AP[25]	28	26	–
TagProp-ML[9]	48	25	227
TagProp[9]	46	35	266
FastTag[3]	47	26	280
2PKNN-ML[29]	54	37	278
context-RM	56	24	224
context-RM-B	61	24	242

Table 1. Performance evaluation for IAPR-TC-12 dataset

	%age mean Precision	%age mean Recall	N^+
CRM[16]	29	19	227
MBRM[6]	21	17	218
JEC[20]	23	19	227
Lasso[20]	22	18	225
AP[25]	24	24	–
TagProp-ML[9]	49	20	213
TagProp[9]	39	27	239
FastTag[3]	46	22	247
2PKNN-ML[29]	53	27	252
context-RM	55	21	226
context-RM-B	61	20	234

Table 2. Performance evaluation for ESP-game dataset

7.4. Observations

As explained in section 2, different strategies have been previously explored for the task of image annotation while each strategy has its pros and cons. In the tables 1 and 2, CRM and MBRM refer to two relevance model based annotation techniques which are very efficient computationally and perform moderately well. Our annotation strategy is also based on relevance models but incorporates the *context* estimated through our novel feature-independent strategy. Our annotation prediction framework performs much better than other relevance model based systems. TagProp, FastTag and 2PKNN-ML refer to a few iterative optimization or nearest-neighbor type frameworks which are computationally quite expensive but predict annotations more accurately. Our strategy performs better than such systems in terms of precision of predicted annotations. Performance of our system is comparable in terms of recall of predicted annotations to FastTag and TagProp-ML. The bulk of the computational complexity lies in the pre-processing stage of our system which involves *context* estimation. The rest of our system is computationally efficient. Our strategy also beats greedy algorithms based systems such as JEC and Lasso[20].

8. Conclusion

We proposed a novel strategy for feature-independent *context* estimation for images, employing tensor decomposition. Tensors have been previously suggested as a natural representation scheme for video processing. Our contribution is a unique solution for forming tensors from individual images in a way that each tensor encodes useful information regarding *context* of images. The proposed *context* estimation strategy is feature-independent. We employed the estimated *context* in the process of automatic image annotation, a problem that usually suffers from *semantic gap* between visual features and textual descriptions. The performance of our annotation strategy provides evidence of the effectiveness of our *context* estimation process. In future, we intend to explore tensor decomposition, with rank more than one, for *context group* formation as well as *context* estimation.

References

[1] B. W. Bader, M. W. Berry, and M. Browne. Discussion tracking in enron email using parafac. In *Survey of Text Mining II*, pages 147–163. Springer, 2008.

[2] B. CHEN, Z. LI, and S. ZHANG. On tensor tucker decomposition: The case for an adjustable core size.

IAPR	High recall	counter, fielder, root, advertising, minibus, steel, neck, block, junction, bookshelf, sky, concrete
	Low recall	hair, canoe, wood, monkey, writing, grassland, green, cape, finish, face
ESP	High recall	Haryana, Europe, visa, Punjab, fortune, station, university, vegetable, Mars
	Low recall	Swing, surf, wood, crystal, cartoon, Chinese, stick, airplane, bark

Table 3. Sample of words with low and high recall values

[3] M. Chen, A. Zheng, and K. Q. Weinberger. Fast image tagging, 2013.

[4] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[5] L. De Lathauwer, B. De Moor, and J. Vandewalle. On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1324–1342, 2000.

[6] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[7] Y. Feng and M. Lapata. Automatic image annotation using auxiliary text information. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, 2008.

[8] Y. Feng and M. Lapata. Topic models for image annotation and text illustration. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010.

[9] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *IEEE 12th International Conference on Computer Vision*, 2009.

[10] Z. Hao, L. He, B. Chen, and X. Yang. A linear support higher-order tensor machine for classification. *IEEE Transactions on Image Processing*, 22(7):2911–2920, 2013.

[11] M. Ishteva, L. De Lathauwer, P.-A. Absil, and S. Van Huffel. Differential-geometric newton method for the best rank-(r 1, r 2, r 3) approximation of tensors. *Numerical Algorithms*, 51(2):179–194, 2009.

[12] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003.

[13] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[14] G. Kuhne, J. Weickert, O. Schuster, and S. Richter. A tensor-driven active contour model for moving object segmentation. In *Proceedings of IEEE International Conference on Image Processing*, 2001.

[15] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[16] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Advances in neural information processing systems*, 2003.

[17] X. Li, C. G. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 11(7), 2009.

[18] Z. Li, Z. Shi, W. Zhao, Z. Li, and Z. Tang. Learning semantic concepts from image database with hybrid generative/discriminative approach. *Engineering Applications of Artificial Intelligence*, 26(9), 2013.

[19] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. A survey of multilinear subspace learning for tensor data. *Pattern Recognition*, 44(7):1540–1551, 2011.

[20] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *Computer Vision–ECCV 2008*. 2008.

[21] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daum III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of Annual Meeting of European Association of Computational Linguistics*, 2012.

[22] S. Moran and V. Lavrenko. Optimal tag sets for automatic image annotation. In *Proceedings of the British Machine Vision Conference*, 2011.

[23] K. Palaniappan, I. Ersoy, G. Seetharaman, S. R. Davis, P. Kumar, R. M. Rao, and R. Linderman. Parallel flux tensor analysis for efficient moving object detection. In *Proceedings of the 14th IEEE International Conference on Information Fusion (FUSION)*, pages 1–8, 2011.

[24] A.-H. Phan, A. Cichocki, and P. Tichavsky. On fast algorithms for orthogonal tucker decomposition. In *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2014.

[25] M. Rubinstein, C. Liu, and W. T. Freeman. Annotation propagation in large image databases via dense image correspondence. In *Computer Vision–ECCV 2012*.

[26] C. Sun, M. Tappen, and H. Foroosh. Feature-independent action spotting without human localization, segmentation, or frame-wise tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[27] A. Tariq and H. Foroosh. Scene-based automatic image annotation. In *IEEE International Conference on Image Processing*, 2014.

[28] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *Computer Vision–ECCV 2002*.

[29] Y. Verma and C. Jawahar. Image annotation using metric learning in semantic neighbourhoods. In *Computer Vision–ECCV 2012*.

[30] Q. Zhao, G. Zhou, L. Zhang, and A. Cichocki. Tensor-variate gaussian processes regression and its application to video surveillance. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.