Event Fisher Vectors: Robust Encoding Visual Diversity of Visual Streams

Nagel, M.; Mensink, T.; Snoek, C.G.M.

DOI
10.5244/C.29.178

Publication date
2015

Document Version
Final published version

Published in
Proceedings of the British Machine Vision Conference 2015: BMVC 2015: 7-10 September, Swansea, UK

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

Abstract

In this paper we focus on event recognition in visual image streams. More specifically, we aim to construct a compact representation which encodes the diversity of the visual stream from just a few observations. For this purpose, we introduce the Event Fisher Vector, a Fisher Kernel based representation to describe a collection of images or the sequential frames of a video. We explore different generative models beyond the Gaussian mixture model as underlying probability distribution. First, the Student's-*t* mixture model which captures the heavy tails of the small sample size of a collection of images. Second, Hidden Markov Models to explicitly capture the temporal ordering of the observations in a stream. For all our models we derive analytical approximations of the Fisher information matrix, which significantly improves recognition performance. We extensively evaluate the properties of our proposed method on three recent datasets for event recognition in photo collections and web videos, leading to an efficient compact image representation which achieves state-of-the-art performance on all these datasets.

1 Introduction

The goal of this paper is to design an effective representation for event recognition in visual streams, such as photo collections [2, 7] and video clips [9, 19]. This is challenging, since a compact (vectorial) representation is preferred for efficient recognition, while this representation should still capture the visual semantics and temporal diversity of the stream from only a few samples.

We are inspired by the success of Fisher Vectors [24] for the encoding of images [24, 25] and videos [21, 27]. In the Fisher Vector, local patches from a single image or trajectories from a video are encoded using the Fisher Kernel [8] with a Gaussian mixture model (GMM) as underlying generative probability function. A stream of visual imagery, however, behaves significantly different than local patches or trajectories. Most notably, streams may consist of just tens to hundreds of images, while dense sampling methods for patches and trajectories extract 10K-100K local observations per image or video. Moreover, in contrast to low-level local descriptors, *e.g.* SIFT [14] or MBH [27], an image in a stream can be described by

Pages 178.1-178.12
DOI: https://dx.doi.org/10.5244/C.29.178

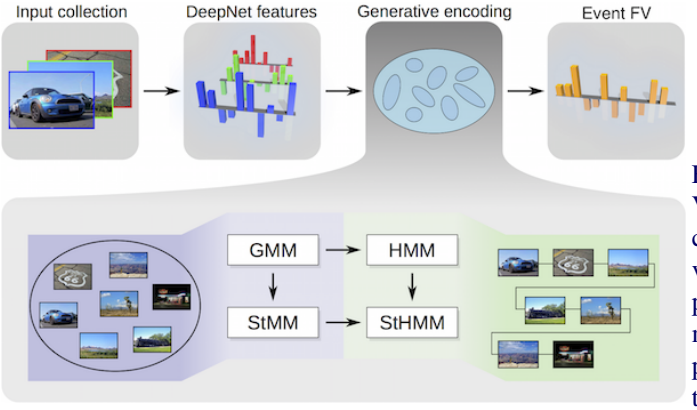


Figure 1: Event Fisher Vector pipeline for a collection of images or video-stream. We explore four generative models as underlying probabilistic distribution of the stream.

more discriminative features, *e.g.* based on pre-trained DeepNets [10, 30]. Finally, the temporal structure of an image collection might be less well defined than the explicit sequential ordering of frames of a video.

Our main contribution is a Fisher Kernel encoding for a visual stream, either videos or collections of still images. Our encoding extracts a single representation per collection, is independent of the number of images in the stream, and is agnostic to the underlying input features. This is advantageous, since it allows efficient learning of event classifiers, and leveraging of discriminative DeepNet features [10, 30]. We coin our encoding the Event Fisher Vector, and illustrate its pipeline in Figure 1.

As our second contribution, we propose alternatives beyond the GMM as the underlying distribution of a visual stream: *i)* We replace the GMM by a Student's-*t* mixture model (S*t*MM), which is more robust for our small sample size. *ii)* We explicitly encode the sequential ordering of a stream using Hidden Markov Models, with both the GMM and the S*t*MM as the emission probability function. As a third contribution, we derive analytical approximations of the Fisher information matrix for all of these models.

The remainder of this paper is organised as follows. We first summarise related work on event recognition in image and video streams in Section 2. We present our Event Fisher Vector and the proposed generative models in Section 3. We evaluate extensively the properties the Event Fisher Vector and its generative models on three challenging event recognition datasets in Section 4 and demonstrate a new state-of-the-art in all cases.

2 Related Work

In this section we discuss some of the most relevant work on representing photo collections, event recognition in videos, and the Fisher Vector representation.

Image Collections Recognition of events in streams of images is commonly achieved by a representation consisting of simple averaging of images features [7] or similarly by a majority vote of single image classification scores [16]. We deem it unlikely that a simple average can capture the variability of the visual semantics of an image stream. A notable exception is the Stopwatch Hidden Markov Model approach by Bossard *et al.* [2], proposed in conjunction with the challenging Personal Event Collection (PEC) benchmark. They propose a

discriminative hidden Markov model, that models the transitions between states as a function of the time gap between consecutive images in a collection of personal photos. This allows to model the sequential nature of the image stream, an advantageous property which we adopt in our representation as well. However, their final model requires evaluation of several per image features and computing HMM potentials per event and per collection, which is computationally and memory inefficient. Our proposed method extracts a single vectorial representation per collection, allowing efficient event recognition. For fair comparison, we evaluate our method also on the PEC collection, with the author provided features.

Video Event Recognition Similarly to encoding collections of images, also videos have predominantly been encoded as the mean visual feature of the sampled frames [13, 15]. Alternatively, using a Fisher Vector over several low-level video descriptors, such as SIFT, STIP and HOG, has been used [17, 18, 26]. An notable exception is the proposed method of Lai *et al*., where a video is treated as a bag-of-frames, and event recognition is handled as a multi-instance learning problem [11].

The current state-of-the-art video representation is the Fisher Vector over motion boundary histograms from improved dense trajectories [27]. For few example recognition on challenging web videos from the TrecVID multimedia event detection dataset [19], the approach by Wang and Schmid [27] was further improved by learning a compact semantic embedding [6] from auxiliary data. The latter method has been shown mainly effective for recognition using only 10 examples, the performance difference decrease for recognition using more examples. While we also obtain a compact representation, we do not require additional training data to learn the embedding. Moreover, our approach can model the variation over time explicitly.

Besides reporting results on TrecVID MED [19], we also report on the frequently used Columbia Consumer Video dataset introduced by [9]. To the best of our knowledge these are the largest publicly available video corpora in the literature for event recognition containing user-generated videos with a large variation in quality, length and content.

Fisher Vector Representation The Fisher Vector representation was introduced as an alternative to the Bag-of-Words image representation [22, 24]. Subsequently, it has been used for video classification, for which it is currently the state-of-the-art representation [27]. Over the years, many extensions have been proposed to the Fisher Vector framework, from which we highlight a few directions. First, the idea of using multiple layers of Fisher Vectors [21, 25] is similar in spirit to the proposed Event-FV. Indeed, while we use DeepNet features as input for our Event-FV, we could equally well have used a sequence of Fisher Vectors as input. Second, the idea of encoding the spatial coordinate of local patches [23] is similar to encoding the temporal axis of a stream. In its most basic form it would induce using $d-1$ dimensions for the PCA projection, and adding a temporal scalar. We have conducted preliminary experiments according to this temporal coordinate coding scheme, but the results where disappointing. We believe this is caused by the combination of the higher dimensional features (even after PCA projection we use a few hundreds of dimensions), with the fewer mixture components used in our models compared to [23]. Finally, the idea of using non-iid generative models has been explored for image classification in [4]. While we also use non-iid models, we base them on the temporal structure of our visual streams.

3 Event Fisher Vector

In this section we describe the Event Fisher Vector (Event-FV) encoding for a visual stream. Our encoding is based on the Fisher Kernel [8] and bears a resemblance to the Fisher Vector image representation [24]. We assume that the images of a stream can be modelled by a probability density function $p(\cdot|\theta)$ with parameters θ. Let $X = \{x_1, \ldots, x_n\}$ denotes a stream of n images, where each image i is described by the image descriptor x_i. Then, the *Fisher score* of the stream is given by the gradient of the log-likelihood w.r.t. the parameters:

$$G_\theta^X = \nabla_\theta \log p(X|\theta) \tag{1}$$

and describes how the parameters of the model contribute in the generative process. These Fisher scores can be used in linear classifiers, after transformation with the *Fisher Information Matrix* (FIM), $F_\theta = \mathbb{E}_{X\sim\theta}[G_\theta^X G_\theta^{X\top}]$:

$$\mathscr{G}_\theta^X = F_\theta^{-\frac{1}{2}} \nabla_\theta \log p(\mathbf{X}|\theta), \tag{2}$$

This transformation ensures invariance w.r.t. re-parametrisation of the probabilistic model.

The Fisher Vector image representation [24] encodes local SIFT features using a GMM as probabilistic model and a closed-form approximation of the FIM. In contrast, we focus on encoding a visual stream of images, where each image is described by a single feature, in our case extracted from a DeepNet [10, 30]. Moreover, we explore different generative encodings *and* provide analytical approximations of the FIM for these probabilistic models.

3.1 Student's-*t* Mixture Model

A problem with the Gaussian mixture model is that it is highly effected by the presence of (a small number of) outliers. While this is not a problem when encoding a set of approximate 10K local descriptors, as in the Fisher Vector image encoding, it becomes a problem when encoding an image stream consisting of around 50 images each. In order to make our model more robust against outliers, we replace the Gaussian distribution with a Student's-*t* distribution, which is known for its heavier tails [1, 20], see also Figure 2.

The Student's-*t* mixture model for observation $x_i \in \mathbb{R}^D$ is defined as:

$$p(x_i|\theta) = \sum_k \pi_k \, St(x_i|\theta_k), \tag{3}$$

where π_k is the mixing weight and the Student's-*t* distribution is parametrised by $\theta_k = \{\mu_k, \sigma_k, \nu_k\}$, with μ representing the mean, σ the diagonal co-variance matrix and ν the degrees-of-freedom. The Student's-*t* distribution for component k is defined by:

$$St(x_i|\theta_k) = Z_k \left(1 + \tfrac{1}{\nu_k}\delta_k(x_i)\right)^{-\frac{\nu_k+D}{2}}, \tag{4}$$

with Mahalanobis distance $\delta_k(x_i) = \sum_d (x_{id} - \mu_{kd})^2/\sigma_{kd}^2$ and Z_k is the normalisation factor:

$$Z_k = \frac{\Gamma(\frac{\nu_k+D}{2})}{\Gamma(\frac{\nu_k}{2})\,(\pi\nu_k)^{\frac{D}{2}}\,|\sigma_k^2|^{1/2}}, \tag{5}$$

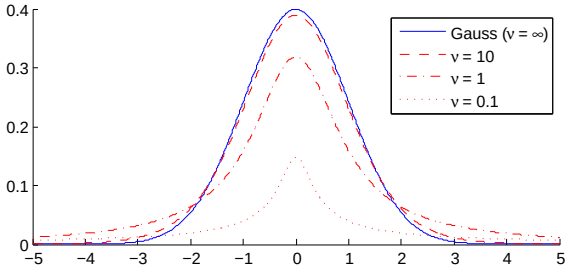


Figure 2: Illustration of the Gaussian and Student's-*t* distribution: showing the heavy tails of the *probability density function* for various values of ν.

with the gamma function Γ, and the mathematical constant π. The *Fisher score* w.r.t. the mean μ_k is given by[1]:

$$G^X_{\mu_\mathbf{k}} = \frac{\nu_k + D}{\sigma_k^2\,\nu_k} \sum_{i=1}^{n} \gamma_i(k)\, \frac{(x_t - \mu_k)}{1 + \frac{1}{\nu_k}\delta_k(x_i)}, \tag{6}$$

where $\gamma_i(k) = \frac{\pi_k\ St(x_i|\theta_k)}{\sum_k \pi_k\ St(x_i|\theta_k)}$ is the *responsibility* value of the k-th Student's-*t* mixture component [1]. Similar to the GMM this *Fisher score* is also a weighted average of the observations, where each observation is weighted by its responsibility $\gamma_i(k)$. However, two important differences are (i) that the responsibility values γ are now based on the Student's-*t* mixture model, and (ii) each observation is also weighted by the Mahalanobis distance w.r.t. component k and the degrees-of-freedom ν_k.

3.1.1 Analytical Approximation of the Fisher Information Matrix

Perronnin *et al.* [22, 24] showed that the Fisher information matrix (FIM) for the GMM is diagonal, under the following assumptions:

Hard-assignment assumption. When all patches are sharply peaked around a single component k (*i.e.* $\forall i\ \exists k\ \gamma_k(i) \approx 1$), the off-diagonal entries in the FIM are zero if they involve mean or variance parameters from different mixture components, or if they involve the mixing weight parameter and a mean or variance parameter.

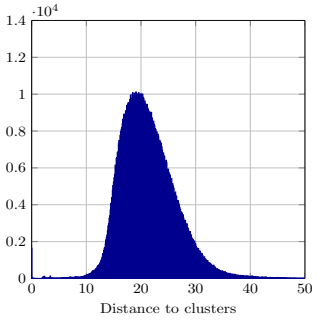
Diagonal covariance assumption. Using diagonal covariance matrices in the GMM makes the Fisher scores independent per dimension and therefore the cross-terms in the FIM for the mean and variance of the same component are zero.

The resulting analytical approximation yields, for the mean parameter: $F_{\mu_k}^{-\frac{1}{2}} = \sigma_k\ \pi_k^{-\frac{1}{2}}$ [24].

In contrast, the dimensions of the Fisher scores of the Student's-*t* model are interdependent due to the Mahalanobis distance $\delta_k(x_i)$ in Eq. (6). 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, which makes the Fisher score per dimensional independent. This assumption is based on the *concentration of distances* theorem [31] 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. We illustrate the distances for our visual data in Figure 3.

[1] For clarity of presentation we address only Fisher scores w.r.t. the mean parameter.



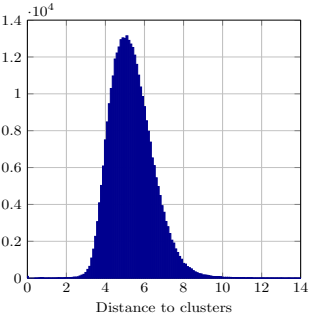


Figure 3: Histograms showing the distance between datapoints and cluster centres on the two used video-datasets. 99% of the data is within a distance ratio of 4, confirming the appropriateness of the *constant distance* assumption.

This results in the following analytical approximation for the mean component of the FIM:

$$F_{\mu_k}^{-\frac{1}{2}} = \sigma_k \left(\pi_k \frac{\nu_k}{\nu_k + D}\right)^{-\frac{1}{2}}. \tag{7}$$

Which is somehow intuitive, since it is similar to the approximation for the GMM, but includes a weighing with the degrees-of-freedom ν_k. 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.

3.2 Sequential Modelling of a Visual Stream

In this section we propose to model the temporal relationship among images in a stream, using a Hidden Markov Model (HMM). While the independent mixture models, discussed above, ignore the temporal structure of the visual stream and treat each image as an independent observation, the HMM models encode the temporal relation explicitly.

The temporal relation in the HMM 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 probability of a sequence X is given by:

$$p(X|\theta) = \sum_Z \prod_{i=1}^{n} p(z_i|z_{i-1}) p(x_i|z_i), \tag{8}$$

where $p(z_i|z_{i-1})$ models the *transitional* probability, parametrised by the transition matrix $A \in \mathbb{R}^{k\times k}$, and the vector $\pi_k \in \mathbb{R}^{1\times k}$ for the initial state distribution; and $p(x_i|z_i)$ models the *emission* probability, for which we use either the Gaussian distribution $\mathcal{N}(x_i|\theta_k)$ or the Student's-*t* distribution $St(x_i|\theta_k)$.

The Fisher score w.r.t. the mean μ_k, using the Gaussian distribution is given by:

$$G_{\mu_k}^X = \sum_Z p(Z|X)\, \nabla_{\mu_k} \log p(X, Z|\theta) \tag{9}$$

$$= \sum_{i=1}^{n} \gamma_i(k)\, \nabla_{\mu_k} \log p(x_i|\theta_k) = \sum_{i=1}^{n} \gamma_i(k) \left(\frac{x_i - \mu_k}{\sigma_k^2}\right), \tag{10}$$

this is identical to the Fisher score of the GMM [24], except that the responsibility values $\gamma_i(k)$ are now computed by:

$$\gamma_i(k) = \frac{p(x_1, \ldots, x_i, z_i = k)\; p(x_{i+1}, \ldots, x_n | z_i = k)}{p(X)}, \tag{11}$$

which reflects the dependence among the images in the collection. These responsibility values can be efficiently computed using the forward-backward algorithm [1]. Similarly, the Fisher scores for the HMM with the Student's-t emission distribution are given by Eq. (6), using the HMM responsibility values $\gamma_i(k)$.

Approximating the Fisher Information Matrix The Fisher scores in the hidden Markov models differ from the independent image mixtures only by the definition of the responsibilities $\gamma_i(k)$. Resorting again on the hard-assignment assumption, *i.e.* all observations are sharply peaked around a single state k, we obtain the same closed form approximations as used in the independent GMM and StMM models. To the best of our knowledge, we are the first to propose an analytical approximation of the FIM for HMMs. We are aware that we use a very crude approximation, the HMM even explicitly models the observations as interdependent. However, our experimental evaluation, see Section 4.2, shows that this approximation is sufficient and clearly outperforms the commonly used identity and empirical approximations. Note, again, that the assumptions are only used to derive analytical approximations of the FIM, not for computation of the Fisher scores.

4 Experimental Evaluation

4.1 Datasets and Setup

Photo Event Collection (PEC) [2] This dataset was introduced in 2013 as benchmark for event classification from Flickr photo collections. It consists of 14 social event classes, *e.g.* Birthday, Christmas, Hiking, Halloween, and 807 photo collections with in total over 61K photos. We use the suggested experimental setup: per event 30 collections are selected for training and 10 for testing. Performance is evaluated using mean class accuracy (MCA).

TrecVID Media Event Detection (MED13) [19] This dataset was part of the 2013 TrecVID benchmark task on Media Event Detection. We follow the *100Ex evaluation procedure* in our experiments. With over 10K training and 27K testing videos this is one of the biggest datasets for event detection in video. In our experiments we focus on the visual aspect of the videos and therefore use only visual frame-based features. Performance is evaluated using mean average precision (MAP) over the 20 events.

Columbia Consumer Video (CCV) [9] This dataset contains over 9K user-generated videos from YouTube, with an average video length of 80 seconds. The dataset comes with video level ground-truth annotations for 20 semantic categories, 15 of which are events while the other 5 are objects or scenes classes. We use the split suggested by the authors which consists of 4,659 training videos and 4,658 test videos. Performance is evaluated using mean average precision (MAP) over the 20 categories.

Extracting Event Fisher Vectors For each image in a stream we extract visual features from a pre-trained DeepNet [10, 30], we use an in-house implementation of [30], trained on 15K ImageNet classes from the fall 2012 release [5]. As is common practice, we use the output of the final fully connected layer of the DeepNet, resulting in a 4K dimensional vector which is whitened per dimension (*i.e.* to obtain zero-mean an unit variance). For the PEC

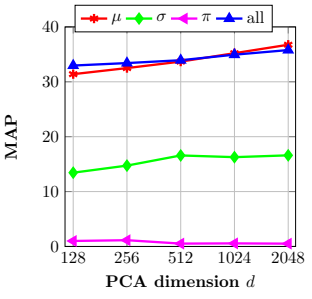
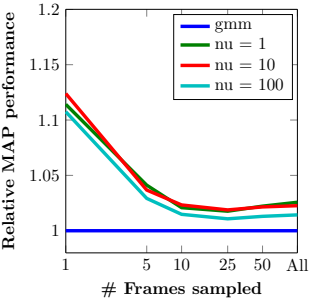


Figure 4: *Left:* comparison of using Event-FV w.r.t. to weight, mean and variance parameters of the GMM on MED13 for $k = 8$. *Right:* relative improvement (in MAP) of S*t*MM over GMM for various number of sampled frames per videos on CCV.



FIM	GMM	S*t*MM	HMM	S*t*HMM
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**

	GMM	S*t*MM	HMM	S*t*HMM	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

Table 1: Overview of different approximations for the Fisher Information Matrix on MED13 (*left*), and comparison of the proposed generative encodings for all datasets, including the oracle 'best per event' (BPE) score (*right*)

dataset we use all photos belonging to a collection, while for the video streams we sample a frame every 2 seconds.

For the Event-FV we obtain the parameters of our generative models by maximum likelihood estimation on the training set. We fix the values for the degrees-of-freedom to $\nu = 10$ and use standard expectation maximisation algorithms. Following common practice we apply power-normalisation and ℓ_2 normalisation before training SVM classifiers.

PCA and Mixture Components Two key parameters of any Fisher Vector representation is the number of PCA dimensions and the number of mixture components. We have performed a rather exhaustive search in preliminary experiments. In general it holds that PCA is helpful, unsurprisingly since it decorates the input space which matches the diagonal covariance assumption used. Moreover, there seems to be a correlation with the dataset size and the number of dimensions and components. For all experiments we use PEC (256,8), CCV (512,16), and MED13 (2048,8). The number of mixture components used here is much lower than used for image classification [24], we believe this is because of the highly discriminative DeepNet features and the relatively high number of PCA dimensions used.

Moreover, we have experimentally validated that using the Event-FV w.r.t. the mean only has the best performance vs dimensionality ratio, see Figure 4 (*left*).

4.2 Approximations of the Fisher Information Matrix

In this experiment we study the influence of the Fisher information matrix approximations, and compare the proposed analytical approximation to two comment FIM approximations: (i) the identity matrix, since asymptotically the Fisher information matrix is immaterial [8]; (ii) the diagonal empirical approximation, which results in a whitening of the signal for the diagonal [1], *i.e.* each dimension will have zero-mean and unit-variance.

The results are presented in Table 1 (*left*), for all four different generative models on the MED13 dataset, using $k = 8$ and $d = 2048$. First, we observe that for all generative models

the behaviour is very similar; approximating the Fisher information matrix with the identity matrix performs worst (MAP $\sim 27\%$), the empirical estimation brings a strong improvement to $\sim 34\%$ and finally the analytical approximation increases performance to $\sim 37\%$. For the GMM model, our results are in line with the findings of [24], where the analytical approximation also obtains best performance. While we have made strong additional assumptions for the closed form approximations, the resulting approximation works well in practice.

Furthermore, we note that the sequential encoding models (HMM and S*t*HMM) outperform the independent image models when the identity approximation is used. Unfortunately, these improvements vanish when the empirical or analytical approximations are used. Likely, this is caused by the fact that the diagonal Fisher information approximations are not valid in the sequential models, due to the dependencies between the frames (or images).

4.3 Robustness against Number of Samples

One of the main motivations of using the S*t*MM model is it robustness towards outliers in the case that the FV is extracted from just a few observations. In this experiment we aim to testify this hypothesis, by varying the number of frames sampled from a video. We randomly subsample a given number of frames from each video of the CCV dataset and then employ the Event-FV using the S*t*MM as probabilistic model, using different values for ν. The results are given in Figure 4 (*right*), where the relative performance w.r.t. to the GMM is shown. Note that the absolute MAP performance of the GMM drops from ~ 67 when sampling all video frames (on average the dataset contains 51 frames per video) to around ~ 45 MAP when sampling just a single frame per video. However , the relative performance increase of the S*t*MM model confirms the robustness.

4.4 Performance of Different Generative Models

In Table 1 we evaluate the performance of the different encoding methods for all three datasets. In all cases the S*t*MM performs best, showing that it is important to model the heavy tails of the emission probability. For the two video datasets (CCV and MED13) the GMM model is second best, while for the PEC dataset the HMM models is second best.

We observe, that the performance of the different encodings per event varies quite significant, see also the per class AP scores in the supplementary material. Therefore, we also evaluate the performance if an oracle could give the best model per event. From the results in Table 1, we observe that for all datasets the S*t*MM performance is close to the performance of the oracle model. Indicating it is not worth the effort to try to predict which model to use for each event, instead we will use the S*t*MM model to compare to the state-of-the-art.

4.5 Comparison with State-of-the-Art

In this last set of experiments we compare our proposed method to alternative event recognition approaches, including the Mean DeepNet (MDN) baseline, which averages the DeepNet descriptors of the images from a stream. We use the Event-FV using the S*t*MM generative model, and, inspired by [21, 25], also the concatenation of Event-FV and MDN; as well as temporal pyramids (TP), inspired by the spatial pyramids [9, 17] by extracting Event-FVs over the whole stream, and over three non-overlapping subsequent parts of the stream. This results in a total of four Event-FVs, each of which is power- and ℓ_2-normalised.

Method	MCA	F_1
Baseline [2]	41.4	38.9
Stopwatch [2]	55.7	56.2
S*t*MM (features [2])	60.0	60.5
MDN	79.3	80.3
S*t*MM	**85.7**	**85.8**
+ MDN	**85.7**	**85.8**
+ TP	82.9	82.9

(a) PEC Dataset

Method	MAP
Ye *et al.* [29]	64.0
Liu *et al.* [13]	69.5
Wu *et al.* [28]	70.6
MDN	66.3
S*t*MM	69.0
+ MDN	71.4
+ TP	**71.7**

(b) CCV Dataset

Method	100Ex MAP	10Ex MAP
Baseline MBH [27]	31.5	17.4
Habibian *et al.* [6]	32.0	19.6
MDN	28.6	16.6
S*t*MM	37.2	21.3
+ MDN	37.7	21.7
+ TP	**38.6**	**21.8**

(c) MED13 100Ex and 10Ex

Table 2: Comparison of the Event-FV performance to current state-of-the art methods.

For the PEC dataset, the overview is presented in Table 2a. Since the PEC dataset is rather new, we can only compare to methods of [2], and we include an experiment using their provided features. From the results, we first observe that our MDN baseline outperforms all of the methods proposed by [2]. Second, we note that even when using the same image features, our Event-FV outperforms their results significantly.

For the CCV dataset, the overview is presented in Table 2b. On this dataset plenty of methods have been evaluated, using the provided features of [9], therefore we show only the highest performing ones. While our MDN is a decent baseline, the Event-FV and the Event-FV + TPs outperform any of the previous methods.

For the MED13 dataset, the overview is presented in Table 2c, where we also show the results for the *10Ex* task, with just 10 positive train examples per event. Again, our MDN is a decent baseline, and our Event-FV clearly outperforms the MBH performance and the VideoStory encoding [6], for both the 10Ex and the 100Ex task. Finally, for this dataset extracting TPs increases performance further by about 1% MAP.

5 Conclusion

In this paper we have introduced the Event-FV to represent visual streams for event recognition. We have argued that the Student's-*t* mixture model is more appropriate for a small set of observations, than the commonly used GMM. For the S*t*MM model we have derived a closed form approximation for the Fisher information matrix for this model, which experimentally outperforms the identity or empirical approximation. Finally, we also have explored Hidden Markov models which explicitly model the sequential order of the stream.

We have conducted experiments on three recent datasets and showed that the analytical approximations of the FIM outperform an identity or empirical approximation by a large margin, that the S*t*MM model has a slight edge over the other explored models, and that it results in state-of-the-art performance. This indicates the advantage of using appropriate probabilistic models within the Fisher vector and to derive their analytical FIM approximations. We conclude, for the task of visual event classification capturing the heavy tails of the small sample size is more beneficial than modelling the temporal relation of the stream.

Acknowledgements

This research is supported by the STW STORY project and the Dutch national program COMMIT.

References

[1] C. Bishop. *Pattern recognition and machine learning*. Spinger-Verlag, 2006.

[2] L. Bossard, M. Guillaumin, and L. Van Gool. Event recognition in photo collections with a stopwatch hmm. In *ICCV*, 2013.

[3] J. Choi, W. Jeon, and S.-C. Lee. Spatio-temporal pyramid matching for sports videos. In *ICMR*, 2008.

[4] G. Cinbis, J. Verbeek, and C. Schmid. Image categorization using Fisher kernels of non-iid image models. In *CVPR*, 2012.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[6] A. Habibian, T. Mensink, and C. Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *MM*, 2014.

[7] N. Imran, J. Liu, J. Luo, and M. Shah. Event recognition from photo collections via pagerank. In *MM*, 2009.

[8] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, 1999.

[9] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ICMR*, 2011.

[10] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[11] K.-T. Lai, F. Yu, M.-S. Chen, and S.-F. Chang. Video event detection by inferring temporal instance labels. In *CVPR*, 2014.

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

[13] D. Liu, K.-T. Lai, G. Ye, M.-S. Chen, and S.-F. Chang. Sample-specific late fusion for visual category recognition. In *CVPR*, 2013.

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. of Computer Vision*, 60(2):91–110, 2004.

[15] A. Ma and P. Yuen. Reduced analytic dependency modeling: Robust fusion for visual recognition. *Int. J. of Computer Vision*, 2014.

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

[17] I. Mironica, J. Uijlings, N. Rostamzadeh, B. Ionescu, and N. Sebe. Time matters! capturing variation in time in video using Fisher kernels. In *MM*, 2013.

[18] D. Oneata, J. Verbeek, and C. Schmid. Action and event recognition with Fisher vectors on a compact feature set. In *ICCV*, 2013.

[19] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. Smeaton, and G. Queenot. An overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID Workshop*, 2013.

[20] D. Peel and G. McLachlan. Robust mixture modelling using the t distribution. *Statistics and computing*, 2000.

[21] X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked Fisher vectors. In *ECCV*, 2014.

[22] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[23] J. Sánchez, F. Perronnin, and T. de Campos. Modeling the spatial layout of images beyond spatial pyramids. *Pattern Recognition*, 2012.

[24] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the Fisher vector: Theory and practice. *Int. J. of Computer Vision*, 2013.

[25] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Fisher networks for large-scale image classification. In *NIPS*, 2013.

[26] C. Sun and R. Nevatia. Large-scale web video event classification by use of Fisher vectors. In *Workshop on Applications of Computer Vision*, 2013.

[27] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

[28] Z. Wu, Y.-G. Jiang, J. Wang, J. Pu, and X. Xue. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In *MM*, 2014.

[29] G. Ye, D. Liu, I.-H Jhuo, and S.-F. Chang. Robust late fusion with rank minimization. In *CVPR*, 2012.

[30] M. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

[31] A. Zimek, E. Schubert, and H.-P. Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 2012.