Combining Image Similarity Metrics for Semantic Image Annotation

Bart Jansen[1,2,*], Tran Duc Toan[1,2], and Frederik Temmermans[1,2]

[1] Vrije Universiteit Brussel, Dept. of Electronics and Informatics, Pleinlaan 2, 1050 Brussels, Belgium
bjansen@etro.vub.ac.be
http://www.etro.vub.ac.be

[2] Interdisciplinary Institute for Broadband Technology (IBBT), Dept. of Future Media and Imaging (FMI), Ghent, Belgium
http://www.ibbt.be

Abstract. This paper describes automated image annotation as an image retrieval problem, in which the distance metric used to express similarity among images is learnt from available distance metrics on several image descriptors. Rather than describing the problem as an optimization problem, we study it as a regression problem. On a limited dataset of images of buildings taken in the city center of Brussels, we illustrate the superior performance of the combined distance metrics over any of the considered individual distance metrics in automated image annotation.

Keywords: automated image annotation, distance metric learning.

1 Introduction

The domain of automated semantic image annotation studies how to automatically assign one or more semantic labels to images in a given database. Typically, these labels originate from a predefined ontology which is specific to the application. Automated image annotation (AIA) [8] is based on the retrieval of images similar to the given image from a database of already annotated images. As such, it relies on the assumption that *similarity* in the image domain relates to *similarity* in the semantic domain (see for instance [3] for a discussion on *the semantic gap*). Therefore, progress in the domain of AIA largely depends on advancements obtained in content based image retrieval (CBIR). In CBIR, the main challenges are to define image descriptors that are relevant for identifying similarity among images and to define methods to combine various descriptors. Image descriptors are typically based on interest point detectors (SIFT, SURF, ...) or on color, shape or texture (of objects present in) the images.

Although various methods exist for combining multiple descriptors, a common approach is to use machine learning techniques to train a classifier for each of the classes. One of the disadvantages of the classifier based approach is that a

[*] Bart Jansen is currently funded by Innoviris.

P. Herrero et al. (Eds.): OTM 2012 Workshops, LNCS 7567, pp. 294–297, 2012.

separate and independent classifier is most of the time trained for each class. Consequently, the importance and complexity of a second classification system for aggregating the votes of the individual classifiers grows with an increasing number of classes [8]. Such approaches are typically evaluated in CBIR tasks but far less often in AIA tasks.

2 Algorithms

Several methods exist for learning a distance metric based on a weighted combination of several distance metrics in a supervised learning scheme. We adopt the notation of [6] and adapt it to the specific case of CBIR and AIA.

Suppose there is a set of n single labeled images $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ where $x_i \in X$ and $y_i \in \{1, 2, ..., c\}$ and c is the number of unique classes. Each image x_i is represented by a vector $[x_i^1...x_i^m]$ where $x_i^l \in X_l, l = 1...m$ are different descriptors computed from the image x_i. Thus, $x_i \in X = X_1 \times ... \times X_m$. On each of the X_l spaces a distance metric D_l is defined. If D is defined as $D(x_i, x_j) = [D_1(x_i^1, x_j^1), ..., D_m(x_i^m, x_j^m)]^T$, then

$$D_A^2(x_i, x_j) = D(x_i, x_j)^T \, A \, D(x_i, x_j) \tag{1}$$

can be defined. In order to ensure that D_A is a valid pseudometric, A should be positive semi-definite, so $A = W^T W$.

Now, define S_S to be the set of all image pairs annotated with the same class and S_D to be the set of all image pairs with a different class. $S_S = \{(x_i, x_j) \mid y_i = y_j\}$ and $S_D = \{(x_i, x_j) \mid y_i \neq y_j\}$, then the metric learning problem in a supervised setting consists of an optimization problem:

$$\min_A F_A(S_S, S_D, D_A^2) \tag{2}$$

Depending on the definition of F, the within class distances are minimized, the between class distances are maximized, or combination of both, often resulting in complex optimization problems [7,2].

If we assume that A is a diagonal matrix, then the distance $D_A(x_i, x_j)$ reduces to a simple weighted sum of the distances $D_l(x_i, x_j)$ and the problem is reduced to the learning of the weights a_{ll}. By specifying that $D_A(x_i, x_j)$ should be 0 for all $(x_i, x_j) \in S_S$ and that $D_A(x_i, x_j)$ should be 1 for all $(x_i, x_j) \in S_D$, the weights can be learnt in a supervised binary classification problem. In this paper weights are learnt using linear regression and by back-propagation learning using a single layer neural network (NN).

3 Results

The proposed method is evaluated in a task of automated annotation of images of buildings in the city of Brussels. The database contains pictures of 28 different buildings in the city center of Brussels. Four images per class were selected for training (112 in total) and 1 image per class was used as an independent test set. This dataset is similar - but still smaller - than the ZuBuD dataset [1], often used in content based image retrieval tasks. A difference with ZuBuD is the rather big difference in views of the same buildings and the big differences in viewing distance.

On each image, the following MPEG-7 descriptors [4] were computed: Color Layout Descriptor (CLD), Color Structure Descriptor (CSD), Edge Histogram Descriptor (EHD), Homogenous Texture Descriptor (HTD) and Scalable Color Descriptor (SCD). Additionally, we complemented these image descriptors with SURF [1]. On each of the six interest point descriptors, we used an appropriate distance metric as specified in [4] and [5].

Results of this experiment are summarized in Table 1, in which the accuraies of several weighted distance metric combinations are listed. Rows 1 to 6 show the accuracies of the isolated use of each individual distance metric. Although all descriptors result in recognition rates far above random (which is only 1/28), there is as expected a clear performance difference between SURF and the MPEG-7 descriptors in this object recognition task. However, SURF in itself is not resulting in perfect recognition. This confirms our assumption that this dataset has more variety in viewing angles etc. compared to the ZuBuD dataset, on which SURF results in 100% correct recognition. Rows 7 and 8 list the results for the case in which a weighted combination of the different distance metrics are learned using linear regression (row 7) and NN (row 8), resulting in accuracies of 82.14% and 85.74% respectively, resulting in an increase of 4% to 7% when only using SURF. In both cases, a much higher weight was learned for SURF compared to the other descriptors, confirming the importance of SURF in object recognition tasks.

However, SURF is not a good descriptor for object class recognition tasks, where the kind of object rather than the exact object needs to be recognized. Therefore, it was investigated whether a combined metric can be learned only

Table 1. Experimental results

Nr	CLD	CSD	EHD	HTD	SCD	SURF	Accuracy
1	1	0	0	0	0	0	42.86
2	0	1	0	0	0	0	53.57
3	0	0	1	0	0	0	42.86
4	0	0	0	1	0	0	35.71
5	0	0	0	0	1	0	35.71
6	0	0	0	0	0	1	78.57
7	0.17	0.35	0.15	0	0	0.28	82.14
8	0.13	0.37	0.07	0.05	0.07	0.26	85.74
9	0.15	0.45	0.14	-0.01	0.04	0	71.42
10	0.21	0.40	0.20	0	0.03	0	75.00

[1] http://www.vision.ee.ethz.ch/showroom/zubud/

based on the MPEG-7 descriptors. Results are listed in rows 9 and 10 for linear regression and NN and show that although the combined metrics clearly outperform the individual metrics, performance is still slightly lower than only using SURF, i.e. 71.42% and 75.00% using all MPEG-7 descriptors against 78.57% for SURF only.

Given the low number of training images per class, the limited number of classes and the limited evaluation of the proposed methods, comparison results need to be interpreted with care. However, the initial experiments show the validity of the approach, which needs to be confirmed by further and more elaborate experiments.

4 Conclusion

This paper explored the learning of a combined distance metric on several image descriptors to improve annotation accuracy on a database of images of buildings in the city of Brussels. The problem was described as a supervised learning problem based on a dataset of annotated images. Results showed that both using linear regression and neural networks superior performance could be obtained compared to using any of the individual distance metrics.

References

1. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
2. Globerson, A., Roweis, S.: Metric learning by collapsing classes. In: Weiss, Y., Schölkopf, B., Platt, J. (eds.) Advances in Neural Information Processing Systems 18, pp. 451–458. MIT Press, Cambridge (2006)
3. J.S. Hare, P.A.S. Sinclair, P.H. Lewis, K. Martinez, P.G.B. Enser, and J. S. Christine. Bridging the Semantic Gap in Multimedia Information Retrieval: Top-Down and Bottom-Up Approaches. In 3rd European Semantic Web Conference (ESWC 2006). LNCS, vol. 4011. Springer (2006)
4. Sikora, T.: The MPEG-7 visual standard for content description-an overview. IEEE Transactions on Circuits and Systems for Video Technology 11(6), 696–702 (2001)
5. Temmermans, F., Jansen, B., Deklerck, R., Schelkens, P., Cornelis, J.: The Mobile Museum Guide: Artwork Recognition with Eigenpaintings and SURF (2011)
6. Woznica, A., Kalousis, A., Hilario, M.: Learning to combine distances for complex representations. In: ICML 2007: Proceedings of the 24th International Conference on Machine Learning, Corvalis, Oregon, pp. 1031–1038. ACM (2007)
7. Xing, E., Ng, A., Jordan, M., Russell, S.: Distance Metric Learning with Application to Clustering with Side-Information. In: Advances in Neural Information Processing Systems. MIT Press, Cambridge (2003)
8. Zhang, D., Monirul Islam, M., Lu, G.: A review on automatic image annotation techniques. Pattern Recognition 45(1), 346–362 (2012)