Location recognition on lifelog images via a discriminative combination of generative models

Alessandro Perina
alessandro.perina@iit.it
Matteo Zanotto
matteo.zanotto@iit.it
Baochang Zhang
baochang.zhang@iit.it
Vittorio Murino
vittorio.murino@iit.it

Pattern Analysis and Computer Vision (PAVIS)
Istituto Italiano di Tecnologia
Genova, Italy

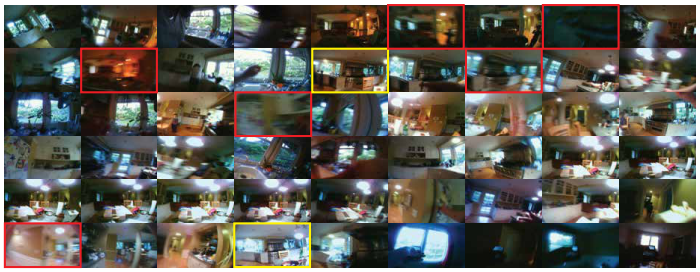

Figure 1: The first 54 images of a lifelog. Notice the high blur (red boxed images) and the dramatic changes in illumination (yellow box)

It is a common belief that in the near future, wearable technology will be the next computing revolution. Such wearable systems are intended to be used in a seamless way like a piece of clothing and they are at the basis of "lifelogging". Among all wearable sensors, the first lifelogging cameras are recently becoming available for a large number of people to use: all of them use a passive record-it-all approach, automatically shooting a photo every 10-30 seconds. However, the soon-to-be enormous amount of images must be organized in order to be useful, and simply using temporal arrangement of the shots is totally unsatisfactory. This paper represents a first step towards this goal: we focused on location recognition and we propose the use of a combination of heterogeneous generative models, each one able to capture the different aspects that characterize each location. Our approach of combining evidence outperforms each individual model as well as other advanced techniques.

Challenges. Lifelog images represent a serious challenge for computer vision researchers. Cameras are usually worn around the neck or attached to clothes and this causes non-linear and unpredictable motion which causes blur and rapid changes in the scene. Figure 1 shows 54 consecutive images spanning a period of ∼15 minutes over which the bearer changes location few times (kitchen, living room, garage). Notice how most of the frames are blurred, while few are highly blurred and difficult to understand even for a human. Moreover, the illumination exhibits dramatic changes over short time periods even when the bearer stays in the same location. Another intrinsic characteristics of lifelogs is that, in a real scenario, the labeled data available to accomplish a classification task are inherently scarce: most of the images, in fact, can only be labeled by the bearer of the camera and crowd-sourcing is difficult, if not impossible.

Motivations. This paper focuses on location recognition. It exploits several recent and classical generative models used for scene understanding to propose a framework able to learn a discriminative combination of weights dealing with the several complexities of multiple heterogeneous models for each location. This choice is motivated by an intuitive and a theoretical reason:

1. The locations one visits are so different that it does not exists a single model able to fit well everywhere. Our favorite grocery store, could nicely be modeled by a full bag-of-word approach like LDA, whereas locations like kitchen or living room are probably well recognized by looking at the objects that contain, and finally contained environments like our work cubicle or our car may well be modeled by an exemplar based-method or by a panoramic reconstruction method like the epitome.
2. When none of the models in an ensemble is the true data generator (TDG) model, there usually exists a combination that can replicate the behavior of the TDG more closely than any individual model on its own.

Overview of the proposed approach. Instead of searching for the best model, or for a combination that can more closely replicate the true data generator model behavior than any individual model on its own, we looked for a discriminative combination of weights. Furthermore, we computed it per-class as, in general, different combinations of models could be better suited for different classes.
Working in a one-vs-all setting, for each class l, we propose to compute the weights π^l which maximize the margin between the average conditional ensemble log-likelihood ratio **A-CLLR** of positive samples and that of negative samples (e.g., belonging to all the other classes). The average conditional log-likelihood of a set of bags of features $\mathbf{c}^t$, is defined as follows

$$A-CLL = \frac{1}{T}\sum_{t=1}^{T} \log p(l^t = l|\mathbf{c}^t) \tag{1}$$

where t indexes a sample, and l^t its class. The likelihood of the ensemble $\mathcal{E}$ is the likelihood of a mixture model whose components are the K individual models $\mathcal{M}_k$ themselves

$$p(l^t = l|\mathbf{c}^t, \mathcal{E}) \propto \sum_k \pi_k^l \cdot p(\mathbf{c}^t|\mathcal{M}_k^l) \tag{2}$$

Our technique allows to exploit all the data in both the generative and discriminative steps. This is crucial as lifelogs cannot have a lot of training data and standard methods could overtrain.

Results. We considered the SenseCam-32 dataset, a portion of lifelog where the dataset authors highlighted 32 recurrent classes visited by the camera bearer over a period of 21 days. We compared our approach with generative combination methods like Bayesian model averaging, discriminative fusion methods and kernel methods built from the log-likelihood of the individual models.
A snapshot of the results is reported in Fig.2. As visible, our combination method always outperforms each individual model in the ensemble, even with a very limited number of training images.
Further results are reported in the paper, where we also exploited the weak temporal relationships between lifelog images and tested the framework on the 67-indoor scene dataset.

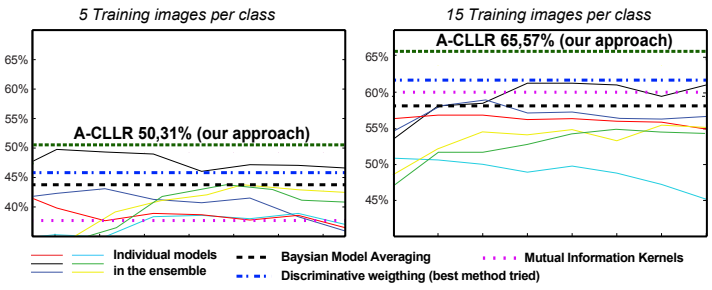

Figure 2: Model combination results on the SenseCam-32 dataset. On the x-axis the K complexities of each model $\mathcal{M}$; on the y-axis the classification accuracy over the 32 classes. See the paper for details.