Incorporating Temporal Context with Content for Classifying Image Collections

Matthew Boutell
Department of Computer Science
University of Rochester
boutell@cs.rochester.edu

Jiebo Luo
Research and Development Laboratories
Eastman Kodak Company
jiebo.luo@kodak.com

Abstract

Semantic scene classification is an open problem in image understanding, especially when information purely from image content (i.e., pixels) is employed. However, in applications involving image collections, surrounding images give each image a temporal context. We present a probabilistic approach to scene classification, capable of integrating both image content and temporal context. Elapsed time between images can be derived from the timestamps recorded by digital cameras. Our temporal context model is trained to exploit the stronger dependence between images captured within a short period of time, indicated by the elapsed time. We demonstrate the efficacy of our approach by applying it to the problem of indoor-outdoor scene classification and achieving significant gains in accuracy. The probabilistic temporal context model can be applied to other scene classification problems.

1. Introduction

Determining the semantic classification (e.g., indoor, beach, birthday party) of an arbitrary image has been studied much in recent years (e.g., [8,9,11]). It has many useful applications, including content-sensitive image enhancement (e.g., color-balancing indoor scenes differently than sunsets) and content-based image organization [9]. Most classifiers use low-level features derived from the image *content* only (e.g., colors, textures, edges) only and achieve some success on constrained image sets (e.g., Corel). However, on unconstrained home photos, which vary more greatly in content, scene classification is still very much an open problem.

While *individual* home photos are more difficult to classify than professional stock photos, they usually appear as *collections* (e.g., film scans or digital image uploads), a fact that can be exploited in the classification process. Each photograph has a *temporal context* consisting of the surrounding photos in the collection. Furthermore, one expects the context to be more closely related if the photos were captured within a short time, and image timestamps can be used to predict the strength of the context relationships.

In this paper, we exploit temporal context for the purpose of automatic scene classification of home photographs, focusing on the problem of indoor-outdoor scene classification. We first classify photos solely using content and then impose a temporal context model consisting of entire sequences of photos. The temporal model helps improve classification accuracy by correcting mistakes made by the content-based classifier, with little computational overhead.

The value of elapsed time between photographs has been shown in two non-classification applications: clustering and retrieval. Elapsed time can be used to help cluster images into events [2,6], complementing content-based strategies. By definition, CBIR (*content-based* image retrieval) uses only image content, but recently, Mulhem and Lim used the classification of groups of images to improve retrieval performance [4]. Our proposed approach differs from previous research in that it is for classification. While clustering could be used as a preprocessing step before classification, our principled approach combines context and content in a single step, thus avoiding errors made by a clustering algorithm and the associated computational overhead. Finally, our probabilistic approach differs greatly from the rule-based scheme in [4].

The main contribution of our work is a probabilistic model that integrates image content with elapsed time between images for classification of image collections. In particular, the transitions between images are functions of the elapsed time. The use of temporal context in our work is more sophisticated than that in video [12], in which the elapsed time between images is constant.

2. Temporal Context Model

Inspired by the speech recognition paradigm, we model sequences of images using a graphical model in the form of a Hidden Markov Model (HMM), representing the class of each image (a random variable) as a node in the graph and the temporal dependencies between nodes as edges.

Our goal is to maximize the probability of classifying the sequence of images correctly, given the content-based evidence (the output of a per-image classifier such as a Support Vector Machine (SVM)) and the contextual relationships between images. In the terminology of HMMs, the classes are the (hidden) states and the evidence is the observed output.

Let C_i be the true semantic classification (label) of image i, and $C = \{C_1, C_2, \dots, C_N\}$ be a vector of the classifications of a sequence of N images arranged in temporal order. Let E_i be the evidence (based on image i alone) for the class of image i, and $E = \{E_1, E_2, \ldots, E_N\}$ be the corresponding evidence vector. We call E *content-based* evidence, because it is originates in the content (e.g. color, texture) of the image.

In this formalism, our goal is to maximize:

$$\begin{aligned} P(C \mid E) &\propto P(E \mid C)P(C) \\ &= P(E \mid C)P(C_n \mid C_1 \ldots C_{n-1})P(C_1 \ldots C_{n-1}) \\ &= P(E \mid C)P(C_n \mid C_1 \ldots C_{n-1}) \ldots P(C_2 \mid C_1)P(C_1) \\ &= \left(\prod_{i=1}^{n} P(E_i \mid C_i)\right)\left(\prod_{i=1}^{n} P(C_i \mid C_{i-1})\right) \end{aligned} \quad (1)$$

The derivation is standard for HMMs [3]. The first line follows from Bayes' rule. The second and third lines follow from successive applications of the chain rule. Finally, the last line utilizes two independence assumptions: (1) the content-based classifier's output for image *i* depends only on image *i* (and upon no other image) and (2) a locality (Markov) assumption in the form of the bigram model, typically used for on-line speech recognition: node *i*'s context consists only of node *(i-1)*. We define $P(C_1|C_0) = P(C_1)$ for notational convenience.

We note that our assumptions mirror those used in image analysis problems in which the underlying content of a scene is to be inferred from an image [1]. Figure 1 is our graphical model; a chain topology captures our assumptions for image collections. We discuss learning $P(C_i|C_{i-1})$ and $P(E_i|C_i)$ in the following section.

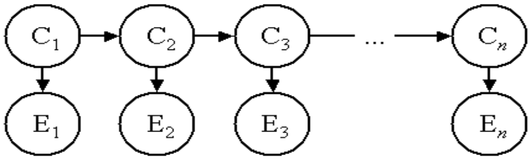

Figure 1. An appropriate graphical model for temporally-related images is a Hidden Markov Model.

3. Learning Time-dependent Transitions

The graphical model requires two types of probabilities, transition probabilities and output probabilities, each of which can be either set by an expert or learned from data. We choose to learn them (but understand that they are intuitive enough to be fine-tuned by experts).

3.1 *Timestamp-Dependent* Transition Probabilities

Transition probabilities, e.g., $P(C_i|C_{i-1})$, model temporal context between images, providing the relative likelihoods of a scene having a class label, given the labels of surrounding scenes. For example, if an image is both preceded by and succeeded by indoor scenes, we would certainly expect that it, too, is of an indoor scene. Transition probabilities govern the *strength* of class relationships between neighboring images. However, the elapsed time between the two images affects this strength greatly (e.g., two neighboring images captured within seconds are much more likely to be related than those captured days apart): transition strength between two images *decreases* as the elapsed time between them *increases*.

We learn these probabilities by counting transition pairs in a training set and then normalizing to obtain the appropriate probabilities, shown in Table 1 (for indoor-outdoor scene classification, discussed in Section 5). We used the time intervals shown because we hypothesized that the strength of the transition drops off exponentially and to mitigate the fact that the bins contain a monotonically decreasing number of images per bin. Note that the indoor and outdoor probabilities are not symmetric because they have different priors. Furthermore, the probabilities approach the class priors in the limit, although the data is noisy for large time intervals.

Table 1. Elapsed-time dependent transition probabilities learned from data set D1. Note a trend towards the prior probability as the elapsed time increases.

Time Interval	nImages	$P(C_i\text{=in}\mid C_{i-1}\text{=in})$	$P(C_i\text{=out}\mid C_{i-1}\text{=out})$
0-30 sec	1040	0.994	0.985
30-90 sec	528	0.973	0.974
90 sec-5 min	260	0.933	0.952
5-15 min	145	0.847	0.849
15-45 min	109	0.897	0.863
45 min-2 hrs	98	0.803	0.757
2-6 hrs	62	0.697	0.621
6+ hrs	314	0.707	0.455

3.2 Output Probabilities

Output probabilities, e.g., $P(E_i|C_i)$, give the relative likelihoods of observing the evidence (the output of the content-based classifier), given the true scene class. We shape the real-valued output from an SVM classifier by a sigmoid function into a pseudo-probability that image *i* belongs to class 1 [10]. We interpret this as virtual evidence with probability $P(E_i)$, and use the same treatment as Pearl (see [5, p. 44]); when computing the MAP solution, only the likelihood of the evidence is important.

4 Inference Using the Viterbi Algorithm

To maximize Equation 1 using a *brute force* algorithm requires exponential complexity, $O(M^N)$, where N is the sequence length and M is the number of scene labels. However, because we assume that the class of an image depends only on the class of the previous image in the sequence, an efficient dynamic programming algorithm can be used to perform the optimization [3]. The Viterbi algorithm iterates through the sequence in a forward manner, keeping track of the optimal path (maximal probability) to each state from the start, then reading that optimal path in a backward pass. The efficiency is gained because the optimal path to any state S_i *must* contain the optimal path to one of the S_{i-1}, allowing local computations [3]. The Viterbi algorithm has complexity of only $O(M^2N)$. In scene classification, M is expected to be small (e.g., $M=2$ in indoor-outdoor classification), yielding time linear in the sequence length.

We obtain our ***elapsed time-dependent transition probabilities*** for inference as follows. Each pair $(i, i+1)$ of adjacent images in a test sequence has an associated elapsed time T_i between them. We discretize T_i using the same bins used in training, thus obtaining the transition matrices, $P(C_{i+1}|C_i)$, by indexing into the appropriate row of Table 1. Note that other options may be more helpful depending on the data: interpolating between rows (if large differences exist between transition probabilities in adjacent bins) or fitting a continuous function to the transitions (if the data is noisy).

5. Indoor-Outdoor Scene Classification

Due to the difficult nature of the classification of photographs, hierarchical methods have been proposed [11] in which indoor-outdoor classification is the first step. Therefore, it is an important problem addressed by a number of researchers (e.g., Refs [8,9,11]).

Our baseline content-based classifier is a Support Vector Machine using block-based color histogram and wavelet (texture) features, designed in a similar way to [8,9]. We trained it on an independent set of film and digital images not used elsewhere in this study. The output of the SVM is shaped via a sigmoid function into a pseudo-probability [10]. Balance between indoor- and outdoor-image recall can be obtained by thresholding the probability.

Our image database consists of 24,000 pictures with a wide variety of image content; 56 photographers from three U.S. cities took pictures over the course of 12 months. We randomly sampled a subset, D1, of 132 image sequences (containing a total of 2049 images) from the database and labeled them for many semantic traits (including indoor-outdoor). Of these, 1205 were indoor, giving a prior probability, $P(indoor) = 0.588$.

We learned transition probabilities from D1 (see Table 1). As expected, the strength of the probabilities diminished as the elapsed time between images increased. We evaluated the system performance on D1 as well, mainly because of a shortage of labeled data. However, using a single set of images to derive simple statistics (transition probabilities) only for probabilistic models and to test the system does not give overly biased results (as opposed to classifiers like neural networks, which can memorize the training set). We had trained our content-based classifier on a completely independent set of images, so this was not a concern.

For testing, the content-based classifier classified each image, the transition probabilities were computed from the elapsed times, and the Viterbi algorithm was used to determine the optimal classification of the entire sequence. As a result, images with weak content-based cues tended to be influenced by the classification of surrounding images, while the classifications of those with strong content-based cues or those captured in relative "temporal isolation" usually remained unchanged. Figure 2 shows a comparison of the temporal model (content + context) against the baseline (content-based classification only). The curve is obtained by varying the bias of the baseline classifier. While not a true ROC curve, it shows the balance between indoor and outdoor recall at different operating points. The gain due to imposing the temporal context model is evident, leading to substantial improvements regardless of the underlying classifier bias. Our probabilistic temporal context model increased the expected accuracy on the data set by correcting many errors made by the baseline classifier, while only causing a few images to be classified incorrectly.

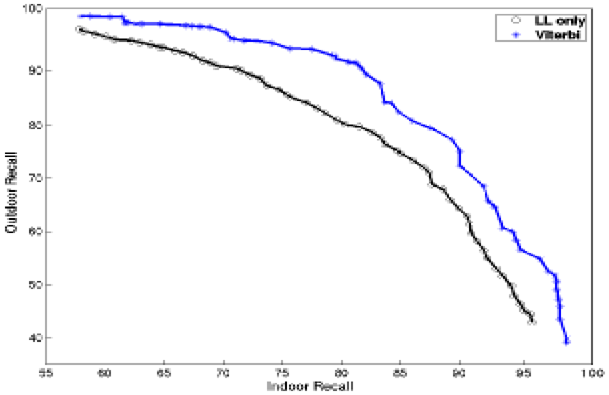

Figure 2. Comparison of the baseline content-based indoor-outdoor classifier with that enhanced with the elapsed-time temporal model.

Figure 3 gives examples of images for which the context model changed its classification. The first two sequences are examples in which the temporal model helped correct the baseline classifier's errors. Considering each image in isolation, it is understandable why the color-texture classifier made errors. However, the short elapsed time between images made the baseline classifier's predictions unlikely. Conversely, the model biased against the rare cases in which the photographer moved from outdoors to indoors and back again quickly (e.g., sequence 3), but these errors are a small price to pay for the gains the model makes.

6. Conclusions and Future Work

We have demonstrated that a temporal context model based on *timestamp-dependent* transitions between images is valuable for improving upon the performance of content-only scene classifiers. Our context model capitalizes on coherence in semantics between neighboring images, especially if captured within a short length of time, because photographers can only move between scenes so fast. Temporal inference adds little overhead to the baseline classifier, and can be utilized whenever the output from the content-based classifier is probabilistic in nature.

In the future, we plan to extend the system to handle multi-class problems such as general outdoor scene classification (e.g., [11]).

Acknowledgments

We would like to thank Chris Brown of the University of Rochester and Bill Freeman of MIT for helpful discussions.

References

1. W.T. Freeman, E.C. Pasztor, and O.T. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 40(1):24-57, October 2000.
2. A.C. Loui and A. Savakis. Automatic event segmentation and quality screening for albuming applications. *Proceedings of IEEE International Conference on Multimedia and Expo*, 2000.
3. C. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.
4. P. Mulhem and J-H Lim. Home photo retrieval: time matters. *Lecture Notes in Computer Science*, 2728:321-330. 2003.
5. J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann Publishers, 1988.
6. J. Platt. AutoAlbum: Clustering digital photographs using probabilistic model merging. *Proeedings of IEEE Workshop on Content-based Access of Image and Video Libraries*, 2000.
7. Y. Song and A. Zhang. Analyzing scenery images by monotonic tree. *Multimedia Systems Journal*, 8(6):495-511, 2002.
8. N. Serrano, A. Savakis, and J. Luo. A computationally efficient approach to indoor/outdoor scene classification. *Proceedings of International Conference on Pattern Recognition*, 2002.
9. M. Szummer and R. W. Picard. Indoor-outdoor image classification. *Proceedings of IEEE Workshop on Content-based Access of Image and Video Databases*, 1998.
10. D. Tax and R. Duin. Using two-class classifiers for multi-class classification. *Proceedings of International Conference on Pattern Recognition*, 2002.
11. A.Vailaya, M.Figueiredo, A.Jain, and H.-J. Zhang. Content-based hierarchical classification of vacation images. *Proceedings of International Conference on Multimedia Computing and Systems*, 1999.
12. N. Vasconcelos and A. Lippman. Statistical models of video structure for content analysis and characterization. *IEEE Trans. Image Processing,* 9(1):3-19, 2000.

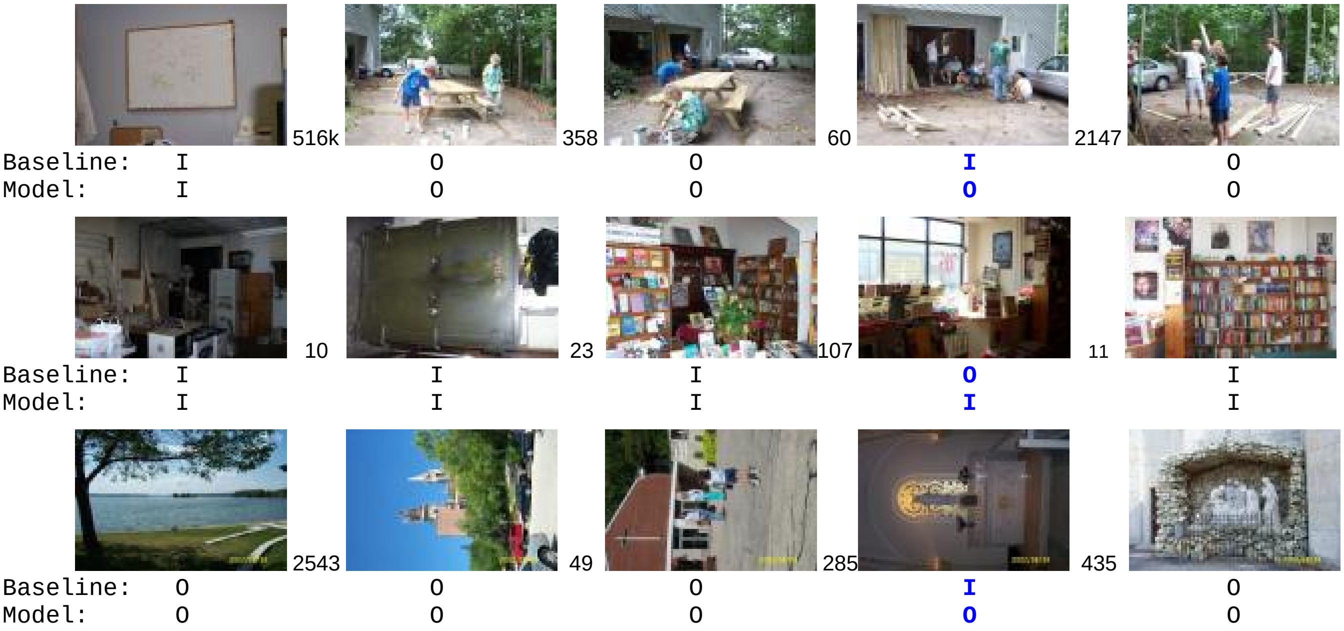

Figure 3. Image sequences affected by the context model. Elapsed times (in seconds) between images are shown. The first two sequences show examples in which the model corrected errors made by the baseline classifier. The third sequence shows a conceptual error: a rare case where the photographer walks into a room briefly between taking two outdoor photos.