Dynamic Eye Movement Datasets and Learnt Saliency Models for Visual Action Recognition

Stefan Mathe[1,3] and Cristian Sminchisescu[2,1]

[1] Institute of Mathematics of the Romanian Academy (IMAR)
[2] Faculty of Mathematics and Natural Science, University of Bonn, Germany
[3] Department of Computer Science, University of Toronto, Canada

Abstract. Systems based on bag-of-words models operating on image features collected at maxima of sparse interest point operators have been extremely successful for both computer-based visual object and action recognition tasks. While the sparse, interest-point based approach to recognition is not inconsistent with visual processing in biological systems that operate in "saccade and fixate" regimes, the knowledge, methodology, and emphasis in the human and the computer vision communities remains sharply distinct. Here, we make three contributions aiming to bridge this gap. First, we complement existing state-of-the art large-scale dynamic computer vision datasets like Hollywood-2[1] and UCF Sports[2] with human eye movements collected under the ecological constraints of the visual action recognition task. To our knowledge these are the first massive human eye tracking datasets of significant size to be collected for video (497,107 frames, each viewed by 16 subjects), unique in terms of their *(a) large scale and computer vision relevance, (b) dynamic, video stimuli, (c) task control, as opposed to free-viewing.* Second, we introduce novel *dynamic consistency and alignment models*, which underline the remarkable stability of patterns of visual search among subjects. Third, we leverage the massive amounts of collected data in order to pursue studies and build automatic, end-to-end trainable computer vision systems based on human eye movements. Our studies not only shed light on the differences between computer vision spatio-temporal interest point image sampling strategies and human fixations, as well as their impact for visual recognition performance, but also demonstrate that human fixations can be accurately predicted, and when used in an end-to-end *automatic* system, leveraging some of the most advanced computer vision practice, can lead to state of the art results.

1 Motivation

Recent progress in computer-based visual recognition, in particular image classification, object detection and segmentation or action recognition heavily relies on machine learning methods trained on large scale human annotated datasets. The level of annotation varies, spanning a degree of detail from global image or video labels to bounding boxes or precise segmentations of objects[3]. Such annotations have proven invaluable for performance evaluation and have also supported fundamental progress in designing models and algorithms. However, the

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part II, LNCS 7573, pp. 842–856, 2012.

annotations are somewhat subjectively defined, primarily by the high-level visual recognition tasks generally agreed upon by the computer vision community. While such data has made advances in system design and evaluation possible, it does not necessarily provide insights or constraints into those intermediate levels of computation, or deep structure, that are perceived as ultimately necessary in order to design highly reliable computer vision systems. This is noticeable in the accuracy of state of the art systems trained with such annotations, which still lags significantly behind human performance on similar tasks. Nor does existing data make it immediately possible to exploit insights from an existing working system–the human eye–to potentially derive better features, models or algorithms.

The divide is well epitomized by the lack of matching large scale datasets that would record the workings of the human visual system, in the context of a visual recognition task, at different levels of interpretations including neural systems or eye movements. The human eye movement level, defined by image fixations and saccades, is potentially the less controversial to measure and analyze. It is sufficiently 'high-level' or 'behavioral' for the computer vision community to rule-out, to some degree at least, open-ended debates on where and what should one record, as could be the case, for instance with neural systems in different brain areas. It can potentially foster links with the human vision community, in particular researchers developing biologically plausible models of visual attention, who would be able to test and quantitatively analyze their models on common large scale datasets[4, 5].

Apart from linking the human and computer vision communities, human eye movement annotations offer pragmatic potential: fixations provide a sufficiently high-level signal that can be precisely registered with the image stimuli, for testing hypotheses and for training visual feature extractors and recognition models quantitatively. Some of the most successful approaches to action recognition employ bag-of words representations based on descriptors computed at spatial-temporal video locations, obtained at the maxima of an interest point operator biased to fire over non-trivial local structure (space-time 'corners' or spatial-temporal interest points[6]). More sophisticated image representations based on objects and their relations, as well as multiple kernels have been employed with a degree of success[7–9], although it appears still difficult to detect a large variety of useful objects reliably in challenging video footage. The dominant role of sparse spatial-temporal interest point operators as front end in computer vision systems raises the question whether computational insights from a working system like the human visual system can be used to improve performance. The sparse approach to computer visual recognition is actually consistent to the one in biological systems, but the degree of repeatability and the effect of using human fixations in conjunction with computer vision algorithms in the context of action recognition have not been yet explored.

In this paper we make the following contributions:

1. We undertake a significant effort of data recording and analysis of human eye movements in the context of computer vision-based *dynamic visual action*

recognition tasks for two existing computer vision datasets, Hollywood 2[1] and UCF-Sports[2]. This dynamic data, obtained by a significant, capture and processing effort, is made publicly available to the research community at `http://www.imar.ro/clvp/datasets/eyetracking`.

2. We introduce a number of novel consistency models and algorithms, as well as relevance evaluation measures adapted for video. Our findings (see §3) suggest a remarkable degree of dynamic consistency–both spatial and sequential–in the fixation patterns of human subjects but also underline a less extensive influence of task on dynamic fixations than previously believed, at least within the class of the datasets and actions we studied.
3. By using our large-scale training set of human fixations and by leveraging static and dynamic image features based on color, texture, edge distributions (HoG) or motion boundary histograms (MBH), we introduce novel saliency detectors and show that they can be trained effectively to predict human fixations as measured under both average precision (AP), and Kullblack-Leibler spatial comparison measures. See §4.1 and table 1 for results.
4. We show that training an end-to-end automatic visual action recognition system based on our learned saliency interest operator (point 3), and using advanced computer vision descriptors and fusion methods, leads to state of the art results in the Hollywood-2 action dataset. This is, we argue, one of the first demonstrations of a successful symbiosis of computer vision and human vision technology, within the context of a very challenging dynamic visual recognition task. It shows the potential of interest point operators learnt from human fixations for computer vision. We describe models and experiments in §4.3 and give results in table 2.

2 Related Work

Datasets containing human gaze pattern annotations of images have emerged from studies carried out by the human vision community, some of which are publicly available[10, 4, 11, 12] and some that are not[13]. Most of these datasets have been designed for small quantitative studies, consisting of at most a few hundred images or videos, usually recorded under free-viewing, in sharp contrast with the data we provide, which is large-scale, dynamic, and task controlled. These studies[10, 5, 13, 11, 12, 14–16] could however benefit from larger scale natural datasets, and from studies that emphasize the task, as we pursue. See [17] for alternative work, published in this conference, on eye movement data collection and saliency models for action recognition.

The problem of visual attention and the prediction of visual saliency have long been of interest in the human vision community[16, 10, 18]. Recently there was a growing trend of training visual saliency models based on human fixations, mostly in static images (with the notable exception of [13]), and under subject free-viewing conditions[4, 19]. While visual saliency models can be evaluated in isolation under a variety of measures against human fixations, for computer vision, their ultimate test remains the demonstration of relevance within an

end-to-end automatic visual recognition pipeline. While such integrated systems are still in their infancy, promising demonstrations have recently emerged for computer vision tasks like scene classification[20] or verifying correlations with object (pedestrian) detection responses[21, 11]. An interesting early biologically inspired recognition system was presented by Kienzle et al.[13], who learn a fixation operator from human eye movements collected under video free-viewing, then learn action classification models for the KTH dataset with promising results.

In contrast, in the field of 'pure' computer vision, interest point detectors have been successfully used in the bag-of-visual-words framework for action classification[6, 1, 7, 22], but a variety of other methods exists, including random field models[23]. Currently the most successful systems remain the ones dominated by complex features extracted at interesting locations, bagged and fused using advanced kernel combination techniques[1, 7].This study is driven, primarily, by our computer vision interests, yet leverages data collection and insights from human vision. While in this paper we focus on bag-of-words spatio-temporal computer-based action recognition pipelines, the scope for study and the structure in the data are far broader. We do not see this investigation as a terminus, but rather as a first step in explaining some of the most advanced data and models that human vision and computer vision can offer at the moment.

3 Human Eye Movement Data Collection in Video. Static and Dynamic Consistency Analysis

An objective of this work is to introduce additional annotations in the form of large-scale eye movement recordings for two popular video datasets for action recognition: Hollywood-2[1] and UCF Sports[2]. The datasets span approximately 21 hours of video (around 500k frames) and cover 12, respectively 9 action classes. Our study includes two subject groups, an active group (12 subjects), asked to perform an action recognition task and a free viewing group (4 subjects), which was simply instructed to watch the video clips. More details on the recording protocol and environment can be found in our companion report[24]. Studies in the human vision community[25, 18] have advocated a high degree of agreement between human gaze patterns for subjects queried under static stimuli. We now investigate whether this effect extends to dynamic data.

Static Consistency Among Subjects: In this section, we investigate how well the regions fixated by human subjects agree on a frame by frame basis, by generalizing the procedure used in [11] for static stimuli to video data.

Evaluation Protocol: For the task of locating people in a static image, one can evaluate how well the regions fixated by a particular subject can be predicted from the regions fixated by the other subjects on the same image[11]. This measure is however not meaningful in itself, as part of the inter-subject agreement is due to bias in the stimulus itself (*e.g.* shooter's bias) or due to the tendency of humans to fixate more often at the center of the screen[18]. One can address

this issue by checking how well the fixations of a subject on one simulus can be predicted from those of the other subjects on a different, unrelated, stimulus. We expect fixations to be better predicted on similar compared to different stimuli.

We generalize this protocol for video, by evaluating inter-subject correlation on randomly chosen frames. We consider each subject in turn. For a video frame, a probability distribution is generated by adding a Dirac impulse at each pixel fixated by the other subjects followed by blurring with a Gaussian kernel. The probability at the pixel fixated by the subject is taken as the prediction for its fixation. For our control, we repeat this process for pairs of frames chosen from different videos and predict the fixation of each subject on one frame from the fixations of the other subjects on the other frame. Differently from[11], who consider several fixations per subject for each exposed image, we only consider the single fixation, if any, that our subject made on that frame, since our stimulus is dynamic and the spatial positions of future fixations are influenced by changes in the stimulus itself. We set the width of the Gaussian blur kernel to match a visual angle span of 1.5° and draw 1000 samples for both similar and different stimulus predictions. We disregard the first 200ms from the beginning of each video to remove the bias due to the initial central fixation.

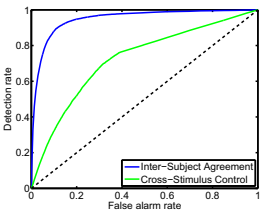
consistency measure	agreement	control
static consistency	94.8%	72.3%
temporal AOI alignment	70.8%	51.8%
AOI Markov Dynamics	70.2%	12.7%

Hollywood2

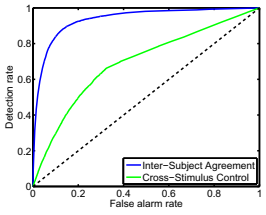
consistency measure	agreement	control
static consistency	93.2%	69.2%
temporal AOI alignment	65.4%	30.4%
AOI Markov Dynamics	55.5%	12.9%

UCF Sports

(a)



Hollywood2



UCF Sports

(b)

Fig. 1. (a) Static and dynamic inter-subject eye movement agreement for the Hollywood-2 and UCF Sports datasets. (b) ROC curves used to evaluate static inter-subject consistency. Fixations of one subject are predicted using data collected from the other subjects on the same video frame (blue) or on a frame coming from a different video (green) randomly selected from the dataset.

Findings: The ROC curves for inter-subject agreement and cross-stimulus control are shown in fig.1b. For the Hollywood-2 dataset, the area under the curve (AUC) is 94.8% for inter-subject agreement and 72.3% for cross-stimulus control. For UCF Sports, we obtain values of 93.2% and 69.2%. These values are consistent with the results reported for static stimuli by [11], with slightly higher cross-stimulus control. This indicates that shooter's bias is stronger in artistic datasets (movies) than in consumer photographs.

The Influence of Task on Eye Movements: Next, we evaluate the impact of task on eye movements from our data. For a given video frame and for each free viewing subject, we compute the p-statistic at the fixated location with respect

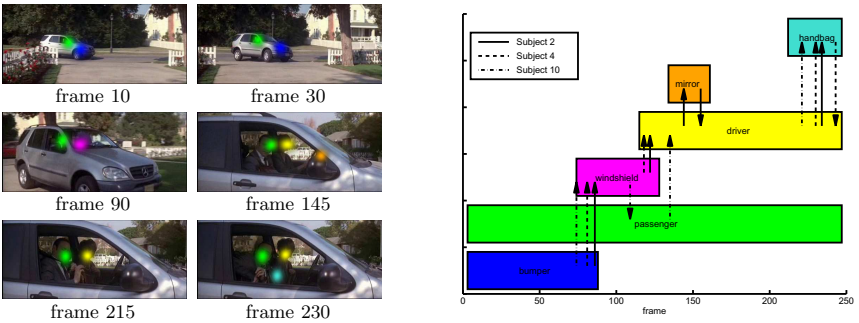


Fig. 2. Illustration for our automatic AOI generation method. Areas of interest are obtained *automatically* by clustering the fixations of subjects. **Left:** Heat maps illustrating the assignments of fixations to AOIs. The colored blobs have been generated by pooling together all fixations belonging to the same AOI and performing a gaussian blur with $\sigma = 1^{\circ}$ visual angle. **Right:** Scan path through automatically generated AOIs for three subjects. The horizontal axis denotes time. Colored boxes correspond to the AOIs generated by the algorithm. Arrows illustrate saccades which landed in a different AOI than the fixation preceeding them. Drawn according to scale. Semantic labels have been manually assigned, for visualization purposes. This figure illustrates the existence of cognitive routines centered at semantically meaningful objects.

to the probability distribution derived using fixation data from our active subjects. We repeat this process for 1000 randomly sampled frames and compute the average p-value for each subject. Somewhat surprisingly, we find that fixation patterns of our free viewers do not deviate significantly from those of active subjects ($p = 0.65$ for Hollywood-2 and $p = 0.58$ for UCF Sports). Since in the Hollywood-2 dataset several actions can be present in a video, either simultaneously or sequentially, this rules out initial habituation effects and further neglect (free viewing) to some degree.[1]

Dynamic Consistency Among Subjects: Our static inter-subject agreement analysis shows that the spatial distribution of fixations in video is highly consistent across subjects. It does not however reveal whether there is significant consistency in *the order* in which subjects fixate among these locations. To our knowledge, there are no existing agreed upon dynamic consistency measures in the community at the moment. In this section, we propose two metrics that are sensitive to the temporal ordering among fixations and evaluate consistency under these metrics. We first model the scanpath made by each subject as a

[1] Notice that our findings do not assume or imply that free-viewing subjects may not be recognizing actions. However we did not ask them to perform a task, nor where they aware of the purpose of the experiment, or the interface presented to subjects given a task. While this is one approach to analyze task influence, it is not the only possible. For instance, subjects may be asked to focus on different tasks (e.g. actions versus general scene recognition), although this type of setting may induce biases due to habituation with stimuli presented at least twice.

sequence of discrete symbols and show how this representation can be produced *automatically*. We then define two metrics, *AOI Markov dynamics* and *temporal AOI alignment*, and show how they can be computed for this representation. After we define a baseline for our evaluation we conclude with a discussion of the results.

Scanpath Representation: Human fixations tend to be tightly clustered spatially at one or more locations in the image. Assuming that such regions, called *areas of interest* (AOIs), can be identified, the sequence of fixations belonging to a subject can be represented discretely by assigning each fixation to the closest AOI. For example, from the video depicted in fig.2-left, we identify six AOIs: the bumper of the car, its windshield, the passenger and the handbag he carries, the driver and the side mirror. We then trace the scan path of each subject through the AOIs based on spatial proximity, as shown in fig.2-right. Each fixation gets assigned a label. For subject 2 shown in the example, this results in the sequence [*bumper*, *windshield*, *driver*, *mirror*, *driver*, *handbag*]. Notice that AOIs are semantically meaningful and tend to correspond to physical objects. Interestingly, this supports recent computer vision strategies based on object detectors for action recognition[7, 9, 8, 26].

Automatically Finding AOIs: Defining areas of interest manually is labour intensive, especially in the video domain. Therefore, we introduce an *automatic* method for determining their locations based on clustering the fixations of all subjects in a frame. We start by running the k-means algorithm with 1 cluster and we successively increase their number until the sum of squared errors drops below a threshold. We then link centroids from successive frames into tracks, as long as they are closely located spatially. For robustness, we allow for a temporal gap during the track building process. Each resulting track becomes an AOI, and each fixation is assigned to the closest AOI at the time of its initiation.

AOI Markov Dynamics: In order to capture the dynamics of eye movements, we represent the transitions of human visual attention between AOIs by means of a Markov process. We assume that the human visual system transitions to the next AOI conditioned on the previously visited AOIs. Due to data sparsity, we restrict out analysis to a first order Markov process. Given a set of human fixation strings $\mathbf{f}_i$, where the j^{th} fixation of subject i is encoded by the index $f_i^j \in \overline{1, A}$ of the corresponding AOI, we estimate the probability $p(s_t = b \mid s_{t-1} = a)$ of transitioning to AOI b at time t given that AOI a was fixated at time $t-1$ by counting transition frequencies. We regularize the model using Laplace Smoothing to account for data sparsity. The probability of a novel fixation sequence g under this model is $\prod_j p(s_t = g^j \mid s_{t-1} = g^{j-1})$ assuming the first state in the model, the central fixation, has probability 1. We measure the consistency among a set of subjects by considering each subject in turn, computing the probability of his scanpath with respect to the model trained from the fixations of the other subjects and normalizing by the number of fixations in his scanpath. The final consistency score is the average probability over all subjects.

Temporal AOI Alignment: Another way to evaluate dynamic consistency is by measuring how pairs of AOI strings corresponding to different subjects can be

globally aligned. Although not modeling transitions explicitly, a sequence alignment has the advantage of being able to handle gaps and missing elements. An efficient algorithm having these properties due to Needleman-Wunsch[24] uses dynamic programming to find the optimal match between two sequences $f^{1:n}$ and $g^{1:m}$, by allowing for the insertion of gaps in either sequence. It recursively computes the alignment score $h_{i,j}$ between subsequences $f^{1:i}$ and $g^{1:j}$ by considering the alternative costs of a match between f^i and g^j versus the insertion of a gap into either sequence. The final consistency metric is the average alignment score over all pairs of distinct subjects, normalized by the length of the longest sequence in each pair. We set the similarity metric to 1 for matching AOI symbols and to 0 otherwise, and assume no penalty is incurred for inserting gaps in either sequence. This setting gives the score a semantic meaning: it is the average percentage of symbols that can be matched when determinig the longest common subsequence of fixations among pairs of subjects.

Baselines: In order to provide a reference for our consistency evaluation, we generate 10 random AOI strings per video and compute the consistency on these strings under our metrics. We note however that the dynamics of the stimulus places constraints on the sampling process. First, a random string must obey the time ordering relations among AOIs (*e.g.* the passenger is not visible until the second half of the video in fig.2). Second, our automatic AOIs are derived from subject fixations and are biased by their gaze preference. The lifespan of an AOI will not be initiated until at least one subject has fixated it, even if the corresponding object is already visible. To remove some of the resulting bias from our evaluation, we extend each AOI both forward and backwards in time, until the image patch at its center has undergone significant appearance changes, and use these extended AOIs when generating our random baselines.

Findings: For the Hollywod-2 dataset, we find that the average transition probability of each subject's fixations under AOI Markov dynamics is 70%, compared to 13% for the random baseline (fig.1a). We also find that, across all videos, 71% of the AOI symbols are successfully aligned, compared to only 52% for the random baseline. We notice similar high gaps in the UCF Sports dataset. These results indicate a high degree of consistency in the dynamics of human eye movements across the two datasets.

4 Learnt Saliency Models for Visual Action Recognition

In this section, we show that it is possible to train an effective human fixation detector on our dataset and present a state-of-the-art end-to-end automatic visual action recognition system based on the saliency maps generated by our detector.

4.1 Human Saliency Map Prediction

We first show that we can effectively predict saliency maps, using several features, both static and motion based. Our analysis includes many features derived directly from low, mid and high level image information. In addition, we train our

own detector that fires preferentially at the locations fixated by the subjects, using the vast amount of eye movement data available in the dataset. We evaluate all these features and their combinations on our dataset under two metrics, AUC and KL divergence, and find that our detector gives the best performance under the KL divergence measure, which, we argue, is better suited for recognition.

Human Saliency Map Predictors: We run several saliency map predictors on our dataset, which we describe below.

Baselines: We provide three baselines for saliency map comparison. First, the uniform saliency map, which assigns the same probability to each pixel of the video frame. Second, we consider the center bias (CB) feature, which assigns each pixel with the distance to the center of the frame. This feature is intended to capture both screen and shooter's bias effects. At the other end of the spectrum lies the human saliency predictor, which computes the saliency map derived from the fixations made by half of our human subjects. This predictor is always evaluated with respect to the rest of the subjects, as opposed to the entire group.

Static Features (SF): We consider several features used by the human vision community for saliency prediction in the image domain[4], which can be classified into three categories: low, mid and high-level. The four low level features used are color information, steerable pyramid subbands, the feature maps used as input in[10] and the output of the saliency model described in [27, 28]. We run a Horizon detector[27] as our mid-level feature. Object detectors are used as high level features, which comprise faces[29], persons and cars[30].

Motion Features (MF): We include five novel feature maps, in the context of saliency prediction, which are derived from motion or space-time information.

Flow: We extract optical flow from each frame and compute the magnitude of the flow at each location. We do this in order to investigate whether regions with significant optical changes attract human gaze.

Pb with Flow: We run the Pb edge detector with both image intensity and the flow field as inputs. This detector fires both at intensity boundaries and at motion boundaries, where both image and flow discontinuities arise.

Flow Bimodality: We wish to investigate how often people fixate on motion edges, where the flow field typically has a bimodal distribution. We design a feature that measures optical flow bimodality around each location (for details see[24]).

Harris: This feature encodes the spatio-temporal Harris cornerness measure[6].

HoG-MBH Detector: The models for saliency considered so far access higher level image structure by means of pre-trained object detectors. That approach does not prove effective on our dataset, due to high variability in pose and illumination. On the other hand, our dataset provides a rich set of human fixations. Our analysis suggests that the image regions fixated by human subjects often contain semantically meaningful objects or object parts (see fig.2). Inspired by this insight, we exploit the structure present at these locations and train a detector for human fixations, which uses both static (HoG) and motion (MBH)

descriptors centered at fixations. We build a saliency map from the confidence of our HoG-MBH detector when run over each video in a sliding windows fashion.

Feature Combinations: We also linearly combine various subsets of our feature maps in pursuit of better saliency prediction. We investigate the predictive power of static features and motion features alone and in combination, with and without central bias. Details can be found in [24].

Experimental Protocol: When training our HoG-MBH detector, we use 10^6 training examples, half of which are positive and half of which are negative. At each of these locations, we extract spatio-temporal HoG and MBH descriptors, with various grid configurations. These descriptors are concatenated and the resulting vector is lifted into a higher dimensional space by employing an order 3 χ^2 kernel approximation[31]. We train a linear SVM to obtain our detector.

Findings: We first evaluate our saliency predictors under the AUC metric (table 1a), which interprets saliency maps as predictors for separating fixated pixels from the rest. Combining predictors always improves performance under this metric. As a general trend, low-level features are better predictors than high level ones. Low level motion features provide similar performance to static low-level features. On the other hand, our HoG-MBH detector is comparable to the best static feature, the horizon detector.

Table 1. Evaluation of individual feature maps and their combinations for the problem of human saliency prediction. Two metrics are shown, area under the curve (AUC) and KL divergence. Notice that AUC and KL induce different saliency map rankings, but for visual recognition measures that emphasize spatial localization are essential (see also table 2 for action recognition results and [24] for visual illustrations).

baselines		
feature	**AUC** (a)	**KL divergence** (b)
uniform baseline	0.500	18.63
central bias (CB)	0.840	15.93
human	0.936	10.12
static features (SF)		
color features [4]	0.644	17.90
subbands [32]	0.634	17.75
Itti&Koch channels [10]	0.598	16.98
saliency map [27]	0.702	17.17
horizon detector [27]	0.741	15.45
face detector [29]	0.579	16.43
car detector [30]	0.500	18.40
person detector [30]	0.566	17.13

our motion features (MF)		
feature	**AUC** (a)	**KL divergence** (b)
flow magniture	0.626	18.57
pb edges with flow	0.582	17.74
flow bimodality	0.637	17.63
Harris cornerness	0.619	17.21
HOG-MBH detector	0.743	**14.95**
feature combinations		
SF [4]	0.789	16.16
SF + CB [4]	0.861	15.96
MF	0.762	15.62
MF + CB	0.830	15.97
SF + MF	0.812	15.94
SF + MF + CB	**0.871**	15.89

We also use the Kullback-Leibler (KL) divergence measure for comparing predicted saliency maps to the human ground truth. Interestingly, under this metric, the ranking of the saliency maps changes (table 1) and in this setting the HoG-MBH detector performs best. The only other predictor which performs significantly higher than central bias is the horizon detector. Under KL, the linear combination method does not always improve performance, as it optimizes pixel-level classification accuracy, and is not able to account for the inherent spatial competition due to image-level normalization. We conclude that fusing our

Table 2. Columns a-e: Action recognition performance on the **Hollywood-2** dataset when interest points are sampled randomly across the spatio-temporal volumes of video from various distributions (b-e), with the Harris corner detector as baseline (a). Average precision is shown for the uniform (b), central bias (c) and ground truth (e) distributions, and for the output (d) of our HoG-MBH detector. All pipelines use the same number of interest points per frame. **Columns f-i**: Significant improvements over the state of the art approach[22] (f) can be achieved by augmenting their method with channels derived from the predicted saliency map (g) and ground truth saliency (h), but not when using the classical uniform sampling scheme (e).

action	interest points						trajectories + interest points		
	Harris corners (a)	uniform sampling (b)	central bias sampling (c)	predicted saliency sampling (d)	ground truth saliency sampling (e)	trajectories only (f)	uniform sampling (g)	predicted saliency sampling (h)	ground truth saliency sampling (i)
AnswerPhone	16.4%	21.3%	23.3%	23.7%	**28.1%**	**32.6%**	24.5%	25.0%	32.5%
DriveCar	85.4%	92.2%	92.4%	92.8%	**57.9%**	88.0%	93.6%	93.6%	**96.2%**
Eat	59.1%	59.8%	58.6%	**70.0%**	67.3%	65.2%	69.8%	**75.0%**	73.6%
FightPerson	71.1%	74.3%	76.3%	76.1%	**80.6%**	81.4%	79.2%	78.7%	**83.0%**
GetOutCar	36.1%	47.4%	49.6%	54.9%	**55.1%**	52.7%	55.2%	**60.7%**	59.3%
HandShake	18.2%	25.7%	26.5%	**27.9%**	27.6%	**29.6%**	29.3%	28.3%	26.6%
HugPerson	33.8%	33.3%	34.6%	**39.5%**	37.8%	**54.2%**	44.7%	45.3%	46.1%
Kiss	58.3%	61.2%	62.1%	61.3%	**66.4%**	65.8%	66.2%	66.4%	**69.5%**
Run	73.2%	76.0%	77.8%	82.2%	**85.7%**	82.1%	82.1%	84.2%	**87.2%**
SitDown	54.0%	59.3%	62.1%	**69.0%**	62.5%	62.5%	67.2%	**70.4%**	68.1%
SitUp	26.1%	20.7%	20.9%	29.7%	**30.7%**	20.0%	23.8%	**34.1%**	32.9%
StandUp	57.0%	59.8%	61.3%	**63.9%**	58.2%	65.2%	64.9%	**69.5%**	66.0%
Mean	49.1%	52.6%	53.7%	57.6%	57.9%	58.3%	58.4%	61.0%	**61.7%**

trained HoG-MBH detector with our static and dynamic features gives the best results under AUC metrics, while the HoG-MBH detector is the best predictor of visual saliency from our candidate set, under the probabilistic measure (KL) matching the spatial distribution of human fixations.

4.2 Visual Action Recognition Pipeline

We next present an automatic end-to-end action recognition system based on predicted saliency maps, together with several baselines. Both our recognition system and our baselines use the same processing pipeline, which we will describe upfront. It consists of an interest point operator, descriptor extraction, bag of visual words quantization and an action classifier.

Interest Point Operator: We experiment with various interest point operators, both computer vision based and biologically derived (see §4.3). Each interest point operator takes as input a video and generates a set of spatio-temporal coordinates, with associated spatial and temporal scales.[2]

Descriptors: We obtain features by extracting descriptors at the spatiotemporal locations returned by of our interest point operator. For this purpose, we use the spacetime generalization of the HoG descriptor as well as the MBH descriptor computed from optical flow. We consider 7 grid configurations, fixed throughout

[2] Our studies indicate that the spatial distribution of Harris interest points differs sharply from that of human fixations[24].

our experiments, and extract both types of descriptors for each configuration. We end up with 14 features for classification.

Visual Dictionaries: We cluster the resulting descriptors using k-means into vocabularies of 4000 visual words. For computational reasons, we use only 500k randomly sampled descriptors as input to the clustering step. We then represent each video by the L_1 normalized histogram of its visual words.

Classifiers: From histograms we compute kernel matrices using the RBF-χ^2 kernel and combine them by means of a Multiple Kernel Learning (MKL) framework. We train one classifier for each action label in a one-vs-all fashion. We determine the kernel scaling parameter, the SVM loss parameter C and the regularization parameter σ of the MKL framework by cross-validation. We report the average precision on the test set for each action class, in line with [1, 22].

4.3 Visual Action Recognition Studies

We now configure our action recognition pipeline with an operator that randomly samples interest points from the saliency map predicted by our HoG-MBH detector. We choose this map because it best approximates the ground truth saliency map spatially in a probabilistic sense, *i.e.* under a KL divergence measure. We also provide 4 performance baselines. In order to compare with classical action-recognition approaches, we consider the spatiotemporal Harris corner operator [6] and an operator that randomly chooses interest point locations from the uniform distribution spanning spatio-temporal volume of the video. Our third baseline is an operator that samples interest points using the central bias saliency map, which was also shown to approximate to a reasonable extent human fixations under the less intuitive AOI measure (table 1). Fourth, we consider an operator that samples locations from the ground truth saliency map, computed from human fixations by blurring their locations with a Gaussian spatio-temporal kernel. This last baseline assumes that fixations are available at test time and provides an upper bound on recognition performance. Finally, we also investigate whether our end-to-end recognition system can be combined with the state-of-the-art trajectory approach of [22] for better performance.

Experimental Protocol: We run our HoG-MBH detector over the entire data set and obtain our automatic saliency maps, from which we sample interest points. For our first baseline, we run the Harris corner operator on the input video, while for the other baselines we compute uniform, central bias and ground truth saliency maps from which we randomly sample interest points. All our random sampling operators use the same number of interest points per frame as generated by the Harris corner detector. In a final experiment, we also run the pipeline of [22] and combine the 4 kernel matrices they compute with the ones we obtain for our 14 descriptors, sampled from the saliency maps, using MKL.

Results: We note only a small drop in performance when comparing our automatic pipeline to the one obtained by sampling from ground truth saliency maps (table 2d,e). Average precision is also markedly superior to that of a pipeline

sampling interest points uniformly (table 2c). Although central bias is a relatively close approximation of the ground truth saliency map under AOI measures, it performs significantly worse when used for recognition. This further indicates that approximations produced by the HoG-MBH detector are qualitatively different from a central bias distribution, focusing on local image structure that frequently occurs in its training set of fixations, while at the same time being agnostic (as it should be, modulo dataset bias) to image location.

Even though our pipeline is sparse, it achieves near state of the art performance compared to a pipeline that uses dense trajectories. When the sparse descriptors obtained from our automatic pipeline are combined with the kernels associated to the dense trajectory features in [22] using MKL, we are able to go beyond the state-of-the-art. This demonstrates that an end-to-end automatic system incorporating both human and computer vision technology can deliver top performance in a challenging problem such as visual action recognition.

5 Conclusions

We have presented experimental and computational modelling work at the incidence of human visual attention and computer vision, with emphasis on action recognition in video. Inspired by earlier psychophysics and visual attention findings, not validated quantitatively at large scale until now and not pursued for video, we have collected, and made available to the research community, a set of comprehensive human eye-tracking annotations for Hollywood-2 and UCF Sports, some of the most challenging, recently created action recognition datasets in the computer vision community. Besides the collection of large datasets, we have performed quantitative analyses and introduced novel models for evaluating the static and the dynamic consistency of human fixations across different subjects, videos and actions. We have found good inter-subject fixation agreement but, perhaps surprisingly, only moderate evidence of task influence.

We have also presented a large scale analysis of automatic visual saliency models and end-to-end automatic visual action recognition systems. Our studies are performed with particular focus on computer vision techniques and interest point operators and descriptors. In particular we show that accurate saliency operators can be effectively trained based on human fixations. Finally, we show that such automatic saliency predictors can be used within end-to-end computer-based visual action recognition systems to achieve state of the art results in some of the hardest benchmarks in the field.

We hope that our work will foster further communication, benchmarks and methodology sharing between human vision and computer vision, and ultimately lead to improved end-to-end artificial visual action recognition systems.

Acknowledgements. This work was supported by CNCS-UEFICSDI, under PNII RU-RC-2/2009, PCE-2011-3-0438 and CT-ERC-2012-1.

References

1. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR (2009)
2. Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: CVPR (2008)
3. Everingham, M., Gool, L.V., Williams, C., Winn, J., Zisserman, A.: The Pascal visual object classes (VOC) challenge. IJCV (2010)
4. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV (2009)
5. Larochelle, H., Hinton, G.: Learning to combine foveal glimpses with a third-order boltzmann machine. In: NIPS (2010)
6. Laptev, I.: On space-time interest points. In: IJCV (2005)
7. Han, D., Bo, L., Sminchisescu, C.: Selection and context for action recognition. In: ICCV (2009)
8. Prest, A., Schmid, C., Ferrari, V.: Weakly supervised learning of interactions between humans and objects. PAMI (2011)
9. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR (2010)
10. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research 40 (2000)
11. Ehinger, K.A., Sotelo, B., Torralba, A., Oliva, A.: Modeling search for people in 900 scenes: A combined source model of eye guidance. Visual Cognition 17 (2009)
12. Judd, T., Durand, F., Torralba, A.: Fixations on low resolution images. In: ICCV (2009)
13. Kienzle, W., Schölkopf, B., Wichmann, F.A., Franz, M.O.: How to Find Interesting Locations in Video: A Spatiotemporal Interest Point Detector Learned from Human Eye Movements. In: Hamprecht, F.A., Schnörr, C., Jähne, B. (eds.) DAGM 2007. LNCS, vol. 4713, pp. 405–414. Springer, Heidelberg (2007)
14. Fei-Fei, L., Iyer, A., Koch, C., Perona, P.: What do we perceive in a glance of a real-world scene? Journal of Vision (2007)
15. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: ICCV (2007)
16. Itti, L., Rees, G., Tsotsos, J.K. (eds.): Neurobiology of Attention. Academic Press (2005)
17. Vig, E., Dorr, M., Cox, D.: Space-Variant Descriptor Sampling for Action Recognition Based on Saliency and Eye Movements. In: Fitzgibbon, A., Lazebnik, S., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VII. LNCS, vol. 7578, pp. 84–97. Springer, Heidelberg (2012)
18. Land, M.F., Tatler, B.W.: Looking and Acting. Oxford University Press (2009)
19. Torralba, A., Oliva, A., Castelhano, M., Henderson, J.: Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. Psychological Review 13 (2006)
20. Borji, A., Itti, L.: Scene classification with a sparse set of salient regions. In: ICRA (2011)
21. Elazary, L., Itti, L.: A Bayesian model for efficient visual search and recognition. Vision Research 50 (2010)
22. Wang, H., Klaser, A., Schmid, C.: Liu, C.: Action recognition by dense trajectories. In: CVPR (2011)
23. Li, W., Zhang, Z., Liu, Z.: Expandable data-driven graphical modeling of human actions based on salient postures. In: IEEE TCSVT, vol. 18 (2008)

24. Mathe, S., Sminchisescu, C.: Actions in the eye: dynamic gaze datasets and learnt saliency models for visual recognition. Technical report, Institute of Mathematics of the Romanian Academy and University of Bonn (February 2012)
25. Yarbus, A.: Eye Movements and Vision. Plenum Press, New York (1967)
26. Hwang, A., Wang, H., Pomplun, M.: Semantic guidance of eye movements in real-world scenes. Vision Research (2011)
27. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV (42) (2001)
28. Rosenholtz, R.: A simple saliency model predicts a number of motion popout phenomena. Vision Research (39) (1999)
29. Viola, P., Jones, M.: Robust real-time object detection. IJCV (2001)
30. Felzenswalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
31. Li, F., Guy, L., Sminchisescu, C.: Chebyshev approximations to the histogram chi-square kernel. In: CVPR (2012)
32. Simocelli, E., Freeman, W.: The steerable pyramid: A flexible architecture for multi-scale derivative computation. In: ICIP (1995)