Event Fisher Vectors: Robust Encoding Visual Diversity of Visual Streams

Markus Nagel[1]
mail@markusnagel.com
Thomas Mensink[1]
thomas.mensink@uva.nl
Cees G.M. Snoek[1,2]
cgmsnoek@uva.nl

[1] Intelligent Systems Lab Amsterdam
University of Amsterdam
[2] Qualcom Research Netherlands
Amsterdam

The goal of this paper is to design an effective representation for event recognition in visual streams, such as photo collections and video clips. We are inspired by the success of Fisher Vectors for the encoding of images and videos, where a Gaussian mixture model (GMM) is used as generative probability density function to model 10K-100K local observations per image or video. A visual stream, however, behaves significantly different than local patches or trajectories. Most notably, streams may consist of just tens to hundreds of images, each image in a stream can be described by more discriminative DeepNets, and the temporal structure of the stream can be modelled explicitly.

In order to be more robust against outliers in the small set of observations, we replace the GMM with a Student's-t mixture model (StMM), known for its heavier tails. For observation $x_i \in \mathbb{R}^D$, the StMM is defined as:

$$p(x_i|\theta) = \sum_k \pi_k \, St(x_i|\theta_k), \tag{1}$$

where π_k is the mixing weight and the Student's-t is defined using parameters $\theta_k = \{\text{mean } \mu_k, \text{var } \sigma_k, \text{dof } \nu_k\}$ as:

$$St(x_i|\theta_k) = Z_k \left(1 + \tfrac{1}{\nu_k}\delta_k(x_i)\right)^{-\frac{\nu_k+D}{2}}, \tag{2}$$

with Mahalanobis distance $\delta_k(x_i)$ and normalisation factor Z_k. The *Fisher score*, *i.e.* the derivative w.r.t. the parameters of $\log p(\cdot;\theta)$ is than:

$$G^X_{\mu_k} = \frac{\nu_k + D}{\sigma_k^2 \, \nu_k} \sum_{i=1}^{n} \gamma_i(k) \, \frac{(x_t - \mu_k)}{1 + \frac{1}{\nu_k}\delta_k(x_i)}, \tag{3}$$

where $\gamma_i(k)$ is the *responsibility* value of the k-th mixture component. This score function is, similar to the GMM case, a weighted average of the observations. However, two differences are: (i) the responsibility values $\gamma_i(k)$ are now based on the StMM, and (ii) each observation is weighed by the degrees-of-freedom ν_k and the Mahalanobis distance $\delta_k(x_i)$.

For image classification and retrieval, these Fisher scores are transformed with the *Fisher Information Matrix* (FIM), $F_\theta = \mathbb{E}_{X\sim\theta}[G^X_\theta G^{X\top}_\theta]$ to ensure invariance w.r.t. re-parameterizations of the model. Perronnin *et al.* derived an analytical approximation of the FIM for the GMM, by using the following two assumptions:

Hard-assignment assumption. All patches are sharply peaked around a single component k (*i.e.* $\forall i \, \exists k \; \gamma_k(i) \approx 1$).

Diagonal covariance assumption. Using diagonal covariance matrices in the GMM yields the Fisher scores independent per dimension.

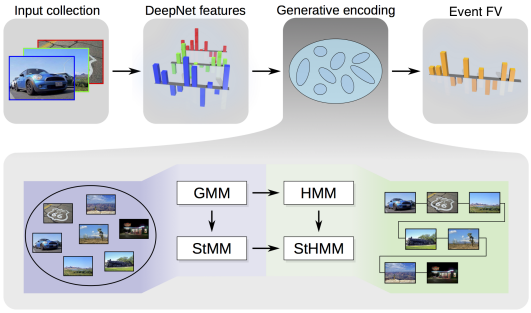
In contrast, the dimensions of the Fisher scores of the StMM are inter-dependent due to the Mahalanobis distance $\delta_k(x_i)$. In order to derive an analytical approximation, we propose the following assumption:

Constant distance assumption. We assume that in expectation, the Mahalanobis distance $\delta_k(x_i)$ becomes a constant factor. This assumption is based on the *concentration of distances* theorem, which states that for high dimensional data the proportional distance difference between any point and the mean of all data points vanishes. Intuitively, this theorem states that the distance differences $\delta_k(x_i)$, for $k = \{1,\ldots,K\}$ for a specific data point x_i are immaterial.

This results in the following approximation for the mean component:

$$F^{-\frac{1}{2}}_{\mu_k} = \sigma_k \left(\pi_k \, \tfrac{\nu_k}{\nu_k + D}\right)^{-\frac{1}{2}}. \tag{4}$$

To the best of our knowledge, this is the first closed-form approximation of the Fisher information matrix for the Student's-t mixture model. Note that the assumptions above are only used to derive the analytical approximation of the FIM, not for computation of the Fisher scores.



To model the temporal structure of a visual stream, we also propose to use Hidden Markov Models (HMM). The temporal relation in the HMMs is modeled by the latent state z_i, which depends not only on the observation x_i, but also on the latent variable z_{i-1} of the previous observation. The Fisher score of an HMM model w.r.t. the mean μ_k is given by:

$$G^X_{\mu_k} = \sum_{i=1}^{n} \gamma_i(k) \, \nabla_{\mu_k} \log p(x_i|\theta_k), \tag{5}$$

this is identical to the independent models, except that the responsibility values $\gamma_i(k)$ are now computed by:

$$\gamma_i(k) = \frac{p(x_1,\ldots,x_i,z_i{=}k) \; p(x_{i+1},\ldots,x_n|z_i{=}k)}{p(X)}, \tag{6}$$

which reflects the dependence among the images in the collection. For the analytical approximation of the FIM we use once more the crude hard-assignment assumption and obtain the same analytical FIM approximations as for the independent GMM and StMM models.

Different FIM approximations on MED13

FIM	GMM	StMM	HMM	StHMM
Identity	27.6	27.2	28.2	28.5
Empirical	34.7	34.6	34.5	33.8
Closed form	**36.8**	**37.2**	**36.6**	**36.8**

Different models (incl. oracle best per event)

	GMM	StMM	HMM	StHMM	BPE
PEC	80.7	**85.7**	83.6	80.7	88.6
MED13	36.8	**37.2**	36.6	36.8	38.0
CCV	67.4	**69.0**	66.2	66.5	69.1

We evaluated our Event-FV on three recent datasets of photo and video events: Photo Event Collection (PEC, event classification from Flickr collections), TrecVID Media Event Detection (MED13 benchmark with 100 examples per event for training), and Columbia Consumer Video (CCV, Youtube videos of events). For each image in a collection (or sampled frame from a video) we extract the final fully connected layer (4K) from a pre-trained DeepNet (trained on 15K ImageNet classes).

The conducted experiments show that the analytical approximations of the FIM outperform an identity or empirical approximation by a large margin, that the StMM model has a slight edge over the other probabilistic models, and that it results in state-of-the-art performance (when combined with the Mean DeepNet feature and Temporal Pyramids, see paper). This indicates that it is worth to explore more appropriate probabilistic models for the input data within the Fisher vector framework and to derive their analytical FIM approximations. For the task of visual event classification the conclusion is that capturing the heavy tails of the small sample size is more beneficial than modelling the temporal relation of the stream.