Space-Variant Descriptor Sampling for Action Recognition Based on Saliency and Eye Movements

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

[1] The Rowland Institute at Harvard, Cambridge, MA 02142, USA
{vig,cox}@rowland.harvard.edu
[2] Schepens Eye Research Institute, Harvard Medical School, Boston, MA 02114, USA
michael.dorr@schepens.harvard.edu

Abstract. Algorithms using "bag of features"-style video representations currently achieve state-of-the-art performance on action recognition tasks, such as the challenging Hollywood2 benchmark [1,2,3]. These algorithms are based on local spatiotemporal descriptors that can be extracted either sparsely (at interest points) or densely (on regular grids), with dense sampling typically leading to the best performance [1]. Here, we investigate the benefit of space-variant processing of inputs, inspired by attentional mechanisms in the human visual system. We employ saliency-mapping algorithms to find informative regions and descriptors corresponding to these regions are either used exclusively, or are given greater representational weight (additional codebook vectors). This approach is evaluated with three state-of-the-art action recognition algorithms [1,2,3], and using several saliency algorithms. We also use saliency maps derived from human eye movements to probe the limits of the approach. Saliency-based pruning allows up to 70% of descriptors to be discarded, while maintaining high performance on Hollywood2. Meanwhile, pruning of 20-50% (depending on model) can even improve recognition. Further improvements can be obtained by combining representations learned separately on salience-pruned and unpruned descriptor sets. Not surprisingly, using the human eye movement data gives the best mean Average Precision (mAP; 61.9%), providing an upper bound on what is possible with a high-quality saliency map. Even without such external data, the Dense Trajectories model [2] enhanced by automated saliency-based descriptor sampling achieves the best mAP (60.0%) reported on Hollywood2 to date.

Keywords: action recognition, saliency maps, eye movements, bag of features, descriptor pruning.

1 Introduction

Action recognition performance critically depends on the choice of video representation. Bag of Features (BoF) video representations [4] view each image sequence as a collection of space-time descriptors extracted at certain locations

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part VII, LNCS 7578, pp. 84–97, 2012.

in the video. These descriptors capture spatial appearance and motion properties, and are usually sampled either densely over the entire scene or at space-time interest points. Despite the initial success of well-known interest point detectors, such as the Harris3D corner detector [5], the Cuboid [6], or the Hessian, the current trend is towards dense sampling [1,2,3]. Indeed, compared to the typically small number of stable interest points, dense sampling offers a richer description of the scene and also captures contextual information from which discriminative methods (operating on these representations) greatly benefit. However, when processing the entire scene everywhere in detail, serious challenges must be faced: a huge amount of — possibly irrelevant and distracting — data needs be processed often in real-time and with restricted computing resources. This may turn such brute-force dense sampling approaches computationally intractable.

As a way to control the combinatorial explosion inherent in an in-depth processing of the visual environment, the human visual system has evolved highly efficient mechanisms that restrict the visual processing to behaviorally relevant scene locations only. Such *space-variant* processing involves a fast coarse stage, in which potentially important, *salient* scene areas are identified; the visual information of only these regions is then processed exhaustively. The prediction of visually salient areas, stored in topographical *saliency maps*, has a long history in the vision sciences (e.g. [7,8,9]). Saliency-based approaches have been successfully applied to various computer vision tasks, too, such as image compression [10], quality assessment, and object recognition [11,12,13]. In these studies, saliency maps are used to emphasize task-relevant regions by filtering out irrelevant parts of the image.

In this paper, we investigate the potential of such space-variant saliency-based processing for the task of action recognition. We propose to *prune* the set of densely extracted descriptors based on a saliency mask of the underlying video. We consider several different saliency models (see Fig. 1) that range in complexity from simple, single-parameter ones, such as a central mask, to more complex, biologically more plausible ones. The optimal predictor of visual saliency, however, is still based on data from human viewers. Hence, in order to build a ground truth "empirical" saliency measure, we collected eye movements on the training and test video samples of the Hollywood2 benchmark [14].

We demonstrate the advantages of an attentional selection stage for three competitive action recognition algorithms that employ either hand-designed descriptors [1,2] or features learned in an unsupervised manner [3]. Several interesting findings that are remarkably similar for all three algorithms emerge from this analysis. First, we show that action recognition performance can be maintained with as little as 30% of the densely extracted descriptors selected at random. Second, we find that a more modest pruning of the descriptors based on visual saliency improves recognition performance, even beyond the currently best published results. On the professionally-edited videos of the challenging Hollywood2 benchmark, a simple central mask proves to be highly beneficial: with the Dense Trajectories model [2] and a central mask, we obtain a mean average precision (mAP) of 60.0% (vs. the best previously reported mAP of

58.3% [2]). Conversely, we find that masks including only peripheral information reduce recognition performance.

To explore the limits of saliency-based masking, we measured the eye movements of human observers watching the benchmark videos, and used this data to produce an "empirical" saliency map. As expected, the resulting saliency map enabled higher performance with even sparser feature representations. Because human observers during the eye-tracking experiment performed the same action recognition task as the computer vision algorithms, it is reasonable to assume that they looked at those image regions that are most informative for the task. Because of human top-down knowledge, performance of the empirical saliency map cannot be compared directly; however, these data are useful to probe the limits of saliency-guided descriptor pruning and, ultimately, the limits of the state-of-the-art algorithms. The eye tracking data from these experiments are publicly available.[1]

Finally, we combine the pruned saliency-based representations described above with the original, unpruned representations, via feature concatenation and Multiple Kernel Learning blending. This step is motivated by the observation that humans often rely on low-resolution, full-field "gist" information in addition to high-resolution information at the center of gaze. Such blending yields even better performance, at the expense of somewhat higher computational cost.

For alternative work, published in this conference, on eye movement data collection and saliency models for action recognition see [15].

2 Salience-Based Masking of Descriptors

Here, we investigate how much action recognition models benefit from an attentional filtering phase incorporated in the standard Bag of Features architecture [4]. This architecture treats each video sample as an orderless collection of local descriptors that are extracted either at regular grid points or at space-time interest points. Next, the descriptors are quantized into codebook-frequency histograms according to a pre-learned dictionary. The dictionary is derived by clustering the descriptors of training video samples into fewer codebook vectors (usually with k-means). Finally, a non-linear Support Vector Machine with a χ^2 kernel is trained on the codebook-frequency histograms and is used to classify human actions.

2.1 Action Recognition Algorithms

An attentional masking stage is incorporated into three state-of-the-art action recognition algorithms that employ a common processing pipeline (based on the BoF framework) described by Wang et al. [1]. All three models sample their descriptors densely and differ only in the feature extraction stage.

[1] `http://www.coxlab.org/resources/hw2_eye_movement/`

When combined with dense sampling, **(1) HOGHOF** [16] descriptors provide good results on a variety of action recognition benchmarks [1]. HOGHOF characterizes both static appearance (through Histograms of Oriented Gradients) and motion properties (with Histograms of Optical Flow).

The algorithm of Wang et al. [2], dubbed **(2) Dense Trajectories**, extracts several different descriptors along trajectories of densely sampled interest points. Optical flow fields are used to extract the dense trajectories, and each trajectory is described by four types of descriptors: the shape of the trajectory, HOG, HOF, and Motion Boundary Histograms (MBH).

Many existing algorithms for object and action recognition rely on manually-designed features such as those mentioned above. More recently, unsupervised feature learning has been shown to deliver highly competitive results in various computer vision tasks (e.g. [17]). For action recognition, Le et al. [3] introduced a two-layered Independent Subspace Analysis (ISA) algorithm — the **(3) Stacked Convolutional ISA** — that learns spatiotemporal features of interest points from unlabeled videos. Deep learning techniques, such as convolution and stacking, were employed to scale the algorithm to large images and learn hierarchical representations.

2.2 Saliency Masks

We here investigate three different saliency models of varying biological plausibility, and further investigate three deliberately unbiological control models.

Central Masking. Filmmakers often place the subject of interest in the center of the video (e.g. [18,19]). Therefore, a simple — but surprisingly highly predictive — saliency map is one that considers merely the spatial distance of each pixel to the video center.

Analytical Saliency Mask. To compute an analytical saliency mask, we consider a simple but powerful model built on the *structure tensor* and its *geometric invariants*. The structure tensor has been used extensively in image processing e.g. for interest point detection (Harris detector [5]), motion and orientation estimation [20], and occlusion detection. For a video viewed as a function of space (x and y) and time t ($f : \mathbb{R}^3 \to \mathbb{R}$), the structure tensor is defined as

$$\mathbf{J} = \int_{\Omega} \begin{bmatrix} f_x^2 & f_x f_y & f_x f_t \\ f_x f_y & f_y^2 & f_y f_t \\ f_x f_t & f_y f_t & f_t^2 \end{bmatrix} \mathrm{d}\Omega \ , \tag{1}$$

where the integral over Ω can be implemented as a spatiotemporal Gaussian smoothing function and f_i stand for the first order partial derivatives.

Recently, $\mathbf{J}$'s geometric invariants have been shown to predict well visual saliency and eye movements on videos [9]. With the help of the invariants one can characterize typical video structures in terms of the *intrinsic dimensionality* of the local video signal. The invariants are defined as

$$\begin{aligned} H &= 1/3 \text{ trace}(\mathbf{J}) \\ S &= M_{11} + M_{22} + M_{33} \\ K &= |\mathbf{J}| \end{aligned} \tag{2}$$

where M_{ij} are minors of $\mathbf{J}$. We here only consider the invariant S, which encodes transient edges and corners and stationary corners; S has been shown to predict human gaze better than H, and is less sparse than K [9]. Since we are more interested in relevant regions, rather than (the only few) interest points, pooling (with Ω) is done over a large spatiotemporal neighborhood.

Empirical Saliency Mask. While analytical saliency maps (derived from low-level image properties) have proven useful for predicting relevant scene regions (e.g. [7,8,9]), a "ground truth" saliency map for a given video can be determined by measuring where human observers actually look in the video. To create such a map, eye trackers are usually employed to collect eye movements of human viewers. When dealing with gaze data, one needs to compensate for possible imprecisions in both the eye tracking and the human visual system. Also, eyes are typically directed at particular regions of interest, not at single points. In vision science, *fixation density maps* [21] have been proposed as means to convert the raw fixation data into an *empirical saliency map*. A fixation density map is constructed by the superposition of spatiotemporal Gaussians centered at each gaze sample of all subjects, normalized to a probability density map in which regions of interest of human observers have higher probabilities.

Peripheral Masking. Capturing rich contextual information is one of the benefits of the BoF video representation. However, encoding irrelevant descriptors can also add noise and thus degrade recognition performance. Under the assumption that Hollywood directors put objects of interest roughly in the center of the screen, we created a control condition by inverting the central mask, thus preserving descriptors only from the periphery.

Random Uniform Sampling. To test whether performance can be preserved with significantly fewer descriptors sampled uniformly from the whole scene rather than locally at salient locations, we picked descriptors from the dense sampling grid at random.

Randomly-Offset Empirical Saliency Mask. As a third control condition, we introduced random spatial offsets of the empirical saliency map for each movie sample (with 2D toroidal wrapping around the borders). Such randomly-offset empirical saliency maps simulate random gaze patterns: instead of uniformly sampling from the whole scene, only small patches are sampled densely. This condition preserves the rough spatiotemporal statistics of human eye movements, while disrupting their relationship with the content of the movies.

2.3 Descriptor Pruning

In this paper, we extend the standard BoF pipeline to allow for an attentional filtering of the irrelevant scene regions. Once descriptors are extracted on a dense grid, we prune the descriptor set based on one of the above saliency masks of the underlying video.

To this end, we sample descriptors with a probability given by the cumulative distribution function (CDF) of the Weibull distribution

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

where $x \in [0, 1]$ is the raw saliency value, and $k > 0$ and $\lambda > 0$ are the shape and scale parameters of the Weibull function. The coverage of the mask can be increased or decreased by varying the parameter λ. We use the Weibull distribution because its parameters can flexibly modify the probability with which descriptors are sampled from video regions of different saliency (as opposed to only thresholding the saliency map).

2.4 Feature Combination

As mentioned earlier, one disadvantage of extracting descriptors only at salient locations is the possible loss of contextual information. We therefore also investigate the effect of not pruning, but complementing the original, densely extracted BoF representations with saliency-based descriptor sets. Feature combination is performed in two ways:

1. An extended BoF representation is generated by *concatenating* the two codebook-frequency histograms. A standard SVM with χ^2 kernel operates on these new feature vectors of doubled dimensionality.
2. Adaptive feature combination is performed with Multiple Kernel Learning (MKL) techniques [22,23]. MKL optimizes jointly over a linear combination of Q kernels $K^* = \sum_{q=1}^{Q} \beta_q K_q$ and the SVM parameters $\alpha \in \mathbb{R}^n$ and $b \in \mathbb{R}$. In our case $Q = 2$, i.e. separate χ^2 kernels are assigned to the BoF representations derived from densely sampled and saliency-filtered descriptor sets.

3 Experimental Setup

3.1 Data Set

To comply with the common evaluation scheme of the three considered models, we evaluate the proposed pruning methods on the challenging Hollywood2 benchmark [14]. For this action recognition set, almost 6 hours of video data were collected from 69 different Hollywood movies and split into 823 training and 884 test video clips. The set contains 12 actions and video clips may have

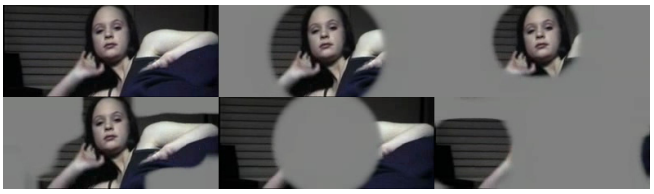


Fig. 1. Example screenshot of five (out of six) different masks used to prune the densely sampled descriptor set. From left to right, top row: unmasked ("sit up" action), center mask, empirical saliency mask (fixation density map based on eye traces of three viewers), bottom row: analytical saliency mask (invariant S), peripheral mask, and offset empirical saliency mask. Not shown is a random sampling of descriptors. For video examples, see supplementary material.

multiple labels. Hollywood2 captures well the many challenges in action recognition: it is characterized by high intra-class variability, high degree of clutter and camera motion, and clips are often ambiguously labeled.

Binocular eye movements were recorded on this video set at 1000 Hz from three subjects using a commercially available video-oculographic eye tracker (SR Research EyeLink 1000). Subjects were seated 75 cm away from a ViewSonic VX2265wm TFT monitor running at 120 Hz, and their heads were stabilized in a head rest. At this distance, the entire screen with a spatial resolution of 1680 by 1050 pixels spanned 35 by 22 degrees of visual angle; videos were scaled to cover the full screen (in "letterbox" format where necessary). At the beginning of each recording session, the eye tracker was calibrated with the manufacturer's five-point calibration routine. Then, gaze was recorded while video clips were shown in random order. Typical recording sessions lasted about 15 to 20 minutes; regular drift corrections are not necessary with this eye tracker, as was validated in previous experiments. Subjects were familiar with the action classes, and their task was to 'identify the action(s)'.

3.2 Implementation

Executables that are available online and default toolbox parameters were used to obtain HOGHOF and Dense Trajectory (DT) descriptors. In case of the Stacked ISA model, for the full analysis (including codebook generation, SVM evaluation, etc.), we used the original implementation and default parameter settings. For a full listing of parameters also for invariant S, see supplementary material. Examples of the different masks, overlaid on individual video frames, are rendered in Fig. 1. For the evaluation, we also strictly followed the processing pipeline of Wang et al. [1].

4 Experimental Results

For all three action recognition models, we systematically investigated how the various masks and different mask sizes affect action recognition performance on

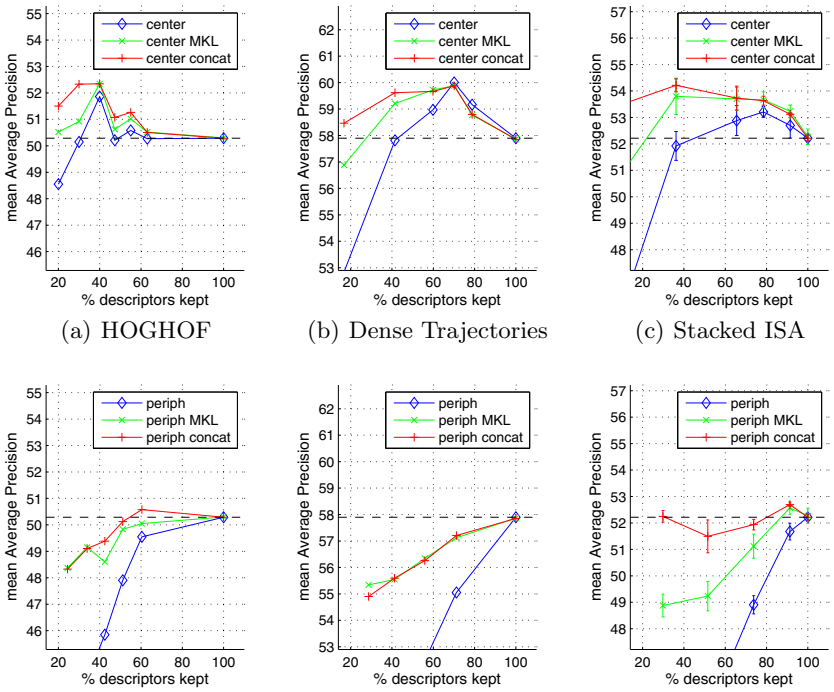
Table 1. Best mean Average Precision on Hollywood2 for the three state-of-the-art action recognition algorithms and the different masks (used either exclusively or to emphasize relevant regions)

	HOGHOF	Dense Trajectories	Stacked ISA
literature	47.7%	58.3%	53.3%
baseline (reproduced)	50.3%	57.9%	52.3%
central mask	52.3%	60.0%	54.2%
peripheral mask	48.2%	57.2%	52.7%
analytical saliency	52.1%	59.4%	54.4%
random sampling	51.0%	58.0%	52.9%
empirical saliency	54.5%	61.9%	55.2%
offset emp. saliency	50.8%	57.4%	52.6%

the Hollywood2 benchmark. By varying the value of the Weibull scale parameter λ, we gradually increased mask coverage from retaining as little as 10-20% of descriptors to preserving all 100%, which corresponds to the baseline (reproduced) performance. Our results are summarised in Table 1. As we shall see below, several interesting patterns clearly emerge from this analysis.

Central and Peripheral Masking. Recognition performance for central and peripheral masking is shown in Fig. 2. There, we plot the mean Average Precision (mAP, i.e. the mean of the AP of 12 binary classifiers) for various mask sizes (blue curves). Whereas a central mask (top row) achieves baseline performance — dashed horizontal line — with as little as 30-40% of the descriptors, peripheral masking drastically deteriorates recognition (blue curves below the dashed line in bottom row of Fig. 2). Moreover, a more modest filtering of the descriptors with the central mask leads to substantially better recognition than that obtained with the unpruned descriptor set (see blue peaks in the upper row of Fig. 2). These observations are highly consistent among the three models; note, however, the different mAP ranges along the vertical axis due to the different predictive power of the models. With 58.3% mAP reported, Dense Trajectories is, to our knowledge, the best performing algorithm in the literature, and our reproduced baseline (57.9% mAP) comes close to this value. However, with only about 70% of the Dense Trajectories descriptors preserved with a central mask, we obtain 60.0% mAP and thus outperform all previously published results. The good performance of a simple center mask on Hollywood2 can be explained by the filmmaker's bias to place actors and objects of interest near the center of the video.

Next, we examine the effect of complementing the BoF representations obtained for full descriptor sets with the ones of the pruned sets. Feature combination turns out to be beneficial: both simple descriptor concatenation (red lines) and MKL (green lines) deliver improved performance compared to the "mask only" condition (blue line). There are, however, three additional interesting patterns. First, the improvement is more prominent with an "aggressive" selection



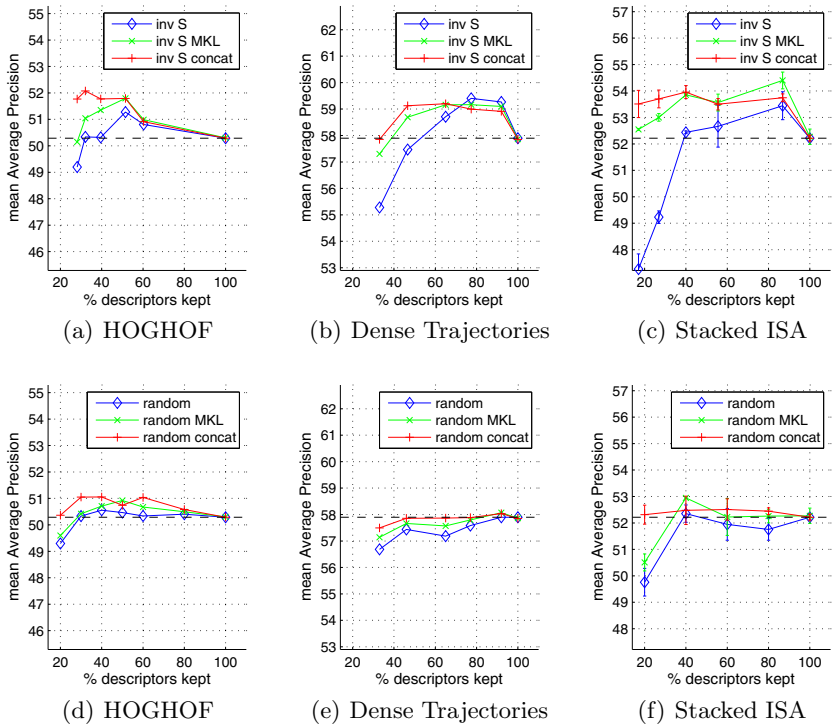
(a) HOGHOF (b) Dense Trajectories (c) Stacked ISA

(d) HOGHOF (e) Dense Trajectories (f) Stacked ISA

Fig. 2. Mean Average Precision for three action recognition models (dashed horizontal lines show reproduced results — without masking — as baseline) when descriptors are pruned either with a *central* (top row) or *peripheral* (bottom row) mask of systematically varied size. For all models, using only central descriptors (blue line, top) improves performance (up to a mAP of 60%), while peripheral descriptors (blue line, bottom) impair performance. When unpruned and saliency-based descriptor sets are combined (red lines - concatenation, green lines - multiple kernel learning), results get robustly better for central and less impaired for peripheral masking

of descriptors, and decreases as the mask coverage increases, i.e. the codebooks of the unmasked and pruned sets become more similar. Second, even with feature combination, peripheral masking is still inferior to the baseline, indicating that the irrelevant peripheral descriptors only add noise to the classifier. Third, training a single SVM on the concatenated codebook-histogram representations derived from unpruned and masked descriptor sets performs slightly better than MKL.

Note that compared to our 50.3% baseline, a lower mAP of 47.7% was reported for HOGHOF in the literature [1]. The improvement is the result of separate codebook generation for HOG and HOF descriptors, as suggested by [2]. Error bars for Stacked ISA are based on three runs automatically performed by the authors' toolbox. For the other two models it would be computationally



(a) HOGHOF (b) Dense Trajectories (c) Stacked ISA

(d) HOGHOF (e) Dense Trajectories (f) Stacked ISA

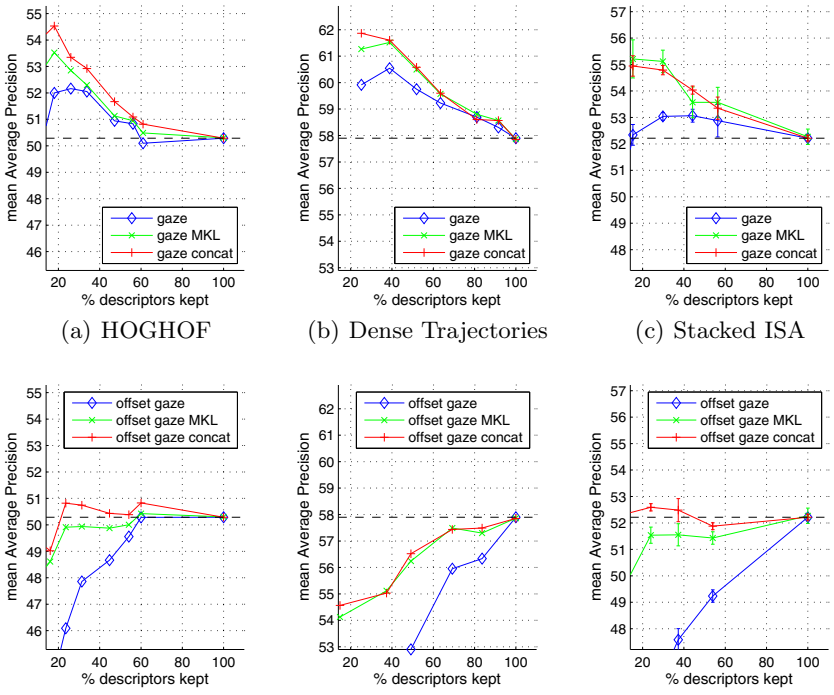
Fig. 3. Results for masking with an analytical saliency map (geometric invariant S; top row) and for various random subsampling factors (bottom row). While a small number (30-40%) of randomly picked descriptors is enough to maintain baseline level (dashed line), recognition can be improved by a more moderate pruning based on saliency.

prohibitive to run the full analysis multiple times. In previous studies [1], a standard deviation of about 0.5% was reported.

Analytical Saliency Mask. Results for masking with invariant S are shown in the upper row of Fig. 3. Note that temporal border effects reduce the number of all descriptors to around 65% in case of HOGHOF, hence the lack of results in the right half of the HOGHOF figures.

Based on S, recognition is improved beyond the baseline using fewer descriptors, but — surprisingly — not beyond what we obtain with a simple central mask. Low-level feature-based analytical saliency algorithms, however, only model bottom-up attentional processes and therefore cannot capture semantic information, which in Hollywood movies typically is found in the video center.

Random Uniform Sampling. We find that randomly discarding up to 60-70% of densely sampled descriptors does not substantially impair baseline performance (Fig. 3, bottom row). Given that descriptors are extracted from a



(a) HOGHOF (b) Dense Trajectories (c) Stacked ISA

(d) HOGHOF (e) Dense Trajectories (f) Stacked ISA

Fig. 4. Mean Average Precision when masking with an empirical saliency mask (top row) and the randomly offset version of it (bottom row). 20-40% of descriptors picked from the empirical saliency map give optimal results (blue curve peaks), and when combined with baseline features the highest performance boost is obtained. Randomly offset gaze masks, on the other hand, degrade recognition performance.

dense grid *independently* on several spatiotemporal scales, a random global subsampling will probably still include most regions of the video. Still, given that the three algorithms were optimized separately for grid density, this result is nevertheless striking.

Empirical Saliency Mask. Fig. 4 shows results for the empirical saliency mask and its randomly offset version. Highest recognition is obtained with only 20-40% of the descriptors, based on regions that human observers looked at while they performed the action recognition task. When concatenated to the baseline, the performance boost is even more pronounced: it adds up to 4%, which for Dense Trajectories leads to a mAP of about 62%. Offsetting the saliency maps corresponds to densely sampling from quasi-random regions with a random scanpath. Similar to peripheral masking, such a descriptor selection is detrimental to recognition performance.

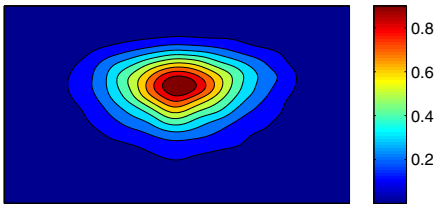


Fig. 5. Distribution of gaze on the Hollywood2 training and test videos. The normalized probability map is the result of the superposition of Gaussians at each gaze sample (of three viewers) and subsequent normalization. Eye movements on professionally-edited Hollywood movies are highly centered.

To further investigate the central bias in the Hollywood2 video set, in Fig. 5, we plotted the average empirical saliency map from all 1707 movie clips. Our results reinforce previous findings in the vision literature (e.g. [19,24,18]): eye movements on professionally-captured and edited Hollywood videos are highly centered. In addition to the cinematographer's bias for placing salient content near the center of the frame, frequent scene cuts were also found to trigger reorienting eye movements towards the center of the screen.

Saliency-Based Pruning vs. Saliency-Based Weighting. There is considerable across-run variability in the models induced by random subset selection and k-means initialization. However, it would be too time-consuming to run models repeatedly in order to evaluate whether, for any given number of descriptors, results are better for the saliency-based descriptors alone or a combination of all and salient descriptors through concatenation or multiple kernel learning. Over all models and mask types, feature vector concatenation (red curves) yields significantly higher performance than pruned descriptor sets based on saliency (blue curves; paired Wilcoxon signed rank test, 84 tuples, $p < 10^{-15}$).

5 Discussion and Conclusions

Despite rapid progress in computer vision algorithms, human performance is still unrivaled for scene understanding and complex action recognition. The human visual system uses visual attention to separate task-relevant from irrelevant regions, which greatly reduces computational demands. Earlier computer vision attempts to mimic this efficient strategy have limited processing to a small number of interest points; recently, however, dense sampling approaches have achieved better recognition performance albeit at much higher computational cost. In this paper, we reconcile these paradigms. We use several saliency models and state-of-the-art action recognition algorithms to investigate how performance is affected when either uninformative regions are discarded or salient regions are represented in more detail. Our findings were highly similar for the different recognition algorithms and confirmed that a large amount of densely

extracted descriptors is indeed unnecessary and may be even harmful for action recognition. Using only relevant descriptors is particularly critical on vast data sets such as Hollywood2 on which, with the optimal Dense Trajectories model [2] about 200 GB of descriptor data are extracted. The manipulation of such huge amounts of data is time-consuming even on large computing clusters.

Performance can be maintained with as little as 30% of descriptors picked at random, which is similar to using a sparser regular grid [1]. With a less aggressive, saliency-guided selection of the descriptors, however, recognition performance increased, even beyond the best results reported for this data set before. Nevertheless, the risk remains that image feature-based saliency algorithms may fail to identify all task-relevant, semantically meaningful image regions. To probe the limits of saliency-guided descriptor pruning, we therefore collected eye movements from human observers on the Hollywood2 videos and make this data set publicly available here. Although not a "fair" comparison, such a "ground truth" saliency map is nonetheless interesting and indeed turned out to be a superior *pruning* mask.

Probing the abilities of action recognition algorithms is only possible with large and challenging video sets. However, using readily available, professionally-made videos as in Hollywood2 is likely to have introduced a number of biases into the data set. Directors deliberately place the action in the center of the video, and thus a simple central mask provided surprisingly competitive performance. We expect that this effect would vanish in videos that lack such guided "framing," such as those acquired by surveillance cameras or autonomously navigating robots.

Finally, scene context may still provide additional information. Performance significantly increased when the saliency-based representations *complemented* the original, unpruned ones. Salient, task-relevant scene areas were represented by more codebook vectors and thus emphasized. In biological terms, such feature combination could be interpreted as having separate representations for the coarse scene *gist* and a detailed, *foveal* view of a scene.

Acknowledgments. This work was funded by the NSF (IIS 0963668) and by a Google Research Award. EV was supported by a fellowship in the Postdoc-Programme of the German Academic Exchange Service (DAAD, D/11/41189). MD was supported by the NIH (EY018664 and EY019281). Additional computer resources were provided by the INB Lübeck.

References

1. Wang, H., Ullah, M.M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC, p. 127 (2009)
2. Wang, H., Kläser, A., Schmid, C., Cheng-Lin, L.: Action recognition by dense trajectories. In: CVPR, pp. 3169–3176 (2011)
3. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: CVPR, pp. 3361–3368 (2011)

4. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: ICPR, vol. 3, pp. 32–36 (2004)
5. Laptev, I., Lindeberg, T.: Space-time interest points. In: ICCV, Nice, France (2003)
6. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72 (2005)
7. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. on PAMI 20, 1254–1259 (1998)
8. Bruce, N., Tsotsos, J.: Saliency based on information maximization. In: Advances in NIPS, vol. 18, pp. 155–162 (2006)
9. Vig, E., Dorr, M., Martinetz, T., Barth, E.: Intrinsic dimensionality predicts the saliency of natural dynamic scenes. IEEE Trans. on PAMI 34, 1080–1091 (2012)
10. Geisler, W.S., Perry, J.S.: A real-time foveated multiresolution system for low-bandwidth video communication. In: Rogowitz, B.E., Pappas, T.N. (eds.) Human Vision and Electronic Imaging: SPIE Proceedings, pp. 294–305 (1998)
11. Rutishauser, U., Walther, D., Koch, C., Perona, P.: Is bottom-up attention useful for object recognition. In: CVPR, pp. 37–44 (2004)
12. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust object recognition with cortex-like mechanisms. IEEE Trans. on PAMI 29, 411–426 (2007)
13. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to detect a salient object. IEEE Trans. on PAMI 33, 353–367 (2011)
14. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR, pp. 2929–2936 (2009)
15. Mathe, S., Sminchisescu, C.: Dynamic Eye Movement Datasets and Learnt Saliency Models for Visual Action Recognition. In: Fitzgibbon, A., Lazebnik, S., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VII, vol. 7573, pp. 842–856. Springer, Heidelberg (2012)
16. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR, pp. 1–8 (2008)
17. Pinto, N., DiCarlo, J.J., Cox, D.D.: How far can you get with a modern face recognition test set using only simple features? In: CVPR (2009)
18. Dorr, M., Martinetz, T., Gegenfurtner, K., Barth, E.: Variability of eye movements when viewing dynamic natural scenes. Journal of Vision 10, 1–17 (2010)
19. Tseng, P., Carmi, R., Cameron, I., Munoz, D., Itti, L.: Quantifying center bias of observers in free viewing of dynamic natural scenes. Journal of Vision 9 (2009)
20. Mota, C., Aach, T., Stuke, I., Barth, E.: Estimation of multiple orientations in multi-dimensional signals. In: ICIP, pp. 2665–2668 (2004)
21. Pomplun, M., Ritter, H., Velichkovsky, B.: Disambiguating complex visual information: Towards communication of personal views of a scene. Perception 25, 931–948 (1996)
22. Bach, F.R., Lanckriet, G.R.G., Jordan, M.I.: Multiple kernel learning, conic duality, and the SMO algorithm. In: ICML, New York, USA, pp. 6–13 (2004)
23. Sonnenburg, S., Rätsch, G., Schäfer, C., Schölkopf, B.: Large scale multiple kernel learning. J. Mach. Learn. Res. 7, 1531–1565 (2006)
24. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV, pp. 2106–2113 (2009)