SALIENCY-BASED SELECTION OF SPARSE DESCRIPTORS FOR ACTION RECOGNITION

Eleonora Vig[1], *Michael Dorr*[2], *and David D. Cox*[1]

[1] The Rowland Institute at Harvard
[2] Schepens Eye Research Institute
Harvard University
Cambridge, MA

ABSTRACT

Local spatiotemporal descriptors are being successfully used as a powerful video representation for action recognition. Particularly competitive recognition performance is achieved when these descriptors are densely sampled on a regular grid; in contrast to existing approaches that are based on features at interest points, dense sampling captures more contextual information, albeit at high computational cost. We here combine advantages of both dense and sparse sampling. Once descriptors are extracted on a dense grid, we prune them either randomly or based on a sparse saliency mask of the underlying video. The method is evaluated using two state-of-the-art algorithms on the challenging Hollywood2 benchmark. Classification performance is maintained with as little as 30% of descriptors, while more modest saliency-based pruning of descriptors yields improved performance. With roughly 80% of descriptors of the Dense Trajectories model, we outperform all previously reported methods, obtaining a mean average precision of 59.5%.

Index Terms— Action Recognition, Saliency Maps, Space-time Image Descriptors, Sparse Representations

1. INTRODUCTION

In recent years, "Bag of Features" (BoF) approaches to representing videos have been applied with increasing success to the problem of human action recognition in video sequences. To provide such a representation, current approaches (e.g. [1, 2, 3]) rely on the extraction of local feature descriptors at certain locations in the video and use discriminative methods to tell one action from the others based on this description (for a review see [4]). Such local descriptors can be sampled either densely across the entire scene or at spatiotemporal interest points only. However, the number of stable interest points (obtained, for example, with the Harris 3D corner detector) is typically small, and therefore current approaches typically employ dense sampling. Such sampling facilitates robust recognition, but it also presents serious challenges: processing such a large number of — often irrelevant — features is computationally intractable, making it hard to meet real-time constraints with limited resources. An ideal approach would thus use only "the right" subset of densely sampled descriptors.

In order to combat the enormous computational load associated with producing a dense, in-depth analysis of visual scenes, the human visual system has evolved a highly effective strategy of *space-variant* processing. In this strategy a coarse processing stage is employed to identify potentially relevant ("salient") scene regions, and only these regions are processed in full detail. A variety of techniques inspired by biological attentional processing have been proposed (e.g. [5, 6, 7]); these methods attempt to encode relevant scene locations from incoming inputs and can thus be used as a "mask" to filter out irrelevant parts of a video.

In the present work, we explore the potential role of sampling sparsity and saliency masking in activity recognition by densely extracting feature descriptors and applying a saliency mask derived from the underlying video to prune the set of all descriptors. To compute a saliency mask, we use a generic saliency model, which is based on the geometrical invariants of the structure tensor. In Sec. 2, we give a summary of its main computational steps. We demonstrate the advantage of utilizing a pruned descriptor set over a complete one on the challenging Hollywood2 benchmark [8] and for two state-of-the-art action recognition algorithms. Surprisingly, classification performance can be roughly maintained after discarding as much as 70% of the descriptors at random. With less aggressive selection of descriptors based on saliency, we improve recognition beyond the currently best published results.

2. SALIENCY BY INTRINSIC DIMENSIONALITY

To generate saliency masks for a video, we here employ a generic yet powerful model for bottom-up visual attention derived from the simple assumption that the degree of local intensity variation is correlated with the informativeness (or *saliency*) of a video region [7]. The concept of *intrinsic dimensionality* (iD) measures this degree and yields a basic description of how a multidimensional signal may change. Typical video structures can be characterized based on the geo-

metrical invariants H, S, and K:

$$\begin{aligned} H &= 1/3\ \mathrm{trace}(\mathbf{J}) \\ S &= M_{11} + M_{22} + M_{33} \\ K &= |\mathbf{J}| \end{aligned} \tag{1}$$

of the structure tensor **J**

$$\mathbf{J} = \int_{\Omega} \nabla f \otimes \nabla f \ \mathrm{d}\Omega = \int_{\Omega} \begin{bmatrix} f_x^2 & f_x f_y & f_x f_t \\ f_x f_y & f_y^2 & f_y f_t \\ f_x f_t & f_y f_t & f_t^2 \end{bmatrix} \mathrm{d}\Omega \ , \tag{2}$$

where the integral over Ω is a Gaussian smoothing function and f_x, f_y, and f_t represent partial derivatives of the video intensity function $f(x, y, t)(f : \mathbb{R}^3 \rightarrow \mathbb{R})$. M_{ii} in Eq. 1 stand for the minors of **J**. The geometric invariants describe uniform regions ($H = 0, iD = 0$), spatiotemporal edges ($H > 0, iD = 1$) or corners ($S > 0, iD = 2$), and transient corners ($K > 0, iD = 3$).

The above formulation can be readily extended to incorporate multispectral information. When combined with machine learning methods that operate on multiscale representations, this model outperforms more complex saliency models in predicting eye movements in naturalistic videos [7]. Moreover, video regions of higher intrinsic dimensionality (those encoded by S or K) have been shown to be more predictive for human gaze [7].

3. ACTION RECOGNITION WITH A PRUNED DESCRIPTOR SET

We demonstrate the advantage of pruning the descriptor set and salient masking on two competitive action recognition algorithms that employ the same BoF framework and processing pipeline described by Wang et al. [1]. Both approaches implement dense descriptor sampling and only differ in the descriptor types they extract. In the following, we briefly review these descriptors and the common processing pipeline.

3.1. HOGHOF with dense sampling

Among existing space-time features, HOGHOF [9] descriptors combined with dense sampling have shown to provide good results for action recognition [1]. HOGHOF captures both static appearance (Histograms of Oriented Gradients) and motion (Histograms of Optical Flow) information. We use the online available executables [1] and default toolbox parameters to obtain such descriptors: the 3D patch around an interest point is divided into a $3 \times 3 \times 2$ space-time grid, and for each grid cell 4-bin HOG and 5-bin HOF descriptors are extracted and concatenated. Descriptors are computed on a dense grid (50% overlap, the minimum grid spacing is 18×18 pixels and 10 frames), on 6 spatial and 2 temporal scales.

[1] `http://www.irisa.fr/vista/Equipe/People/Laptev/download.html#stip`

3.2. Dense Trajectories

The second method extracts Dense Trajectories [2] by tracking densely sampled points with the help of an optical flow field. Along the trajectory, four different descriptors are computed: the trajectory shape (represented by normalized relative point coordinates), HOG, HOF, and Motion Boundary Histograms (MBH). MBH contains the gradients of the horizontal and vertical optical flow components (quantized in 8-bin histograms), and thus encodes the relative pixel motion and is more robust to camera motion. The online available code [2] and default parameters are used to extract such descriptors: a trajectory length of 15 frames, dense grid spacing of 5 pixels, and the 3D patch around the interest point subdivided into a $2 \times 2 \times 3$ grid.

3.3. Bag of Features (BoF) representation

To test whether the original recognition performance can be maintained or even increased with a carefully pruned descriptor set, we employ the original evaluation framework of Wang et al. (HOGHOF model) [1]. The second algorithm deviates only marginally from this baseline. Wang et al.'s action classification setting is based on the standard Bag of Features (BoF) video representation [10]. A subset of the training descriptors (100k) is first clustered into codebooks using k-means (8-times random seed initialization). As opposed to the original approach, and consistent with the Dense Trajectories algorithm, we construct codebooks for each descriptor type (HOG, HOF, MBH, etc.) separately. In the case of Dense Trajectories, this modification improves prediction performance from 53% mean Average Precision to 58%. The number of codebooks is set to 4000. Descriptors are then quantized according to the obtained dictionary and a video is represented by the resulting codebook-frequency histograms. A non-linear SVM with a χ^2 kernel is finally used to classify different human actions.

3.4. Descriptor selection based on a salient mask

We extend the standard BoF framework by incorporating an attentional filtering phase. After descriptor extraction and prior to codebook generation, we filter all descriptors based on a saliency mask of the underlying video. In our formulation, the geometrical invariants H, S, and K serve as simple but generic saliency models. In the implementation of the geometrical invariants, we chose 15-tap spatiotemporal binomials for Ω. Pooling over such a large spatiotemporal neighbourhood is desirable because we here are not interested in relevant points, but regions. This, however, also leads to quite large spatiotemporal borders; in the current model, spatial borders get extended by copying the nearest valid pixel, but temporal borders are left blank. We compute **J** for a spatially downsampled version of the input video (subsampling

[2] `http://lear.inrialpes.fr/software`

by a factor of two). Output invariants are normalized to $[0, 1]$ by taking the eighth root and linearly scaling the maximum in a frame to 1.

Finally, we prune the descriptor set based on the saliency mask. We sample descriptors with a probability that is a function of raw saliency values $x \in [0, 1]$ using the cumulative distribution function (CDF) of the Weibull distribution

$$F(x; k, \lambda) = 1 - e^{-(x/\lambda)^k}, \quad (3)$$

with $k > 0$ shape parameter and $\lambda > 0$ scale parameter. Varying λ enables us to increase or decrease the coverage of the mask. k is set to a high number to ensure relatively hard margins between salient and non-salient video regions.

To distinguish between the effects of a reduced set of descriptors and the effects of selecting descriptors only at salient locations, we also randomly selected subsets of descriptors of varying size.

4. RESULTS

To allow for a direct comparison of results reported in the original papers, we evaluate both discussed algorithms (with and without descriptor pruning) on the challenging Hollywood2 benchmark [8]. The Hollywood2 action recognition data set consists of more than 5 hours of video data collected from 69 different Hollywood movies and split into 823 training and 884 test video clips. The set contains 12 actions (AnswerPhone, DriveCar, Eat, FightPerson, GetOutCar, HandShake, HugPerson, Kiss, Run, SitDown, SitUp and StandUp) and video clips may be multiply labeled. The performance measure reported on this data set is the mean Average Precision (mAP), computed by taking the mean of the AP of 12 binary classifiers.

By varying the parameter λ in Eq. 3, we systematically evaluate recognition performance for a range of different mask sizes. For a summary of results, see Table 1. Mask coverage is systematically increased from preserving only 20% of descriptors to keeping all 100%. Figure 1 shows the resulting mean AP values for both algorithms. In case of the HOGHOF descriptors (Figure 1(a)), recognition performance can be maintained with as little as 30% of descriptors for random pruning and all three types of saliency masks. With only 50% of the descriptors selected through salient masking, recognition performance improves by 2.5% beyond that obtained with the full descriptor set (see Table 1). Note that in [1], Wang et al. reported a lower mAP of 47.7% for the densely sampled HOGHOF compared to our 50.0%. Our improvement is due to separate codebook generation for HOG and HOF descriptors, as suggested by [2]. Also note that temporal border effects in the saliency maps reduce the number of all descriptors in the data set to around 65%. This explains the lack of results (for saliency masking) in the right half of Figure 1(a). The rather large spatiotemporal kernels (for Ω) and low spatial scale make the qualitative results for the three geometrical invariants differ only marginally. Therefore, we only consider the slightly better performing invariant S in the remainder of this analysis.

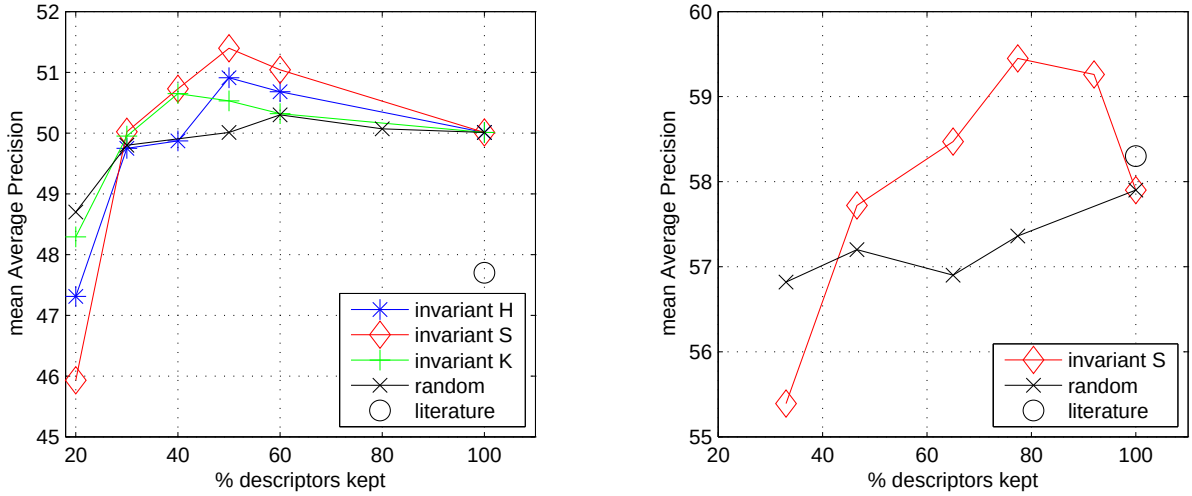
In Figure 1(b), a similar trend can be observed for the Dense Trajectories model. Recognition performance is maintained with less than half of the descriptors even with random sampling. Moreover, a more moderate pruning of the descriptor set based on saliency leads to better recognition than that obtained with the full descriptor set. With 78% of descriptors of the Dense Trajectories model, we obtain a mean Average Precision of 59.5% and thus outperform previously published results.

Table 1. Mean Average Precision on the Hollywood2 data set. Reproduced recognition performance can be maintained with 30-45% of the descriptors pruned either randomly or through a saliency mask ("sparse" case). Moreover, recognition is improved upon when using a saliency mask and more moderate pruning (i.e. by preserving 50-78% of the descriptors; "best (saliency)" case).

	HOGHOF	Dense Trajectories
literature	47.7	58.3
reproduced	50.0	57.9
sparse (saliency)	50.0 (30%)	57.7 (45%)
sparse (random)	49.8 (30%)	57.2 (45%)
best (saliency)	**51.4** (50%)	**59.5** (78%)
best (random)	50.3 (60%)	57.3 (78%)

5. DISCUSSION AND CONCLUSION

Often only small parts of the visual scene contain information relevant to action recognition. Thus, processing the whole scene everywhere in full detail (corresponding to dense sampling) leads to a large amount of irrelevant and possibly distracting data. Biological evolution solved this problem for primates — which are highly skilled at interpreting natural scenes — by developing a small high-resolution center (*fovea*) of the retina that is moved around several times per second by the oculomotor system. Inspired by this, we have here used a saliency algorithm to reduce the processing only to the most informative parts of the visual scene. Indeed, our results show that a large proportion of the descriptors are not necessary for classification. Even a random pruning by as much as 70% of densely sampled descriptors did not substantially impair performance, which is surprising given the optimized sampling grid density [1]. This is especially critical on large and challenging data sets such as Hollywood2. For the Dense Trajectories model [2], about 200 GB of descriptor data must be stored and processed, and manipulation of such a large volume of data is time-consuming even on large computing clusters.



(a) HOGHOF descriptors

(b) Dense Trajectories

Fig. 1. Mean Average Precision for various saliency mask sizes and random subsampling factors. The trend is consistent for both the HOGHOF descriptors and Dense Trajectories: only few (30-40%) space-time descriptors are enough to maintain recognition level and a more moderate pruning based on saliency improves recognition over the baseline (i.e. when no descriptors were discarded).

When pruning was guided by a saliency mask, recognition performance even increased for moderate amounts of pruning. It is important to note that we varied the coverage of the saliency mask, but kept the original sampling density in salient regions. The results of random pruning indicate that a further reduction in the number of descriptors might be obtainable by a lower sampling density even in the salient regions.

However promising our results are, a simple bottom-up saliency algorithm cannot be expected to fully predict optimal attentional orienting in complex natural scenes. Future work exploring more sophisticated saliency algorithms holds the potential to improve results even further.

6. ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation (IIS 0963668). Eleonora Vig was supported by a fellowship within the Postdoc-Programme of the German Academic Exchange Service (DAAD, D/11/41189).

7. REFERENCES

[1] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2009, p. 127.

[2] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin, "Action Recognition by Dense Trajectories," in *IEEE CVPR*, 2011, pp. 3169–3176.

[3] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *IEEE CVPR*, 2011, pp. 3361–3368.

[4] R. Poppe, "A survey on vision-based human action recognition," *Image Vision Comput.*, vol. 28, pp. 976–990, 2010.

[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on PAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[6] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in NIPS 18*. 2006.

[7] E. Vig, M. Dorr, T. Martinetz, and E. Barth, "Intrinsic dimensionality predicts the saliency of natural dynamic scenes," *IEEE Trans. on PAMI*, 2011, (in press).

[8] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *IEEE CVPR*, 2009, pp. 2929 –2936.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE CVPR*, 2008, pp. 1 –8.

[10] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *ICPR*, 2004, vol. 3, pp. 32–36.