Modeling the Temporality of Saliency

Ye Luo, Loong-Fah Cheong and *John-John Cabibihan

Department of Electrical & Computer Engineering, National University of Singapore
*Mechanical and Industrial Engineering Department, Qatar University

Abstract. Dynamic cues have until recently been usually considered as a simple extension of the static saliency, usually in the form of optic flow between two frames. The evolution of stimuli over a period longer than two frames has been largely ignored in saliency research. We argue that considering temporal evolution of trajectory even for a relatively short period can significantly extend the kind of meaningful regions that can be extracted from videos, without resorting to higher-level processes. Our work is a systematic and principled investigation of the temporal aspect of saliency under a dynamic setting. Departing from the majority of works where the dynamic cue is considered as an extension of the static saliency, our work places central importance on temporality. We formulate both intra- and inter-trajectory saliency to measure relationships within and between trajectories respectively. Our inter-trajectory saliency formulation also represents the first attempt among computational saliency works to look beyond the immediate neighborhood in space and time, utilizing the perceptual organization rule of common fate (temporal synchrony) to make a group of trajectories stand out from the rest. At the technical level, our use of the superpixel trajectory representation captures the detailed dynamics of superpixels as they progress in time. This allows us to better measure changes such as sudden movement or onset compared to other representations. Experimental results show that our method achieves state-of-the-art performance both quantitatively and qualitatively.

1 Introduction

Salient objects capture attention by coming to the foreground in perception. The process is apparently a bottom-up effect that begins at early level in perception. Traditionally, something can only be salient if it is considered unpredictable within some local or global context. When the scene contains strong semantic objects such as faces, texts, or other socially salient contents [1, 2], improved models have been proposed to better predict human fixations by integrating higher-level features such as face, horizon [3–6], etc. The most common approach is to add object-specific detectors but this has an unsatisfactory piecemeal quality given that there are thousands of object categories. Lately deep learning framework has been used to discover non-object-specific features [7].

When we consider saliency in the context of every day dynamic activity (aka dynamic saliency or video saliency), the issues seem more complex but less

well-explored. At the most basic level, reacting quickly to unexpected changes (something with strong temporal contrast) is clearly important. Beyond this most basic level of temporal stimuli, all animals are probably highly sensitive to the difference between animate and inanimate motions (agents vs non-agents), given the importance of this distinction to their survival. At an even higher level, they would also be sensitive to the intent of these moving agents or meaning of these actions, be it in the timeless drama played out between predators and preys, or in a social interaction setting for a social animal like us. Our eyes and brains have evolved in a dynamic visual world, so why should not vision be designed by evolution to exploit this rich source of information that reveals itself through time?

Despite their importance, motion cues have until recently been usually considered as an extension of the static saliency (usually in the form of optic flow between two frames). The evolution of stimuli over a period longer than two frames has been largely ignored in saliency research. Even if some newer datasets contain video clips, they mainly comprise of short video clips strung together by abrupt transitions (jump cuts), in order to avoid high-level influence. This precludes analyzing those attributes of motion cues mentioned in the preceding paragraph. We argue that considering temporal evolution of trajectory even for a relatively short period can significantly extend the kind of meaningful regions that can be extracted from videos, without resorting to higher-level processes.

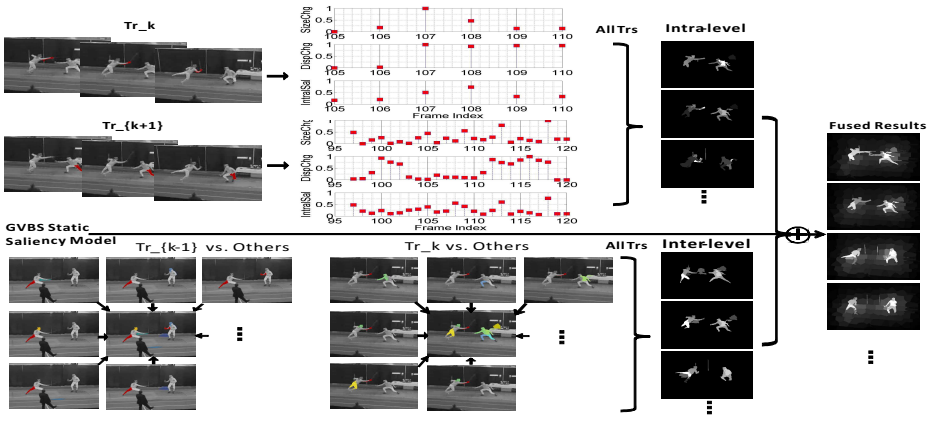

Fig. 1. Illustrations of the proposed video saliency estimation method. The upper part depicts saliency estimation at the intra-trajectory level (from left to right): the depiction of the thrust movement of the corresponding superpixels k and k+1 (in red), the intra-trajectory saliency profile of the two superpixels, with the peak indicating a thrust movement, and the intra-trajectory saliency maps of three of the frames, with brighter values indicating higher saliency. The lower part depicts saliency estimation at the inter-trajectory level (from left to right): the superpixel k-1 and k (in red) found to have strong correlations with 7 other superpixels depicted in different colors, and the inter-trajectory saliency maps of three of the frames. The middle line represents the static image saliency model, and the rightmost figure shows some overall video saliency results after fusion from all three levels. Best viewed in color.

This is akin to the development of object-level attributes in image saliency works: one looks beyond the immediate spatial neighborhood of a pixel to compute mid-level visual cues such as convexity, surroundings, symmetry, etc.[8, 9]. In the temporal domain, there exist similar perceptual organization cues such as common fate or temporal synchrony. Indeed, there is abundant psychophysical evidence that the brain can exploit these temporal structures so that certain features stand out as a group (for review, see [10]). In this paper, we propose a principled hierarchical framework that jointly utilizes low-level temporal and spatial cues to define a more comprehensive range of salient objects in videos.

At the most basic level of this hierarchy, we have the Harel and Koch's graph-based visual saliency model (GVBS) [1] with its static spatial features. In addition to this static level, we have two further levels to incorporate dynamic cues: the intra-trajectory and finally the inter-trajectory levels. We then propose a simple scheme to naturally integrate these various levels together. Fig. 1 illustrates the ideas of our framework and some results are shown in the rightmost of the figure. In the following, we will briefly explicate the information extracted from the intra-trajectory and the inter-trajectory levels, and discuss the motivations behind some of our designs.

1.1 Intra-trajectory level

There is much information residing in a single trajectory that is related to the distinction between agents and non-agents, and between entities capable of intentionality or not. Of cues that make the object's movements appear goal-directed include sudden direction and speed change, rational interactions with spatial contexts and other objects, apparent violations of Newtonian mechanics [11]. For this work, we wish to eschew the use of high-level semantics and non-visual cue such as gravity direction. Thus, we only adopt the "sudden direction and speed change" cue to model these intentionality attributes. In addition, we also model human's sensitivity to onset and offset (when a particular spatial region appears or disappears over time). Specifically, we look out for any sudden change in the size and the displacement of a superpixel. Referring to the upper part of Fig. 1, where the temporal evolution of the intra-trajectory saliency of the fencer's right hand has been depicted as a plot, it is clear that the right hand catches our attention when it makes the sudden cut and thrust movement.

1.2 Inter-trajectory level

There are motions that might not be considered particularly meaningful individually, but when they exhibit temporal synchrony with other motions, they become salient. These well-synchronized movements might be between various body parts of the same person or even from different persons. At the coarsest level, they alert us to the presence of purposive behaviors and encode causality. At a more fine-grained level, it could signify something socially relevant and govern our interaction with others, or it could even be maneuvers perceived as threatening (either in real physical combats or in sports). To detect temporal

correlation, we use mutual information between each pair of trajectories. Using fencing as an example again (Fig. 1, lower part), we can see that the pair of fencers are more salient than the judge (especially during the cut and thrust movement); our scheme captures the fact that we feel in the coordinated movements of the hands and the legs a sense of purpose (threatening in this case), and in the coordinated offense and defense movements a sense of cause and effect.

In sum, the main contribution of our work is a systematic and principled investigation of the temporal aspect of saliency under a dynamic setting. Departing from the majority of works where dynamic cue is considered as an extension of the static saliency, our work places central importance on temporality, as it is mainly through time that intentionality (clearly salient to us as social beings) is expressed. Being able to detect regions that carry meaningful actions has implications for the design of action recognition algorithms; one can use the dynamic saliency proposed here to drive the pooling step [12] as it has a more intrinsic relationship with the semantics of the actions. At the technical level, our use of the superpixel trajectory representation captures the detailed dynamics of superpixels as they progress in time. This allows us to better measure changes such as sudden movement compared to other representations such as video cube [13] or site entropy [14]. Our inter-trajectory saliency formulation also represents the first attempt among computational saliency works to look beyond the immediate neighborhood in space and time, utilizing the perceptual organization rule of common fate to make a group of trajectories stand out from the rest.

2 Related Work

Despite a spate of recent works on dynamic visual saliency (e.g. [15–17, 14, 18, 19]), they do not depart from the various traditional notions used in image saliency works. These works are either based on extending center-surround saliency [20, 13, 18, 19, 21], and those with an information-theoretic slant [22, 17]. The center-surround scheme with optic flow as one of the feature channels was first proposed by Itti in [20]. This basic idea has since been implemented in various different ways for video: the statistical likelihood of a voxel to its near surroundings [13], the error of reconstructing a patch from its spatial and temporal surrounding patches [21], and the contrast between the center and the surround regions [18, 19]. Then there are those measures which are rooted in an information-theoretic interpretation of perception, such as the mutual information which is maximized to discriminate the salient and the non-salient classes [17], and saliency regarded as a kind of maximum information sampling [23, 22].

While the first two levels proposed in our framework are similarly based on the notion of distinctiveness, our trajectory representation substantially deviates from the above approaches in terms of implementation and allows us to capture much more of the temporal structure. Various dynamic saliency works [18, 19] utilize the motion between a pair of frames (e.g. optical flow) as one of the low-level features and compute the local distinctiveness of the flow in a spatial neighborhood. The flow's variation in time is ignored. Works such as [24, 14] look

at the feature evolution in time at a site (pixel or patch) or globally [23]. Unlike our trajectory representation, these measurements are rooted either at a site or global, and hence they do not track the motion characteristics of a specific point or region over a longer interval of time.

The most important difference with the aforementioned computational saliency works lies in the third level of saliency cue proposed in our framework. As far as we are aware, it is the first attempt to formulate saliency based on temporal synchrony, which is not rooted in the traditional concept of conspicuity or distinctiveness. Furthermore, by favoring those synchronous trajectories that exhibit goal-directedness, our work also represents the first attempt to encode movements that are likely to be socially salient in our interaction with other animate agents.

Our work is also related to the "objectness" works [25, 26], especially those that also include saliency. Specifically, objectness can be viewed as a mid-level concept that should include classical perceptual grouping cues such as convexity, symmetry, etc. (besides the enclosedness cue used in [26]). Similarly, our work can be considered as a kind of perceptual grouping but based on temporal cues. Due to the different grouping cues used, our temporal grouping may yield different "objects" from those of spatial grouping. For instance, a group of objects interacting together will be regarded as temporally salient due to their synchrony. This could include a person and the object he or she holds, say, a handphone, even though the latter might not be salient in the spatial sense. Conversely, in a sport video with multiple people, the objectness approach might return all people as objects, even though not all are salient, whereas our approach will only return those with strong dynamic interaction. We argue that it is such dynamic saliency rather than objectness per se that is more appropriate in a dynamic video setting.

In a similar vein, video segmentation works [27–29] might appear related but their objective is quite different. They focus on dividing the video into motion layers, and in the simple case (e.g. planar scenes or rigid motions), such motion layers yield foreground objects and the background. However, even if this simple scenario holds, the distinction between objects and saliency as objective mentioned in the preceding paragraph still holds. Another important difference lies in that our work attempts to capture general temporal synchrony, not just the specific form of synchrony arising from rigid motions. Thus two persons shaking hands would be regarded as an ensemble exhibiting dynamic saliency due to the correlation in their movements. We also favor those movements that exhibit goal-directedness because they are socially salient; these aspects are what distinguish us from pure video segmentation works.

In the psychophysics community, alternative models of gaze allocation in complex dynamic scenes are emerging (for a review, see [30]). This is because conspicuity-based models are found to lack explanatory power in the context of dynamic vision under natural viewing. So far, such deviation of viewing behavior from the conspicuity-based theoretical models are primarily explained as coming from higher level factors, such as the influence of tasks [2, 30]. While there might

have been a few works that explore low-level dynamic cues such as flicker and motion contrast [30], on the whole, there has been a lack of systematic investigation of how various facets of low-level dynamic cues can be used to better account for how we distribute attention in a dynamic environment.

3 The Proposed Method

In this section, the details of the proposed framework for video saliency are introduced, with emphasis given to the dynamic part. For the static part, it suffices to note that given a video clip V with T frames, for the $t^{th}(t \in [1,T])$ frame, we obtain its image saliency map S_I^t by the well-known GVBS algorithm [1]. Since GVBS is pixel based, we take the average saliency value within a superpixel as the saliency value of the superpixel in S_I.

To better describe the long term motion cues, we first employ [31] to obtain the so-called temporal superpixels. We denote the i^{th} superpixel trajectory as a sequence of superpixel locations:

$$Tr_i = \{(x_i^k, y_i^k, t_i^k), k = t_i^s \cdots t_i^e\}, \quad i = 1 \cdots n, \tag{1}$$

where (x_i^k, y_i^k, t_i^k) is the spatiotemporal position of the centroid of the i^{th} superpixel R_i^k at frame k, t_i^s and t_i^e are the start and the end time indices of Tr_i, with $[t_i^s, t_i^e]$ being an interval inside $[1,T]$, and n is the number of detected trajectories in V. Based on this temporal superpixel representation, we can now proceed to estimate the intra-trajectory and the inter-trajectory components of the dynamic saliency.

3.1 Intra-trajectory level

At this level, we want to first capture any significant change in the size of the superpixel, including outright appearance and disappearance, as a measurement of the onset/offset phenomenon. We also want to capture any sudden direction or speed change in the superpixel displacement. For the former, we describe the size change of a superpixel i between two consecutive frames k and $k-1$ as $\Delta R_{sz}^k = abs\left(|R_i^k| - |R_i^{k-1}|\right)$, where $|R_i^k|$ is the cardinality of the superpixel R_i^k, and $abs()$ returns the absolute value. For the latter, we describe the displacement change as $\Delta R_{disp}^k = d\left(R_i^k, R_i^{k-1}\right)$, where $d()$ returns the Euclidean distance between the centroids of R_i^k and R_i^{k-1}. The intra-trajectory saliency for the i^{th} trajectory at frame k (or equivalently, R_i^k) can then be estimated as follows, with both the ΔR_{sz}^k and ΔR_{disp}^k weighted equally with a suitable normalization:

$$S_{intra}(R_i^k) = \begin{cases} \frac{1}{2}\left(\frac{\Delta R_{sz}^k}{\Delta R_{sz}^{Max}} + \frac{\Delta R_{disp}^k}{\Delta R_{disp}^{Max}}\right) & t_i^s < k < t_i^e \\ 1 & k = t_i^s \text{ or } k = t_i^e. \end{cases} . \tag{2}$$

Here ΔR_{sz}^{Max} and ΔR_{disp}^{Max} are the maximum size and displacement change over all the trajectories in the current video clip V. The second condition represents

the instant when the i^{th} superpixel appears or disappears (onset and offset respectively), during which we give maximum intra-saliency. Note that we do not want to consider the appearance and disappearance of superpixels at the image boundary as salient in that it is simply an artificial onset/offset caused by the image boundary. Furthermore, sudden change in speed or direction is also difficult to ascertain at the image boundary. Thus, in addition to the above, we also remove all those trajectories currently lying close to the image boundaries from consideration. Saliency estimation results at the intra-trajectory level for a football video are shown in Fig. 2. As can been seen from Fig. 2, the superpixels with significant changes stand out from others and are estimated with large values in the intra-trejecory level saliency maps.

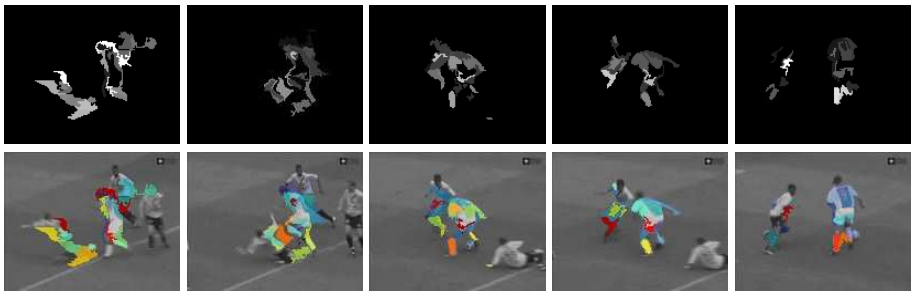

Fig. 2. Intra-trajectory saliency estimation for a football video. The intral-trajectory saliency maps and their corresponding heat maps are shown in the first and second rows, respectively. In the heat map, warm colors indicate large saliency values. Best viewed in color.

3.2 Inter-trajectory level

Two trajectories $Tr_i = \{(x_i^k, y_i^k, t_i^k), k = t_i^s \cdots t_i^e\}$ and $Tr_j = \{(x_j^k, y_j^k, t_j^k), k = t_j^s \cdots t_j^e\}$ are potentially interesting to us if they are temporally synchronized. We use mutual information (MI) to measure the synchronization between these two trajectories over the time interval during which they overlap. We denote this overlapping time interval between Tr_i and Tr_j by $[t^s, t^e] = [t_i^s, t_i^e] \cap [t_j^s, t_j^e]$, assuming $[t^s, t^e] \neq \emptyset$. For simplicity, we use the Gaussian distribution to model the probability of motion vectors from a trajectory. That is $(v_x^i, v_y^i) \sim N(\mu_i, \Sigma_i)$, where $\mu_i = [\mu_x^i \quad \mu_y^i]^T$ and $\Sigma_i = C_{ii} = diag(\sigma_x^i, \sigma_y^i)$. Similarly, for Tr_j, we have another Gaussian $N(\mu_j, \Sigma_j)$, where $\mu_j = [\mu_x^j \quad \mu_y^j]^T$ and $\Sigma_j = C_{jj} = diag(\sigma_x^j, \sigma_y^j)$. The mutual information between Tr_i and Tr_j can then be estimated as [32]:

$$MI(Tr_i, Tr_j) = \begin{cases} \frac{1}{2} \log \frac{|C_{ii}| \cdot |C_{jj}|}{|C|} & Tr_j \notin \mathcal{N}(Tr_i) \text{ and } |\{t^s, \cdots, t^e\}| \geq 3 \\ 0 & \text{Otherwise} \end{cases}, \quad (3)$$

where $|.|$ is the determinant of a matrix, $C = \begin{bmatrix} C_{ii} & C_{ij} \\ C_{ji} & C_{jj} \end{bmatrix}$, and $C_{ij} = C_{ji}^T$ is the

between-sets covariance matrix computed as $C_{ij} = \begin{bmatrix} cov(v_x^i, v_x^j) & cov(v_x^i, v_y^j) \\ cov(v_y^i, v_x^j) & cov(v_y^i, v_y^j) \end{bmatrix}$. $\mathcal{N}(Tr_i)$ in the first condition is the spatial-temporal neighborhood of Tr_i used to enforce a mutual inhibition zone: the reason being that we should be allocating more attention only if the temporally synchronous trajectories are not originating from superpixels immediately adjacent to one another (immediately adjacent superpixels exhibiting synchrony would be less surprising). More specifically, $\mathcal{N}(Tr_i)$ is defined as all the trajectories which are spatially connected to Tr_i at some point in time. An example can be seen in Fig. 3, in which the spatial-temporal neighbors of Tr_5 originating from frames k and $k+1$ are illustrated, i.e. $\mathcal{N}(Tr_5) = \left\{ \cdots, \underbrace{Tr_1, Tr_2, Tr_3, Tr_6, Tr_7, Tr_8}_{\text{from frame } k}, \underbrace{Tr_9, Tr_{10}}_{\text{from frame } k+1}, \cdots \right\}$. The condition $|\{t^s, \cdots, t^e\}| \geq 3$ aims to measure MI only for those trajectories which have temporal intersection of at least three frames.

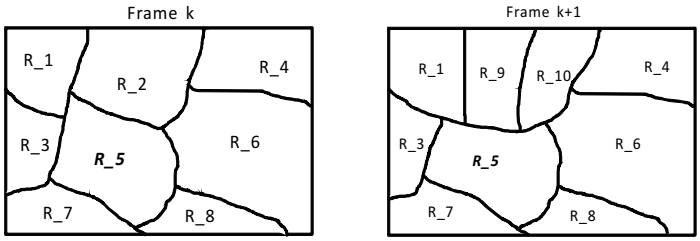

Fig. 3. The spatial-temporal neighbors of Tr_5 at frame k and frame $k+1$.

From the MI computed between all pairwise trajectories, a mutual information matrix $G \in R^{n \times n}$ with $G(i,j) = MI(Tr_i, Tr_j)$ can be obtained between all trajectories. The inter-trajectory saliency of Tr_i should then be the maximum MI values in row i of G. However, we also want to put into context the value of this MI. For instance, the temporal synchrony exhibited between two ballet dancers involved in complex *pas de deux* sequence should have higher value than that between two parallel linear trajectories. Thus we use the entropy of motion vectors from Tr_i itself to weigh the inter-trajectory level saliency as:

$$S_{inter}(Tr_i) = max_j\left(G(i,j)\right) \times H_i, \tag{4}$$

where $H_i = \sum_{k=t_i^s}^{t_i^e}(-p_k \log(p_k))$ is the entropy of motion vectors of Tr_i, and p_k, the probability of the motion vector at frame k, can be obtained from $N(\mu_i, \Sigma_i)$. This saliency value is defined at the level of trajectory; thus all superpixels on the trajectory Tr_i are assigned the same value. Fig. 4 shows the inter-trajectory saliency results for a skiing video clip. As can be seen, the strong movement correlations among different parts of the skier make him stand out from the background.

3.3 Fusion and others

Thus far, we have obtained saliency values from all three levels, with the value of the static and intra-trajectory saliency already normalized to between $[0, 1]$

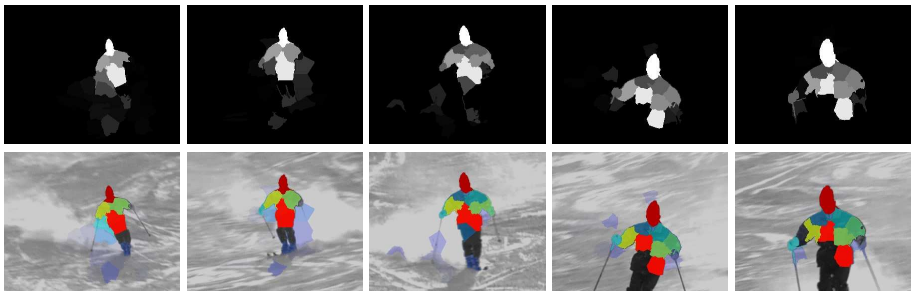

Fig. 4. Inter-trajectory saliency estimation for a skiing video. The inter-trajectory saliency maps and their corresponding heat maps are shown in the first and last rows, respectively. In the heat map, warm colors indicate large saliency values. Best viewed in color.

on a per image and per-video basis respectively. To recap, the maximum value used for normalization in the intra-trajectory level is sought over all values in a particular video. We now also normalize the inter-trajectory level saliency to $[0, 1]$ on a per video basis for the final fusion step. While there might be reasons to perform normalization over the entire video corpus, we stick to the aforementioned simple scheme, in keeping with the normalization practice for static saliency computation (whereby the normalization for each feature channel is usually done on a per-image basis, not over the entire image dataset).

Without any particular reason to favor the saliency values of one level over the other, we perform a simple weighted combination of the normalized saliency values of all three levels, with the weights equal to $\frac{1}{3}$:

$$SM(R_i^k) = \frac{1}{3}\left(S_I(R_i^k) + S_{intra}(R_i^k) + S_{inter}(R_i^k)\right). \tag{5}$$

where $SM(R_i^k)$ is the fused saliency map value indexed by the i^{th} superpixel at the k^{th} frame.

Several other points should be noted. Firstly, the background motion induced by camera movement could significantly affect both the intra- and the inter-trajectory saliency computation. Thus, we first estimate the background model with a simple homography model, using RANSAC to mark out the outliers (i.e. the objects of interest). The background motion is then removed before the intra- and the inter-trajectory saliency are computed.

Secondly, our work is meant to capture the salient aspects of trajectories over a relatively short period of time. Clearly, there must be some upper limit to the length of the video clips T, in accordance with human's short-term memory. In practice, we did not split our videos into shorter clips, as the datasets used in our experiments usually consist of video clips between 5 to 12 seconds, which we consider to be short enough. An exception is some of the long surveillance video clips. But even in the latter, objects do not appear in the videos for long duration (unless they are not moving, in which case their saliency would be attenuated by their high entropy).

Lastly, for simplicity, we did not consider the inhibition-of-return (IOR) mechanism in our model. In fact, under a dynamic setting, it is not even clear if there is a transient inhibition at attended locations for humans[30].

4 Experimental Results and Analysis

4.1 Datasets

We conduct experiments on three public datasets and one additional dataset compiled by us: respectively, UCF-Sports dataset [33], ASCMN [15], Ten-video dataset [34] and Interaction dataset. As the name implies, our own Interaction dataset contains video clips depicting human-human interaction, which has been not the primary focus of those currently available public datasets but is of interest to us.

1. UCF-Sports dataset consists of 150 video sequences of 10 different sports action classes. The averaged length of videos is between 5 to 12 seconds. Four subjects (2 males and 2 females) were asked to freely view videos and the eye tracker data recorded include the fixation points and the saccade movements.
2. ASCMN dataset contains 5 classes of videos: videos with abnormal motions, surveillance videos, videos with crowd motions, videos with moving camera and videos with sudden motions. Several participants were asked to freely view all the videos and the gaze points were recorded and then blurred by a low-pass Gaussian filter, the output of which serves as the ground truth. Due to the difficulties to extract trajectories from videos with crowd motions, videos except for ones with crowd motions are used on this dataset.
3. Ten-Video-Clips dataset contains 10 short video clips of 5 to 10 seconds each. Every video clip has the camera focused on one major object in the scene. The ground truths are taken to be the manually defined object masks.
4. Interaction dataset (to be released later) consists of 8 video clips with 2 fencing videos, 2 boxing videos, 1 ice dancing videos, 1 American football videos and 2 soccer videos. We manually define the ground truths in terms of object masks. While there might be multiple persons interacting in these clips, we choose a maximum of 3 objects for masking, governed by the accepted view that human short-term memory has a capacity of 3-4 items. This choice also helps to reduce the arbitrariness of the ground truth creation process, in that it is usually the pair of persons directly involved in the interaction (e.g. the forward and the immediate defender in a football video), and often a target object (e.g. the ball) that are selected.

4.2 Evaluation Metrics

Various measures can be used to evaluate a particular saliency model against some ground truth data. Each measure has its own strengths and drawbacks depending on the form of the ground truths [35, 36].

For the UCF-Sport dataset and the ASCMN dataset where the ground truths are given in terms of the eye fixations, the Normalized Scanpath Saliency (NSS) [37], the Linear Correlation Coefficients (CC) [38] and the Area under the Receiver Operating Characteristics Curve (AUC-ROC) [39] are employed for evaluation. Readers are referred to the references for details of these measures, but

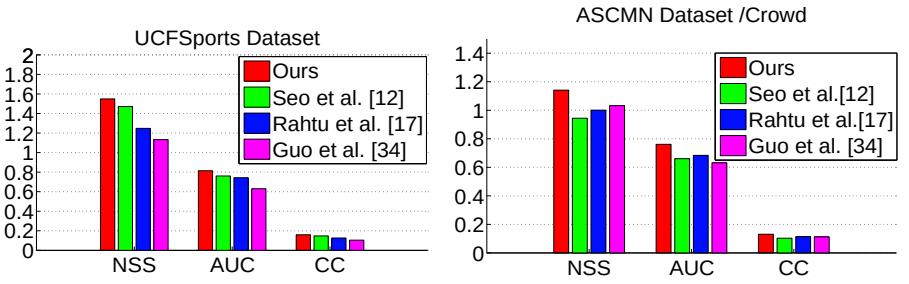

Fig. 5. Results on UCF-Sports dataset and ASCMN dataset respectively.

basically NSS is the average of the response values obtained by using the fixation points to index into the estimated saliency map, and CC measures the linear correlation between the estimated saliency map and the Gaussian-smoothed fixation map. For both measures, larger values indicate better prediction of the saliency model. We adopt the implementations in [36] for all these three measurements.

For the Ten-Video-Clips dataset and our Interaction dataset, where human-labeled masks of the attended regions are provided, the ROC curve is more appropriate for comparison. The ROC curve is generated by plotting the true positive rate against the false positive rate for different values of threshold.

4.3 Results Comparison

We compare the performance of our proposed method with three state-of-the-art methods (Seo et al. [13], Rahtu et al. [18] and Guo et al. [40]) in the four video datasets mentioned above. We did not compare with [19] since the algorithm works only on videos without camera motions. Further experiments are also performed to analyse the different components of our proposed method. In our method, the number of temporal superpixels for the initial frame of each video is set as 100 — as a result, more than 500 trajectories can be extracted for a video. The threshold for the RANSAC algorithm is empirically set, its values ranging from 0.01 to 0.001 (pixel unit in the normalized coordinate), depending on the motion magnitudes of the video. We deem a trajectory as belonging to the background if the optical flows along the trajectory are grouped into the background for more than half of its duration. For other methods chosen for comparison, we use the codes released by the authors as well as their default parameter settings.

We first show the results of UCF-Sports dataset and ASCMN dataset in Fig. 5. As can be seen, our proposed method outperforms the other methods on all three metrics employed: NSS, CC and AUC-ROC. The estimated video saliency maps for one of the clips from UCF-Sports dataset are also provided in the top half of Fig 9, from which it can be seen that our results are qualitatively closest to the human fixations among all methods. Both riders have been successfully detected by our method as salient objects due to their high correlation in movement, whereas all other methods primarily focus on the more conspicuous (static-feature-wise) rider on the left.

Next, we show in Fig. 6 the results of Ten-Video-Clips dataset and Interaction dataset, plotted in terms of the ROC curve (or the hit-miss curve) for different threshold values and matched against the manually specified ground truth information. Direct comparisons with the AUC values of all methods on the two databsets can been in Table 1. The estimated saliency maps for one of the clips in the Interaction dataset are also shown in the bottom half of Fig. 9. As can be seen, the strong movement correlations among two fencers make them stand out from the judge and others in background while the other methods either detect the judge (i.e. Rahtu et al. [18] on the fourth row) or do not have the shapes of object's bodies clearly detected (i.e. Seo et al. [13] on the third row and Guo et al. [40] on the fifth row).

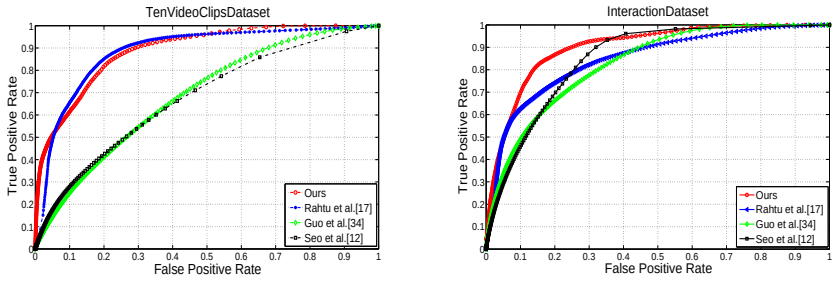

Fig. 6. Results on the Ten-Video-Clips dataset and the Interaction dataset respectively.

Table 1. AUC values comparisons on Ten-Video-Clips dataset and Interaction dataset.

AUC	Our Method	Rahtu et al.[18]	Guo et al.[40]	Seo et al.[13]
Ten-Video-Clips	0.8903	0.8861	0.6870	0.6768
Interaction	0.9007	0.8472	0.8284	0.8485

The remaining experiments aim to shed light on the different components of our proposed method. First we show in Fig. 7 the performance of the proposed method on the UCF-Sports dataset in terms of their individual components, and various ways of combining these components. From the figure, it can be seen that when the intra- and the inter-level saliency are individually combined with the static saliency (i.e. intra + static, inter + static), they seem to only marginally improve the performance of the resulting algorithm. However, when all three levels are combined together, there is a substantial increase in performance. This means that while there are clips in which the individual components are diagnostic, there are also other clips in which these individual components may not be useful or even counter-productive (thus resulting in only a marginal improvement). However, the intra- and the inter-level seem to complement each other well so that when all three components are fused together, there is a significant improvement with regarding to all three metrics.

Last but not least, we analyse the effect of clip length on the performance of our method. We use the twelve *horse-riding* videos from UCF-Sports dataset

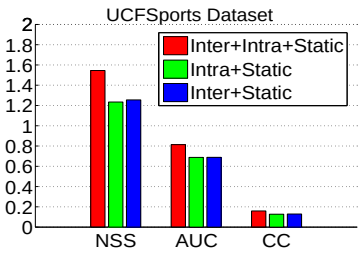

Fig. 7. Results of our method on UCF-Sports dataset at the individual levels, and those obtained from fusion of selected and all the components.

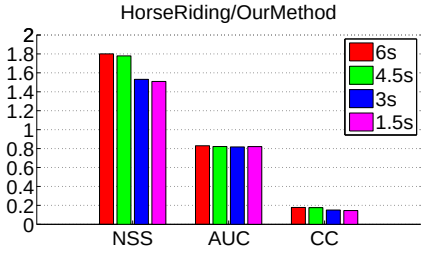

Fig. 8. Results for the proposed method on 12 *horse-riding* videos from UCF-Sports dataset at the individual time lengths.

as an example. For each video, the frame rate is 10 frames per second and the total length of each video is uniformly 6s. For each clip, we take the first $\frac{1}{4}$, first $\frac{1}{2}$, first $\frac{3}{4}$ of the clip, as well as the full clip, resulting in four videos with the lengths of $\{1.5s, 3s, 4.5s, 6s\}$ respectively. We then run our algorithm on these videos and the average results over the 12 videos for each time duration are shown in Fig. 8.

From Fig. 8, it can be noted that longer time durations generally improves the performance, at least up to maximum time length tested in this experiment. It is beyond the scope of this paper to determine if human can keep track of these temporal correlations for an indefinite amount of time, but it suffices for the purpose of this paper to note that the beneficial effect of temporal consideration, even for a relatively short period of time of 1.5s or 3s, is already quite evident.

Acknowledgement. This work was partially supported by the Singapore PSF grant 1321202075 and the NUS AcRF grant R-263-000-A21-112.

5 Conclusion

In this paper, we have investigated the temporality aspect of saliency estimation. A principled method based on three levels of saliency has been proposed: the intra-trajectory level, the inter-trajectory level and the static level. Experimental results validate the concepts put forth in the paper, as well as characterizing the effects of time, and the contributions made by individual levels. Comparisons with three state-of-the-art methods on four datasets with different forms of ground truth demonstrate the superiority of the proposed method.

References

1. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: NIPS. (2007) 545–552
2. Einhauser, W., Spain, M., Perona, P.: Objects predict fixations better than early saliency. Journal of Vision **8** (2008) 1–26

Fig. 9. A qualitative comparison of our method with four other video saliency estimation methods, using clips from the UCF sports dataset (top half) and the Interaction dataset (bottom half) respectively. For each half, from top to bottom: the original frames, ground truth (Gaussian-smooth eye fixations overlaid on the original frames for the top half and object masks for the bottom half), our results, results from Seo et al. [13], Rahtu et al. [18] and Guo et al. [40] respectively. In the heat map, warm colors indicate large saliency values.

3. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV. (2009) 2106–2113
4. Cerf, M., Harel, J., Einhaeuser, W., Koch, C.: Predicting human gaze using low-level saliency combined with face detection. In: NIPS. (2008) 241–248
5. Zhao, Q., Koch, C.: Learning a saliency map using fixated locations in natural scenes. Journal of vision **11** (2011) 1–15
6. Cerf, M., Koch, C.: Faces and text attract gaze independent of the task: Experimental data and computer model. Journal of vision **9** (2009) 1–15
7. Shen, C., Mingli, S., Zhao, Q.: Learning high-level concepts by training a deep network on eye fixations. In: Deep Learning and Unsupervised Feature Learning Workshop, in conjunction with NIPS. (2012)
8. Fowlkes, C.C., Martin, D.R., Malik, J.: Local figure-ground cues are valid for natural images. Journal of vision **7** (2007) 1–9
9. Xu, J., Jiang, M., Wang, S., Kankanhalli, M.S., Zhao, Q.: Predicting human gaze beyond pixels. Journal of Vision **14** (2014) 1–20
10. Blake, R., Lee, S.H.: The role of temporal structure in human vision. Behavior and Cognitive Neuroscience Reviews **4** (2005) 21–42
11. Gao, T., Scholl, B.: Chasing vs. stalking: Interrupting the perception of animacy. Journal of Experimental Psychology **37** (2011) 669–684
12. Ballas, N., Yang, Y., Lan, Z.Z., Delezoide, B., Preteux, F., Hauptmann, A.: Space-time robust representation for action recognition. In: ICCV. (2013) 2704–2711
13. Seo, H.J.J., Milanfar, P.: Static and space-time visual saliency detection by self-resemblance. Journal of Vision **9** (2009) 1–27
14. Wang, W., Wang, Y., Huang, Q., Gao, W.: Measuring visual saliency by site entropy rate. In: CVPR. (2010) 2368–2375
15. Riche, N., Mancas, M., Culibrk, D., Crnojevic, V., Gosselin, B., Dutoit, T.: Dynamic saliency models and human attention: A comparative study on videos. In: ACCV. (2012) 586–598
16. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. T-PAMI **35** (2013) 185–207
17. Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in dynamic scenes. T-PAMI **32** (2010) 171–177
18. Rahtu, E., Kannala, J., Salo, M., Heikkilä, J.: Segmenting salient objects from images and videos. In: ECCV. (2010) 366–379
19. Zhou, F., Kang, S.B., Cohen, M.F.: Time-mapping using space-time saliency. In: CVPR. (2014)
20. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. T-PAMI **20** (1998) 1254–1259
21. Li, Y., Zhou, Y., Xu, L., Yang, X., Yang, J.: Incremental sparse saliency detection. In: ICIP. (2009) 3093–3096
22. Zhang, L., Tong, M.H., Cottrell, G.W.: SUNDAy: Saliency using natural statistics for dynamic analysis of scenes. In: the Thirty-first Annual Cognitive Science Society Conf. (2009) 1–6
23. Hou, X., Zhang, L.: Dynamic visual attention: Searching for coding length increments. In: NIPS. (2008) 681–688
24. Itti, L., Baldi, P.: A principled approach to detecting surprising events in video. In: CVPR. (2005) 631–637
25. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. T-PAMI **34** (2012)
26. Bergh, M.V.D., Roig, G., Boix, X., Manen, S., Gool, L.V.: Online video seeds for temporal window objectness. In: ICCV. (2013) 377–384

27. T.Brox, J.Malik: Object segmentation by long term analysis of point trajectories. In: ECCV. (2010) 282–295
28. P.Ochs, T.Brox: Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions. In: ICCV. (2011) 1583–1590
29. Zhang, D., Javed, O., Shah, M.: Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In: CVPR. (2013) 628–635
30. Tatler, B.W., Hayhoe, M.M., Land, M.F., Ballard, D.H.: Eye guidance in natural vision: Reinterpreting salience. Journal of vision **11** (2011) 1–23
31. Chang, J., Wei, D., III, J.W.F.: A video representation using temporal superpixels. In: CVPR. (2013) 2051–2058
32. Borga, M.: Learning Multidimensional Signal Processing. PhD thesis, Linköping University, Sweden, SE-581 83 Linköping, Sweden (1998)
33. Shapovalova, N., Raptis, M., Sigal, L., Mori, G.: Action is in the eye of the beholder: Eye-gaze driven model for spatio-temporal action localization. In: NIPS. (2013) 2409–2417
34. Fukuchi, K., Miyazato, K., Kimura, A., Takagi, S., Yamato, J.: Saliency-based video segmentation with graph cuts and sequentially updated priors. In: Proc. of International Conference on Multimedia and Expo (ICME). (2009) 638–641
35. Riche, N., Duvinage, M., Mancas, M., Gosselin, B., Dutoit, T.: Saliency and human fixations: State-of-the-art and study of comparison metrics. In: ICCV. (2013)
36. Borji, A., Sihite, D.N., Itti, L.: Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. TIP (2012) 55–69
37. Peters, R.J., Iyer, A., Itti, L., Koch, C.: Components of bottom-up gaze allocation in natural images. Vision research **45** (2005) 2397–2416
38. Jost, T., Ouerhani, N., von Wartburg, R., Muri, R., Hugli, H.: Assessing the contribution of color in visual attention. CVIU **100** (2005) 107 – 123
39. Green, D.M., Swets, J.A.: Signal detection theory and psychophysics. Wiley, New York (1966)
40. Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. TIP **57** (2010) 1856–1866