

In the sight of my wearable camera: Classifying my visual experience

Alessandro Perina
Microsoft Research
One Microsoft Way, Redmond
alperina@microsoft.com

Nebojsa Jojic
Microsoft Research
One Microsoft Way, Redmond
jojic@microsoft.com

1. Introduction

We introduce and we analyze a new dataset which resembles the input to biological vision systems much more than most previously published ones. Our analysis led to several important conclusions. First, it is possible to disambiguate over dozens of visual scenes (locations) encountered over the course of several weeks of a human life with accuracy of over 80%, and this opens up possibility for numerous novel vision applications, from early detection of dementia to everyday use of wearable camera streams for automatic reminders, and visual stream exchange. Second, our experimental results indicate that, generative models such as Latent Dirichlet Allocation [4] or Counting Grids [1], are more suitable to such types of data, as they are more robust to overtraining and comfortable with images at low resolution, blurred and characterized by relatively random clutter and a mix of objects.

2. Data Acquisition

To gather the data, a subject wore a SenseCam¹ during all waking hours for three weeks. The camera was rarely turned off, except during potentially sensitive moments. The SenseCam snapshots² are automatically triggered by sudden changes in the visual field, or by default every 45s. On average the snapshots were taken every 20s or so. This translated into $\sim 2k$ images a day with a resolution of 640×480 , for a total of 43522 images.

We selected a random 10% of the data, where we found that the recurrent types of scenes fell into 32 classes. About 15% of images in this random selection belong to spurious types of scenes with only one or two examples. With the help of the original subject, we started from a few examples of each of the 32 recurring scenes and then we manually labeled the rest of the selected frames. In case of some classes, this procedure did not yield enough images for proper training and testing, and for these classes

we looked at the whole dataset again and extracted more examples of each of these classes for both testing and training. This process yielded to a total of 3959 labeled images. Some images for each class are shown in Fig.1.

During acquisition, each image is time-tagged, enabling us to illustrate in Fig.2a the number of different days that each class was seen, while in Fig.2b we show the distribution over time of day (morning/afternoon/evening/night) for each class.

We also labeled the images coming from two whole days to test if the timestamp information can help recognition. The number of images of each day is, respectively 2043 and 1703.

To download the dataset, please send an email to the authors³.

3. Experiments

In the following experiments, we used SIFT features [6], extracted from 16×16 patches spaced 8 pixels apart, clustered in $Z=200$ visual words or gist descriptors [8], extracted on 4 scales with 8 orientations per scale. We compared the performances of generative, discriminative and epitome-like methods. As discriminative methods we employed a Support Vector Machine with linear kernel on gist [8] and on quantized sift histograms and the Spatial Pyramid Kernel (SPK, 3 Levels) [5]. As generative models we considered Fei-Fei's semi-supervised Latent Dirichlet Allocation (LDA) [4], Counting Grids (CG) [1], and a mixture of Dirichlet distributions over the quantized sift histograms. The last approach is similar in spirit to [7] where a mixture of gaussians over gist descriptors is learnt for each class. Finally we also tried epitome-like approaches: Structural Epitome [3], Epitomic Location Recognition⁴ [9] and FFT's Counting Grids [2] (an hybrid between epitomes and Counting Grids).

¹<http://viconrevue.com/>

²<http://profs.sci.univr.it/~perina/sensecam.htm>

³http://profs.sci.univr.it/~perina/download/SenseCam_ECCV.zip

⁴We only used SIFT



Figure 1. From top left to bottom right, we show three images from each class of *SenseCam-32*. The 32 classes are: *Bathroom Home, Bedroom, Biking, Cafeteria, Car, Classroom, Conference Room, Corridors Work, Dining Room, Bakery, Garage, Atrium, Entry, Hiking Trail, Ice Palace, Kids Bedroom, Game Room, Kitchen, Living Room, Lounge, Home Office, Campus, Parking Work, Patio, Playground, Restroom Work, Small Bathroom Home, Small Home Office, Tennis Court, Food Court, Grocery Store, Work Office*.

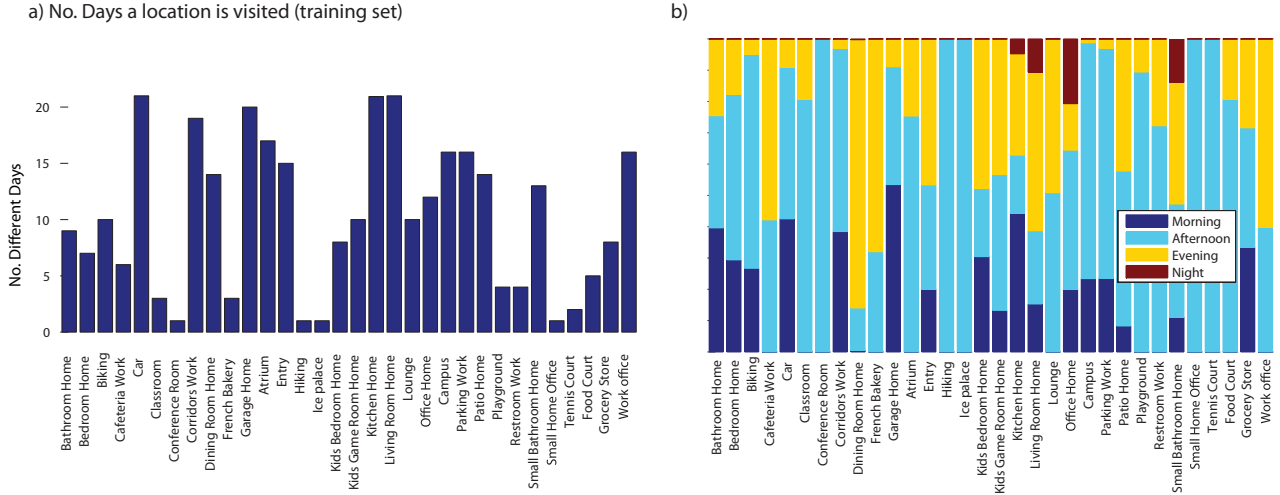


Figure 2. Statistics on SenseCam Dataset

3.1. Scene Classification

In this section we provide some baseline on the scene/place classification task. Since some categories have much more images than others, we used 15 images per class for training and at most 15 images of each class as test set. We repeated the experiments for 5 times, averaging the results. Classification accuracies are shown in Tab. 1, where we report the best result obtained by each method.

Generative classifiers are built learning a model for each class with the training data, and assigning to a test sample the class that produces higher likelihood. As visible in Tab.1 they reach the best results, moreover their performance doesn't vary much for a "reasonable" choice of the parameter setting (Topic number of LDA, Capacity for

Counting Grids). The poor accuracies of discriminative methods [5, 8], are clearly do to overtraining but with this type of datasets we must expect scarce labeled data. Interestingly, the methods based on pixelwise comparisons fail on this extended dataset as they cannot capture well the geometric transformations. Being a hybrid between epitomes and counting grids, [2] reaches decent results, but it also finds it hard to generalize with so little labeled data.

3.2. Where was I?

As second test we asked how many places we could correctly guess using *i)* each of the generative models considered in the previous section, and *ii)* an HMM over possible classes c_t , capturing transition constraints such as living

Table 1. **Scene Classification Results.** We reported the best result obtained.

<i>Discriminative Methods</i>		
Gist + SVM [8]	SIFT + SVM	SPK [5]
38,10%	49,94%	52,76%
<i>Generative Methods</i>		
LDA [4]	Dirichlet Mixt. [7]	CGs [1]
58,05%	49,19%	54,43%
<i>Epitome-like Methods</i>		
Stel Ep. [3]	CG-fft [2]	Feature Ep. [9]
15,02%	39,40%	21,32%

room - kitchen, or office - corridor - atrium, etc.

We considered labeled images from two days. During each day, the camera bearer visited roughly 20 of the labeled locations and the two days share only 12 locations. Nevertheless we trained models with all the 32 classes as a-priori we cannot know the locations visited during a day. Our goal is to compute the place posterior probabilities at the instant t , given all the previous images $P(c_t = k|x_{1:t})$. We used the forward-backwards procedure to recursively compute it, in formulae:

$$P(c_t = k|x_{1:t}) \propto p(x_t|c_t = k) \cdot P(c_t = k|x_{1:t-1}) =$$

$$p(x_t|c_t = k) \cdot \sum_c P(c_t = k|c_{t-1} = c) \cdot P(c_{t-1}|x_{1:t-1})$$

We fixed HMM's observation loglikelihood (e.g., $p(x_t|c_t = k)$) to the negative free energy of the generative model in hand, while we used EM estimate the transition matrix $A_{k|c} = P(c_t = k|c_{t-1} = c)$, the place prior $\pi_k = P(c_1 = k)$ and the place posteriors an unsupervised way, simply fitting the likelihood to the day's images. We used a strong dirichlet prior over self transitions to favor the stay in the same state/location. Finally since the observation likelihood terms, often dominate the effects of the transition prior, we adopt the standard solution of re-scaling the likelihood terms.

This approach is very similar to [7], therefore the reference provides a natural point of a comparison. Besides a similar use of the HMM the idea of [7] is to learn a mixture of model for each class to eventually compute the observation likelihood.

We used at most 30 images per class to learn the models. Results are reported in Tab.2: as expected, the recognition accuracy rises respectively when we "turn on" the HMM (Eq. 1). For sake of completeness we have also implemented the original method of [7] using their descriptors from the whole images and within the four sectors. Their performances were lower (below 50%).

Table 2. **Where was I?**
"HMM off"

LDA [4]	Dirichlet Mixt. \sim [7]	CGs [1]
62.21%	54.68%	66.76%
<i>"HMM on"</i>		
LDA [4]	Dirichlet Mixt. \sim [7]	CGs [1]
76.80%	70.37%	81.21%

4. Conclusion

In this extended abstract we presented a large dataset which differently from others is totally natural as it represents the visual input of a subject. Using our labels, it would be easier to analyze the full data collected in [3]. We also showed how temporal information can be used to reach compelling accuracies on classification of all the locations a subject visits during a day.

References

- [1] N.Jojic, A.Perina Multidimensional counting grids: Inferring word order from disordered bags of words UAI 2011: 547-556
- [2] A.Perina, N.Jojic Image analysis by counting on a grid CVPR 2011: 1985-1992
- [3] N.Jojic, A.Perina, V.Murino Structural Epitome: a way to summarize one's visual input NIPS 2010: 1027-1035
- [4] Fei-Fei Li, P.Perona A Bayesian Hierarchical Model for Learning Natural Scene Categories CVPR (2) 2005: 524-531
- [5] S. Lazebnik, C.Schmid, J.Ponce Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories CVPR (2) 2006: 2169-2178
- [6] D.Lowe Distinctive Image Features from Scale-Invariant Keypoints *Int. J. of Computer Vision*, 60, 2004.
- [7] A.Torralba, K.P.Murphy, W.T.Freeman and M.A.Rubin Context-based vision system for place and object recognition ICCV 2003: 273-280
- [8] A.Oliva and A.Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. of Computer Vision*, 42, 2001
- [9] K.Ni, A.Kannan, A.Criminisi and J.Winn Epitomic Location Recognition *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(12), 2009