Improving Tag Transfer for Image Annotation using Visual and Semantic Information

Sergio Rodriguez-Vaamonde¹, Lorenzo Torresani², Koldo Espinosa³, Estibaliz Garrote¹
¹Computer Vision Area – ICT/ESI Division, TECNALIA, Zamudio, Spain
²Visual Learning Group – CS Department, Dartmouth College, Hanover, USA
³Multimedia Group – Electronics and Telecommunications Department – University of the Basque Country, Bilbao, Spain
Corresponding: sergio.rodriguez@tecnalia.com

Abstract— This paper addresses the problem of image annotation using a combination of visual and semantic information. Our model involves two stages: a Nearest Neighbor computation and a tag transfer stage that collects the final annotations. For the latter stage, several algorithms have been implemented in the past using labels’ information or including implicitly some visual features. In this paper we propose a novel algorithm for tag transfer that takes advantage explicitly of semantic and visual information. We also present a structured training procedure based on a concept we have called Image Networking: all the images in a training database are “connected” visually and semantically, so it is possible to exploit these connections to learn the tag transfer parameters at annotation time. This learning is local for the test image and it exploits the information obtained in the Nearest Neighbor computation stage. We demonstrate that our approach achieves state-of-the-art performance on the ImageCLEF2011 dataset.

Keywords— Image annotation, Image indexing, tag transfer, multi-modal information fusion

I. INTRODUCTION

In this work we address the problem of automatic image annotation. Given a test image, an annotation algorithm selects a set of tags or labels describing the input photo content.

In the last decade this has been a very active field of research. Prior work falls in three main categories [10]: generative models, discriminative models and neighbor-based models. The first two models have been widely explored, especially the discriminative models where an individual classifier is trained for each possible tag. However, some studies show that neighbor-based models are more appropriate for image annotation when dealing with hundreds of tags [9][1].

A Nearest Neighbor model is defined by two main stages: first, a nearest neighbor method collects the training images that are most similar to the test image. Each of the training images has a set of tags so in a second stage a tag transfer algorithm selects the relevant tags for the test image.

There is prior relevant work on using Nearest Neighbor models for image annotation. In [8] each image is described using several visual characteristics (i.e. color and edges histograms) and then the Joint Equal Contribution (JEC) distance is used to measure distances between images. In the tag transfer stage, a greedy transfer is proposed: the final tags are selected from the nearest image and also from other similar images based on the frequency and co-occurrence of the tags.

This approach has been influential and has spurred further improvements, such as TagProp [4], where a richer set of visual features is used to describe images. In the tag transfer stage, the distances between the nearest neighbors are used as weights for discriminative models trained for each tag. These models encode the presence/absence of keywords in images and they are trained to explicitly exploit the dependencies between tags in the training database. In addition, TagProp also includes word-specific sigmoidal modulation to increase the recall of the rare words and it uses metric learning to improve the final tag prediction.

A more recent work [2] argues that training individual models, as done in [4], is not scalable for larger datasets so a different method to retrieve the nearest neighbors is proposed: a Random Forest trained with a) visual information from the images and b) semantic information from the tags to improve the relevancy of the nearest neighbor images. The tag transfer stage adopts a conventional tf-idf scheme to rank the tags from the nearest neighbor set.

We can identify a certain evolution in the tag-transfer algorithms: the first algorithms (e.g. [8]) only made use of semantic information of the nearest neighbor tags. The following ones (e.g. [4]) demonstrated that using some visual information in the creation of the tag models improves the performance of the annotation. Another example is [2] where both semantic and visual information is used to retrieve the similar images but this explicit information is not used during the tag transfer. This trend is similar to other areas on information retrieval were this combination also gives superior performance (e.g. [11]). In our work we follow this approach proposing a novel algorithm that integrates explicitly visual information and semantic information during the tag transfer process.

The two main contributions of this paper are:

• A novel algorithm for tag transfer based on visual and semantic information that can be used with any arbitrary nearest neighbor technique
A local-learning technique that gives useful information to the tag transfer algorithm during the annotation stage.

The rest of the paper is structured as follows: in section II we describe the proposed algorithm and how to train it efficiently; in section III.A we define the experimental setup and section III.B shows the results of the experiments. Finally we discuss these results and future work in section IV.

II. PROPOSED ALGORITHM

A. Tag Transfer Algorithm

Given a test image that needs to be annotated, a Nearest Neighbor model first obtains a set S of training images (|S|=K) that are visually similar to the test image. Each image \( s_i \in S \) is annotated with a set of tags \( T_i = \{ t_{i1}, t_{i2}, \ldots, t_{in_i} \} \) / i=1\ldots K, so during the tag transfer stage we have a combined set of tags \( T = \{ T_1, T_2, \ldots, T_K \} \) with size \( N = \sum_K N_i \).

We are going to define a function \( f \) for a query image \( q \), a set of similar images \( S \), a set of tags \( T \) and configuration parameters \( W \):

\[
\bar{y} = f(T, W, q, S) \tag{1}
\]

Where \( \bar{y} \) is a vector containing a real value in the range \([0,1] \) indicating the presence/absence of each tag \( t \in T \) (in the train set 0 indicates absence and 1 presence). If a tag \( t \) is more relevant to the test image \( q \) than another tag \( t \in T \), then \( f(t, W, q) > f(t, W, q) \). With this condition we define the following function \( f \):

\[
f(T, W, q, S) = W \cdot x^t \tag{2}
\]

Where \( W \in M_x \times m(\mathbb{R}) \) contains the learned configuration parameters and \( x \in M_N \times m(\mathbb{R}) \) contains a representation \( \tilde{x}_{ij} \in \mathbb{R}^m \) of each tag \( t \) in the set \( T \). This representation maps each tag to a Euclidean space where it is represented by a feature vector. This vector is composed by semantic information about the tag \( t \) and visual information about the query image \( q \) and the image \( s_i \):

\[
\tilde{x}_{ij} = g(q, t_{ij}, s_i) \tag{3}
\]

Where \( g \) denotes that \( \tilde{x}_{ij} \) depends on the test image \( q \), the train image \( s_i \) and the tag \( t_{ij} \). So for each triplet, we have empirically defined \( m=6 \) features in the vector \( \tilde{x}_{ij} \), which are:

- **Distance Test-train Images** (\( \tilde{x}_{ij}^{(1)} \)): Visual distance between the test image \( q \) and the neighbor \( s_i \). This distance can be computed with any arbitrary distance metric between the mathematical representations of both images. In the present work, that representation is the concatenation of the color and edge feature vectors.
- **Color Distance Test-train Images** (\( \tilde{x}_{ij}^{(2)} \)): Color has been shown to have an important role when classifying images, so this distance measures the color similarity between the feature vectors of the test image \( q \) and the neighbor \( s_i \).
- **Edges Distance Test-train Images** (\( \tilde{x}_{ij}^{(3)} \)): Edges are also an important part for classification. This distance measures the edge similarity between the feature vectors of test image \( q \) and the neighbor \( s_i \).
- **Tag cooccurrence** (\( \tilde{x}_{ij}^{(4)} \)): This semantic feature measures the aggregate number of times that a tag \( t_{ij} \) appears in the training set with the rest of keywords in \( T \).
- **Tag frequency** (\( \tilde{x}_{ij}^{(5)} \)): This feature measures the number of times the tag \( t_{ij} \) appears in the set \( T \).
- **Tag probability** (\( \tilde{x}_{ij}^{(6)} \)): This is the probability of appearance of the tag \( t_{ij} \) in the training database images.

Given this definition, if two training images in the set \( S \) are annotated with the same tag, the feature vector can be different because it encodes visual information related to the images and not only the semantic information. The rationale behind such definition is that the proposed machine learning algorithm can learn the relation between the pair test-training image visual features (\( x_{ij}^{(1-3)} \)) and the semantic information of them tags (\( x_{ij}^{(4-6)} \)).

B. Training: Image Networking

Given the formulation (2) and the representation of each tag in the set \( T \) (3), it is needed to learn \( W \). Differing from other ideas that learn \( W \) for the whole training space, the proposed approach makes use of the results of the nearest neighbor computation to learn local weights.

We propose to compute \( W \) for each test image. So, given a query image \( q \), we can obtain their \( K \) nearest neighbors, so it is possible to locate \( q \) in the visual Euclidean space (Fig 1 middle column). \( W \) will be learned locally for this region using the concept that we call *Image Networking*: all the images in the training set are "connected".

![Image Networking for training \( W \)](image)

In the Fig 1 we show an example of one query image \( q \), several nearest neighbor images in the train database and \( K \) similar images to a train image \( i \). This means that given \( i \)-th nearest neighbor \( s_i \in S \) of \( q \), it is possible to get their \( K \) nearest neighbors (the set \( S^i \)). Given \( S^i \) we also have a tag set \( T^i \). Using (1) we obtain the set \( Y^i \), which are the annotations of the \( s_i \) image. In this case, because this \( i \)-th image is a training image, we know that \( Y^i = T_i \). Doing this for all the \( K \) similar images of \( q \) we have this set of equations:

\[
T_i = f(T^{(i)}, W, s_i, S^{(i)}) \quad | i = 1 \ldots K \tag{4}
\]
Due to our requirement $f(t_{np}, W, q, S) = f(t_{np}, W, q, S)$, we look for $W$ where the maximum number of inequalities is fulfilled in the training subset. This problem is NP-Hard, but it is possible to solve it by relaxing the inequalities. In this case, it is possible to obtain $W$ thanks to the following convex optimization [6]:

$$
\min_{w, \xi_{i,j}, k} \frac{1}{2} \|W\|^2 + C \sum_i \xi_{i,j,k}
$$

s.t. $W : \{g(q, t_{np}, s_i) - g(q, t_{np}, s_j)\} > 1 - \xi_{i,j,k}$ for all pair $((t_{ij}, s_i), (t_{ij}, s_j))$

This problem is solved efficiently by using the cutting-plane algorithm proposed in [7].

III. EXPERIMENTS

A. Experimental Settings

In this paper, we propose a tag transfer algorithm, so we need a procedure to compute the similar images which are the input of our algorithm. We have selected the widely accepted baseline proposed in [8]. In this baseline, images are described using color and edges histograms. In our work color and edges description are represented by the Scalable Color Descriptor and the Edge Histogram Descriptor, as defined in MPEG-7 standard. Performed tests show the same results than the original visual features proposed in [8] but using 98% less memory. The basic distances selected in our implementation are L1 for both descriptors as suggested in the MPEG-7 standard ($\tilde{d}_{ij}(2)$ and $\tilde{d}_{ij}(3)$ in our feature vector). Additionally, as specified in the baseline, we combine the normalized distance of the color descriptor and edges descriptor using the JEC distance [8] ($\tilde{d}_{ij}(1)$ in our feature vector), and the final number of tags selected per image is five [8].

B. Influence of the visual and semantic features in the proposed algorithm

We build our first experiment over the IAPR TC-12 database [3] using the same subset and annotations as previous works [4][2][8]. It contains 17,665 images for training, 1,962 images for testing and 291 possible tags having an average of 4.7 tags per image. For image annotation task a better precision is preferred over recall [9], so in Figure 2 we show the average precision of our algorithm using three feature vectors.

Fig. 2. Average precision for three configurations of the proposed tag transfer algorithm.

This experiment shows the influence of the images’ visual information and the tags’ semantic information. We evaluate three possible configurations of our algorithm: tag transfer based on semantic information ($\tilde{d}_{ij}(1-6)$), on visual information ($\tilde{d}_{ij}(1-3)$) and on a combination of both ($\tilde{d}_{ij}(1-6)$).

Fig. 2 shows that our algorithm performs better when aggregating semantic and visual information. Moreover, the best result is achieved using smaller neighborhoods (K=10 and K=50). So the proposed tag transfer algorithm has better results when training the algorithm with local information instead of using the entire training database. This is consistent with our local-learning technique based on the concept of Image Networking.

C. State-of-the-art comparison

It has been demonstrated the advantage of the joint use of visual and semantic information in our tag transfer stage. Now it is needed to compare its precision against state-of-the-art algorithms. The first algorithm is the baseline proposed in [8], where the tag transfer algorithm is called GreedyTransfer. Another top performing approach is the proposed in [4], where its tag transfer algorithm is called TagProp. In this comparison we have selected two configurations that do not make use of metric learning in the Nearest Neighbor stage. Additionally, we also compare one of the latest published algorithms, named as RF_count [2].

This paper is centered on the tag transfer stage so in order to have fair comparison we need to establish a new test bed. To this end, we have selected a common Nearest Neighbor technique as the input of these tag transfer algorithms using the authors’ provided code. The common Nearest Neighbor configuration is the baseline [8], presented in the section III.A, with K=10 as defined in the baseline. Furthermore, we have selected a database that has not been used in [8], [4] or [2] to avoid any bias towards one algorithm. The selected database is the MIRFlickr database [5], with the dataset provided by the ImageCLEF2011 competition. This dataset contains 8,000 train images, 10,000 test images and 99 concepts. This dataset has a bigger test set than IAPR TC-12 so the results will be more significant. To compare the different tag transfer algorithms we have measured the precision and recall for each test image.

<table>
<thead>
<tr>
<th>TABLE I. RESULTS IN IMAGECLEF2011 DATASET</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ImageCLEF 2011 dataset</strong></td>
</tr>
<tr>
<td><strong>Precision (%)</strong></td>
</tr>
<tr>
<td>------------------------------------------</td>
</tr>
<tr>
<td>Greedy Transfer [8]</td>
</tr>
<tr>
<td>TagProp vSD [4]</td>
</tr>
<tr>
<td>TagProp SD [4]</td>
</tr>
<tr>
<td>RF_count [2]</td>
</tr>
<tr>
<td>Proposed algorithm</td>
</tr>
</tbody>
</table>

In this experiment, all tests are done using the same Nearest Neighbor computation technique, so we can compare the tag transfer stage. First, we see tag transfer algorithms of [2] and [4] are not better than the baseline [8]. This could be caused because both works focused on improving the Nearest Neighbor phase rather than the tag transfer algorithm. In this scenario for a fixed K=10 our approach has the best performance, obtaining more than 6% of precision improvement over [8], indicating that the combination of visual
and semantic features in the tag transfer stage has a significant impact in the final result.

In order to corroborate these results, we have chosen a second measurement to evaluate the algorithms. Using the official script provided in the ImageCLEF2011 competition, we have computed the Mean Average Precision (MAP). MAP measures the average precision for all tags. Note that we were not able to run the script for GreedyTransfer because the algorithm doesn’t provide probability estimations, as required by the script. Additionally we have computed the precision for different neighborhood sizes and results are presented in Fig 3.

Fig. 3. MAP for four tag transfer algorithms using different sizes of neighborhood (K)

Independently of the size of the neighborhood our approach outperforms the rest of the algorithms, and it has a 2.3% better MAP than \(RF\_count^2\) (22.9% vs. 20.6%) in the best configuration (K=50). Table I and Figure 3 show that our algorithm predicts correctly rare tags (higher MAP per tag) and it also predicts correctly the tags present in an image (higher average precision per image).

Table II show several returned labels for two random images and we compare our approach against [4] and [2]. Our algorithm returns more tags which are present in the ground truth (these labels are marked as bold and underscored words).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral, Illumination, No, Blur, No, Persons, cute, calm, bodyspart</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a novel algorithm for the tag propagation stage in a Nearest Neighbor based image annotation model. This algorithm incorporates semantic information from candidate tags and visual information from test and train images. The algorithm has been compared against four algorithms using a medium size database (ImageCLEF2011). Using two different measurement metrics (average precision per image and mean average precision per tag) our approach performs best.

In addition we have proposed a local-learning technique, Image Networking, which trains the algorithm at query time. We have demonstrated in the IAPR TC-12 database that this structured learning technique gives useful local information to the tag transfer algorithm instead of giving global and general information like other approaches present in the literature.

Finally, our method can be used with any arbitrary Nearest Neighbor technique, so the results can be improved by improving the technique (e.g. using metric learning). Moreover, the speed of the annotation can also be increased by using an approximated nearest neighbor computation. For that reason, future work will include testing the proposed tag transfer algorithm using different Nearest Neighbor methods.

ACKNOWLEDGMENT

We would like to thank anonymous reviewers and Hao Fu for the \(RF\_count^2\) code. MPEG7 code was obtained from LIRE.

REFERENCES
