Abstract

Predicting attention is a popular topic at the intersection of human and computer vision, but video saliency prediction has only recently begun to benefit from deep learning-based approaches. Even though most of the available video-based saliency data sets and models claim to target human observers’ fixations, they fail to differentiate them from smooth pursuit (SP), a major eye movement type that is unique to perception of dynamic scenes. In this work, we aim to make this distinction explicit, to which end we (i) use both algorithmic and manual annotations of SP traces and other eye movements for two well-established video saliency data sets, (ii) train Slicing Convolutional Neural Networks (S-CNN) for saliency prediction on either fixation- or SP-salient locations, and (iii) evaluate ours and over 20 popular published saliency models on the two annotated data sets for predicting both SP and fixations, as well as on another data set of human fixations. Our proposed model, trained on an independent set of videos, outperforms the state-of-the-art saliency models in the task of SP prediction on all considered data sets. Moreover, this model also demonstrates superior performance in the prediction of “classical” fixation-based saliency. Our results emphasize the importance of selectively approaching training set construction for attention modelling.

1. Introduction

Saliency prediction has a wide variety of applications, be it in computer vision, robotics, or art [6], ranging from image and video compression [20, 21] to such high-level tasks as video summarization [38], scene recognition [52], or human-robot interaction [44].

As to how we view the world around us, there is evidence to suggest significant differences between the eye movement types of fixation and smooth pursuit (SP; tracking a target while keeping it continuously foveated) [33]. A work by Land [31] for instance, considered taking eye tracking into the real world – whilst driving, making tea or doing sports. It discusses the role and usefulness of SP for these various activities, but SP is difficult to distinguish from fixations (especially automatically) in a noisy wearable eye-tracking scenario. Human perception itself is different during SP and fixations: For instance, subjects demonstrate higher chromatic contrast sensitivity during pursuit than during fixation [49], and visual motion prediction is enhanced [55]. Furthermore, SP requires a target and needs to be actively initialized and maintained, whereas subjects will necessarily “fixate” (keep their eyes still at times) even in the absence of a stimulus. SP is also being increasingly used for interaction with devices [58, 17, 48].

We can therefore hypothesize that SP is more selective, which is also corroborated by the highly varying share of time an average observer spends performing SP vs fixations,
depending on video content (see Figure 1, which shows data for a subset of 50 randomly selected videos from the Hollywood2 data set [39, 40]). Thus the pursued targets need to be more special or more characteristic to the viewed video.

And yet available data sets ignore this important distinction, and saliency models naturally follow the same logic, e.g. [32, 25]. In fact, not one of the video saliency models we came across mentions the tracking of objects in scenes performed via SP, and the only data set we found to purposefully attempt such a separation is GazeCom [14], which aims to discard pursuit samples to have cleaner fixations.

Recent advances in automated SP detection [5], however, allow us to analyze large-scale sets of eye tracking recordings, big enough to constitute a training set for deep learning models. In this paper, we target predicting the locations of the input video where SP will occur. We call this problem (i.e., the prediction of maps that correspond to how likely an input video location is to induce SP) supersaliency prediction, due to the higher selectivity of this eye movement type.

We approach this problem using the video analogue to patch-based image CNN processing (e.g., [28] for image saliency, and [12] for individual video frames) – subvolume-based processing. Similar subvolumes were used in [13] for unsupervised feature learning. This way, we are still able to capture local motion, while maintaining a relatively straightforward binary classification-based architecture. Unlike [12], we do not extract motion information explicitly, but rely on the network architecture entirely without any further input processing.

The main contributions of this work consist of (i) introducing the concept of supersaliency – smooth pursuit-based saliency, and pointing out necessary adjustments in evaluation procedures, (ii) demonstrating on three data sets that training exclusively for supersaliency estimation not only allows predicting occurrences of smooth pursuit, but improves traditional, fixation-based saliency, for which our model is not explicitly trained, beyond the state of the art.

2. Related work

Predicting saliency for images has been a very active research field. A widely accepted benchmark is represented by the MIT300 data set [27, 8], which is currently dominated by deep learning solutions. Saliency for videos, however, lacks an established benchmark. It generally is a challenging problem, because in addition to larger computational cost, in a dynamic scene, the objects of interest may be displayed only for a limited time and in different positions and contexts, so attention prioritisation is more crucial.

2.1. Saliency prediction

A variety of algorithms has been introduced to deal with human attention prediction [6]. Video saliency approaches broadly fall into two groups: Published algorithms mostly operate either in the original pixel domain [24, 20, 32, 62] and its derivatives (such as optic flow [64] or other motion representations [66]), or in the compression domain [25, 34, 65]. Transferring expert knowledge from images to videos in terms of saliency prediction is consistent with pixel-domain approaches, and the mounting evidence that motion attracts our eyes contributed to the development of compression-domain algorithms.

Traditionally, from the standpoint of perception, saliency models are also separated into two categories based on the nature of the features and information they employ. Bottom-up models focus their attention (and assume human observers do the same) on low-level features such as luminance, contrast, or edges. For videos, local motion can also be added to the list, and with it the video encoding information. Hence, all the currently available compression-domain saliency predictors are essentially bottom-up.

Top-down models, on the contrary, use high-level, semantic information, such as concepts of objects, faces, etc. These are notoriously hard to formalize. One way to do so would be to detect certain objects in the video scenes, as was done in [40], where whole human figures, faces, and cars were detected. Another way would be to rely on developments in deep learning and the field’s endeavour to implicitly learn important semantic concepts from data. In [12], either RGB space or contrast features are augmented with residual motion information to account for the dynamic aspect of the scenes (i.e., motion is processed before the CNN stage in a handcrafted fashion). The work in [5] uses a 3D CNN to extract features, plus an LSTM network to expand the temporal span of the analysis.

While using a convolutional neural network in itself does not guarantee the top-down nature of the resulting model, its multilayer structure fits the idea of hierarchical computation of low-, mid-, and high-level features. A work by Krizhevsky et al. [29] pointed out that whereas the first convolutional layer learned fairly simplistic kernels that target frequency, orientation and colour of the input signal, the activations in the last layer of the network corresponded to a feature space, in which conceptually similar images are close, regardless of the distance in the low-level representation space. Another study [10] concluded that, just like certain neural populations of a primate brain, deep networks trained for object classification create such internal representation spaces, where images of objects in the same category get similar responses, whereas images of differing categories get dissimilar ones. Other properties of the networks discussed in that work indicate potential insights into the visual processing system that can be gained from them.

2.2. Video saliency data sets

A good overview of existing data sets was given in [63]. Here we dive into the aspect particularly relevant
to this study – the identification of “salient” locations of the videos, i.e., how did the authors deal with dynamic eye movements. For the most part, this issue is not being consistently addressed. The majority of the data sets either make no explicit mention of SP (ASCNN [46], SFU [22], two Hollywood2-based sets [40] [60]), or rely on the event detection built into the eye tracker, which in turn does not differentiate SP from fixations (TUD [4], USC CRCNS [11], CITIUS [32]). IRCCyN/IVC (Video 1) [7] makes no mention of any eye movement types at all, while IRCCyN/IVC (Video 2) [16] mentions SP by name only.

There are two notable exceptions from this logic. First, DIEM [42], which comprises video clips from a rich spectrum of sources, including amateur footage, TV programs and movie trailers, so one would expect a hugely varying fixation–pursuit balance. The respective paper touches on the properties of SP that separate it from fixations, but in the end only distinguishes between blinks, saccades, and non-saccadic eye movements, referring to the latter as generic foveations, which combine fixations and SP.

GazeCom [14] on the other hand, explicitly acknowledges the difficulty of distinguishing between fixations and smooth pursuit in dynamic scenes. The used fixation detection algorithm employed a dual criterion based on gaze speed and dispersion. However, the recently published manually annotated ground truth data [15] shows that these coarse thresholds are insufficient to parse out SP episodes.

Part of this work’s contribution is, therefore, to establish a saliency data set that builds on the existing works, but explicitly separates SP samples from fixations, thus laying the groundwork for SP-based supersaliency and cleaner fixation-based saliency prediction.

3. Saliency and supersaliency

In this section we describe the methodology behind our approach to predicting saliency and supersaliency with a particular emphasis on data sets and model training.

3.1. Data sets and their analysis

We selected three data sets for this work: (i) GazeCom [14], as it is the only one that provides full manual annotation [15], (ii) Hollywood2 [40], as its size allows to use it as a training set for our deep learning solution, and (iii) CITIUS, because of its recent use for a large-scale evaluation of the state of the art in [32].

GazeCom contains eye tracking data for 54 subjects, with 18 dynamic natural scenes used as stimuli, around 20 seconds each, for over 4.5 total hours of viewing time. The clips contain almost no camera motion (major camera viewpoint translation in just one video). A high number of observers and the hand-labelled eye movement type information make this a suitable benchmark data set. Figure 2 displays an example scene from one of the data set clips, together with its empirical saliency maps for both fixations and smooth pursuit, and the same frames in saliency maps predicted by different models.

Hollywood2 contains about 5.5 hours of video data, already split into training and test sets, viewed by 16 subjects. The movies have all types of camera movement, including translation and zoom, as well as scene cuts. The diversity of the movies themselves, as well as the sheer amount of eye tracking recordings, make this data set attractive as a source of training data. For testing all the models, we randomly selected 50 clips from the test subset (same as in Figure 1).

For these first two data sets, we detect SP using a clustering-based approach [3], the implementation of which [15] also labels fixation samples. Its main assumption is that multiple observers would often pursue the same target, thus increasing the robustness against false SP detections compared to looking at individual gaze traces in isolation. This method has demonstrated state-of-the-art performance [2] and, unlike manual labelling, does not require human supervision (which would be practically impossible for such large data sets as Hollywood2).

CITIUS [32] was recently used for evaluation of one of the state-of-the-art models (namely, AWS-D). It is comprised of both real-life and synthetic video sequences, split into subcategories of static and moving camera. For this work, we will use the real-life part, CITIUS-R, for comparison to state-of-the-art models, since our training data lies in this domain, too. Only fixation data for this data set
is provided, not separated into individual observers’ sets, so SP analysis was impossible.

Since both the SP detector and ground truth data for GazeCom label individual gaze samples rather than episodes, we will also evaluate fixation prediction against fixation samples, where possible (to add symmetry to evaluation of SP and fixation prediction). To conform with previously published approaches, and to provide a fair comparison for models that have been optimized for such data sets, we evaluated prediction of fixation onsets as well, either detected by a standard algorithm (for GazeCom and Hollywood2, as described in [14]), or provided with the data set (CITIUS). An overview of the different aspects of our evaluation data sets can be found in Table 1.

3.2. Slicing CNN saliency model

We adopted the slicing convolutional neural network (S-CNN) architecture from [51]. It takes a different approach to extracting motion information from a video sequence. There, instead of handcrafted motion descriptors [12], 3D convolutions [26], or recurrent structures [5], temporal integration is achieved by rotating the feature tensors after initial individual frame-based feature extraction. This way, time (frame index) is one of the axes of the network’s subsequent convolution operations. The whole network would consist of three branches, in each of which the rotation performed is different, and the ensuing convolutions are performed in the planes xy (equivalent to no rotation, and movement information being processed through temporal pooling only), xt or yt (branches are named respectively). The xt branch of the model is outlined in Figure 3.

The architecture is based on VGG-16 [53], with the addition of dimension swapping operations and temporal pooling. All the convolutional layers were initialized with pretrained VGG-16 weights, and the fully-connected layers were initialized randomly. We simply changed the size of the last (output) layer to 2 (instead of 96 in the original paper [51]) to adapt the network for binary classification: (super)salient vs non-salient volumes (as input we use video RGB subvolumes 128 px × 128 px × 15 frames around the pixel that is being classified). Since the model is rather big, only one branch could be trained at a time. We decided to use the xt-branch for our experiments, since it yielded the best individual results in [51], and the horizontal axis seems to be more important for human vision [47, 59].

3.3. Training details

Out of 823 training videos in Hollywood2, 90% (741 clips) were used for training and 10% for validation. Before extracting the subvolumes centred around positive or negative locations of our videos, these were rescaled to 640 × 360 pixels size, and mirror-padded to reduce boundary effects. To assess the influence of the eye movement type in the training data, we trained the same model twice, with two different notions of just what a positive label in our data entails. First, we trained the model S-CNN SP, which is aimed at predicting supersaliency, so the positive locations of the videos are those where SP has occurred. Analogously, for the S-CNN FIX model, which predicts saliency in its regular sense, we attributed positive labels to those subvolumes of the input video where human observers fixated. It is important to note that for the Hollywood2 data set no manual annotations are available, so a state-of-the-art eye movement classification tool [15] was used to detect pursuit and fixation locations used for model training.

The Hollywood2 training set contains 4,520,813 unique SP locations and 10,448,307 unique fixated locations. For both S-CNN SP and S-CNN FIX, the training set consisted of 100,000 subvolumes, half of which were positives (as described above, randomly sampled from the respective eye movement locations in the training videos), half negatives (randomly selected in a uniform fashion to match the number of positive samples per video, non-overlapping with the positive set). For validation, 10,000 subvolumes were used, same procedure as for the training set. We used a batch size of 5, and trained both models for 50,000 iterations with stochastic gradient descent (with momentum of 0.9, learning rate starting at 10^-4 and decreasing 10 times after every 20,000 iterations), at which point both loss and accuracy leveled out.

3.4. Saliency map generation

To generate saliency maps for any video, for each pixel and its respective surrounding video subvolume, we took the probability for the positive class at the soft-max layer of the network. To reduce computation time, we only did this for every 10th pixel along both spatial axes. We then upsampled the resulting low-resolution map to the desired dimensions. For GazeCom and Hollywood2, we generated...
on the bias map, which is blurred with a Gaussian filter (σ distribution. At this location, a single unit pixel is placed the centre of the frame, but the centre of mass in the saliency borán [50] and AWS [19]. The latter was extended by Le-MSM-SM [43], PNSP-CS [18], and a range of OBDL-MAM [37], PIM-ZEN [1], PIM-MCS [54], MCSDM [35], in height, and proportionally scaled width for PMES [36], et al. pipeline and provided source of Khatoonabadi 4.1. Models videos of the three data sets introduced in section 3.1. we use for comparison. All tests were performed for the 4. Evaluation bias, and 0.6 for the original frame, as used in [64]). with its respective bias map (with a weight of 0.4 for the frame of the video saliency map was then linearly mixed est saliency value of the currently processed frame. Each frame of the video saliency map was then linearly mixed with its respective bias map (with a weight of 0.4 for the bias, and 0.6 for the original frame, as used in [64]).

4. Evaluation Here we elaborate on the models, baselines and metrics we use for comparison. All tests were performed for the videos of the three data sets introduced in section 3.1

4.1. Models For the models in compression domain, we followed the pipeline and provided source of Khatoonabadi et al. [25], generating the saliency maps for all videos at 288 pixels in height, and proportionally scaled width for PMES [36], MAM [37], PIM-ZEN [11], PIM-MCS [54], MCSDM [35], MSM-SM [43], PNSP-CS [18], and a range of OBDL-modes [25], as well as pixel-domain GBVS [24, 23], STSD [50] and AWS [19]. The latter was extended by Leborán et al. in [32] to AWS-D, which operates on both videos and images. This model we ran via the provided obfuscated Matlab code (for GazeCom, after downscaling to 640 × 360 px due to memory constraints, other data sets – at their original resolution). Additionally we compared our models to the three invariants (H, S, and K) of the structure tensor [61] at fixed temporal and spatial scales (second and third, respectively). For Hollywood2, the approach of Mathe [41], combining static (low-, mid- and high-level) and motion features, was evaluated as well.

4.2. Baselines The set of baselines was inspired by the works of Judd et al. [27, 8]. All the random baselines were repeated five times per video of each data set. Here, ground truth saliency map refers to a map that is obtained via superimposing spatio-temporal Gaussians at every attended location of all the considered observers. The two spatial sigmas are set to one degree of visual angle. The temporal sigma is set to a frame count equivalent of 1/3 of a second (so that the effect would be mostly contained within one second’s distance).

Chance baseline randomly assigns saliency values from a uniform distribution in all pixels. Shuffled baseline uses the ground truth saliency map of another randomly selected video of the same data set. Centre baseline scales a square Gaussian to fit the frame aspect ratio.

The two baselines listed below cannot be computed unless raw gaze data or fixation onsets for each individual observer are available (i.e. not possible for CITIUS).

“One human” baseline: A randomly selected observer’s ground truth saliency map is used to predict the salient locations of all other observers in this video.

“Infinite humans” baseline: An approximation of how well an infinite set of human observers can predict another infinite set of human observers. Since finding the direct answer is not feasible, Judd et al. [27] proposed computing this as a limit: For a finite N, two groups of observers, N in each, are randomly sampled from the observer pool without replacement. The ground truth saliency map of one set is evaluated against the eye movement data of the other. The procedure is repeated several times for each video. N iterates through a range of configurations (going as high as possible for a given data set). This baseline’s final value is computed via fitting a power function \( f(n) = a \cdot n^{-b} + c \), and the limit is therefore assumed to be equal to c (provided that b is positive). If the sequence seems to diverge or converges outside the theoretical range, the limit is substituted with the best theoretical value for this statistic, thus accounting for the saturation effect.

4.3. Metrics To avoid confusion, we must note that the term salient in this subsection is used broadly, and will depend on the eye

| Data set    | Number of clips | Total video time | Number of observers | Fixation onsets | Raw gaze data | Fixation detected | Pursuit detected | Fixation ground truth | Pursuit ground truth |
|-------------|-----------------|------------------|---------------------|-----------------|---------------|-------------------|-------------------|-----------------------|---------------------|
| GazeCom     | 18              | 00:05:59         | 54                  | +               | +             | +                 | +                 | -                     | -                   |
| Hollywood2  | 1707            | 05:28:39         | 16                  | +               | +             | +                 | +                 | -                     | -                   |
| CITIUS-R    | 45              | 00:07:01         | 22                  | +               | -             | -                 | -                 | -                     | -                   |

Table 1. Properties and availability of different aspects of the data sets in use. “Fixation detected” and “Pursuit detected” refer to sample-level labels for respective eye movements. As long as raw gaze data was provided, this analysis could be ran. “Ground truth” for SP and fixations refers to sample-level manual annotations for the respective data set.
movement type considered, as well as on the type of data source (i.e. ground truth, or detected by an algorithm). For each condition, we end up with a set of locations, where the given eye movement type has occurred (according to the data source), which are considered salient for the respective video clip.

For a thorough evaluation, we took a broad spectrum of metrics, mostly based on [9] and [30]. For GazeCom and Hollywood2, we fixed all saliency maps’ resolution to $360 \times 360$ pixels during evaluation, either for memory constraints, or for symmetric evaluation in case of differently shaped videos. For CITIUS, the native resolution of $320 \times 240$ pixels was maintained.

A set of AUC metrics evaluates the order of the saliency scores. For these metrics, the perfect score is achieved when all the positive locations get a higher score than all of the negative locations. The only difference that arises between them is how to sample the negative locations (if random, we averaged results over 100 repetitions). The positive ones are always the full set of salient locations.

For AUC-Judd, all the rest of the pixels of the video are considered as negatives. Especially for videos, with viewing time per frame of a fraction of a second, this metric is vulnerable to artefacts as it is highly dependent on the proportion of positive vs negative samples. AUC-Borji solves this problem by uniformly sampling negative locations (as many as positives) from all the video pixels (regardless of whether or not they are being used as positive samples at the sample time).

Shuffled AUC (sAUC) tries to compensate for centre bias (or, for that matter, any other stimulus-independent biases) of the human gaze by sampling the negative locations (as many as positives) from salient locations of other videos. In our implementation, we first rescaled the temporal axes of all these other video clips to fit with the evaluated clip, and then sampled not just spatial (like e.g. Leborán et al. in [32]), but also temporal coordinates of the negative samples from the pool. This preserves the temporal structure of the bias: e.g. the first fixations after stimulus display tend to be more heavily centre biased than subsequent ones both in static images [56] and dynamic scenes [57].

To augment the picture with a simpler metric that reflects similar properties of the models, but does not require integrating the ROC curve, we used balanced accuracy. Positive and negative samples sets are the same as for AUC-Borji. To select a threshold for binary classification, we pick a point on the ROC curve where the number of errors on both classes is the same. All locations with saliency values above this threshold are considered to be classified as positives, all below – as negatives. Accuracy is the only metric that is directly linked with our training process.

In addition to various AUC metrics, we employed normalized scanpath saliency (NSS) [45]. With the saliency maps treated as three-dimensional probability distributions, expressed as histograms with an individual bar placed in each pixel of each frame, we also computed histogram similarity (SIM), correlation coefficient (CC) and Kullback-Leibler divergence (KLD) between the predicted saliency map and the empirically constructed ground truth map.

Information gain (IG), measured in bits per salient location, as introduced in [30], is used to assess the information gain of the predicted saliency map over using an image- or video-independent baseline map (also referred to as prior). As proposed in [50] for image fixations, we used all the salient locations on other videos of the respective data set (the same rescaling procedure as for sAUC is applied) to create a baseline saliency map via Gaussian blurring (same temporal and spatial $\sigma$ values were used as for ground truth saliency map generation).

Due to its selectivity (i.e. observers can decide not to pursue anything) SP is sparse, and highly unbalanced between videos (see Figure 1). Using simple mean across all videos of the data set for many metrics will introduce artefacts. For AUC-based metrics, for example, there exists a “perfect” aggregated score, which could be computed by aggregating the data over all the videos before computing the metric, i.e. combining all positives and all negatives beforehand. This is, however, not always possible, as many models use per-video normalization as the final step of processing, either to allow for easier visualization, or to use the full spectrum of the 8-bit integer range, if the result is stored as a video.

We conducted several experiments to show that averaging per-video AUC scores is a significantly ($p \ll 0.01$, see supplementary material for a detailed explanation) poorer approximation of this “perfect” score than taking a weighted mean of the individual scores, where the weights are proportional to the number of positive samples for each video.

We will therefore present the weighted mean results for SP prediction. Since fixations do not suffer from this issue as greatly, this adjustment is not essential here. However, whereas in GazeCom the fixation samples’ share varies between 60% and 84% across videos, the same statistic in Hollywood2 starts at 30%, so we would generally recommend to use weighting for fixation evaluation too, but present the conventional mean results for fixations in this work for comparability with the literature (the weighted results reveal a quantitatively similar picture).

Another point we raise in our evaluation is directly distinguishing SP-salient from fixation-salient pixels based on the saliency maps. To this end, we introduced cross-AUC (sAUC): the AUC is computed for the positive samples’ set of all pursuit-salient locations, with an equal number of randomly chosen fixation-salient location of the same stimulus used as negatives. The baselines’ performance on this metric will be indicative of how well the targets for these two eye movements can be separated (in comparison to the
separation of salient and non-salient locations). If a model scores above 50% on this metric, it on average favours pursuit-salient locations over fixation-salient ones. For the purpose of distinguishing the two eye movement types, the scores of 70% and 30% are, however, equivalent.

5. Results and discussion

We tested the predicted saliency maps, produced by 25 published models, as well as our own four slicing CNN predictors – trained for pursuit and fixation prediction, both either with or without the additional post-processing step of gravity centre bias. Full tables with 9 individual metrics (plus xAUC) for 34 evaluated baselines and models for 9 data set/eye movement type combinations can be found in the supplementary material. For a summary, Table 2 presents the average model rankings for the top-10 models.

On average, our pursuit prediction model, combined with adaptive centre bias (S-CNN SP + Gravity CB), performs best, always making it to the first or the second position. Remarkably, this holds true both for the prediction of smooth pursuit and the prediction of fixations in each data set, despite training exclusively on SP-salient locations as positive examples. The success of our pursuit prediction approach in predicting fixations can be potentially attributed to humans pursuing and fixating similar targets, but the relative selectivity of SP allows the model to focus on the particularly interesting objects in the scene. Even without the gravity centre bias, both our saliency S-CNN FIX and supersaliency S-CNN SP models outperform the models from the literature, with their average rank at least 4 positions better than that of the next best model (video-GBVS).

Table 3 demonstrates, for a subset of metrics (shuffled AUC, NSS, KLD, and information gain), the gap between our best model’s performance (S-CNN SP + Gravity CB) and the best results across all other models. Because the best-performing literature model varies for the different metrics, this gap serves as a lower bound of the performance gains over any individual model. In Hollywood2, due to its strong spatial bias, all models achieve a negative IG score for fixations; nevertheless, our model is the closest one to 0. For SP, only our models achieve positive information gain. Overall, for most metrics our model delivers an improvement over the state of the art in attention modelling.

As noted in [32] for images, our deep learning solution to saliency prediction, however well it generalizes to other realistic video data sets (GazeCom and CITIUS-R), is a surprisingly poor performer when it comes to simulating human visual system for pop-out effect videos (CITIUS-S).

In the task of distinguishing SP- and fixation-salient locations (the xAUC metric), most models yield a result above 0.5 on GazeCom, which means that they still, by chance or by design, assign higher saliency values to SP locations (unlike e.g. the centre baseline with xAUC score of 0.44, which implies that fixations on this data set are more centre biased than pursuit). Probably due to their emphasis on motion information, the top of the chart with respect to this metric is heavily dominated by compression-domain approaches (top-7 non-baseline models for both manually annotated and algorithmically labelled data in GazeCom, top-4 in Hollywood2). Even though in the limit (infinite humans baseline) this metric’s weighted average can be confidently above 0.9, the best model’s (MSM-SM [43]) result is just below 0.74 for GazeCom, and below 0.6 for Hollywood2. This particular aspect needs more investigation.

6. Conclusion

We introduced the concept of supersaliency. In order to predict smooth pursuit-based attention, we proposed a rather simple deep learning model with S-CNN architecture that targets this problem through binary classification of video subvolumes. Our solution outperforms all other 25 tested models for SP prediction, both in algorithmically detected SP episodes (similar to what the model was trained on) and manually annotated data (confirming the validity of using large-scale automatic analysis of gaze recordings instead of hand-labelling for training set construction).

Our model seems to generalize exceedingly well: Trained on an independent set of movies (Hollywood2, training set), it not only performs well on the test part of the same data set, but also shows excellent results on two other data sets (GazeCom and CITIUS) without any modifications or fine-tuning.

Additionally, with our supersaliency model, we improve the state of the art in fixation prediction on the majority of metrics, in all three considered data sets, even in the cases where SP and fixations are not as meticulously separated (as is the case with established pipelines). Training our model for fixation locations instead reveals similar (practically on-par, depending on the data set) performance for fixation prediction, but significantly lower pursuit prediction capability.

Acknowledgements

Supported by the Elite Network Bavaria, funded by the Bavarian State Ministry for Research and Education.
| Model                        | GazeCom | Hollywood2 (50 clips) | CITIUS-R | avg. rank |
|-----------------------------|---------|-----------------------|----------|-----------|
|                             | SP-GT*  | SP-det*               | FIX-GT   | FIX-det*  | FIX-ons  | SP-GT*  | SP-det* | FIX-det* | FIX-ons  | FIX-ons  |         |
| Infinite Humans Baseline    | 1.0     | 1.0                   | 1.0      | 1.0       | –         | 1.0     | 1.0     | 1.0      | 1.0     | 1.0     | 1.0     |
| S-CNN SP + Gravity CB       | 4.8     | 3.9                   | 2.9      | 2.9       | 3.0       | 3.7     | 3.7     | 2.9      | 3.4     | 3.4     | 1.0     |
| S-CNN SP                    | 2.9     | 2.7                   | 4.3      | 4.2       | 4.1       | 4.7     | 5.8     | 5.7      | 3.9     | 4.2     | –       |
| S-CNN FIX + Gravity CB      | 12.1    | 9.0                   | 2.8      | 2.8       | 4.0       | 3.1     | 3.0     | 3.1      | 4.7     | 4.7     | –       |
| S-CNN FIX                   | 8.9     | 6.7                   | 4.4      | 4.4       | 4.6       | 5.9     | 4.9     | 5.0      | 4.3     | 5.5     | –       |
| GBVS                        | 11.1    | 10.7                  | 10.0     | 10.0      | 9.2       | 9.7     | 9.4     | 9.2      | 6.3     | 9.5     | –       |
| OBDL-MRF-O                  | 13.8    | 13.4                  | 11.2     | 11.3      | 10.8      | 11.8    | 11.3    | 9.8      | 8.4     | 11.3    | –       |
| AWS-D                       | 14.7    | 13.4                  | 6.8      | 6.4       | 6.4       | 22.0    | 16.0    | 16.3     | 5.8     | 12.0    | –       |
| Centre Baseline             | 28.8    | 27.7                  | 15.7     | 16.0      | 15.4      | 7.8     | 8.4     | 8.3      | 9.2     | 15.3    | –       |
| One Human Baseline          | 18.3    | 16.9                  | 18.4     | 18.4      | 21.8      | 10.3    | 9.0     | 9.7      | –       | 15.4    | –       |
| AWS                         | 25.4    | 25.3                  | 13.1     | 13.2      | 13.0      | 24.7    | 21.3    | 21.2     | 13.7    | 19.0    | –       |
| Mathe                       | –       | –                     | –        | –         | –         | 18.9    | 19.9    | 20.1     | –       | 19.6    | –       |
| Invariant-K                 | 11.1    | 11.4                  | 19.7     | 19.2      | 18.7      | 29.1    | 23.8    | 23.8     | 19.8    | 19.6    | –       |
| STSD                        | 17.7    | 11.3                  | 21.0     | 21.1      | 21.0      | 26.2    | 24.2    | 23.9     | 17.1    | 20.4    | –       |
| OBDL                        | 22.0    | 23.4                  | 22.4     | 22.0      | 21.6      | 21.0    | 21.2    | 21.1     | 19.9    | 21.6    | –       |
| PMES                        | 11.4    | 13.6                  | 26.8     | 26.6      | 26.1      | 20.0    | 26.0    | 26.4     | 26.0    | 22.5    | –       |
| PIM-ZEN                     | 13.3    | 16.8                  | 25.4     | 25.3      | 25.6      | 22.2    | 25.6    | 26.1     | 25.8    | 22.9    | –       |
| Permutation Baseline        | 32.4    | 32.4                  | 28.3     | 28.6      | 28.4      | 26.1    | 21.8    | 20.7     | 23.8    | 27.0    | –       |
| Chance Baseline             | 30.3    | 30.0                  | 27.2     | 27.2      | 27.2      | 31.4    | 30.6    | 30.8     | 27.2    | 29.1    | –       |

Table 2. Evaluation results on three data sets, presented as the mean of 1-based rank values for all the metrics we compute, except sAUC. In column names, GT, “det” and “ons” refer to “ground truth”, detected (by the tools in [15]) and onsets (detected as in [14]), respectively. Where marked with a * sign, weighted average was used for ranking (all SP columns), with regular average considered otherwise. The rows are sorted by the average rank on all sets (last column), i.e. mean of all columns. Top-3 model results in each category are boldified. Results here only include the top-10 models (when sorted by average rank) beside our own, and the baselines – rows with gray background. For the sake of diversity, we only present the best score achieved by an OBDL-model, all other seven OBDL-* results would be positioned between AWS-D and AWS, with OBDL-S being the only one below One Human Baseline. For full evaluation tables, see supplementary material.

| Data set         | Smooth pursuit* | Fixation samples | Fixation onsets |
|------------------|-----------------|------------------|----------------|
|                  | sAUC | NSS | KLD | IG | sAUC | NSS | KLD | IG | sAUC | NSS | KLD | IG |
| GazeCom (GT)     | S-CNN | 0.76 | 1.53 | **2.47** | 3.31 | 0.68 | **1.36** | **1.06** | **0.52** | –   | –   | –   | –   |
|                  | Δ    | -0.03 | -0.12 | -0.12 | 0.07 | -0.01 | 0.32 | -0.13 | 0.2 | –   | –   | –   | –   |
| GazeCom (detected)| S-CNN | 0.78 | **1.62** | **2.49** | **3.91** | 0.68 | **1.38** | **1.05** | **0.55** | 0.68 | **1.33** | **1.03** | **0.46** | –   | –   | –   | –   | –   | –   | –   |
|                  | Δ    | -0.01 | 0.07 | -0.16 | 0.13 | -0.01 | 0.31 | -0.14 | 0.21 | -0.02 | 0.3 | -0.13 | 0.18 | –   | –   | –   | –   | –   | –   | –   |
| HW2-50           | S-CNN | 0.74 | **2.17** | **2.13** | **0.16** | 0.72 | **1.99** | **1.75** | **-0.13** | 0.73 | **2.01** | **1.69** | **-0.16** | –   | –   | –   | –   | –   | –   | –   |
|                  | Δ    | 0.05 | 0.84 | -0.25 | 0.38 | 0.04 | 0.71 | -0.21 | 0.34 | 0.04 | 0.69 | -0.2 | 0.31 | –   | –   | –   | –   | –   | –   | –   |
| CITIUS-R         | S-CNN | –   | –   | –   | –   | –   | –   | –   | –   | –   | –   | –   | –   | –   | 0.71 | **1.78** | **1.13** | **0.28** | –   | –   | –   | –   |
|                  | Δ    | –   | –   | –   | –   | –   | –   | –   | –   | –   | –   | –   | –   | -0.06 | 0.06 | -0.2 | 0.16 | –   | –   | –   | –   |

Table 3. Quantitative differences in individual metric results between our S-CNN SP + Gravity CB model (S-CNN rows) and the otherwise best ones, selected separately for each metric (Δ-rows). The * sign indicates where weighted averaging was employed. For sAUC, NSS and IG, positive Δ-values (boldified) mean that the score obtained by S-CNN SP + Gravity CB model was better than that of any of the other 25 models we evaluated; for KLD the same is indicated by a negative Δ (lower KLD is better). In row names, HW2-50 stands for the 50-clip subset of Hollywood2, GT refers to manually annotated “ground truth”.
References

[1] G. Agarwal, A. Anbu, and A. Sinha. A fast algorithm to find the region-of-interest in the compressed MPEG domain. In Multimedia and Expo, 2003. ICME ’03. Proceedings. 2003 International Conference on, volume 2, pages II–133–6 vol.2, July 2003.

[2] I. Agtzidis, M. Startsev, and M. Dorr. In the pursuit of (ground) truth: A hand-labelling tool for eye movements recorded during dynamic scene viewing. In 2016 IEEE Second Workshop on Eye Tracking and Visualization (ETVIS), pages 65–68, Oct 2016.

[3] I. Agtzidis, M. Startsev, and M. Dorr. Smooth pursuit detection based on multiple observers. In Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications, ETRA ’16, pages 303–306, New York, NY, USA, 2016. ACM.

[4] H. Alers, J. A. Redi, and I. Heynderickx. Examining the effect of task on viewing behavior in videos using saliency maps. In Human Vision and Electronic Imaging, page 82910X, 2012.

[5] L. Bazzani, H. Larochelle, and L. Torresani. Recurrent mixture density network for spatiotemporal visual attention. CoRR, abs/1603.08199, 2016.

[6] A. Borji and L. Itti. State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1):185–207, Jan 2013.

[7] F. Boulos, W. Chen, B. Parrein, and P. L. Callet. Region-of-interest intra prediction for H.264/AVC error resilience. In 2009 16th IEEE International Conference on Image Processing (ICIP), pages 3109–3112, Nov 2009.

[8] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. MIT saliency benchmark.

[9] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? arXiv preprint arXiv:1604.03605, 2016.

[10] C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLOS Computational Biology, 10(12):1–18, 12 2014.

[11] R. Carmi and L. Itti. The role of memory in guiding attention during natural vision. Journal of Vision, 6(9):4, 2006.

[12] S. Chaabouni, J. Benoist-Pineau, O. Hadar, and C. B. Amar. Deep learning for saliency prediction in natural video. CoRR, abs/1604.08010, 2016.

[13] G. Chen, D. Clarke, M. Giuliani, A. Gaschler, D. Wu, D. Weikersdorfer, and A. Knoll. Multi-modality Gesture Detection and Recognition with Un-supervision, Randomization and Discrimination, pages 608–622. Springer International Publishing, Cham, 2015.

[14] M. Dorr, T. Martinetz, K. R. Gegenfurtner, and E. Barth. Variability of eye movements when viewing dynamic natural scenes. Journal of Vision, 10(10):28, 2010.

[15] M. Dorr, M. Startsev, and I. Agtzidis. Smooth pursuit. http://michaeldorr.de/smoothpursuit/.

[16] U. Engelke, R. Pepion, P. L. Callet, and H.-J. Zepernick. Linking distortion perception and visual saliency in H.264/AVC coded video containing packet loss. In Proceedings of Visual Communications and Image Processing. SPIE, 2010.

[17] A. Esteves, E. Velloso, A. Bulling, and H. Gellersen. Orbits: Gaze interaction for smart watches using smooth pursuit eye movements. In Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology, UIST ’15, pages 457–466, New York, NY, USA, 2015. ACM.

[18] Y. Fang, W. Lin, Z. Chen, C. M. Tsai, and C. W. Lin. A video saliency detection model in compressed domain. IEEE Transactions on Circuits and Systems for Video Technology, 24(1):27–38, Jan 2014.

[19] A. García-Díaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil. Saliency from hierarchical adaptation through decorrelation and variance normalization. Image and Vision Computing, 30(1):51 – 64, 2012.

[20] C. Guo and L. Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Transactions on Image Processing, 19(1):185–198, Jan 2010.

[21] H. Hadizadeh and I. V. Bajic. Saliency-aware video compression. IEEE Transactions on Image Processing, 23(1):19–33, Jan 2014.

[22] H. Hadizadeh, M. J. Enriquez, and I. V. Bajic. Eye-tracking database for a set of standard video sequences. IEEE Transactions on Image Processing, 21(2):898–903, Feb 2012.

[23] J. Harel. A saliency implementation in MATLAB.

[24] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In Advances in Neural Information Processing Systems, pages 545–552, 2007.

[25] S. Hossein Khatoonabadi, N. Vasconcelos, I. V. Bajic, and Y. Shan. How many bits does it take for a stimulus to be salient? In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[26] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1):221–231, Jan 2013.

[27] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. 2012.

[28] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In 2009 IEEE 12th International Conference on Computer Vision, pages 2106–2113, Sept 2009.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 1097–1105. Curran Associates, Inc., 2012.

[30] M. Kämmerer, T. S. A. Wallis, and M. Bethge. Information-theoretic model comparison unifies saliency metrics. Proceedings of the National Academy of Sciences, 112(52):16054–16059, 2015.

[31] M. F. Land. Eye movements and the control of actions in everyday life. Progress in Retinal and Eye Research, 25(3):296–324, 2006.
[32] V. Leborán, A. García-Díaz, X. R. Fdez-Vidal, and X. M. Pardo. Dynamic whitening saliency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):893–907, May 2017.

[33] R. J. Leigh and D. S. Zee. Smooth pursuit and visual fixation. In *The neurology of eye movements*, volume 90, pages 188–240. Oxford University Press, USA, 2015.

[34] Y. Li and Y. Li. A fast and efficient saliency detection model in video compressed-domain for human fixations prediction. *Multimedia Tools and Applications*, pages 1–23, Dec 2016.

[35] Z. Liu, H. Yan, L. Shen, Y. Wang, and Z. Zhang. A motion attention model based rate control algorithm for H.264/AVC. In *2009 Eighth IEEE/ACIS International Conference on Computer and Information Science*, pages 568–573, June 2009.

[36] Y.-F. Ma and H.-J. Zhang. A new perceived motion based shot content representation. In *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)*, volume 3, pages 426–429 vol.3, 2001.

[37] Y.-F. Ma and H.-J. Zhang. A model of motion attention for video skimming. In *Proceedings of International Conference on Image Processing*, volume 1, pages I–129–I–132 vol.1, 2002.

[38] S. Marat, M. Guironnet, and D. Pellerin. Video summarization using a visual attention model. In *2007 15th European Signal Processing Conference*, pages 1784–1788, Sept 2007.

[39] M. Marszałek, I. Laptev, and C. Schmid. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2929–2936, June 2009.

[40] S. Mathe and C. Sminchisescu. Dynamic gaze movement datasets and learnt saliency models for visual action recognition. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part II*, ECCV’12, pages 842–856, Berlin, Heidelberg, 2012. Springer-Verlag.

[41] S. Mathe and C. Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(7):1408–1424, July 2015.

[42] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 3(1):5–24, Mar 2011.

[43] K. Muthuswamy and D. Rajan. Salient motion detection in compressed domain. *IEEE Signal Processing Letters*, 20(10):996–999, Oct 2013.

[44] Y. Nagai and K. J. Rohlfing. Computational analysis of motionese toward scaffolding robot action learning. *IEEE Transactions on Autonomous Mental Development*, 1(1):44–54, May 2009.

[45] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397 – 2416, 2005.

[46] N. Riche, M. Mancas, D. Culibrk, V. Crnojevic, B. Gosselin, and T. Dutoit. Dynamic Saliency Models and Human Attention: A Comparative Study on Videos, pages 586–598. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[47] K. G. Rottach, A. Z. Zivotofsky, V. E. Das, L. Averbuch-Heller, A. O. Discenna, A. Poonyathalang, and R. Leigh. Comparison of horizontal, vertical and diagonal smooth pursuit eye movements in normal human subjects. *Vision Research*, 36(14):2189 – 2195, 1996.

[48] S. Schenk, P. Tiefenbacher, G. Rigoll, and M. Dorr. SPOCK: A smooth pursuit oculomotor control kit. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA ’16, pages 2681–2687, New York, NY, USA, 2016. ACM.

[49] A. C. Schütz, D. I. Braun, and K. R. Gegenfurtner. Eye movements and perception: A selective review. *Journal of Vision*, 11(5):9, 2011.

[50] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):15, 2009.

[51] J. Shao, C.-C. Loy, K. Kang, and X. Wang. Slicing convolutional neural network for crowd video understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[52] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):300–312, Feb 2007.

[53] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[54] A. Sinha, G. Agarwal, and A. Anbu. Region-of-interest based compressed domain video transcoding scheme. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages iii–161–4, May 2004.

[55] M. Spering, A. C. Schütz, D. I. Braun, and K. R. Gegenfurtner. Keep your eyes on the ball: Smooth pursuit eye movements enhance prediction of visual motion. *Journal of Neurophysiology*, 105(4):1756–1767, 2011.

[56] B. W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):4, 2007.

[57] P.-H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, 9(7):4, 2009.

[58] M. Vidal, A. Bulling, and H. Gellersen. Pursuits: Spontaneous interaction with displays based on smooth pursuit eye movements and moving targets. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp ’13, pages 439–448, New York, NY, USA, 2013. ACM.

[59] E. Vig, M. Dorr, and E. Barth. Contribution of Spatio-temporal Intensity Variation to Bottom-Up Saliency, pages 469–474. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[60] E. Vig, M. Dorr, and D. Cox. Space-Variant Descriptor Sampling for Action Recognition Based on Saliency and Eye Movements, pages 84–97. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
[61] E. Vig, M. Dorr, T. Martinetz, and E. Barth. Intrinsic dimensionality predicts the saliency of natural dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1080–1091, June 2012.

[62] J. Wang, H. R. Tavakoli, and J. Laaksonen. Fixation prediction in videos using unsupervised hierarchical features. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2225–2232, July 2017.

[63] S. Winkler and R. Subramanian. Overview of eye tracking datasets. In *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 212–217, July 2013.

[64] Z. Wu, L. Su, Q. Huang, B. Wu, J. Li, and G. Li. Video saliency prediction with optimized optical flow and gravity center bias. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, July 2016.

[65] M. Xu, L. Jiang, X. Sun, Z. Ye, and Z. Wang. Learning to detect video saliency with HEVC features. *IEEE Transactions on Image Processing*, 26(1):369–385, Jan 2017.

[66] Y. Zhai and M. Shah. Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of the 14th ACM International Conference on Multimedia*, MM ’06, pages 815–824, New York, NY, USA, 2006. ACM.