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.

Automatic image annotation is an important tool for image search, retrieval and archival systems. Automatic annotation process usually suffers from *semantic gap*, i.e., the lack of any correlation between low-level visual features and textual annotations. We propose to tackle this problem through incorporation of *context* of an image as well as its *contents*, in the annotation process. We present a unique strategy for *context* estimation. Our strategy is based on tensor analysis of raw images. Tensors have been employed as a natural representation scheme for videos. Novelty of our approach is to use tensor analysis to extract useful *context* information from raw images. The proposed *context* estimation process is feature-independent, thus avoiding the *semantic gap* problem associated with any form of low-level visual features.

We propose a three-step process for *context* estimation.

- Images are clustered into groups such that
  - Each cluster is a representative of some *context category*, defined by *distinctive* words used in the descriptions of its member images
  - All member images of every cluster have enough visual similarity to each other to be able to define some *visual signature* for the corresponding *context category*.

  We employed hierarchical clustering technique based on cosine similarity between *tfidf* vectors of descriptions of images in training data to construct these *context categories*. Nature of *tfidf* representation ensures that clustering process puts an emphasis on grouping together images with similar *distinctive* words in their descriptions. If word 'sky' is common among image descriptions, it is not a *distinctive* word. If word 'snow' appears in descriptions of a few images, it is a *distinctive* word for those images. High similarity between image descriptions indicates a reasonable level of visual similarity between images.

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

- Images in one cluster/*context category* are resized to fixed height and width, converted to gray-scale, blurred by a Gaussian filter and concatenated together to form a *context tensor*, denoted by $\mathcal{T}_c$. Rank-1 Tucker decomposition generates three vectors and a scalar value for this tensor. The vector R corresponding to the tensor dimension indicating indices of images contains information on how each image relates to other images of the same *context category*. This vector forms a compact *signature* of the corresponding *context category*.

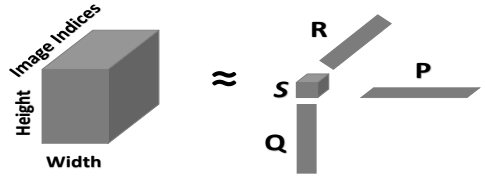

Figure 2: Rank-1 Tucker decomposition: $\mathcal{S}$ is a scalar, P, Q and R are vectors, $R = Image\text{-}indices \times 1$ where 1 represent the single *context group* represented by tensor.

- After formation of *signatures* of *context categories* defined over training data, the next step is to calculate association of a test image I_o to all *context categories* as probability distribution $P(\mathcal{T}_c|I_o)$. Each tensor $\mathcal{T}_c$ is updated by swapping every L^{th} image in the tensor by the test image I_o. Tucker decomposition is applied to calculate updated *signature* vector R'. Insertion of a foreign entity, such as a test image, disturbs entries at and around every L^{th} location in vector R. Amount of disturbance is proportional to the dissimilarity between foreign entity and images in the neighborhood of every L^{th} index. $P(\mathcal{T}_c|I_o)$ is given by

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

  where γ is assumed to be a uniform diagonal matrix whose diagonal entries are equal to an empirically selected constant.

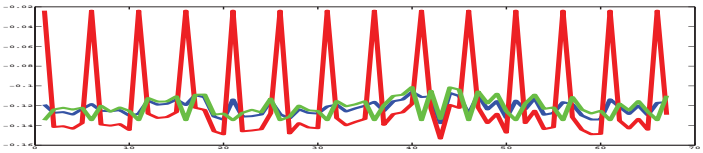

Figure 3: Comparison of rank-1 tucker decomposition with visually similar and dissimilar image inserted into a tensor; Blue curve: original tucker decomposition vector R, Green curve: New tucker decomposition vector R' with an image visually similar to images of the *context group* inserted into *context tensor*, Red curve: New decomposition vector R' with visually dissimilar image inserted into tensor.

We employ an expectation process weighted by estimated *context*, inspired by relevance model from machine translation, that maximizes the following equation

$$P(\mathbf{w},\mathbf{r}|I_o) = \sum_{\mathcal{T}_c\in\mathbb{T}} P(\mathcal{T}_c|I_o)\sum_{I\in\mathbb{I}} P(I|\mathcal{T}_c)\prod_{b\in B} P_{\mathbf{V}_{\mathcal{T}_c}}(w_b|I)\prod_{a\in A} P_{\mathbb{R}}(r_a|I) \quad (2)$$

Each image $I \in \mathcal{I}$ is made up of a few visual unites r_a and words w_b where $\mathcal{I}$ denotes training set. $\mathcal{V}_{\mathcal{T}_c}$ denotes vocabulary set of *context category* $\mathcal{T}_c$. A 'general' *context category* is added to service words which are not *distinctive* words of any *category*. $P(I|\mathcal{T}_c)$ is a step distribution. $P_{\mathbf{V}_{\mathcal{T}_c}}(w_b|I)$ is w_b^{th} component of multiple Bernoulli distribution over $\mathcal{V}_{\mathcal{T}_c}$. $P_{\mathbb{R}}(r_a|I)$ is computed by putting a Gaussian kernel over distance between r_a and corresponding visual unit of I.

We evaluated the proposed *context*-sensitive annotation process over IAPR-TC 12 and ESP game dataset. Our experiments indicate that the proposed scheme achieves far better results than other relevance model inspired methods and greedy algorithms based systems, e.g., CRM, MBRM, JEC, Lasso, etc. The proposed scheme also outperforms various time-intensive iterative optimization based methods such as TagProp, FastTag, etc.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

http://www.imageclef.org/photodata
http://hunch.net/ jl/