Trans Media Relevance Feedback for Image Autoannotation

Thomas Mensink[12]
thomas.mensink@xrce.xerox.com
Jakob Verbeek[2]
jakob.verbeek@inrialpes.fr
Gabriela Csurka[1]
gabriela.csurka@xrce.xerox.com

[1] TVPA - Xerox Research Centre Europe
Meylan, France

[2] LEAR - INRIA Rhône-Alpes
Montbonnot, France

In image auto-annotation the goal is to predict relevant keywords from a finite vocabulary given a new image. These predictions can then be used in tools for clustering, classification, retrieval and visualization. A possible solution to auto-annotation is to train a specific classifier for each keyword. An alternative solution is tag propagation, a test image is tagged with the keywords from similar images from the train data set.

In this paper we are building on the ideas of TagProp [2], a probabilistic nearest neighbor model which additionally allows for integrated metric learning. Our contribution is, that we include Trans-Media Relevance Feedback, to obtain a set of nearest neighbors based on both visual and textual similarity. Even though the test image contains only visual information, we are able to exploit the textual modality in the database to improve the performance. Our experiments show that using the available textual information around the images improves the state-of-the-art baseline of TagProp.

Trans-Media Relevance Feedback is a generalization of the well known pseudo-relevance feedback models. The idea is to extend the initial query with information taken from the top k retrieved documents, where the similarity functions used in the two retrieval steps are based on different modalities. For example, given a query image we select the k most similar visual documents. Then, we use the associated text to re-rank all the documents in the database according to their textual similarity. These models have shown significant improvement on retrieval performance in multi-modal databases [1].

TagProp predicts the presence of a keyword for a query image, by taking weighted sum over the annotations of the nearest neighbours images, defined by a visual similarity. The weight $p(j|i)$ to use image j as neighbor for image i is defined as $p(j|i) \propto \exp(-\mathbf{w}^\top \mathbf{d}_{ij})$, where $\mathbf{d}_{ij}$ is a vector of visual base distances, and $\mathbf{w}$ combines the distances, and controls the exponential decay. The parameter vector $\mathbf{w}$ is estimated in a maximum-likelihood fashion. Instead of using all images in the dataset, we use the J nearest neighbors, and assume all other weights are zero.

Trans-media pseudo-relevance feedback is an extension of the relevance feedback principle to multi-modal databases. The idea is that the first retrieval step is done in one modality, while the second step is performed in another modality. The Linear Transmedia Pseudo Relevance Feedback (**LTP**) distance d^{VT} between image i and j is given by:

$$d_d^{VT}(i,j) = \sum_k \gamma_{dk}\, d_d^V(i,k)\, d^T(k,j), \tag{1}$$

where the subscript d means we used the d^{th} base distance from the vector $\mathbf{d}_{ik}$, and γ_{dk} is used to weight the neighbors of the trans-media pseudo-relevance step. If for all k, $\gamma_{dk} = 1$ is used, we obtain the Trans Media similarity as is used in [1].

To learn non-uniform weights γ_{dk}, we can use the TagProp framework and add for each k an entry to the base distances, which has than $(k+1) \times J$ elements. So, we define $f_{ij}^{VT} = \mathbf{w}_{VT}^\top \mathbf{d}_{ij}^{VT}$, where $\mathbf{w}^{VT}$ is the weighting vector, and $\mathbf{d}$ is a vector representation of the distances. Similar, for the visual distances we define $f_{ij}^V = \mathbf{w}_V^\top \mathbf{d}_{ij}^V$. The probability $p(j|i)$ based on both their visual and cross-modal similarities becomes: $p(j|i) \propto \exp(-(f_{ij}^V + f_{ij}^{VT}))$. This can be replaced in the objective function to be maximized.

Alternatively, we propose Softmax Transmedia Pseudo Relevance Feedback (**STP**), a model which requires only 2 parameters for the Trans-Media distance. We use a softmax function on $d_d^V(i,k)$ to define:

$$d_d^{VT}(i,j) = \sum_k \tilde{d}_d^V(i,k)\, d^T(k,j), \text{ with } \tilde{d}_d^V(i,k) \propto \exp(-\gamma_d\, d_d^V(i,k)). \tag{2}$$

This model requires γ_d to define the softmax, and a single weight in $\mathbf{w}$.

Table 1: Performance on the IAPR dataset, combining 4 different base distances d^V, using $J = 400$, $K = 10$, and different d^T

	LTP				STP			
	MAP	BEP	iMAP	iBEP	MAP	BEP	iMAP	iBEP
TagProp	35.7	36.1	49.0	44.1				
d^T ={Tag}	36.0	36.7	49.6	44.6	35.6	36.1	49.2	44.4
d^T ={Text}	36.4	36.7	49.6	44.3	35.7	35.7	49.5	44.2
d^T ={Tag, Text}	36.2	36.6	49.9	44.8	35.9	36.6	49.8	44.6

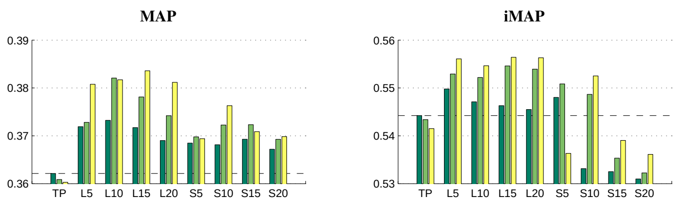

Figure 1: Performance of TagProp (TP), **LTP** (L) and **STP** (S) on COREL image database, with $K = \{5, \ldots, 20\}$, $J = \{200, 400, 1000\}$ (the different bars), and using the tag distance as d^T.

We present a comparative evaluation of our models with the original TagProp [2]. In the experiments, the same visual features as in [2] are used, including local SIFT descriptors, and the global GIST descriptor.

For the d^T distance we use one minus the intersection over union of the set of annotations of two images (tag distance). Besides, we also use a distance based on the full captions for the IAPR database.

To measure the performance as keyword based retrieval system we use mean average precision (**MAP**). For annotation performance, we use **iMAP**, which calculates AP over the ranked tags, and averages over all images. We also include break-even point precision, or R-precision (**BEP** and **iBEP**).

The Corel 5K dataset, contains around 5000 images with manual annotation between 1 to 5 keywords. The images are annotated with for the purpose of keyword-based retrieval. Fig. 1 shows that for most parameter configurations our methods significantly outperform TagProp. Furthermore, the figures show that **LTP** generally outperforms **STP** on this dataset, and that if we increase the neighborhood size J the performances increases.

The IAPR TC12 dataset, contains about 20.000 images accompanied with descriptions, the annotation keywords are extracted nouns of these captions. The TagProp baseline is much harder to beat on this dataset than for Corel 5K dataset.

In Table 1 we show the performance when we use several visual base distances, *i.e.* using the metric learning properties of TagProp. For this experiment we used the 4 distances from TagProp with the highest weights when learned with the 15 described distances. The used features are the GIST, Dense-SIFT, Harris-SIFT, and Dense-SIFT-V3. In this case the **LTP** model outperforms the **STP** model, while without combining the different base distances **STP** outperforms **LTP** on this dataset. Also we can see that using **LTP** with either the Tag or Text distance we improve on all performance measures.

[1] J. Ah-Pine, C. Cifarelli, S. Clinchant, G. Csurka, and J.M. Renders. XRCE's participation to ImageCLEF 2008. In *Working Notes of the 2008 CLEF Workshop*, 2008.

[2] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. TagProp: Discriminative Metric Learning in Nearest Neighbor Models for Image Auto-Annotation. In *ICCV*, 2009.