Weather-Adaptive Distance Metric for Landmark Image Classification

Ding-Shiuan Ding and Wei-Ta Chu

National Chung Cheng University, Taiwan
lzi94u@gmail.com, wtchu@cs.ccu.edu.tw

Abstract. Visual appearance of landmark photos changes significantly in different weather conditions. In this work, we obtain weather information from a weather forecast website based on a landmark photo's geotag and taken time information. With weather information, we adaptively adjust weightings for combining distances obtained based on different features and thus propose a weather-adaptive distance measure for landmark photo classification. We verify the effectiveness of this idea, and accomplish one of the early attempts to develop a landmark photo classification system that resists to weather changes.

Keywords: Weather-adaptive distance metric, landmark classification

1 Introduction

Landmark image classification has emerged as an important research topic due to its potential usage of location-based service and large-scale image retrieval. Famous landmarks such as Eiffel Tower and Statute of Liberty attract millions of visitors every year, who took pictures of the landmarks from unlimited viewpoints in various conditions, and then shared them on social media platforms. Large amounts of landmark photos thus urge the need of efficient retrieval/access as well as effective recognition/classification.

Many studies of landmark classification and its extended variants, i.e., location prediction/recognition, have been widely proposed in recent years. They mainly focus on integrating multimodal features such as geographical information and visual information, or developing classification models based on large-scale datasets. However, the problem of high intra-class variations caused by drastically different visual conditions still remains.

In this paper, we investigate one factor that largely affects visual appearance of landmark images: *weather types*. Through the whole year many people visit Notre Dame, for example, and take photos under various weather conditions. Fig. 1 shows sample photos taken at Notre Dame and Sacre Coeur on sunny and cloudy days, respectively. From this figure we see visual appearances are significantly different in different weathers due to the sky and the intensity of lighting on the building. Such intra-class variations impede accurate image classification. However, the influence of weather types on measuring image similarity was overlooked before. In this work we

adfa, p. 1, 2011.

propose a weather-adaptive distance metric so that better similarity measurement between images can be achieved, and thus better landmark image classification is expected.

When comparing two landmark images, we could calculate their distance from many perspectives, such as texture and local feature points, and then linearly combine distances respectively calculated based on each feature. With weather properties obtained from a weather forecast website, we propose to adjust weightings by formulating this task as an optimization problem. As the first contribution of this work, we consider its analogy to single neuron training and determine the optimal weightings by the gradient method. As the second contribution, more effective features can be discovered and prioritized through the learnt weightings, and more accurate landmark image classification can be achieved.

The rest of this paper is organized as follows. In Section 2 literature of landmark image classification will be surveyed. Details of the weather-adaptive distance metric with weight learning are described in Section 3. Section 4 provides discussion of the proposed metric and performance of landmark image classification, followed by concluding remarks in Section 5.

Fig. 1. Left to right: sample photos of Notre Dame on sunny days, Notre Dame on cloudy days, Sacre Coeur on sunny days, and Sacre Coeur on cloudy days.

2 Related Works

Landmark image classification has been widely studied in the past decade. We briefly review some of them in the following. Zheng et al. [14] built an internet-scale landmark dataset by mining true landmark images from GPS-tagged photos and tour guide web pages. Unsupervised clustering techniques and visual models based on feature points were adopted to build a landmark recognition engine. Yi et al. [11] also built a large-scale dataset and adopted the bag of feature approach associated with multiclass SVM to achieve landmark image classification. They also showed that using textual tags and temporal constraints leads to significant performance improvement over the visual only method. Li et al. [13] combined 2D appearance and 3D constraints to discover iconic views of a landmark, which were later used in landmark recognition. Chen et al. [9] proposed a soft bag-of-visual phrase approach for mobile landmark recognition. Visual phrases were learnt in a category-dependent manner to achieve promising recognition performance. Min et al. [12] proposed an efficient mobile landmark search system where the client uploads compressed images to the server,

and the server recognizes landmark by matching the uploaded image with landmark texture projected from pre-constructed landmark 3D models.

Since the IMG2GPS system proposed in [6], studies of geographical location estimation emerge in recent years. Hays and Efros [6] estimated the geographical location of a query photo based on a data-driven scene-matching approach. Li et al. [8] improved the scene-matching approach by jointly considering visual similarity and geographical proximity to build a ranking method. Lin et al. [7] greatly extended the scene-matching approach by further considering overhead appearance and land cover survey data. A query photo can be localized even if it has no corresponding ground-level images in the database. Fang et al. [5] adopted latent SVM to discover geo-informative attributes from regions in order to facilitate better location recognition and exploration.

Although there have been many works targeting at landmark or location recognition, few of them specially tackled visual variations caused by lighting, editing, or weather change. Shen and Cheng [10] proposed gestalt rule feature points to find visual correspondence between images of different styles (painting vs. photograph, or photographs in different colors) but containing the same semantic meaning. However, methodology or features especially designed to consider visual variations caused by weather conditions are still missing. In this work, we focus on developing a distance metric considering weather conditions.

3 Weather-Adaptive Distance Metric

3.1 Common Distance Metric

Given two images I_p and I_q, assuming that each image can be represented by N types of features, i.e., $I_p = \{\boldsymbol{p}_1, ..., \boldsymbol{p}_N\}$ and $I_q = \{\boldsymbol{q}_1, ..., \boldsymbol{q}_N\}$, the conventional way to integrate distances derived from features is:

$$D(I_p, I_q) = \sum_{i=1}^{N} w_i d_i(\boldsymbol{p}_i, \boldsymbol{q}_i), \tag{1}$$

where $d_i(\boldsymbol{p}_i, \boldsymbol{q}_i)$ is the normalized distance calculated based on the ith feature. Weightings w_i's are often empirically set or simply follows a uniform distribution, i.e., $w_i = 1/N$. However, the integrated distance $D(I_p, I_q)$ often cannot reflect impacts of different features, yielding limited landmark classification performance.

To show the shortage of this simple metric, from Flickr we collect photos of famous landmarks that were captured on sunny days or cloudy days. We then calculate integrated distance $D(I_p, I_q)$ between photos that are randomly selected following four schemes: (1) I_p and I_q are from the same landmark under the same weather type (sunny or cloudy); (2) I_p and I_q are from the same landmark under different weather types (one is sunny and another is cloudy); (3) I_p and I_q are from different landmarks under the same weather type; (4) I_p and I_q are from different landmarks under different weather types. The integrated distance $D(I_p, I_q)$ is obtained by combining individual features respectively derived from Gabor texture features, haze features, bag of visual words, and CNN features. The individual distance $d_i(\boldsymbol{p}_i, \boldsymbol{q}_i)$ is measured by

Euclidean distance. Details of the evaluation dataset and features will be described in Section 4.

Fig. 2 shows distributions of integrated distances $D(I_p, I_q)$ between photos selected based on four different schemes. Comparing the distributions obtained based on the first and the third schemes, under the same weather condition, distances between photos from the same landmark are similar to that from different landmarks. This shows weather properties dominate calculation of distance measure. The first two distributions (obtained based on the first two schemes) are similar to the last two distributions (obtained based on last two schemes). This means the common distance metric cannot reliably describe that photos from the same landmark are similar, while photos from different landmarks are relatively distinct even when they were captured under the same weather condition.

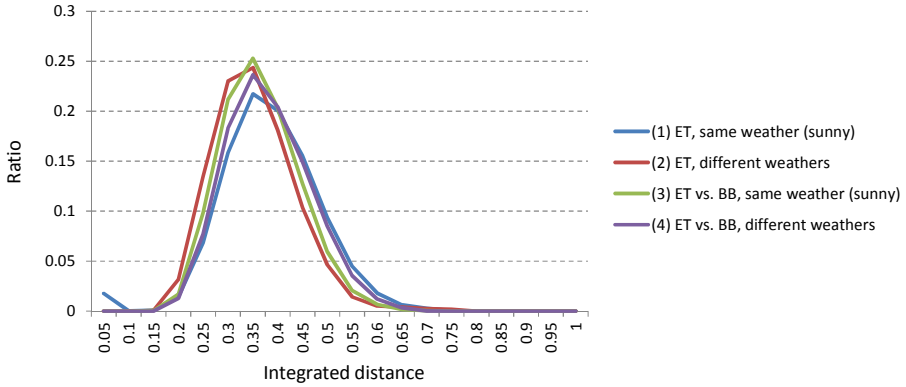

Fig. 2. Distributions of integrated distances calculated based on four settings. ET stands for Eiffel Tower, and BB stands for Big Ben.

3.2 Weather-Adaptive Distance Metric

The characteristics shown in Fig. 2 motivate us to propose a weather-adaptive distance metric for measuring landmark photos. The idea is to adjust weightings for combining individual distances in a systematic manner. Let us model whether two photos I_p and I_q belong to the same landmark based on the integrated distances like this:

$$y = \sum_{i=1}^{N} w_i d_i(\boldsymbol{p}_i, \boldsymbol{q}_i) = \boldsymbol{d}^T \boldsymbol{w}, \tag{2}$$

where $\boldsymbol{w} = [w_1, ..., w_N]^T$ and $\boldsymbol{d} = [d_1(\boldsymbol{p}_1, \boldsymbol{q}_1), ..., d_N(\boldsymbol{p}_N, \boldsymbol{q}_N)]$. The indicator $y = 0$ if I_p and I_q are from the same landmark (no matter whether they were under the same weather condition or not), and $y = 1$ otherwise.

We wish to find the values of weights w_1, …, w_N such that the estimated indication value is as close as the ground truth. Given a training dataset $\{(\boldsymbol{d}_1, y_1), ..., (\boldsymbol{d}_M, y_M)\}$ constituted by randomly selecting M photo pairs from the

collected landmark photo collection, the training problem is formulated as the following optimization problem:

$$\text{minimize} \quad \sum_{j=1}^{M}(y_j - \boldsymbol{d}_j^T \boldsymbol{w})^2, \tag{3}$$

where the minimization is taken over all $\boldsymbol{w} = [w_1, ..., w_N]^T \in R^N$. The term $\boldsymbol{d}_j^T \boldsymbol{w}$ is the estimated indication value, and the objective function represents the sum of squared errors between the desired output y_j and the estimated result $\boldsymbol{d}_j^T \boldsymbol{w}$. We can write the optimization problem in matrix form:

$$\text{minimize} \quad \|\boldsymbol{y} - \boldsymbol{D}^T \boldsymbol{w}\|^2, \tag{4}$$

where $\boldsymbol{D} = [\boldsymbol{d}_1 \cdots \boldsymbol{d}_M]$ and $\boldsymbol{y} = [y_1, ..., y_M]^T$.

We have more training points than the number of weights. Assuming that rank of $\boldsymbol{D}^T$ is N, the objective function is simply a strictly convex quadratic function of $\boldsymbol{w}$. In this work, we utilize a fixed-step-size gradient algorithm [3] that iteratively updates the weighting vector $\boldsymbol{w}$ in the following form:

$$\boldsymbol{w}^{(k+1)} = \boldsymbol{w}^{(k)} + \alpha \boldsymbol{D} \boldsymbol{e}^{(k)}, \tag{5}$$

where α is the predefined step size, and $\boldsymbol{e}^{(k)} = \boldsymbol{y} - \boldsymbol{D}^T \boldsymbol{w}^{(k)}$ is the estimation error at the kth iteration.

Through the process mentioned above, we learn the optimal weighting vector $\boldsymbol{w} = [w_1, ..., w_N]^T$ that causes the minimum estimation error.

4 Experiments

4.1 Experimental Settings

The collected dataset consists of sunny and cloudy photos of five famous landmarks, including Big Ben, Eiffel Tower, Notre Dame, Sacre Coeur and Winsor Castle. Table 1 shows information of the collected dataset. The numbers of sunny and cloudy photos are roughly balanced, and there are totally 1,210 photos in the dataset.

Features. We use Gabor texture features [2], haze features [1], bag of feature points [11], and CNN features [15] to describe an image. For Gabor texture features, image pixels' intensity are transformed into the frequency domain, which is then decomposed into 16 ranges by the Gabor Wavelet functions with four scales and four orientations. Mean and standard deviation of the magnitude of the transform coefficients in each range are used to represent each frequency band, and are then concatenated to form a 32-D texture feature vector.

For haze features, dark channel prior [4] is first calculated for each pixel. An image is partitioned by a spatial pyramid scheme, i.e., uniformly partitioned into 2^2, 4^2, and 8^2 non-overlapping regions to obtain 84 sub-regions. The median values of dark channel intensities in these sub-regions are concatenated as an 84-D haze feature vector [1].

Table 1. Information of the evaluation dataset.

Landmark	Sunny	Cloudy
Big Ben	100	100
Eiffel Tower	138	143
Notre Dame	105	104
Sacre Coeur	168	101
Winsor Castle	141	110
Sum	652	558

Following the single image classification process proposed in [11], we describe an image by a bag of visual words (BoW) model. We utilize the visual vocabulary (with 10,000 visual words) built in Top-SURF [16] to construct an image's 10,000-D BoW representation.

Currently using convolutional neural network (CNN) features largely surpasses hand-crafted features. To extract CNN features, we utilize the MatConvNet package [17] with the pre-trained model obtained based on ImageNet ILSVRC-2012. There are five convolutional layers and three fully-connected layers in the CNN model. The first convolutional layer filters the input image with 64 kernels of size $11 \times 11 \times 3$ with a stride of 4 pixels. The second convolutional layer makes filtering with 256 kernels of size $5 \times 5 \times 64$. The third, fourth, and fifth convolutional layers are connected to one another without any intervening pooling or normalization layers. The third and fourth convolutional layer have 256 kernels of size $3 \times 3 \times 256$, respectively, and the fifth convolutional layer has 4096 kernels of size $6 \times 6 \times 256$. The fully-connected layers have 4096 neurons each. We try to take output of the fifth, sixth, and seventh layers to be CNN features, and found that features from the sixth layer yield the best performance through our preliminary experiments.

Experimental Settings. Based on the dataset, we adaptively adjust weights for measuring distances between photos captured in the same weather condition or in different weather conditions. Particularly, we randomly select pairs of sunny photos to form the training pool $SS = \{I_1^{(s)}, ..., I_M^{(s)}\}$. For each pair in SS, if the two photos I_p and I_q belong to the same landmark, the indicator y is set as 0, and set as 1 otherwise. Initial weighted distances between selected pairs, as defined in eqn. (1), and the associated indicators, are treated as the training data, and the updated procedure described in eqn. (5) is used to adjust weightings specifically for measuring distance between sunny photos. We denote the adjusted weightings as $\boldsymbol{w}_{SS} = \{w_1^{(s)}, ..., w_4^{(s)}\}$. Similarly, we randomly select pairs of cloudy photos to form the set $CC = \{I_1^{(c)}, ..., I_M^{(c)}\}$, and determine the adjusted weightings $\boldsymbol{w}_{CC} = \{w_1^{(c)}, ..., w_4^{(c)}\}$. To appropriately measure distance between two photos that were captured in different weather conditions, we also randomly select M photo pairs, where for each pair one photo is sunny and another is cloudy. Based on the corresponding initial weighted distances and associated indicators, the adjusted weightings $\boldsymbol{w}_{SC} = \{w_1^{(t)}, ..., w_4^{(t)}\}$ are determined.

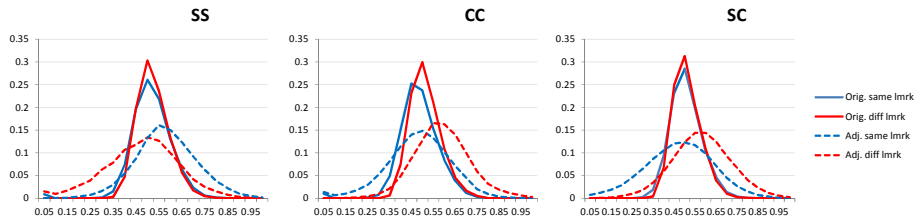

Fig. 3. Distributions of integrated distances before and after weighting adjustment. From left to right: distributions of sunny vs. sunny photos; distributions of cloudy vs. cloudy photos; distributions of sunny vs. cloudy photos.

Because the captured time and geographical information are available for each photo in our database, we can use this information to obtain weather type through the API provided by the Weather Underground website[1]. Overall, given a pair of photos (one may be query, and another may be from the landmark database), we first select appropriate weights from $\boldsymbol{w}_{SS}$, $\boldsymbol{w}_{CC}$, or $\boldsymbol{w}_{SC}$, according to their weather types, and then calculate the weighted distance between them as the foundation for landmark classification or other applications.

4.2 Distributions of Distances

Fig. 2 shows that integrated distance distributions are similar no matter photos in the same landmark or in different landmarks are compared. Through the proposed adjustment, we verify that through the adjusted weightings distances between photos can be more appropriately captured.

Fig. 3 shows distributions of integrated distances between (a) sunny photos, (b) cloudy photos, and (c) one sunny photo and one cloudy photo, in the same landmark (blue curves) or in different landmarks (red curves). From all these three subfigures, we see that before weighting adjustment (solid curves), distance distributions between photos in the same or different landmarks are similar. After adjustment, distance distributions coming from photos at the same landmark move apart from that for different landmarks.

To quantitatively show the effect of weighting adjustment, we calculate the symmetric KL divergence between distance distributions respectively derived for same landmark and different landmarks. Table 2 shows detailed information. We can quantitatively observe that the KL divergence between distance distributions largely increases after weighting adjustment.

Fig. 4 shows absolute values of learnt weights for the three different schemes. We especially notice the relative values of these weights, and observe that BoW and CNN features are consistently more important than the other two features. This conforms to recent studies on image classification, and also shows that the proposed method can effectively learn weights.

[1] Weather Underground, http://www.wunderground.com/

Table 2. KL divergences of distance distributions.

Type	Before adjustment	After adjustment
Sunny-Sunny	0.0830	0.4034
Cloudy-Cloudy	0.2251	0.5134
Sunny-Cloudy	0.0383	0.3717

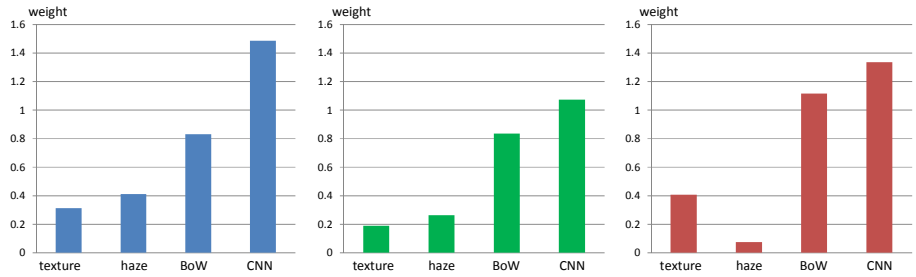

Fig. 4. Absolute weights of different features. Left to right: weights for sunny vs. sunny photos; weights for cloudy vs. cloudy photos; weights for sunny vs. cloudy photos.

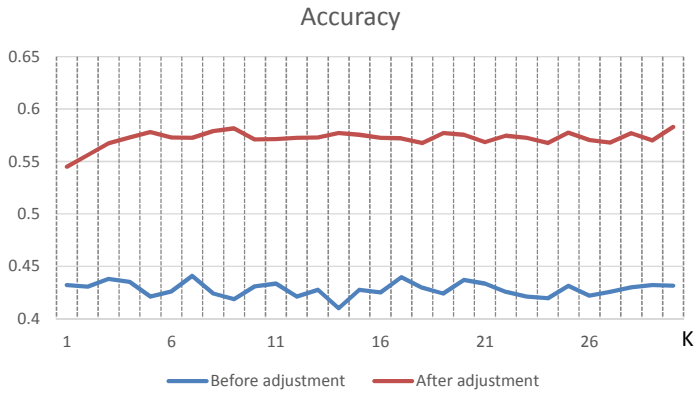

Fig. 5. Accuracy of landmark classification with varied nearest neighbor settings.

4.3 Performance of Landmark Classification

We adopt a simple classification method, i.e., K-nearest neighbor classifier, to more clearly show the effectiveness of weighting adjustment in landmark classification. Given a query photo, we find its K-nearest neighbors based on integrated distance, and classify the query photo as one of the landmark according to majority voting. We compare classification performance obtained based on initial integrated distances (eqn. (1)) with that obtained based on adjusted integrated distances (according to weather types). Fig. 5 shows accuracy of landmark classification with different set-

tings of the number of nearest neighbors (K). From this figure we clearly see the significant improvement given by appropriately adjusting weightings.

5 Conclusion

We have presented a weather-adaptive distance metric that is verified to yield better landmark photo classification based on a pilot database. By considering multiple features, distance between photos is usually calculated by combining individual distance derived from each feature. In this work we advocate that, by further considering weather type of the two compared photos, weightings that can better combine individual features can be learnt. We formulate it as an optimization problem and find the best weighting setting by a gradient algorithm. The reported evaluation verifies that the learnt weightings yield more effective distances between photos and thus improve performance of landmark photo classification increases with adjusted weightings. In the future, a larger-scale evaluation will be conducted, and more elegant methods to combine individual features and the corresponding learning problems will be studied.

Acknowledgements

The work was partially supported by the Ministry of Science and Technology in Taiwan under the grant MOST103-2221-E-194-027-MY3.

Reference

1. C. Lu, D. Lin, J. Jia, and C.-K. Tang, “Two-Class Weather Classification,” Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3718-3725, 2014.
2. B.S. Manjunath and W.Y. Ma, “Texture Features for Browsing and Retrieval of Image Data,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 8, pp. 837-842, 1996.
3. E.K.P. Chong and S.H. Zak, An Introduction to Optimization, 4th edition, Wiley, 2013.
4. K. He, J. Sun, and X. Tang. “Single Image Haze Removal using Dark Channel Prior,” Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1956-1963, 2009.
5. Q. Fang, J. Sang, and C. Xu, “Discovering Geo-Informative Attributes for Location Recognition and Exploration,” ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 11, no. 1s, Article 19, 2014.
6. J. Hays and A.A. Efros, “IM2GPS: Estimating Geographic Information from a Single Image,” Proceedings of IEEE Computer Vision and Pattern Recognition Conference, 2008.
7. T.-Y. Lin, S. Belongie, and J. Hays, “Cross-View Image Geolocalization,” Proceedings of IEEE Computer Vision and Pattern Recognition Conference, pp. 891-898, 2013.
8. X. Li, M. Larson, and A. Hanjalic, “Global-Scale Location Prediction for Social Images Using Geo-Visual Ranking,” IEEE Transactions on Multimedia, vol. 17, no. 5, pp. 674-686, 2015.

9. T. Chen, K.-H. Yap, and D. Zhang, “Discriminative Soft Bag-of-Visual Phrase for Mobile Landmark Recognition,” IEEE Transactions on Multimedia, vol. 16, no. 3, pp. 612-622, 2014.
10. I.-C. Shen and W.-H. Cheng, “Gestalt Rule Feature Points,” IEEE Transactions on Multimedia, vol. 17, no. 4, pp. 526-537, 2015.
11. Y. Li, D.J. Crandall, and D.P. Huttenlocher, “Landmark Classification in Large-Scale Image Collections,” Proceedings of IEEE International Conference on Computer Vision, pp. 1957-1964, 2009.
12. W. Min, C. Xu, M. Xu, X. Xiao, and B.-K. Bao, “Mobile Landmark Search with 3D Models,” IEEE Transactions on Multimedia, vol. 16, no. 3, pp. 623-636, 2014.
13. X. Li, C. Wu, C. Zach, S. Lazebnik, and J.-M. Frahm, “Modeling and Recognition of Landmark Image Collections Using Iconic Scene Images,” International Journal of Computer Vision, vol. 95, no. 3, pp. 213-239, 2011.
14. Y.-T. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven, “Tour the World: Building a Web-Scale Landmark Recognition Engine,” Proceedings of IEEE Computer Vision and Pattern Recognition Conference, pp. 1085-1092, 2009.
15. A. Krizhevsky, I. Sutskever, and G.E. Hinton, “ImageNet Classification with Deep Convolutional Neural Network,” Proceedings of Advances in Neural Information Processing System, 2012.
16. B. Thomee, E.M. Bakker, and M.S. Lew, “TOP-SURF: A Visual Words Toolkit,” Proceedings of ACM International Conference on Image and Video Retrieval, pp. 1473-1476, 2010.
17. A. Vedaldi and K. Lenc, “MatConvNet – Convolutional Neural Networks for Matlab,” arXiv:1412.4564, 2014.