Multimedia Systems (2005)
DOI 10.1007/s00530-005-0202-7

REGULAR PAPER

Matthew Boutell · Jiebo Luo · Christopher Brown

A generalized temporal context model for classifying image collections

Published online: 16 November 2005

Abstract Semantic scene classification is an open problem in computer vision, especially when information from only a single image is employed. In applications involving image collections, however, images are clustered sequentially, allowing surrounding images to be used as temporal context. We present a general probabilistic temporal context model in which the first-order Markov property is used to integrate content-based and temporal context cues. The model uses *elapsed time-dependent* transition probabilities between images to enforce the fact that images captured within a shorter period of time are more likely to be related. This model is generalized in that it allows arbitrary elapsed time between images, making it suitable for classifying image collections. In addition, we derived a variant of this model to use in ordered image collections for which no timestamp information is available, such as film scans. We applied the proposed context models to two problems, achieving significant gains in accuracy in both cases. The two algorithms used to implement inference within the context model, Viterbi and belief propagation, yielded similar results with a slight edge to belief propagation.

Keywords Semantic scene classification · Content-based cues · Temporal context cues · Hidden Markov Model · Camera metadata

M. Boutell
Department of Computer Science, University of Rochester, Rochester, New York, USA; Research and Development Laboratories, Eastman Kodak Company, New York, USA
E-mail: boutell@cs.rochester.edu

J. Luo (✉)
Research and Development Laboratories, Eastman Kodak Company, New York, USA
E-mail: jiebo.luo@kodak.com

C. Brown
Department of Computer Science, University of Rochester, Rochester, New York, USA
E-mail: brown@cs.rochester.edu

1 Introduction

1.1 Motivations

While image and video analyses share many important characteristics, they differ greatly in the availability of temporal context. An image or video frame's *temporal context* includes those images or frames that are temporally adjacent to it, and the timestamps on the image and its neighbors. Because of the lack of such context information in single still images, image understanding in this realm has been recognized as more difficult than that in video streams [20]. However, image *collections* (e.g., photographs from film scans and digital camera downloads) represent a middle ground between the two. In this study, an image collection may contain multiple events, as shown in the examples later in the paper. Individual events, groups of images that are semantically and temporally related (e.g., an indoor party or an outdoor picnic) may be classified with a single class. However, in practice, this requires pre-segmentation of the collection into events. In general, the images in a collection belong to multiple classes.

Drawing an analogy with another genre, image collections are more like *comic strips* (or *storyboards*) than *animated cartoons*, series of images not nearly as tightly coupled as video frames, but still related to one another as part of a photographer's story. If a collection's photos were captured hours apart, this context may be meaningless, but if they were captured mere seconds apart, the content is very likely to be related; digital cameras provide timestamps to distinguish the two situations.

The value of context for recognition has long been appreciated by various research communities. Temporal context is used in speech recognition: humans can understand phone conversations even when some of the syllables or words are muddled by noise, and all successful automatic speech recognizers use temporal context models. In computer vision, spatial context has been shown to improve object recognition [22, 28]. In video processing, researchers also make strong use of temporal coherence for recognition purposes

[14]. However, temporal context in image collections is a largely untapped resource.

In this paper, we develop and exploit temporal context models for the purpose of automatic scene classification of single images in photograph collections. Automatically determining the semantic classification (e.g., mountain, sunset, indoor) of an arbitrary image has many useful applications, such as content-based image enhancement and organization [21].

We employ an integrated approach, first classifying images solely using content embedded in pixels and then imposing a temporal context model (i.e., a Hidden Markov model) consisting of entire image collections. In the Hidden Markov model, the observed states correspond to the output of a classifier based on image content (color and texture features) and the hidden states correspond to the true class of each image. Without temporal context, each pair of observed and hidden states for the same image would be the same. We show that the temporal context model helps improve the overall classification accuracy by correcting mistakes made by the content-based classifier, with little computational overhead. While with some difficulty it is possible to integrate the temporal constraints directly into the classifier, an advantage of this modular approach over integrating the temporal constraints directly into the classifier is that the temporal model can be used in conjunction with any content-based classifier, which itself may be improved upon in the future. Moreover, within this framework, we can accommodate content-based classifiers that are binary or real-valued in nature.

It is noteworthy that the applications themselves are not the central issues of this paper. Instead, we propose a generalized, probabilistic temporal context model that is a function of the elapsed time between images, such as can be derived from digital camera timestamps. When elapsed times are unavailable, as with film scans, this model degenerates to a special case, in which only the relative order of images is exploited. This probabilistic model is an integrated and principled approach to exploit temporal correlation and image content as opposed to an *ad hoc* approach (for example, heuristic rules or temporal smoothing). In some sense, our approach is similar in spirit to edge-preserving image smoothing because it allows for (temporal) discontinuity. We confirm the intuition that using a model with richer (elapsed-time) information increases performance more significantly in the context of two image classification problems. Furthermore, we can treat video classification as a special case (with constant elapsed time between frames) of the proposed temporal context model.

The main contributions of our work are a general (with respect to variable elapsed time), probabilistic temporal context model and its applications to scene classification problems. This model is applicable to scenarios both with and without elapsed time. In addition, our algorithm does not require pre-segmentation of the images into events, where event boundaries would need to be determined manually (labor intensive) or automatically (error prone [19]).

1.2 Related work

Semantic scene classification has been studied extensively in recent years (e.g., Refs. [3, 21, 24, 27, 29]). Most current systems rely on the low-level content (e.g., colors, textures, edges) of the image only and have achieved limited success on constrained image sets (e.g., the Corel database). However, on unconstrained photographs, scene classification is still very much an open problem, especially when only image (i.e., pixel) information is used.

Meanwhile, *relative* time information (elapsed time between photographs) has been used successfully in two non-classification applications. First, *clustering* or grouping photographs by timestamps was used to complement content-based clustering strategies [12, 19]. In [12], Loui and Savakis first use timestamps alone to determine event boundaries, and then rely on a series of heuristic tests to check if the color histograms of the images at event boundaries indeed differ. Similarly in [19], a two-stage process is used by Platt to combine time-based clustering (HMM) and content-based clustering (color histogram), starting with the time-based clusters and then splitting any cluster equal to or larger than 48 images into content-based clusters with an average size of eight. Second, Mulhem and Lim recently proposed, within the context of *image retrieval*, to exploit other images within a temporal cluster [15]. Their metric for relevance between a query Q and a database image D incorporates not only the match between Q and D, but also the match between Q and the best-matching image in the same temporal cluster as D.

The approach proposed here is primarily for supervised classification, distinguishing it from the prior work on unsupervised event clustering [12, 19] and image retrieval [15]. While one could use event clustering as a precursor to image classification, that is not necessarily the best approach because clustering errors would propagate to the classification stage and degrade performance. Because our algorithm operates without the need for clustering, it also avoids the related overhead. In addition, our probabilistic approach differs greatly from the rule-based scheme in [15].

Probabilistic inference, with Hidden Markov Models in particular, has been used extensively for analysis of video streams, because of the strong temporal coherence between adjacent frames [1, 5, 10, 16, 23, 27, 30]. By comparison, image collections pose two unique problems with respect to temporal context. First, the dependencies between images are generally much weaker than between frames, because the elapsed times between images (on the order of minutes) are much greater than those between typical video frames (on the order of milliseconds). Second, the elapsed times between images are highly variable (ranging from seconds to days), while those between frames are constant. Therefore, while our work relates to previous work on video collections, it must use a generalized, more powerful model for temporal context. Indeed, because our model is built upon elapsed-time dependent transition probabilities, it allows for general usage on unconstrained image collections with or

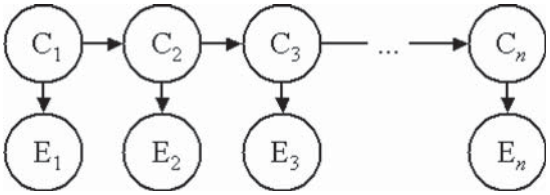

Fig. 1 An appropriate graphical model for temporally related images is a Hidden Markov model

without timestamps; video is essentially a special case where the elapsed time is constant.

2 Probabilistic temporal context model

In this study, we define an image "sequence" as a collection of photos captured in succession, but with *arbitrary* elapsed time between them (as would be the case with a typical amateur photographer's collection). We model image sequences using a Hidden Markov model (HMM), representing the classes of single images (the random variables, C_i) and the observed evidence (the random variables, E_i) as nodes in the graph, and the temporal dependencies between nodes as edges (see Fig. 1).

In image analysis, where the underlying content (e.g., classification) of a scene is to be inferred from an image, two independence assumptions are generally made [4, 7]: scene and observation dependence at each pixel or region, and locality (each scene node depends only on neighboring scene nodes). Our problem differs only in that our observations are for *whole images*, and the locality refers to dependence *across* neighboring images. However, the spirit of our assumptions is the same. First, the output of a content-based classifier on an image depends only on the class of *that image* and no other. Second, the class of an image depends only on the class of the image immediately preceding it (its temporal neighborhood), and upon no other images. This locality assumption is the commonly known Markov property. The *directed* graphical model shown in Fig. 1 captures these assumptions. Later on in Sect. 3.2, we will extend this model to an undirected graph.

We formalize the problem as follows: let C_i be the true semantic classification (label) of image i, and $C = \{C_1, C_2, \ldots, C_n\}$ be a vector of the classifications of a sequence of n images arranged in temporal order. The 1st-order Markov property can be stated formally as $P(C_i|\{C_j\}, j \neq i) = P(C_i|C_{i-1})$. Let M be the number of possible scene labels. 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 for the sequence. We call E *content-based* evidence, because it is derived from the content (e.g., color and texture) of the image. We emphasize that E corresponds to the output of the content-based classifier, and furthermore is a real valued number as opposed to a binary value. In other words, the classifier outputs a value in [0,1] with 0 corresponding to highest confidence in one class (e.g., indoor) and 1 corresponding to highest confidence in the other class (e.g., outdoor). In the terminology of HMMs, the classes are the (hidden) states and the evidence is the observed output.

Our goal is to determine the *maximum a posteriori* (MAP) solution, i.e., maximize the probability of classifying the sequence of images correctly, given both the content-based evidence and the context relationships between images. In this formalism, our goal is:

$$\begin{aligned}\arg\max_C P(C|E) &= \arg\max_C P(E|C)P(C) \\ &= \arg\max_C \left(\prod_{i=1}^{N} P(E_i|C_i)\right) P(C) \qquad (1)\end{aligned}$$

The derivation is standard for HMMs [13], Bayes' rule followed by our first assumption above. Note that $P(E)$ is fixed at run time and thus does not affect the result of MAP estimation. In other words, $P(C|E)$ is proportional to $P(E|C)P(C)$, so when we take the argmax with respect to C, they are equivalent. We discuss learning $P(E_i|C_i)$ and $P(C)$ in Sect. 5.

3 Inference algorithms

We now discuss two inference schemes for imposing the temporal context models, one for directed graphs (HMMs) and one for undirected graphs. The first is the Viterbi algorithm [6], used as in speech recognition, to find the most likely sequence taken as a whole. The second is the max-product version of belief propagation [7], used on an *undirected* version of our temporal context model.

We include the undirected context model to obtain a more comprehensive comparison. The correlation between photographs is not necessarily dependent on time progressing in a forward manner; there is nothing fundamentally wrong with looking at the correlation backwards on our image sequence, for example. Backwards correlation has been exploited in speech recognition [9] where later words in a sentence or syllables in a word disambiguate earlier ones, even though humans are not naturally wired for that. Max-product belief propagation allows us to explore correlation in both directions between photographs. Intuitively, a richer context can be exploited for potential benefit.

The directed and undirected versions each have their own method of expanding and simplifying $P(C)$ using the Markov property.

3.1 Viterbi inference algorithm

In real-time speech recognition, one usually models the prior probability $P(C)$ using a bigram model (in which the context of each state consists of the *previous* state only). Using the chain rule successively and then applying the Markov

property for bigram models gives Eq. (2). In the last line, we define $P(C_1|C_0) = P(C_1)$ for notational convenience.

$$\begin{aligned} P(C) &= P(C_N|C_1 \ldots C_{N-1})P(C_1 \ldots C_{N-1}) \\ &= P(C_N|C_1 \ldots C_{N-1})P(C_{N-1}|C_1 \ldots C_{N-2}) \\ &\quad \ldots P(C_2|C_1)P(C_1) \\ &= \prod_{i=1}^{N} P(C_i|C_{i-1}) \end{aligned} \tag{2}$$

To maximize Eqs. (1) and (2), a brute force algorithm requires complexity that is exponential in the number of images. However, because the conditional probabilities "flow" in one direction only, we may use the Viterbi algorithm, an efficient dynamic programming algorithm, to perform the optimization [6, 13]. The Viterbi algorithm iterates forward through the list, keeping track of, for each state, the optimal path (maximal probability) to that state from the start, then backward, to read the optimal path. The efficiency is gained because the optimal path to any state S_i must contain one of the optimal paths to state S_{i-1}, allowing local computations at each node [13]. The Viterbi algorithm has complexity of only $O(M^2N)$. In scene classification, M is typically small (e.g., $M = 2$ in indoor-outdoor classification), yielding time linear in the sequence length.

3.2 Belief propagation (on an undirected graph)

With an undirected graph, we assume that the class of an image depends only on the classes of the images immediately preceding and succeeding it (its temporal neighborhood), and upon no other images.

Markov networks, in the form of Markov random fields (MRFs), have been used in the computer vision community for problems of inferring scenes from images (e.g., [4, 7]). However, the 2D grid used in those cases contains loops (as opposed to a simple chain), for which no exact efficient inference algorithms exist. Typical inference in this topology is done using greedy [2, 4] or stochastic methods [8, 17]. Belief propagation [7] is a message-passing algorithm that successfully ignores these loops in many graphs (and is exact in topologies without loops, such as ours). It is a generalization of Pearl's algorithm for inference in Bayesian networks [18], in which messages passed in opposite directions do not interfere.

In belief propagation, each node (i.e., image in our study) calculates its own *Maximum a Posteriori* (MAP) estimate of the scene class. The factorization of $P(C)$ is different for each node; its derivation for node i is as follows.

$$\begin{aligned} P(C) &= P(C_i)P(C_1 \ldots C_{i-1}, C_{i+1} \ldots C_N|C_i) \\ &= P(C_i)P(C_1 \ldots C_{i-1}|C_i)P(C_{i+1} \ldots C_N|C_i) \\ &= P(C_i)\left(\prod_{k=2}^{i} P(C_{k-1}|C_k)\right)\left(\prod_{k=i}^{N-1} P(C_{k+1}|C_k)\right) \end{aligned} \tag{3}$$

This corresponds to the following message passing algorithm for the MAP estimate at node j (specific to chain topology, see [7] for derivation of the general case). For chains, the MAP estimate is:

$$\arg \max_{C_j} P(C_j)P(E_j|C_j)L_{(j-1)j}(C_j)L_{(j+1)j}(C_j) \tag{4}$$

$P(C_j)$ is the prior on image j, $P(E_i|C_i)$ are the same output probabilities used in the Viterbi algorithm, and the likelihood messages $L_{kj}(C_j)$ from node k to node j about scene class C_j are calculated using:

$$L_{kj}(C_j) = \max_{C_k} P(C_k|C_j)P(E_k|C_k)\tilde{L}_{lk}(C_k) \tag{5}$$

where the $\tilde{L}$ are the messages from the previous iteration (initially set to 1) and node l is node k's other neighbor. Note that E (evidence) and C (class) in our notation correspond to y and x, respectively, in the notation of [7]. Messages are passed until the MAP estimates converge. Because our topology contains no loops, it is guaranteed to converge to an exact solution in linear time (at most N steps).

4 Elapsed time-dependent transition probabilities

The transition probabilities, $P(C_i|C_j)$, model temporal context between images. For example, if an image is both preceded by and succeeded by images of indoor scenes, we might well expect that it, too, is of an indoor scene. Transition probabilities govern the *strength* of class relationships between neighboring images. However, an integral part of this work rests on the following observation:

Transition strength between two images increases as the elapsed time between them decreases

For example, two adjacent images captured 17 s apart are almost guaranteed to be of the same event, while ones captured three weeks apart are probably unrelated (or even if semantically related, e.g., of the same place, they are not capturing the same event). We formalize the fact that transition probabilities are dependent on elapsed time as follows. Let τ be the elapsed time between two neighboring images, and the probability distribution $e(\tau)$ (e.g., exponential), be the prior on elapsed time. The class of an image is conditioned on the class of the previous image and the elapsed time between them: $P(C_i \mid C_{i-1}, \tau)$, where the C_i are discrete and τ is continuous. This conditional distribution can be specified using $M(M\text{-}1)$ functions (i.e. the number of free parameters, given that probabilities must sum to 1), which we denote as *transition functions*. For example, consider the transition function $P(C_i = c|C_{i-1} = c, \tau)$, corresponding to adjacent images having the same class (Fig. 2), e.g., $c =$ indoor. Ideally this function is expected to decrease monotonically (because the photographer can only move so fast)

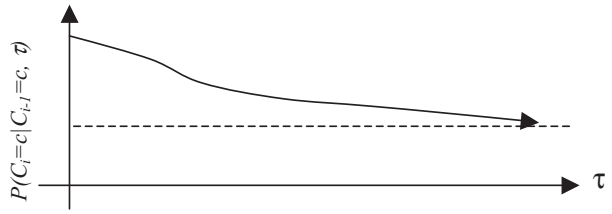

Fig. 2 Transition function for $P(C_i = c|C_{i-1} = c, \tau)$

and approach the prior $P(C_i = c)$in the limit as $\tau \rightarrow \infty$. The transition functions are appropriate for use in situations with or without timestamps.

5 Learning

The graphical model requires two types of probabilities: transition probabilities and output probabilities. Each of these probabilities was learned from data (but understand that they are intuitive enough to be fine tuned by experts); we reference those learned for the indoor-outdoor classification problem throughout our discussion. We start with the transition probabilities, both dependent upon and independent of elapsed times.

5.1 Elapsed time-dependent transition probabilities

Each pair $(i, i{+}1)$ of adjacent images in a test sequence has an associated elapsed time τ_i between them. We have three options regarding how to determine the transition probability with respect to τ_i. The simplest is to discretize τ_i using discrete time slices and use it to index into the discretized transition probability functions (i.e., the slices in Fig. 3). Another is to create these discrete transition functions and then interpolate between them for values of τ_i between the slices. This method may be more precise if there is a large difference between transition probabilities in adjacent bins (assuming that the transition probabilities change smoothly over elapsed time). Third is to model the decrease in transition strength using a continuous function and use τ_i directly, which is helpful if the data is noisy. We chose the first option, due to its simplicity and faithfulness to the data. Moreover, this alleviates the need for large amounts of training data because the number of bins is relatively small.

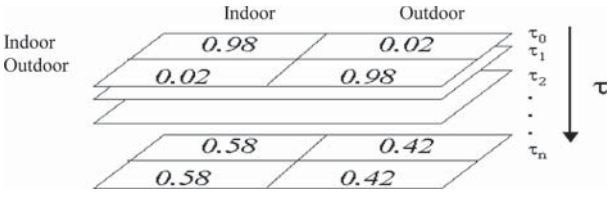

Fig. 3 Elapsed time-dependent temporal context model. The transition probabilities used between two images are a function of the elapsed time between them. As $\tau \rightarrow \tau_n$, the probabilities approach the class priors

We consider a *discrete* slice, $S_j = [\tau_j, \tau_{j+1})$, of elapsed times (e.g., 5–15 min). To determine the transition probabilities for two images with elapsed times in S, we need to calculate the probability:

$$P(C_i|C_{i-1}, \tau \in S_j) = \frac{\int_{\tau_j}^{\tau_{j+1}} P(C_i|C_{i-1}, \tau)e(\tau)\,\mathrm{d}\tau}{\int_{\tau_j}^{\tau_{j+1}} e(\tau)\,\mathrm{d}\tau} \quad (6)$$

This simply means that to find the discretized transition probability for a time slice, we integrate the continuous transition probability function over the interval corresponding to the slice and normalize it by the probability of that interval of elapsed time.

Calculating $P(C_i \mid C_{i-1}, \tau)$ directly is nontrivial. A reasonable discrete approximation can be easily obtained from our training data by partitioning the image pairs in the training set according to elapsed time, counting the number of pairs in each partition, and then normalizing the frequencies to obtain the appropriate probabilities (e.g., Table 1). We used non-overlapping slices that increased exponentially in length, mitigating the fact that the bins contain a monotonically decreasing number of images per bin.

The discretized transition probabilities can be visualized as a 3-D matrix, in which the added dimension is *elapsed* time, τ (see the example of indoor-outdoor transitions in Fig. 3). Each plane of the matrix is a 2-D matrix $P(C_i|C_{i-1}, \tau = \tau_k)$. As stated, the strength of the transitions is expected to drop as elapsed time increases, asymptotically approaching the prior probabilities of each class. The planes are not uniformly spaced with respect to τ for the reasons discussed above.

5.2 Marginalized transition probabilities

When elapsed time information is unavailable, one needs to use the marginal probability over elapsed time. This is obtained by calculating the transition functions (Eq. (7)), over the whole elapsed-time domain ($S_D = [0, \infty)$). This is equivalent to finding the average of the discrete slices S_i (from Table 2), weighted by the proportion, N_{S_j}, of training images in each slice (i.e., the probability distribution of

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

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

Table 2 Transition probabilities learned from data set D1, marginalizing over elapsed time, for the order-only case

	C_i	
C_{i-1}	Indoor	Outdoor
Indoor	0.924	0.076
Outdoor	0.099	0.901

training images over elapsed times):

$$\begin{aligned} P(C_i|C_{i-1}, \tau \in S_D) &= \sum_j N_{S_j} P(C_i|C_{i-1}, \tau \in S_j) \\ &= \sum_j N_{S_j} \frac{\int_{\tau_j}^{\tau_{j+1}} P(C_i|C_{i-1}, \tau) e(\tau)\, \mathrm{d}\tau}{\int_{\tau_j}^{\tau_{j+1}} e(\tau)\, \mathrm{d}\tau} \\ &= \sum_j \int_{\tau_j}^{\tau_{j+1}} P(C_i|C_{i-1}, \tau) e(\tau)\, \mathrm{d}\tau \\ &= \int_0^{\infty} P(C_i|C_{i-1}, \tau) e(\tau)\, \mathrm{d}\tau \end{aligned} \tag{7}$$

The derivation is as follows. We find the probability over the whole time domain by taking the weighted average of the discrete slices (line 1). Substituting Eq. (6) yields line 2. The proportion N_{S_j} is equal to the integral $\int_{\tau_j}^{\tau_{j+1}} e(\tau)\, \mathrm{d}\tau$, causing the two to cancel and yielding line 3. Finally, we can replace the sum of integrals over contiguous, non-overlapping intervals with the integral over the entire interval. The transition probability $P(C_i|C_{i-1}, \tau \in S_D)$ for the indoor-outdoor classification problem is shown in Table 2, obtained by finding a weighted average of the rows in Table 1. Conceptually, this is the same as the weighted average of the slices in Fig. 3. In essence, when the elapsed time is not available, a statistically optimal choice for the transition probability is the expectation of the elapsed time-dependent transition probability function. In this way, we unified the two scenarios within a single temporal context model.

5.3 Output probabilities

Output probabilities $P(E_i|C_i)$ are relative likelihoods of observing the evidence, given the true scene class. They correspond to our confidence in the output of the content classifier. Because an SVM classifier is binary and produces real-valued output, we shape the output using a sigmoid function (as recommended in [26] with an empirically-determined slope) into a pseudo-probability that image i belongs to class 1. When computing the *maximum a posteriori* (MAP) solution, only the likelihood of the evidence is important. We view the SVM output as virtual evidence and use the same treatment as Pearl (see [18, p. 44]). In this case, we consider the content-based evidence to be uncertain or fuzzy input, with probability $P(E_i)$.

6 Experimental results

We demonstrate the effectiveness of our temporal context model on two applications: indoor-outdoor classification and sunset detection.

6.1 Problem 1: Indoor-outdoor classification

Our baseline content-based classifier [21] is a Support Vector Machine using color and texture features, designed in a similar way to and improved over [25], which reported an accuracy of 90% on a different database. We trained it on an independent set of images not used elsewhere in this study.

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 selected a subset, D1, of 130 manually-labeled image sequences randomly sampled from the database (containing a total of 2049 images). Of these, 1205 were indoor, giving a prior probability of $P(\text{indoor}) = 0.588$. Each image sequence consists of the set of photographs taken by a single photographer within a 1-week period. Therefore, it generally contains images from multiple scene classes, and even multiple semantic events (e.g., sports or birthday party). Note that a single semantic event may span multiple scene classes; e.g., a birthday party may include both indoor and outdoor photos.

We learned elapsed-time dependent transition probabilities from D1 (Table 1). D1 contains images taken by different photographers over different months. Therefore, the transition probabilities learned from D1 represent averages and are not specific to individual photographers or seasons. The undirected graphic model, accounting for temporal correlation in both directions, resulted in a slight advantage in performance for belief propagation between the two inference algorithms (Table 3). As expected, the strength of the probabilities diminished as the elapsed time between images increased. Because there were more indoor images than outdoor images, the probabilities are not symmetric. For cases where elapsed time is unavailable, we obtain the bigram transition probabilities shown in Table 2 through

Table 3 Accuracy of the elapsed-time dependent and independent context models using both inference schemes and various cross-validation methods

	Inference method	
Context model	Viterbi	Belief propagation
None (baseline)	81.0%	
Elapsed time		
Learn from D1	86.7%	86.4%
Leave-one-out	86.4	86.5
10-fold	86.0 (0.16)	86.5 (0.16)
5-fold	86.1 (0.39)	86.5 (0.39)
No elapsed time		
Learn from D1	85.1%	85.3%
Leave-one-out	85.0	85.3
10-fold	85.2 (0.18)	85.4 (0.17)
5-fold	85.0 (0.42)	85.3 (0.41)

Both temporal models clearly outperform the baseline. Note that the margin of improvement induced by the elapsed time does not change over different cross-validation schemes. In addition, the differences in accuracy between the two inference algorithms are small (belief propagation seems slightly better). Standard errors are shown in parentheses.

marginalization. Note that indoor and outdoor probabilities are non-symmetric because they have different priors.

Due to the shortage of labeled data, we evaluated the system performance using leave-one-out and various cross-validation schemes. We split the data set *by sequence*, of which there are 130 in D1. Therefore, leave-one-out in this case should be interpreted as leave-one-*sequence*-out. Comparing performance with various cross-validation schemes (Table 3), both temporal models clearly outperformed the baseline. Furthermore, the *elapsed time-dependent* temporal context model outperformed the *elapsed time-independent* model by a statistically significant margin, regardless of the cross-validation schemes. For example, using the Belief Propagation algorithm, the standard error for 10-fold cross-validation is 0.16% for the *elapsed time-dependent* temporal context model and 0.17% for the *elapsed time-independent* model, both much smaller than the difference in the mean accuracies (1.1%). In fact, the largest standard error for either algorithm or cross-validation scheme is 0.42%. Note that the same evaluation is performed for the second application of sunset scene detection.

For testing, each image was classified by the content-based classifier, transition probabilities were calculated based on elapsed time, and then a temporal inference algorithm (Viterbi or belief propagation) was executed on each sequence. Accuracy is shown in Table 3. The two inference algorithms performed approximately the same, with a slight edge in favor of belief propagation, as described above.

Next, we calculated the effects of the elapsed-time model by using the elapsed-time independent transition probabilities (Table 2). Both temporal context models clearly outperformed the baseline (Table 3), more significantly with the elapsed time-dependent model. This is mainly because relations between classes of images captured days apart were (rightfully) not enforced. Figure 4 shows the ROC curves for the baseline and both temporal models using belief propagation (obtained by varying the bias of the baseline classifier).

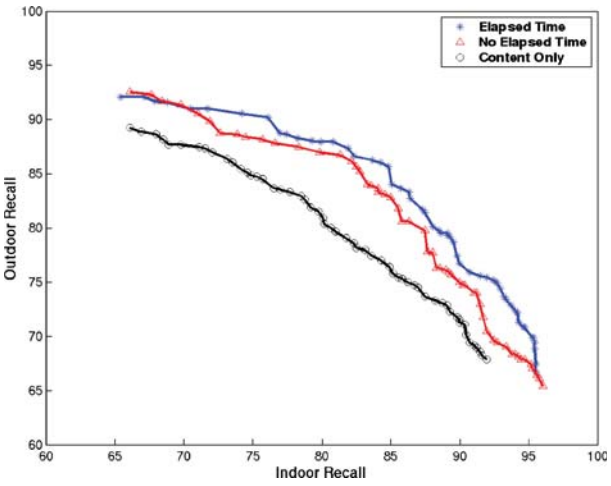

Fig. 4 Comparison of the baseline content-based indoor-outdoor classifier with those improved by the temporal models (with and without elapsed time). Note that this is not a typical ROC curve because we want to show the balance between accuracy on each of the two classes

Our probabilistic temporal context model boosted accuracy by correcting a large number of images classified incorrectly by the baseline content-based classifier. At the same time, it also caused very few images to be classified incorrectly. Figure 5 shows examples of images for which the context model changed its classification. The first three sequences are examples in which the temporal model helped correct the baseline classifier's errors. In each case, if one views each image in isolation, it is understandable why the color-texture classifier made errors (for example, the reading room brightly-lit by sunlight). 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 within seconds (e.g., sequence 3), but such errors are a small price to pay for the gains the model makes. Sequence 4 shows a misclassification that the model cannot correct because the image was captured in relative isolation (between two temporally-distinct events). Sequence 5 shows a misclassification that the model cannot correct because the image was captured in relative isolation (between two temporally-distant events). The last sequence shows a case where no change is desirable, given the long elapsed time between images.

6.2 Problem 2: Sunset detection

To further demonstrate the efficacy of the context model, we also applied it to the problem of sunset detection (we use sunset generically to refer to both sunset and sunrise because most twilight photographs are of sunsets). Similar to [3], our low-level classifier was a Support Vector Machine using spatial color moment features trained on an independent set of images unrelated to this study. For testing data, we used a collection of 4678 personal images containing 191 (4%) sunsets. Figure 6 compares the performance of the content-only classifier to those with the context models. The benefits are again clear.

With a problem such as sunset detection, in which there are many more non-sunsets than sunsets, an operating point with low recall must often be used to keep a small false positive rate. However, the primary benefit of the context model is that for any threshold, many false positives can be rejected, allowing us to bias the baseline detector in favor of sunsets in order to increase the recall while holding a constant false positive rate. Figure 7 shows examples of gains that the context model can make. The first sequence is very typical. Indoor scenes under low incandescent lighting can be mistaken for sunsets by the color-texture classifier. However, those occurring between images with different lighting can easily be corrected. The second sequence shows sunsets with weak colors that were initially missed by the color-texture classifier but were eventually recovered when the context model was used.

Interestingly, from the examples shown for both applications, the classifier's error characteristics are different than

Fig. 5 Image sequences affected by the context model. Elapsed times (in seconds) between images are shown. The first three sequences show examples in which the model corrected errors made by the baseline classifier. The fourth sequence shows a conceptual error: a rare case where the photographer walks into a room briefly between taking two outdoor photos. The last two sequences show examples where long elapsed time causes no change

those of the content-only classifier. Images classified in isolation (based on content) tend to have isolated errors, while the context model can cause errors to cluster. Arguably, in a semi-automatic system (e.g., desktop image albuming software), these errors would be more convenient for a human to correct.

7 Conclusion and future work

We have proposed a general probabilistic temporal context model for scene classification that exploits temporal coherence between neighboring images in the same collection, especially if captured within a short length of time. The

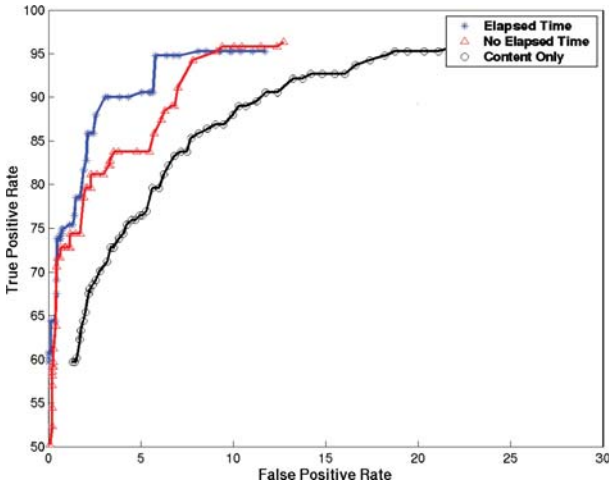

Fig. 6 Comparison of baseline content-based sunset detector performance with those improved by the temporal context models, with and without elapsed time. For any false positive rate, the recall of sunsets can be boosted by 2–10%. Alternately, for a given recall rate, the false positive rate can be reduced by as much as 20% in high recall operating points

loopless topology of the model allows both *exact* and *efficient* calculation of the MAP solution, adding little overhead to the baseline content-based classifiers. The context model is also flexible, allowing for both images with and without timestamps. We have demonstrated the efficacy of the model on a pair of two-class scene classification problems, correcting errors made by low-level content classifiers for indoor-outdoor and sunset detection.

Although using temporal correlation in general and HMM in particular for image and video clustering has been attempted with success in the literature [11, 16, 19, 27, 30], our main contribution is a unified, generalized framework suitable for exploiting weak and variable temporal context for (supervised) image classification and (unsupervised) image clustering. In addition, the temporal context and image content are integrated in a principled framework, which is also amenable to substituting an improved image content classifier in the future. Furthermore, the tightly coupled case of video can be considered as a special case of this model when the elapsed time is extremely short and constant between frames. The effectiveness of this generalized model is demonstrated by two applications, yet the main issues addressed in this paper are not the applications themselves.

In the future, we plan to extend our system to handle multi-class problems such as general scene classification (e.g., sports, party, etc.) once baseline classifiers are developed. Second, we have only exploited the first-order Markov property in this work. However, local interactions can lead to very long (incorrect) runs of images of the same class, even when evidence is propagated along the sequence. To address this, we could exploit the *second-order Markov property*, i.e., approximating $P(C_i|C_1, \ldots C_{i-1})$ with $P(C_i|C_{i-2}, C_{i-1})$. Preliminary results show that while the transition likelihood ratios for the marginalized first-order model are approximately 12:1, the ratios for the second-order model change to either 35:1 or 5:1, depending on whether the two previous classes agree or disagree. Alternately, we could investigate how to incorporate the *expected run length of* scene labels into the classifier. Finally, it may be beneficial to incorporate the absolute value of the timestamp (i.e., time of the day) for image classification (e.g., sunset detection), provided that the photographer set the time correctly.

Acknowledgements We thank Professors Dan Gildea of the University of Rochester and William Freeman of MIT for valuable discussions.

References

1. Assfalg, J., Bertini, M., Marco, C., Del Bimbo, A., Nunziati, W.: Semantic annotation of soccer videos: automatic highlights identification. Comput. Vision Image Understand. **92**(2), 285–305 (2003)

Fig. 7 Two sunset image sequences affected by the context model. In each case, the model corrects an error. The first sequence is typical: indoor images under low incandescent lighting can often be confused as sunsets, but are easy to correct by the temporal model. The second sequence shows a burst of sunset images in which two "weak" (cool-colored) sunsets are missed by the color-texture classifier, but corrected by the model

2. Besag, J.: On the statistical analysis of dirty pictures. J. Royal Stat. Soc. B **48**, 259–302 (1986)
3. Boutell, M., Luo, J., Gray, R.T.: Sunset scene classification using simulated image recomposition. In: Proceedings of IEEE International Conference on Multimedia and Expo (2003)
4. Chou, P.: The Theory and Practice of Bayesian Image Labeling. PhD thesis, University of Rochester, Rochester, NY (1988)
5. Dimitrova, N., Agnihotri, L., Wei, G.: Video classification based on HMM using text and faces. European Conference on Signal Processing, Finland (2000)
6. Duda, R., Hart, R., Stork, D.: Pattern Classification, 2nd Edn. Wiley, New York (2001)
7. Freeman, W.T., Pasztor, E.C., Carmichael, O.T.: Learning low-level vision. Int. J. Comput. Vision **40**(1), 24–57 (2000)
8. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Machine Intell. **6**(6), 721–741 (1984)
9. Glass, J.R., Hazen, T.J., Hetherington, I.L.: Real-Time Telephone-Based Speech Recognition In The Jupiter Domain. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Phoenix AZ (1999)
10. Huang, J., Liu, Z., Wang, Y., Chen, Y., Wong, E.K.: Integration of Multimodal Features for Video Classification Based on HMM, IEEE Workshop on Multimedia Signal Processing, Copenhagen, Denmark (1999)
11. Jaimes, A., Benitez, A.B., Chang, S.-F., Loui, A.C.: Discovering Recurrent Visual Semantics in Consumer Photographs. In: Proceedings of IEEE International Conference on Image Processing, Vancouver, Canada (2000)
12. Loui, A.C., Savakis, A.: Automatic event segmentation and quality screening for albuming applications. In: Proceedings of IEEE International Conference on Multimedia and Expo (2000)
13. Manning, C., Schutze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA (1999)
14. Moore, D., Essa, I., Hayes, M. III.: Exploiting human actions and object context for recognition tasks. In: Proceedings of International Conference on Computer Vision (1999)
15. Mulhem, P., Lim, J.-H.: Home photo retrieval: Time matters. Lect. Notes Comput. Sci. **2728**, 321–330 (2003)
16. Naphade, M.R., Huang, T.S.: A probabilistic framework for semantic video indexing, filtering, and retrieval. IEEE Trans. Multimedia **3**(1), 141–151 (2001)
17. Neal, R.M.: Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto (1993)
18. Pearl, J.: Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann Publishers (1988)
19. Platt, J.: AutoAlbum: Clustering digital photographs using probabilistic model merging. In: Proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries (2000)
20. Sebe, N., Lew, M., Zhou, X., Huang, T., Bakker, E.: The state of the art in image and video retrieval. In: Proceedings of International Conference on Image and Video Retrieval (2003)
21. Serrano, N., Savakis, A., Luo, J.: A computationally efficient approach to indoor/outdoor scene classification. Pattern Recogn. **37**(9), 1773–1784 (2004)
22. Singhal, A., Luo, J., Zhu, W.: Probabilistic spatial context models for scene content understanding. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (2003)
23. Snoek, C., Worring, M.: Multimodal video indexing: A review of the state-of-the-art. Technical Report 2001–20, Intelligent Sensory Information Systems Group, University of Amsterdam (2001)
24. Song, Y., Zhang, A.: Analyzing scenery images by monotonic tree. ACM Multimedia Syst. J. **8**(6), 495–511 (2003)
25. Szummer, M., Picard, R.: Indoor-outdoor image classification. In: Proceedings of IEEE International Workshop on Content-based Access of Image and Video Databases (1998)
26. Tax, D., Duin, R.: Using two-class classifiers for multi-class classification. In: Proceedings of International Conference on Pattern Recognition (2002)
27. Torralba, A., Murphy, K., Freeman, W., Rubin, M.: Context-based vision system for place and object recognition. In: Proceedings of International Conference on Computer Vision (2003)
28. Torralba, A., Sinha, P.: Statistical context priming for object detection. In: Proceedings of International Conference on Computer Vision (2001)
29. Vailaya, A., Figueiredo, M., Jain, A., Zhang, H.-J.: Content-based hierarchical classification of vacation images. In: Proceedings of International Conference on Multimedia Computing and Systems (1999)
30. Vasconcelos, N., Lippman, A.: Statistical models of video structure for content analysis and characterization. IEEE Trans. Image Process. **9**(1), 3–19 (2000)

Matthew Boutell received the BS degree in Mathematical Science from Worcester Polytechnic Institute, Massachusetts, in 1993, the MEd degree from University of Massachusetts at Amherst in 1994, and the PhD degree in Computer Science from the University of Rochester, Rochester, NY, in 2005. He served for several years as a mathematics and computer science instructor at Norton High School and Stonehill College and as a research intern/consultant at Eastman Kodak Company. Currently, he is Assistant Professor of Computer Science and Software Engineering at Rose-Hulman Institute of Technology in Terre Haute, Indiana. His research interests include image understanding, machine learning, and probabilistic modeling.

Jiebo Luo received his PhD degree in Electrical Engineering from the University of Rochester, Rochester, NY in 1995. He is a Senior Principal Scientist with the Kodak Research Laboratories.

He was a member of the Organizing Committee of the 2002 IEEE International Conference on Image Processing and 2006 IEEE International Conference on Multimedia and Expo, a guest editor for the Journal of Wireless Communications and Mobile Computing Special Issue on Multimedia Over Mobile IP and the Pattern Recognition journal Special Issue on Image Understanding for Digital Photos, and a Member of the Kodak Research Scientific Council.

He is on the editorial boards of the IEEE Transactions on Multimedia, Pattern Recognition, and Journal of Electronic Imaging. His research interests include image processing, pattern recognition, computer vision, medical imaging, and multimedia communication. He has authored over 100 technical papers and holds over 30 granted US patents. He is a Kodak Distinguished Inventor and a Senior Member of the IEEE.

Chris Brown (BA Oberlin 1967, PhD University of Chicago 1972) is Professor of Computer Science at the University of Rochester.

He has published in many areas of computer vision and robotics. He wrote COMPUTER VISION with his colleague Dana Ballard, and influential work on the "active vision" paradigm was reported in two special issues of the International Journal of Computer Vision. He edited the first two volumes of ADVANCES IN COMPUTER VISION for Erlbaum and (with D. Terzopoulos) REAL-TIME COMPUTER VISION, from Cambridge University Press. He is the co-editor of VIDERE, the first entirely online refereed computer vision journal (MIT Press).

His most recent PhD students have done research in infrared tracking and face recognition, features and strategies for image understanding, augmented reality, and three-dimensional reconstruction algorithms.

He supervised the undergraduate team that twice won the AAAI Host Robot competition (and came third in the Robot Rescue competition in 2003).