Time Matters! Capturing Variation in Time in Video using Fisher Kernels

Ionuţ Mironică
LAPI, University Politehnica of Bucharest, Romania
imironica@imag.pub.ro

Jasper Uijlings
DISI, University of Trento, Italy
jrr@disi.unitn.it

Negar Rostamzadeh
DISI, University of Trento, Italy
rostamzadeh@disi.unitn.it

Bogdan Ionescu
LAPI, University Politehnica of Bucharest, Romania
bionescu@imag.pub.ro

Nicu Sebe
DISI, University of Trento, Italy
sebe@disi.unitn.it

ABSTRACT

In video global features are often used for reasons of computational efficiency, where each global feature captures information of a single video frame. But frames in video change over time, so an important question is: how can we meaningfully aggregate frame-based features in order to preserve the variation in time? In this paper we propose to use the Fisher Kernel to capture variation in time in video. While in this approach the temporal order is lost, it captures both subtle variation in time such as the ones caused by a moving bicycle and drastic variations in time such as the changing of shots in a documentary.

Our work should not be confused with a Bag of Local Visual Features approach, where one captures the visual variation of local features in both time and space indiscriminately. Instead, each feature measures a complete frame hence we capture variation in time only.

We show that our framework is highly general, reporting improvements using frame-based visual features, body-part features, and audio features on three diverse datasets: We obtain state-of-the-art results on the UCF50 human action dataset and improve the state-of-the-art on the MediaEval 2012 video-genre benchmark and on the ADL daily activity recognition dataset.

Categories and Subject Descriptors

I.4.8 [**Scene Analysis**]: Time-varying imagery

Keywords

Video Classification; Fisher kernels; Variation in Time

1. INTRODUCTION

In video retrieval, an important research problem is how to adequately capture temporal information. Until recently, most video retrieval systems relied mostly on single representative video frames where time is ignored for efficiency reasons [24]. Recent work simply accumulates features over a whole video sequence [4, 8, 22]. Such accumulation may capture more information, but also mixes information, disregarding appearance variation over time. For example, when a car approaches and then turns a corner, there are first straight movements followed by turning movements. It is important that both types of movement do not happen at the same time. We want to have a representation which keeps this distinction.

In this paper we propose a novel video representation method which aggregates frame-based features while retaining their variation in time. Specifically, we propose to use the Fisher representation [15] which was recently introduced for improving a Bag of Local Visual Features approach [2]. Bag of Local Visual Features captures the visual variation in space for images and in both space and time for video. The Fisher Kernel improves over the common k-means vocabulary by modelling the distribution of features within each visual word. In contrast to Local Visual Features, in this paper we apply the Fisher representation on frame-based features, effectively capturing variation *in time only* (as there is no variation in space). Like Bag of Local Visual Features, all ordering is lost but all variation is captured. Using the Fisher representation for modelling variation in time, (1) dissimilar frames will be represented by different mixture components (i.e. clusters), preventing blending of unrelated features while enabling them to co-exist in a single representation. This enables representing videos which consist of dissimilar parts (which may not even have a fixed temporal order) such as news broadcasts that switch between the news-anchor and on-site footage. Furthermore, (2) similar frames that fall in the same mixture component will be modelled with respect to the general distribution of that component, capturing subtle variations in time such as the different appearances of a person walking by.

We test our Fisher-based framework for modelling variation in time on a variety of video benchmarks: genre retrieval on the MediaEval benchmark [21], Human Action Recognition on the UCL50 dataset [17], and daily activity recognition on ADL [14]. Additionally, we employ a variety of frame-based features: global Histogram of Oriented Gradients (HoG) [3], global Histogram of Optical Flow (HoF) [17], global Colour Naming histogram (CN) [28], HoF-

MM'13, October 21–25, 2013, Barcelona, Spain.

ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502183.

based body-part histogram [18], and even on block-based audio features [13]. We show that the Fisher representation consistently and significantly outperforms simple accumulation. Additionally, by explicitly modelling the variation in time we obtain state-of-the-art results or better on all three datasets using a smaller array of simpler features.

To summarise, our main contributions are the following: (1) We introduce a Fisher-based representation for frame-based features in video that captures variation in time. (2) We demonstrate its generality in terms of applications by applying it to genre-recognition, sports-recognition, and daily activity recognition. (3) We demonstrate its generality in terms of features by using audio features, global visual features, and body-part features. (4) We achieve similar or better performance than the state-of-the-art using a smaller array of simpler features.

2. RELATED WORK

Recently, researchers have successfully captured local temporal information in video by using spatio-temporal features, visual features that are measured in the 3D volume spanned by the video frames. These features are extracted either at interest points which are stable in both space and time [4, 5, 8, 10, 27], or at stable trajectories [29]. The specific movement pattern captured by these features yields significant improvements over 2D features. However, these features are accumulated over an entire video sequence ignoring the visual variation at different parts of the video.

Some works include some form of variation in time by using a linear quantization of the video: the video is split into n sequences of an equal number of frames, where for each sequence all features are accumulated [1, 10, 19, 29]. Histograms of the individual sequences are concatenated, leading to good accuracy improvements. In [26] the authors use a linear quantization method for global features, where the features are averaged inside a sequence.

Few works focus directly on modelling the temporal order/variation between frames [24]. There is some work on using Hidden Markov Models [9, 16]. Other work uses temporal rules with high-level concepts [12, 25].

3. MODELLING VARIATION IN TIME

The Fisher Kernel [7] represents a signal as the gradient with respect to the probability density function that is a learned generative model of that signal. Recently, [15] introduced the Fisher Kernel as an improved visual vocabulary for Bag-of-Words. Its success shows that it meaningfully captures the visual variation of local descriptors.

In this paper we employ the Fisher Kernel to capture variation in time in video. We follow [15] and use a Gaussian Mixture Model with diagonal covariance matrices as generative distribution. Specifically, let μ_i and σ_i be the mean and standard deviation of the i-th Gaussian centroid, let $\gamma(i)$ be the soft assignment to the i-th Gaussian of the d-dimensional feature x_t captured at frame t. The gradient of the GMM with respect to μ_i and σ_i are calculated as [15]:

$$\mathcal{G}^x_{\mu,i} = \frac{1}{T\sqrt{\omega_i}} \sum_{t=1}^{T} \gamma(i) \frac{x_t - \mu_i}{\sigma_i} \quad (1)$$

$$\mathcal{G}^x_{\sigma,i} = \frac{1}{T\sqrt{2\omega_i}} \sum_{t=1}^{T} \gamma(i) \left[\frac{(x_t - \mu_i)^2}{{\sigma_i}^2} - 1 \right]. \quad (2)$$

The final Fisher vector is the concatenation of the $\mathcal{G}^x_{\mu,i}$ and $\mathcal{G}^x_{\sigma,i}$ for $i = 1...k$ and has a dimensionality of $2kd$.

Interpreting the formulas in terms of variation in time, Equation 1 averages *related* features over time, which are related as they fall in the same mixture component. Equation 2 models the variation of related features over the video sequence, capturing subtle visual changes (e.g. a car driving by). The different mixture components capture drastic variations in time such as a shot changes.

Important parameters or design choices are: (1) Applying PCA on initial features x_t, reducing dimensions of the final Fisher vector and potentially improving GMM clustering by decorrelation. (2) The number of GMM clusters. (3) Normalization of the Fisher vector. (4) Choice of classifier.

4. EXPERIMENTS

We demonstrate the advantages and generality of our framework on three different datasets using a variety of features. For brevity reasons we will mainly focus on the number of GMM clusters and only touch upon applying PCA. We normalise the Fisher vector by taking the square root followed by the $L2$-norm [15]. In contrast to [15], we use SVMs with RBF-kernels as these performed better than linear SVMs, even at an increased number of clusters for the latter. When combining different types of features we use weighted late fusion, learning weights on our optimization sets.

4.1 Genre Retrieval

We perform genre retrieval on the 2012 MediaEval Genre Tagging Task [21], consisting of 2000 hours in 14,838 videos, labelled according to 26 genres such as art, autos, and comedy. Performance is measured in terms of Mean Average Precision (MAP). We perform all parameter optimization on the training set which we split in two fixed, equally sized parts. We compare with the state-of-the-art using the official training set (5,288 videos) and test set (9,550 videos).

Baseline. We use the following features: (1) *Global Histogram of Oriented Gradients [3] (81 dimensions)* which calculates HoG over the whole frame using a 3x3 spatial division. (2) *Colour Naming histogram [28] (11 dimensions)* of the whole frame. (3) *Audio features [13] (98 dimensions)* which are general purpose audio descriptors extracted over a standard period of 1.28 seconds around the frame using [13]. Results of averaging features over the whole video are presented as the horizontal lines in Figure 1.

Optimizing the Fisher Representation. We ran experiments with PCA dimensionality reduction on the frame-based features, setting the number of cluster centres to 100. We found that for Colour Naming, applying PCA reduces performance. This is because the dimensions are decorrelated and non-redundant by design [28]. For HoG and Audio features the optimal reduction is to keep 80% of the dimensionality, where for HoG accuracy increased a full 5%. We choose these PCA settings for subsequent experiments.

Next, we determine the optimal number of clusters for each feature as shown in Figure 1. First of all, notice the big improvements of the Fisher representation over the baseline which simply averages the features: Even when using only a single (!) centroid, Colour Naming goes up from 0.18 MAP to 0.28 MAP, HoG goes up from 0.22 MAP to 0.38 MAP, and Audio goes up from 0.34 MAP to 0.45 MAP. The modelling of variation in time therefore significantly improves results. Increasing the number of clusters increases performance even

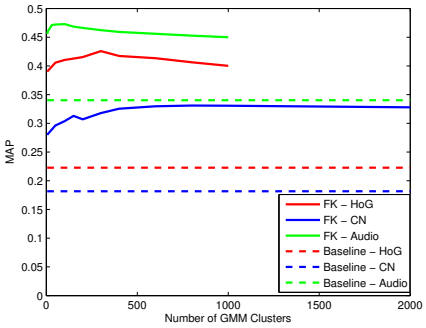

Figure 1: Mean Average Precision (MAP) while varying the number of cluster centres on the MediaEval 2012 training set.

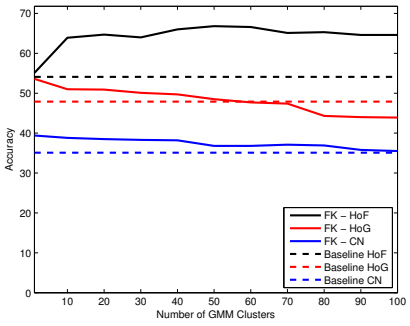

Figure 2: Classification accuracy on half of UCF50 sports while varying the number of cluster centres (8-fold cross-validation).

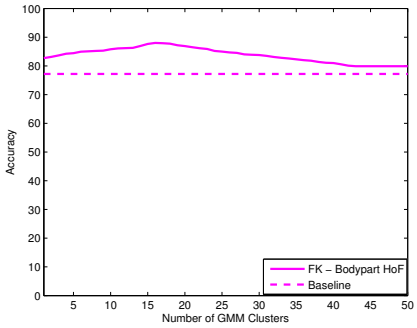

Figure 3: Classification accuracy on ADL daily activity recognition on half the dataset while varying the number of cluster centres.

further: Both Colour Naming and HoG increase an extra 0.05 MAP, reaching 0.33 MAP and 0.43 MAP at 800 clusters and 200 clusters respectively. Audio features increase to 0.47 MAP at 50 clusters. We will use this number of clusters in the next experiment. The final sizes of the Fisher vectors are reasonable at 17,600 for Colour Naming, 42,000 for HoG, and 9,000 for Audio features. Note that performance of both HoG and Audio features go down after the optimal points, likely due to the high dimensionality of the features (i.e. curse of dimensionality).

Comparison to State-of-the-Art. Results on MediaEval 2012 are shown in Table 1. For audio features our results are at 0.47 MAP much better than the best result of 0.19 reported at the MediaEval workshop [6]. For visual features only, at 0.46 MAP we perform significantly better than the best result of 0.35 MAP [23]. Remarkably, our combination of audio and visual features yields with 0.55 MAP a better performance than the use of text from automatic speech recognition and meta-data, which had the highest performance at MediaEval 2012 at 0.53 MAP.

To conclude, using the Fisher kernel to model variation in time significantly improves over a simple averaging of features, yielding much better results than the state-of-the-art on the MediaEval 2012 benchmark.

4.2 Human Action Recognition

We now evaluate our framework on the UCF50 Human Action Recognition dataset [17], which contains 6600 realistic videos from Youtube with large variations in camera motion, object appearance and pose, illumination conditions, scale, etc. It has 50 mutual exclusive categories such as biking, diving, drumming and fencing. Performance is evaluated in terms of classification accuracy. We perform all optimization on half of the dataset, using 8-fold cross-validation. We compare with the state-of-the-art using the standard leave-one-group-out cross-validation on the full dataset [17].

Baseline. We use the following features: (1) *Global Histogram of Oriented Gradients [3] (9, 36, 81, and 144 dimensions)* which calculates HoG over the whole frame using a 1x1, 2x2, 3x3, and 4x4 spatial division. (2) *Global Histogram of Optical Flow [17] (9, 36, 81, and 144 dimensions)* which measures the average velocity of non-stationary pixels over a region in 9 orientations. We use a 1x1, 2x2, 3x3, and 4x4 spatial division. (3) *Colour Naming Histogram [28] (11, 44, 99, and 176 dimensions)* using a 1x1, 2x2, 3x3, and 4x4 spatial division. In all experiments, we combine different spatial divisions for a single feature type using late fusion with equal weights. Results of averaging each feature over the whole video are shown as horizontal lines in Figure 2.

Optimizing the Fisher Representation. We first optimized the dimension reduction using PCA. We found that both the Colour Naming histogram and the Histogram of Optical Flow did not benefit. For HoG we found a good improvement by reducing dimensions to 90% (data not shown).

Next, in Figure 2 we evaluate the performance with respect to the number of GMM clusters, where we use the same number of clusters for all spatial divisions of a single feature type. For Colour Naming and HoG the use of a single cluster improves the baseline with 6% and 5% respectively. More clusters degrade performance as for this dataset the visual changes are subtle and do not require different mixture components. For HoF, using 50 clusters improves the baseline of 54% to 67%, a 13% improvement. Hence the optical flow changes drastically in time which is best captured in multiple clusters. Indeed, for example a baseball pitch has at least three distinct movement patterns: static (before the action), the pitch, and the batting. In the next experiment we use 1 cluster for CN and HoG, and 50 clusters for HoF.

Comparison to State-of-the-Art. We present the state-of-the-art in Table 2. As can be seen, we rank second with 74.7% accuracy after the 76.9% accuracy of Reddy et al. [17]. However, we use only global features whereas all other good performing methods use computationally more expensive Space-Time Interest Points (STIPs). Only the GIST3D entry of [8] does not use STIPs. They use global, frame-based features plus linear quantization. Our performance using the Fisher vector is a significant 9.4% higher.

We conclude that our framework yields similar performance as the state-of-the-art while using simpler features.

4.3 Daily Activities

We report results on daily activity recognition using the ADL dataset [14], consisting of ten human activities such as dialling a phone, peeling banana, and chopping banana. Each activity is performed three times by five people, totalling 150 videos. Performance is measured in accuracy. We do all optimization on half of the dataset and report final results on the full dataset. In both cases we use leave-one-person-out cross-validation [14].

Baseline. As human pose and body-part motion are important for distinguishing the different categories, we extract body-part features [18]. We use the state-of-the-art body-part detector of [31] and extract at every frame for all 18 body-parts a Histogram of Optical Flow in 8 orientations

Table 1: Comparison with State-of-the-Art (SoA) in terms of Mean Average Precision (MAP) on MediaEval 2012.

Feature type	**Summary SoA method MediaEval 2012**	**MAP SoA**	**MAP ours**
Audio	Block Based Audio Features and 5-NN [6]	0.192	0.475
Visual	Visual descriptors (Color, Texture, rgbSIFT) [23]	0.350	0.460
Audio & Visual	-	-	0.550
Metadata & Text ASR	BoW Text ASR & metadata [20]	0.523	-

Table 2: Comparison with State-of-the-art on UCF50 Human Action Recognition.

Method	**Accuracy**
Reddy et al. [17]	76.9%
This paper	**74.7%**
Solmaz et al. [26]	73.7%
Everts et al. [5]	72.9%
Kliper-Gross et al. [8]	72.6%
Solmaz et al. [26]: GIST3D	65.3%

Table 3: Comparison with state-of-the-art on the ADL Daily Activity Recognition dataset.

Method	**Accuracy**
This paper	**97.3%**
Wang et al. [30]	96.0%
Lin et al. [11]	95.0%
Messing et al. [14]	89.0%

(144 dimensions). The result of averaging this feature over the video is shown as the horizontal line in Figure 3.

Optimizing the Fisher Representation. We found no improvements by doing PCA on the bodypart HoF features.

Figure 3 shows accuracy with respect to the number of GMM clusters. Using only a single cluster yields a performance improvement from 77% to 82% accuracy. The best accuracy of 88% is obtained using 17 clusters. Note that the number of clusters is relatively low, likely due to the smaller dataset. At 17 clusters, the final feature has 4,896 dimensions. We use 17 clusters when testing on the full dataset.

Comparison to State-of-the-Art. We compare our work with others in Table 3. As can be seen, our approach yields the highest accuracy of 97.3%. This shows that the Fisher representation is also effective for modelling variation in time using local body-part features.

5. CONCLUSIONS

We propose to use the Fisher kernel to model variation in time for frame-based video features. While the temporal order is lost, the temporal variation is captured at two levels: similar features are grouped together while retaining variation, which enables capturing subtle variations over time such as a exhibited by a moving car. Dissimilar features are kept separate, preventing mixing features from unrelated parts of the video while keeping them in a single representation, which enables capturing different shots in a video.

We demonstrated that our framework is highly general: We showed significant improvements on a wide variety of features, ranging from global visual features, to body-part features, and to audio features. We also demonstrated that our method works on a wide variety of datasets: We obtained state-of-the-art performance on UCF50 using global features instead of the more complex STIPs used in other methods. We improved the state-of-the-art on ADL daily activity recognition. We significantly improved the state-of-the-art on the MediaEval 2012 genre classification task.

In future work we plan to model variation in time using Fisher kernels on more advanced features such as STIPs.

6. REFERENCES

[1] M. Bregonzio, S. Gong, and T. Xiang. Recognising action as clouds of space-time interest points. In *CVPR*, 2009.

[2] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual Categorization with Bags of Keypoints. In *ECCV Workshop on Statistical Learning in CV*, 2004.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, 2005.

[5] I. Everts, J. van Gemert, and T. Gevers. Evaluation of color stips for human action recognition. In *CVPR*, 2013.

[6] B. Ionescu, I. Mironica, K. Seyerlehner, P. Knees, J. Schluter, M. Schedl, H. Cucu, A. Buzo, and P. Lambert. ARF @ mediaeval 2012: Multimodal video classification. In *MediaEval workshop*, 2012.

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

[8] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf. Motion interchange patterns for action recognition in unconstrained videos. In *ECCV*, 2012.

[9] H. Kuehne, D. Gehrig, T. Schultz, and R. Stiefelhagen. On-line action recognition from sparse feature flow. In *VISAPP*, 2012.

[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[11] Z. Lin, Z. Jiang, and L. Davis. Recognizing actions by shape-motion prototype trees. In *ICCV*, 2009.

[12] K. Liu, M. Weng, C. Tseng, Y. Chuang, and M. Chen. Association and temporal rule mining for post-filtering of semantic concept detection in video. *IEEE TMM*, 2008.

[13] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard. YAAFE, an easy to use and efficient audio feature extraction software. In *ISMIR*, 2010.

[14] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, 2009.

[15] F. Perronnin, J. Sanchez, and T. Mensink. Improving the Fisher Kernel for Large-Scale Image Classification. In *ECCV*, 2010.

[16] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, M. Wang, and H.-J. Zhang. Correlative multilabel video annotation with temporal kernels. *ACM TOMCCAP*, 2008.

[17] K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. In *MVAP*, 2012.

[18] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012.

[19] S. Sadanand and J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012.

[20] S. Schmiedeke, P. Kelm, and T. Sikora. TUB @ MediaEval 2012 tagging task: Feature selection methods for bag-of-(visual)-words approaches. In *MediaEval Workshop*, 2012.

[21] S. Schmiedeke, C. Kofler, and I. Ferrané. Overview of MediaEval 2012 genre tagging task. In *MediaEval workshop*, 2012.

[22] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICIP*, 2004.

[23] T. Semela, M. Tapaswi, H. Ekenel, and R. Stiefelhagen. Kit at mediaeval 2012 - content-based genre classification with visual cues. In *MediaEval workshop*, 2012.

[24] C. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 2009.

[25] C. G. Snoek and M. Worring. Multimedia event-based video indexing using time intervals. *IEEE TMM*, 2005.

[26] B. Solmaz, S. M. Assari, and S. Mubarak. Classifying web videos using a global video descriptor. *MVAP*, 2012.

[27] J. Stöttinger, A. Hanbury, N. Sebe, and T. Gevers. Sparse color interest points for image retrieval and object categorization. *TIP*, 2012.

[28] J. Van De Weijer, C. Schmid, J. Verbeek, and D. Larlus. Learning color names for real-world applications. *TIP*, 2009.

[29] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[30] J. Wang, Z. Chen, and Y. Wu. Action recognition with multiscale spatio-temporal contexts. In *CVPR*, 2011.

[31] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.