Event Recognition in Photo Collections with a Stopwatch HMM

Lukas Bossard[1] Matthieu Guillaumin[1] Luc Van Gool[1,2]

[1]Computer Vision Lab
ETH Zürich, Switzerland
lastname@vision.ee.ethz.ch

[2]ESAT, PSI-VISICS
K.U. Leuven, Belgium
vangool@esat.kuleuven.be

Abstract

The task of recognizing events in photo collections is central for automatically organizing images. It is also very challenging, because of the ambiguity of photos across different event classes and because many photos do not convey enough relevant information. Unfortunately, the field still lacks standard evaluation data sets to allow comparison of different approaches. In this paper, we introduce and release a novel data set of personal photo collections containing more than 61,000 images in 807 collections, annotated with 14 diverse social event classes.

Casting collections as sequential data, we build upon recent and state-of-the-art work in event recognition in videos to propose a latent sub-event approach for event recognition in photo collections. However, photos in collections are sparsely sampled over time and come in bursts from which transpires the importance of specific moments for the photographers. Thus, we adapt a discriminative hidden Markov model to allow the transitions between states to be a function of the time gap between consecutive images, which we coin as Stopwatch Hidden Markov model (SHMM).

In our experiments, we show that our proposed model outperforms approaches based only on feature pooling or a classical hidden Markov model. With an average accuracy of 56%, we also highlight the difficulty of the data set and the need for future advances in event recognition in photo collections.

1. Introduction

With the advent of digital photography, we have witnessed the explosion of personal and professional photo collections, both online and offline. The vast amount of pictures that users accumulate raises the need for automatic photo organization. This has initiated extensive research on content-based image retrieval systems such as image indexing based on objects [22], faces [11] or tags [1]. In addition to visual content, EXIF meta data [5], GPS tracks [29] and captions [3] provide excellent cues to reduce the complexity of these tasks. However, these works seldomly exploit the simple fact that online and offline images frequently come in collections: People organize their personal photos in directories, either corresponding to particular contents (persons, things of interest) or particular events. Online photo sharing websites such as Flickr, Panoramio or Facebook adopted this scheme and are organised in albums (examples shown in in Fig. 1). The benefits from recognizing event types are evident: Automatic organisation helps users keep order in their photo collections and also enables the retrieval of similar event types in large photo repositories.

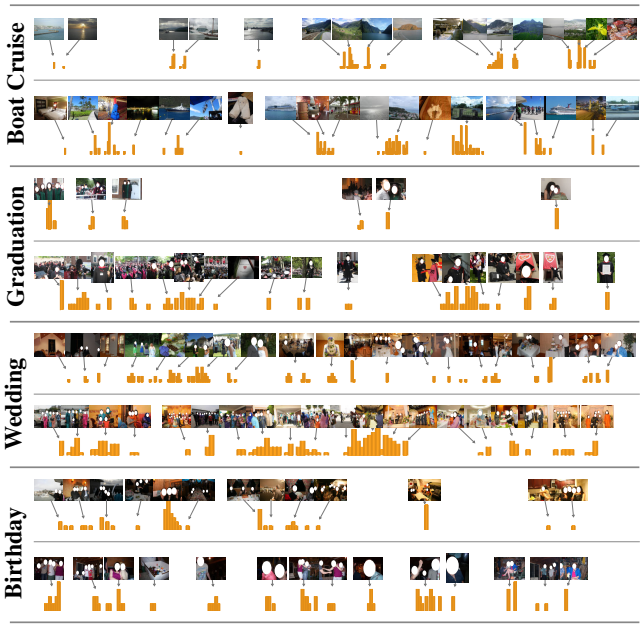


Figure 1: Eight examples of photo collections from four event classes in our data set. The difficulty in predicting the correct class comes from the sparse sampling of images in time, shown with the histograms (number of images within a small time frame), and from the high semantic ambiguity of images (*e.g.*, portraits appear in many event classes).

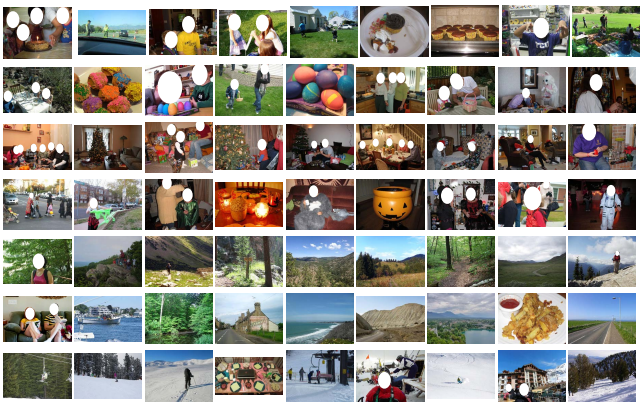


Figure 2: Unordered samples from our data set, where each row corresponds to a class. From top to bottom: *Children's birthday*, *Easter*, *Christmas*, *Halloween*, *Hiking*, *Road Trip*, *Skiing*. Note the high intra and sometimes low inter class variations.

Event and action classification has recently received a lot of attention in the computer vision community when it comes to video [10, 17, 25]. The availability of data sets [17] and difficult challenges [21] explain such profusion. As in videos, discriminative features in photo collection are often outnumbered by many diverse and semantically ambiguous frames that contribute little to the understanding of an event class: portraits, group photos and landscapes all occur in multiple types of events. In contrast to videos where images are sampled at a fixed frame rate, photo collections instead present a very sparse sampling of visual data, such that relating consecutive images is typically a harder task, *c.f*. Fig. 1. A great benefit of photo collections, however, is that the frequency of sampling is itself a measure of the relative importance of photos [8], and that we can exploit this information to distinguish between event classes.

Unfortunately, there are no standard benchmark data set for studying the challenging problem of event recognition for photo collections. In the literature on classifying photo collections [18, 26, 29], only small and private data sets are used. This eventually limits the possibilities to compare different approaches and research new ideas. As a contribution of this paper, we have collected a large data set of more than 61,000 images in 807 collections from Flickr and manually annotated it with 14 event classes as we describe in Sect. 3. These collections correspond to real-world personal photo collections taken by individual photographers. The diversity of depicted events is large: *Birthday party*, *Boat Cruise*, *Concert*, *etc*. as shown in Fig. 2. This data set is available for download with the intention to establish a solid benchmark. As a second contribution, we propose to modify a recent state-of-the-art model [25], initially designed for videos, for event recognition in photo collections. This includes a proper multi-class formulation and a modified hidden Markov model where the transition probabilities depend on observed temporal gaps between images. Hence, we coin this model a Stopwatch Hidden Markov model. We present it and show how to perform inference and learning in Sect. 4. Thirdly, we combine cues from multiple modalities to form image-level and collection-level features (Sect. 5). These cues include low-level visual channels and temporal frequency, as well as higher-level visual information such as scene and human attributes.

We show in our experiments (Sect. 6) that our model outperforms alternative event classification schemes for photo collections based on feature or score pooling or simple hidden Markov models and present our conclusions in Sect. 7. We first discuss related work in Sect. 2 below.

2. Related Work

Our work is related to a large literature on the automatic organization of images. For instance, from an unlabelled set of images, various image similarity measures have been proposed to clusters images based on the objects [27], people [11] or sub-sequences [8] they contain. While these algorithms focus on finding structure in unorganized data, our goal is to exploit the collection structure that is often found in personal and professional photo archives.

Cao *et al*. [6] exploit photo collections to reduce the complexity of propagating labels between images by observing that images within a collection are more likely to depict similar scenes. The authors use a data set of 100 collections and label each image with an event and a scene label. [7] further extends this idea towards a hierarchical model where a photo collection is split in a sub-sequence of so-called "events", composed of images from similar scenes, and exploits additional information such as GPS tracks. GPS tracks make it simpler to distinguish between events such as backyard parties, hikes and road trips [29] because of the difference of their geographical extent, but are still not very common in photo collections. [18] proposes a simple scheme to aggregate the SVM scores of each photo in a collection, and use it for classification into 8 social classes.

Event classification has also been considered for single static photos. For instance, the generative model in [16] allows its authors to integrate cues such as scene, object categories and people to segment and recover the event category in a single image. However, because of the ambiguity between events, a generative approach might not lead to optimal predictive performance. Experiments were performed on a small-scale data set of 8 sport activities with up to 250 images each. Many other works also integrate additional higher-level cues, most often for image classification in a wider sense. [19] exploits user context, location and user-

provided tags and comments on a photo sharing website to improve automatic image annotation.

The most related works to ours deal with event classification in videos [12, 25]. Both works consider the use of latent sub-events in a discriminative learning framework, to maximize predictive performance. However, [12] relies on known sub-events and uses them as an intermediate representation of collections for event classification. Time information is discarded in favor of co-occurrence of sub-events. Instead, we build upon the recent work of [25] and treat sub-events as unobserved latent variables. In [25], these sub-events are associated with explicit durations, and transitions between sub-events can only occur when the previous sub-event has expired. This requires that sub-events and the sub-event boundaries are fully observed. Because of the sparsely sampled photos in our collections, we need to adapt this model. Inspired by discretely observed Markov jump processes [4], we propose a Markov model where transition probabilities are functions of the temporal gap between images as if it were measured by a stopwatch (*c.f.* Sect. 4).

3. Data Set

In this section, we describe our efforts to collect and annotate a large data set of personal photo collections for use as an event recognition benchmark.

We first defined event classes of interest by using the most popular tags on Flickr and Picasa as well as Wikipedia categories that correspond to social events. Because we did not have direct access to large private photo collections we formulated different keyword queries by using variations of the event's name or by adding year numbers to retrieve single images from Flickr. If a returned image was contained in a Flickr set and if we could access the original image and its EXIF meta data, we downloaded the whole photo set. As these sets only loosely correspond to collections, we manually reviewed and discarded those sets that did not consist of a personal album or one single event, had wrong or missing meta data or were heavily retouched. About 60% of the downloaded photo sets had to be discarded.

This led to the choice of 14 event classes as shown in Tab. 1, with in total 807 photo collections which together contain 61,364 photos with EXIF data. We show examples of the resulting data set in Fig. 1 and 2. The data set is available at `http://www.vision.ee.ethz.ch/datasets/pec`.

Class	Collections	Photos	Class	Collections	Photos
Birthday	60	3,227	Graduation	51	2,532
Children Birthday	64	3,714	Halloween	40	2,403
Christmas	75	4,118	Hiking	49	2,812
Concert	43	2,565	Road Trip	55	10,469
Boat Cruise	45	4,983	St. Patrick's Day	55	5,082
Easter	84	3,962	Skiing	44	2,512
Exhibition	70	3,032	Wedding	69	9,953
			Total	807	61,364

Table 1: Statistics of our data set. For each of the 14 classes, we detail the number of photo collections and the total number of images that they contain.

4. The Stopwatch Hidden Markov Model

People usually do not take pictures at fixed intervals when photographing at an event they attend. More often, photos are taken when something interesting happens and thus show a bursty distribution when looking at the time domain. Secondly, events are often composed of different sub-events: At Easter, eggs are hunted and there is often also a joint meal. Weddings also often contain some sort of meal and afterwards people might be dancing. Other events might even expose a more subtle and thus latent substructure. In this work, we assume that the photo bursts act as a proxy for this sub-structure.

Since events of the same type show a very large variety in their temporal composition, it can be difficult even for humans to identify and thus annotate sub-events. This is why we treat the sub-events as latent in this work and learn them while training the event classifier.

Given a photo collection $\mathbf{X} = \{x_0, \dots, x_T\}$ of $T+1$ time ordered images originating from a single event, our goal is to predict the correct event class label y in a set $\mathcal{Y}$ of K possible labels.

We cast this prediction task in the framework of structured-output SVM with latent variables [20, 28], where the output is a multi-class prediction y^* parametrized by Θ:

$$y^* = f_\Theta(\mathbf{X}) = \operatorname*{argmax}_y \max_{\mathbf{Z}} \langle \Theta, \Phi(\mathbf{X}, y, \mathbf{Z}) \rangle \in \mathcal{Y} \quad (1)$$

and where the latent variables $\mathbf{Z} = \{z_0, \dots, z_T\}$ that are associated with the images form a chain.

In the next sections, we first describe our model in detail, explaining the factor graph of the prediction function (Sect. 4.1). Then we discuss in Sect. 4.2 how the solution of Eq. 1 can be efficiently inferred given known parameters Θ. In Sect. 4.3, we detail how we learn the parameters given a set of training photo collections with manual annotations.

4.1. Model

As visible in Fig. 1, events can be described as a series of smaller (visually diverse) sub-events. In this work we model these sub-events explicitly to improve the classification performance. Our model for photo collection classification is based on a hidden Markov model, as commonly done for modelling sequences [14, 24, 25]. Each observed image x_t in the collection is associated with an unobserved latent variable z_t representing its *state* among S possible ones. In the specific context of event recognition, those latent states are often called *sub-events*, to stress their intended semantics.

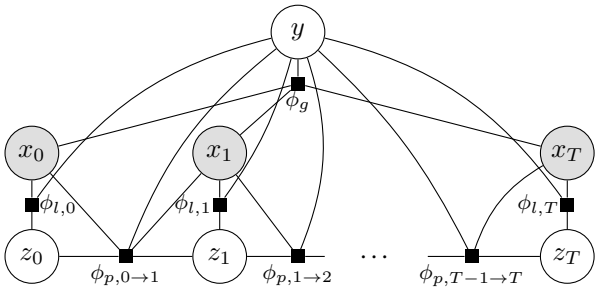


Figure 3: Factor graph corresponding to our photo collection event recognition model. Please refer to the text for notations and explanations.

In our model, the prediction function decomposes as:

$$\begin{aligned}\langle \Theta, \Phi(\mathbf{X}, y, \mathbf{Z}) \rangle &= \langle \Theta_g, \Phi_g(\mathbf{X}, y) \rangle \\ &+ \frac{1}{T+1} \sum_{t=0}^{T} \langle \Theta_l, \Phi_l(x_t, z_t, y) \rangle \\ &+ \frac{1}{T} \sum_{t=0}^{T-1} \theta_{p,z_t,z_{t+1},y} \cdot \phi_p(x_t, x_{t+1}, z_t, z_{t+1}, y) \end{aligned} \quad (2)$$

The feature map $\Phi_g(\mathbf{X}, y)$ allows the integration of *global* cues from the full sequence into the event prediction. The maps $\Phi_l(x_t, z_t, y)$ represent images x_t and their assignments to *latent* sub-events z_t for a particular event class y. Finally, the *pairwise* features $\phi_p(x_t, x_{t+1}, z_t, z_{t+1}, y)$ encode the sub-event transition costs between consecutive images. Denoting $\langle \Theta_g, \Phi_g(\mathbf{X}, y) \rangle$ by ϕ_g, $\langle \Theta_l, \Phi_l(x_t, z_t, y) \rangle$ by $\phi_{l,t}$ and $\theta_{p,z_t,z_{t+1},y} \cdot \phi_p(x_t, x_{t+1}, z_t, z_{t+1}, y)$ by $\phi_{p,t\to t+1}$, Fig. 3 shows the factor graph corresponding to a photo collection.

Unlike most previous modelling of sequential visual data, all these terms depend on the unobserved variable y. This allows to learn sub-events that help discriminate between events in a multi-class setting, whereas [25] only considers binary CRFs. In essence, our CRFs are calibrated to maximize multi-class prediction accuracy.

Note also how the pairwise terms depend on observed data x_t and x_{t+1}. Indeed, inspired by Markov Jump Processes [4], we use the observed time gap $\delta_{t\to t+1} = \tau(x_{t+1}) - \tau(x_t)$ between two consecutive images x_t and x_{t+1} to influence the transition probabilities.

Our Stopwatch Hidden Markov model can model the intuition that the transition matrices for short temporal gaps should typically be close to the identity matrix (*i.e.*, prefers not to change state) while transition matrices for longer temporal gaps should be more distributed as illustrated in Fig. 4.

The model seamlessly integrates the information of the temporal gap $\delta_{t\to t+1}$ between two consecutive images by making the energies for changing sub-event assignments dependent on the probability, that an observed $\delta_{t\to t+1}$ originates from this class. In this particular work, we used

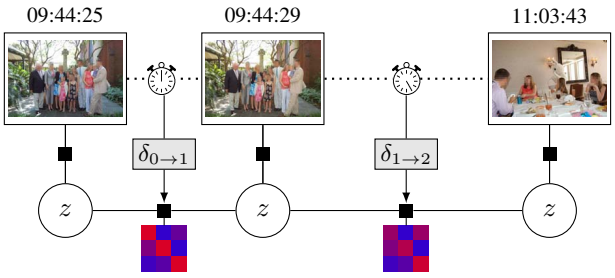


Figure 4: Illustration of our Stopwatch Hidden Markov model. The transition matrix between two consecutive images depends on the temporal gap $\delta_{t\to t+1}$. This allows to model bursts of photos and the typical durations of sub-events.

$$\phi_p(x_t, x_{t+1}, z_t, z_{t+1}, y) = -\log(p(\delta_{t\to t+1}|y))\ 1_{[z_t \neq z_{t+1}]} \quad (3)$$

where $p(\delta_{t\to t+1}|y)$ is estimated by Kernel Density Estimation using a Gaussian kernel. Intuitively, the model "trusts" a transition more, if the observed time-gap is consistent with time-gaps observed for class y.

Inference, *i.e.* estimating the event and sub-event label can be simply done as shown in Sect. 4.2, using the forward-backward algorithm. The learning of the parameters of the structured output SVM with latent variables [28] resembles the EM algorithm, alternating between assigning images to sub-events (using fixed parameters) and optimizing the parameters (under fixed assignments) as we describe in the subsequent Sect. 4.3.

4.2. Inference

Given a photo collection, inferring the event class label and the latent sub-events means to jointly maximize over the latent variables and the class labels as in Eq. 1.

This can be done efficiently by observing that, for a fixed event label y, the problem of inferring over the latent variables $\mathbf{Z}$, *i.e.* solving

$$\mathbf{Z}_y^* = \operatorname*{argmax}_{\mathbf{Z}} \langle \Theta, \Phi(\mathbf{X}, y, \mathbf{Z}) \rangle \,, \quad (4)$$

consists of inferring a chain model. We show such a chain in Fig. 5.

To perform inference in the full model, we therefore simply apply the Viterbi algorithm to infer the latent variables $\mathbf{Z}_y^*$ for each choice of event label y, and then maximize the corresponding prediction function over y:

$$y^* = \operatorname*{argmax}_{y} \left\langle \Theta, \Phi(\mathbf{X}, y, \mathbf{Z}_y^*) \right\rangle . \quad (5)$$

In essence, our model is therefore equivalent to having one chain model per event class, and predicting the class with highest confidence.

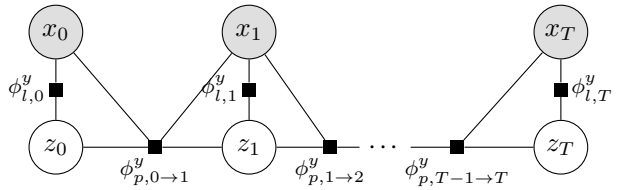


Figure 5: Factor graph corresponding to our photo collection classification model when the event label y is fixed. As the global term becomes constant, it is omitted. The superscripts are added to acknowledge the dependency on y. The Viterbi algorithm can be used directly to infer the latent sub-event variables.

The Viterbi algorithm has a complexity of $O(TS^2)$, therefore the complexity of inferring our full model is $O(KTS^2)$, *i.e.* linear in the number of event classes and size of the photo collection, but quadratic in the number of sub-events.

4.3. Training

In this section, we aim at learning the parameters Θ given a training set $\mathcal{D} = \{(\mathbf{X}_0, y_0), \ldots, (\mathbf{X}_N, y_N)\}$ of N photo collections $\mathbf{X}_i$ with their class labels $y_i \in \mathcal{Y}$. We adopt the Latent Structural SVM framework [28]. The objective function is:

$$\min_{\Theta} \frac{\lambda}{2} \|\Theta\|^2 + \sum_{i=0}^{N} \max_{\hat{y},\hat{\mathbf{Z}}} \left(\left\langle \Theta, \Phi(\mathbf{X}_i, \hat{y}, \hat{\mathbf{Z}}) \right\rangle + \Delta(y_i, \hat{y}) \right) - \sum_{i=0}^{N} \max_{\mathbf{Z}_i} \langle \Theta, \Phi(\mathbf{X}_i, y_i, \mathbf{Z}_i) \rangle , \quad (6)$$

where Δ is the 0/1 margin function that represents the misclassification cost ($\Delta(y, y') = 1_{[y \neq y']}$). Minimizing Eq. 6 consists of finding the best parameters Θ such that the correct class y_i is the minimizer of the margin-augmented prediction function. This is equivalent to wanting y_i to have the most confident score by a margin of 1.

Following [28], we apply the Concave-Convex Procedure (CCCP) [30]. It iterates between the following two optimization problems until convergence:

1. Infer the latent sub-event labels $\mathbf{Z}_i^*$ for the ground-truth labels y_i for fixed parameters Θ. This is precisely solving Eq. 4.
2. Solve the convex problem in Eq. 7 below which is Eq. 6 with fixed latent sub-events $\mathbf{Z}_i^*$.

$$\min_{\Theta} \frac{\lambda}{2} \|\Theta\|^2 + \sum_{i=0}^{N} \max_{\hat{y},\hat{\mathbf{Z}}} \left(\left\langle \Theta, \Phi(\mathbf{X}_i, \hat{y}, \hat{\mathbf{Z}}) \right\rangle + \Delta(y_i, \hat{y}) \right) - \sum_{i=0}^{N} \langle \Theta, \Phi(\mathbf{X}_i, y_i, \mathbf{Z}_i^*) \rangle \quad (7)$$

We optimize the convex objective in Eq. 7 using the Optimized Cutting Plane Algorithm as implemented in the Dlib C++ Library [13]. This implies performing margin-augmented inference which we do simply by adding $\Delta(y_i, y)$ to Eq. 5.

4.4. Initial Sub-events

In this section, we describe how we initialize the sub-event labels. We found this initialization to be much more robust than initializing CCCP with random sub-event assignment. The key is to take again advantage of the photo bursts in the time domain. In this way, we exploit the relative importance given to those photos by the photographer. Our assumption is again that such bursts act as a proxy to latent sub-events. To do so, we segment each photo collection using Hierarchical Agglomerative Clustering using a Gaussian kernel in the time domain $d(t_1, t_2) = \exp\left(-(t_1 - t_2)^2/(2\sigma^2)\right)$.

For each event class, the averaged visual features of each segment are then clustered using K-Means. We use the resulting S clusters as initial sub-event assignments.

5. Features and potentials

In this section, we provide specific details about the global feature vectors ϕ_g and the image-level ones ϕ_l that we use later in our experiments. Global features are functions of the whole photo collection and help capture holistic properties. Sub-event features help capture properties of single photos. Having access to EXIF-data, we also include non-visual cues into our model.

Global Temporal Features. We define different cues based on time and aggregate them over the photo collection in different histograms. Those cues include *time of day*, *day of week*, *month* and the *duration* to help recognize events that show specific patterns in the time domain.

Low-level Visual Features. As visual features, we use densely sampled SURF [2] descriptors and code them into a Bag-of-Words representation using a vocabulary of 1024 words, which is then max pooled. The vocabulary is previously learned using K-Means.

Higher-level Visual Features. To obtain a richer representation of images and improve the semantics of sub-events, we use a number of attribute predictions which have been shown to help classification (*e.g.*, [16]). These attributes consist of the *type of scene* and *type of indoor scene*, the *number of faces*, whether the image is a *portrait*, and a histogram of *facial attributes* over detected faces. To compute the attributes, we pre-trained a set of classifiers from external data. For scene and indoor attributes, a multi-class SVM was trained on the 15 Scenes [15] and the MIT-Indoor [23] data set, respectively. For facial detection and attributes, we use the code of [9] to predict age, gender and presence of sunglasses.

Method	Avg. Acc. [%]	Recall@2 [%]	F_1-Score
Aggregated SVM	41.43	63.57	38.87
Bag of Sub-events	51.43	70.00	50.63
HMM	53.57	68.57	54.61
SHMM (this paper)	55.71	72.86	56.16

Table 2: Different performance measures for the evaluated methods.

Reducing the dimensionality. The high-dimensional BoW vectors are not directly used in the feature map in Eq. 6, as this would make it prohibitively large. Instead we use multi-class SVMs to linearly project the BoW to a space of dimensionality equal to the number of sub-events. To further improve the robustness of these intermediate features, we add a *negative* sub-event containing random images from other classes while training the multi-class SVMs.

6. Experiments

In this section, we evaluate our approach on our novel data set. First we define the experimental protocol in Sect. 6.1. Then we explain in Sect. 6.2 the different baselines and variants of our approach that we compare. In Sect. 6.3, we report and analyze the results.

6.1. Protocol

We start by defining a training, a validation and a testing set. Out of the pool of 807 photo collections, we randomly selected 10 collections for each of the 14 classes as test set, which we use to report our evaluations. We also sampled 6 random collections per class to validate the hyper-parameter. All the remaining collections can be used for learning the parameters of the algorithms for event recognition. Each event class has at least 24 training collections.

We report different performance measures for the evaluated methods: Average accuracy, recall@K and the F_1-score to illustrate the precision/recall characteristics of the evaluated methods. Recall@K is the fraction of test data samples for which the correct class is among the top-K scores.

In the experiments that we report below, we have balanced our training data and used 24 random collections for each event class. For mining the initial sub-events, we set the smoothing window $\sigma = 90$ and clustered them into 5 sub-events for each class. In the subsequent iterations, sub-events that are not assigned to any image are removed. We used the validation data to set the number of outer iterations in our training procedure (*c.f.* Sect. 4.3).

6.2. Approaches for Event Recognition

We now describe the different baselines and variants that we have compared in our experiments.

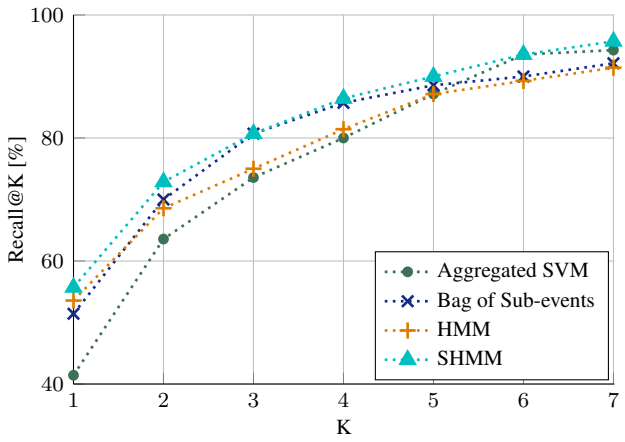


Figure 7: Recall of the different methods when looking at the top K classification scores for each collection.

Aggregated Multi-class SVM: Here, we employ an approach inspired by [18]. We train a multi-class SVM to recognize events in single images, based on visual features alone. At test time, each image is classified on its own and the confidence scores within each collection are averaged to predict the class of the collection.

Bag of Sub-events: In this variant, we adopt a *Bag of Sub-events* view and drop the pairwise connections of our model. This way, the model has no information about transitions and ordering of the photos in the collection. Instead, the latent sub-events are independently assigned to each image to maximize the prediction on the training set. We use the same procedure as described in Sect. 4.3 to learn Θ. In this model, inference becomes trivial.

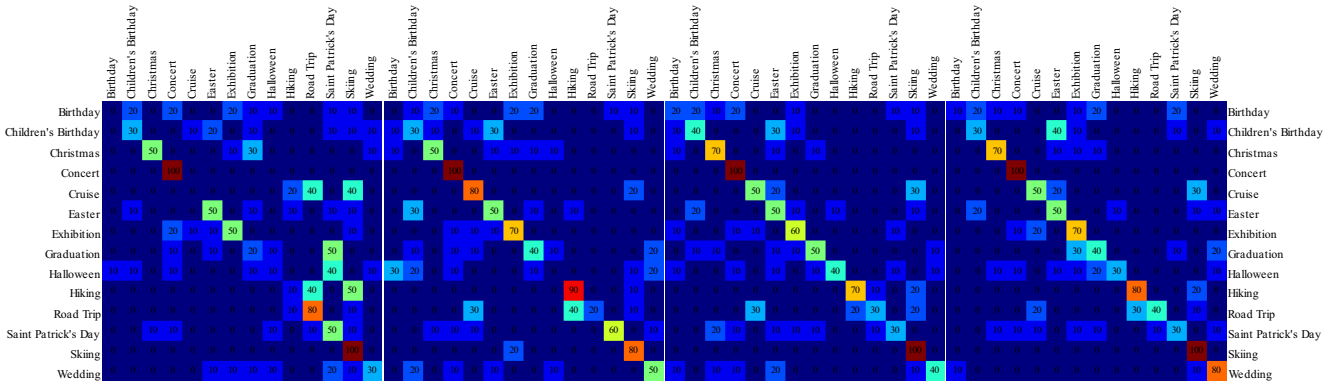
Hidden Markov Model (HMM): To obtain a discriminative HMM, we build on the previous approach and incorporate sub-event transitions. However, we adopt the classical definition of transition matrices in HMMs, which corresponds to $\phi_p(x_t, x_{t+1}, z_t, z_{t+1}, y) = 1$.

Stopwatch Hidden Markov Model (SHMM): We use our full model as described in Sect. 4, including the time-dependent sub-event transition features.

6.3. Results

In Tab. 2, we present the different performance measures obtained on the test set by our four approaches. The corresponding confusion matrices are shown in Fig. 6. As we see, the baseline method of Aggregated SVM scores reaches an average accuracy of 41.43%. Note how events taking place in different scene types can be discriminated properly, but events that have a similar scenery are confused (*e.g.* Hiking vs. Skiing, Fig. 6a).

Switching to the Bag of Sub-events model leads to a significant improvement: 51.43%. This demonstrates how the latent sub-event model can handle the variability within



(a) Aggregated SVM 41.43% (b) Bag of sub-events 51.43% (c) HMM 53.57% (d) SHMM 55.71%

Figure 6: Confusion matrices for the approaches we compare. For each confusion matrix we also show the average accuracy. Please refer to the text for explanations.

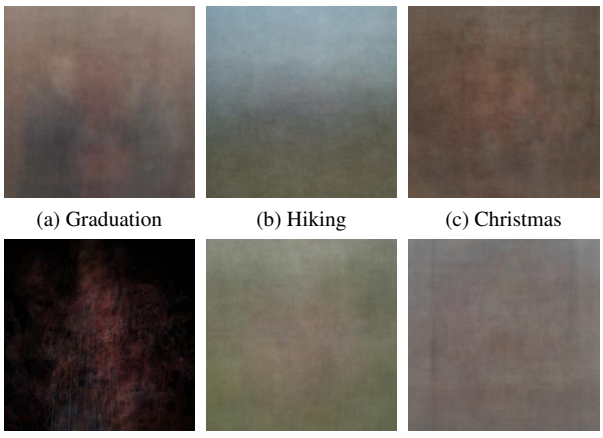
single event classes much better than a single SVM. This is also reflected in the F_1 scores, which increases in the same order. The HMM learnt in a discriminative fashion increases the average accuracy further to 53.57%. As transitions between sub-events can be captured by the model, dependencies between sub-events can be learnt, which helps to recognize classes that failed before (*e.g.* *Halloween* or *Road Trip*, see Fig. 6b and 6c).

Finally, our approach gives the best results both in terms of accuracy and F_1 score. With 55.71% average accuracy, it is 2.14% more accurate than HMMs, 4.28% better than Bags of sub-events, and the performance is as much as 14.28% higher than Aggregated SVM scores. In terms of F_1 scores, the improvements are 1.55%, 5.53% and 17.29%, respectively. Accordingly, we see in Fig. 6d how the confusion was reduced for most classes.

The performances of the different methods as measured by the recall@K are shown in Fig. 7. Our method consistently performs as well as, or outperforms, all the other considered approaches. For instance, the correct event is among the top two predictions for 72.86% of the collections.

Looking at the average of images assigned to sub-events by our SHMM shown in Fig. 8, we can sometimes clearly identify semantic concepts: outdoor view for the *Hiking* class, a typical photo setting for *Graduation*, painting frames for *Exhibitions*. This highlights the benefits of using a latent model for event recognition, as it can provide some additional semantic knowledge that eventually increases the ability to automatically understand, organize and exploit images in photo collections.

We also show in Fig. 9 some examples of photo collections that our approach correctly and incorrectly classified. As can be seen, visually and semantically very similar classes such as *Birthday*, *Children's Birthday*, *Graduation*, *Halloween etc*. are still confused to some extent. This highlights the difficulty of our new data set and the challenge it represents for the future.



(a) Graduation (b) Hiking (c) Christmas

(d) Concert (e) Easter (f) Exhibition

Figure 8: Average images corresponding to sub-events learnt by our model for different classes (best viewed in color on a computer screen).

7. Conclusion

In this paper, we have introduced a novel data set for event recognition in photo collections. We have proposed a model based on hidden Markov models that takes into account the time gap between images to estimate the probability to change state. Our model outperforms several approached based on previously published works. The final accuracy of 56% highlights the sheer difficulty of the data set, which we hope will foster research in this domain.

We believe that semantic hierarchies would help model events as well as complex sub-events, while scaling sublinearly with the number of event classes and sub-events.

Figure 9: Some Classification examples. On the right side, the predicted event class labels are shown and the color indicates if the SHMM correctly predicted it (correct labels shown in braces, only selected subset of images are shown).

Acknowledgments

The authors gratefully acknowledge support from ERC Advanced Grant 273940 VarCity.

References

[1] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *JMLR*, 2003.

[2] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. In *ICCV*, 2006.

[3] T. Berg, A. Berg, J. Edwards, M. Mair, R. White, Y. Teh, E. Learned-Miller, and D. Forsyth. Names and faces in the news. In *CVPR*, 2004.

[4] M. Bladt and M. Srensen. Statistical inference for discretely observed markov jump processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2005.

[5] M. Boutell. Bayesian fusion of camera metadata cues in semantic scene classification. In *CVPR*, 2004.

[6] L. Cao, J. Luo, and T. S. Huang. Annotating photo collections by label propagation according to multiple similarity cues. In *MM*, 2008.

[7] L. Cao, J. Luo, H. Kautz, and T. S. Huang. Image annotation within the context of personal photo collections using hierarchical event and scene models. *IEEE Transactions on Multimedia*, 11, 2009.

[8] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox. Temporal event clustering for digital photo collections. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 1, 2005.

[9] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *CVPR*, 2012.

[10] A. Gaidon, Z. Harchaoui, and C. Schmid. Recognizing activities with cluster-trees of tracklets. In *BMVC*, 2012.

[11] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Face recognition from caption-based supervision. *IJCV*, 2012.

[12] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model. In *ECCV*, 2012.

[13] D. E. King. Dlib-ml: A machine learning toolkit. *JMLR*, 10, 2009.

[14] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[16] L. Li-Jia and L. Fei-Fei. What, where and who? Classifying events by scene and object recognition. In *ICCV*, 2007.

[17] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009.

[18] R. Mattivi, J. Uijlings, F. G. De Natale, and N. Sebe. Exploitation of time constraints for (sub-)event recognition. In *J-MRE*, 2011.

[19] J. McAuley and J. Leskovec. Image labeling on a network: using social-network metadata for image classification. In *ECCV*, 2012.

[20] S. Nowozin and C. H. Lampert. Structured Learning and Prediction in Computer Vision. *Foundations and Trends in Computer Graphics and Vision*, 6(3-4):185–365, 2011.

[21] P. Over, G. M. Awad, J. Fiscus, B. Antonishek, M. Michel, A. F. Smeaton, W. Kraaij, and G. Quénot. Trecvid 2011–an overview of the goals, tasks, data, evaluation mechanisms, and metrics. 2011.

[22] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[23] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

[24] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *PAMI*, 29, 2007.

[25] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *CVPR*, 2012.

[26] S.-F. Tsai, L. Cao, F. Tang, and T. S. Huang. Compositional object pattern: a new model for album event recognition. In *MM*, 2011.

[27] T. Tuytelaars, C. Lampert, M. Blaschko, and W. Buntine. Unsupervised object discovery: A comparison. *IJCV*, 2009.

[28] C.-N. J. Yu and T. Joachims. Learning structural SVMs with latent variables. In *ICML*, 2009.

[29] J. Yuan, J. Luo, H. Kautz, and Y. Wu. Mining GPS traces and visual words for event classification. In *MIR*, 2008.

[30] A. L. Yuille and A. Rangarajan. The Concave-Convex Procedure. *Neural Computation*, 15, 2003.