A SPATIOTEMPORAL SALIENCY FRAMEWORK

Yang Liu, Christos-Savvas Bouganis, Peter Y. K. Cheung

Department of Electrical & Electronic Engineering, Imperial College London

ABSTRACT

This paper presents a novel bio-inspired spatiotemporal saliency framework. The framework incorporates spatial feature detection, feature tracking and motion prediction in order to generate a spatiotemporal saliency map. Experimental results demonstrate its ability and robustness to produce saliency responses to motion pop-up phenomena that are in line with humans responses. Moreover, the limited storage requirements permit real-time implementations of the proposed framework.

Index Terms— Video signal processing, Biological system modeling, Motion analysis, Multidimensional signal detection

1. INTRODUCTION

Saliency detection is a pre-processing technique for further analysis of both image and video. Visual saliency refers to the idea that certain parts of a scene are pre-attentively distinctive and create some form of immediate significant visual arousal. A competent saliency algorithm can improve the effectiveness of systems in applications such as navigation and surveillance.

Saliency of an image refers to spatial saliency. Itti et al. [1] proposed a biological plausible model, which combines intensity, color and orientation information in order to generate a saliency map. There are also an entropy-based model [2] and an eigen-value-based model [3].

Saliency in video refers to spatiotemporal saliency, which requires the incorporation of the temporal information in addition to the spatial information. Laptev et al. [4] proposed the extension of the Harris corner detector in spatial domain to spatiotemporal domain. Saliency occurs where the image values have significant local variations in both space and time. Zhong et al. [5] divided the video into equal length segments and classified the extracted features into prototypes. The segments of the video that can not be matched with any of the computed prototypes are considered as salient. In terms of storage requirement, both spatiotemporal algorithms [4, 5] require accessing stack of video frames and they are not suitable for applications on streaming video input.

Research on human perception and memory suggests that human visual system effectively approximates optimal statistical inference and correctly combines new data with an accurate probabilistic model of the environment [6]. Inspired by this, we propose a novel saliency framework which detects the spatial salient features whose motions contradict a given prediction model. It has also the ability to update the model with new measurements. Moreover, the proposed framework requires only two consecutive frames to be stored, which makes it a good candidate for real-time applications.

2. THE PROPOSED FRAMEWORK

We propose to use the un-predictability in the motion of spatial features as a measure of spatiotemporal saliency. In other words, when the motion path of a spatial feature in the image is predictable, the feature is classified as non-salient, otherwise it is classified as salient. The above definition of spatiotemporal saliency distinguishes the proposed framework from the previous approaches, since it is based on the prediction of the feature’s motion. Because of that, the proposed approach can be thought of working at a higher conceptual level than previous approaches since statistical inference is taken into account. Based on the above new definition of spatiotemporal saliency, a generic framework is proposed that can be built up from numerous established techniques. It is an unsupervised process due to its low-level nature and it operates without any knowledge of the shape or size of the object(s) that may be of interest. The overview of the framework is shown in Figure 1.

![Fig. 1. Overview of the proposed spatiotemporal saliency framework](image)

Two dimensional features are detected in the initialization.
step of the algorithm and every \( n \) frames, where \( n \) is set according to the prior knowledge of the scene’s activity. These are the cues to be both tracked and predicted between consecutive frames and the discrepancies between the tracking and predicting results contribute to the spatiotemporal saliency map. Features can be obtained by any spatial saliency algorithm. The justification of involving spatial saliency algorithm is that in order to be qualified as spatiotemporally salient, they must first be spatially salient. From those features that satisfy the spatial criterion, the temporally salient ones are screened out as spatiotemporal salient points.

The 2D features are then tracked by feature trackers. Let \( \mathbf{u} \) be a feature on frame \( I \), the purpose of feature tracking is to find its corresponding location on frame \( J \).

Each feature is assigned a predictor and is treated separately. Feature coordinates are the measurement inputs of the predictors. Each predictor updates its model on receiving new measurement and predicts the coordinates of the feature in the upcoming frame according to a particular motion model.

At time \( k \), each feature has a predicted coordinates computed at time \( k-1 \) by its predictor and an actual coordinates from the feature tracker. The dissimilarity between the two determines the predictability of the motion behavior of the feature in question, i.e. the temporal saliency of the feature.

We suggest that several spatiotemporally salient features close to each other should make a bigger impact on the spatiotemporal saliency responses than one single feature relatively distant away from other features. The suppressor promotes clusters of salient features which are excited by one another whilst playing down singular features whose silent neighbors inhibit its saliency response.

In terms of storage cost, the framework needs to store only the current frame and the previous frame in order to obtain the spatiotemporal saliency map. Nevertheless, because the predictors update their parameters in a progressive manner, information from earlier frames are implicitly taken into account. Effectively, the spatiotemporal saliency map has temporal support of more than just two frame.

3. REALIZATION OF THE FRAMEWORK

Given the framework structure in Section 2, different algorithms can be applied into the framework. In the current realization of the framework, we have assumed that features with linear motions are classified as not-salient where features with more complicated motions are classified as salient.

3.1. 2D feature detection

Shi’s algorithm [3] has been adopted for 2D feature detection. The choice of this algorithm lies on the fact that can detect strong corners quite robustly and also it has been developed to maximize the quality of tracking, which is important for the subsequent steps of the proposed framework.

The algorithm first calculates the minimal eigenvalue of the \( 3 \times 3 \) neighborhood for every pixel, chooses the local maxima within the \( 3 \times 3 \) neighborhood and then rejects points with minimal eigenvalues less than a predefined threshold. Finally, it ensures that all the detected corners are distanced enough from each other by removing corners that are close to stronger corners.

3.2. Feature tracker

The criteria of a good feature tracker are accuracy and robustness. Small integration window is preferred to avoid smoothing out the details in order to achieve high accuracy. However, the robustness in handling large motion requires having a large integration window. The pyramidal representation of image facilitates a tradeoff between robustness and accuracy.

Such pyramidal implementation of the iterative Lucas-Kanade optical flow algorithm [7] is adopted as the feature tracker in the saliency framework. A feature pixel is tracked in lowest-resolution level within the pyramid and the result is propagated to the next finer level until the original resolution is reached. At each level, a feature pixel \( \mathbf{u} \) in frame \( I \) with coordinates \( [x, y] \) can be tracked in frame \( J \) by finding the spatial gradient matrix \( \mathbf{G} \) defined in (1).

\[
\mathbf{G} = \sum_{x, y \in w} \begin{bmatrix} I_x(x, y) & I_x(x, y) & I_y(x, y) \\ I_y(x, y) & I_y(x, y) & I_y(x, y) \end{bmatrix}
\]

Then the Lucas-Kanade iterative process is applied, where \( k = 1 \rightarrow K \) and \( K \) is the number of iterations.

\[
\delta \mathbf{I}_k(x, y) = \mathbf{I}(x, y) - \mathbf{J}(x + g_x^{k-1} + v_x, y + g_y^{k-1} + v_y)
\]

\[
\mathbf{b}_k = \sum_{x, y \in w} \begin{bmatrix} \delta I_x(x, y) I_x(x, y) \\ \delta I_y(x, y) I_y(x, y) \end{bmatrix}
\]

\[
\mathbf{v}_k = \mathbf{v}_{k-1} + \mathbf{G}^{-1} \mathbf{b}_k, \quad \text{with} \quad \mathbf{v}_0 = [0 \ 0]^T
\]

The final iteration gives the optical flow vector \( \mathbf{v}_K \) of the feature \( \mathbf{u} \) from frame \( I \) to frame \( J \). Thus, the coordinates of the feature \( \mathbf{u} \) in the new frame \( J \) are obtained.

3.3. Prediction

Provided the saliency assumption of linearity, the Kalman filter is a good instrument in predicting the motion. The Kalman filter assumes that the posterior density at every time step is Gaussian and hence exactly and completely characterized by two parameters, its mean and covariance [8]. It can be interpreted as having two parts: update and prediction.

Measurements \( z_k \) at time \( k \) are the coordinates in the current frame computed from the feature tracker. The parameters
of the Kalman filter are updated using equations (2), (3) and (4).

Kalman Gain: \[ K_k = P_k H^T (HP_k H^T + R)^{-1} \] (2)
Update estimate: \[ \hat{x}_k = \hat{x}_k^- + K_k (z_k - H\hat{x}_k^-) \] (3)
Update covariance: \[ P_k = (I - K_k H) P_k^- \] (4)

The future coordinates of each feature are predicted by its dedicated Kalman filter by using the projection equations (5) and (6).

Project state: \[ \hat{x}_k^- = F \hat{x}_{k-1} \] (5)
Project covariance: \[ P_k^- = FP_{k-1} F^T + Q \] (6)

The definitions of symbols are:
\[ \hat{x}_k^- \quad \text{: a priori state vector} \]
\[ \hat{x}_k \quad \text{: a posteriori state vector} \]
\[ P_k^- \quad \text{: a priori estimation error covariance} \]
\[ P_k \quad \text{: a posteriori estimation error covariance} \]
\[ F \quad \text{: transition matrix} \]
\[ H \quad \text{: measurement matrix} \]
\[ Q \quad \text{: process noise covariance} \]
\[ R \quad \text{: measurement noise covariance} \]

We choose transition matrix \( F \), state vector \( x \) and measurement matrix \( H \) to be

\[
F = \begin{bmatrix}
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}, \quad x = \begin{bmatrix}
x \\
y \\
x' \\
y'
\end{bmatrix}
\]

\[
H = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0
\end{bmatrix}
\]

where \( x' \) and \( y' \) are velocity components in \( x \) and \( y \) direction respectively.

3.4. Dissimilarity and Suppression

The dissimilarity metric of this model is the Euclidean distance between the predicted coordinates from the Kalman filter and the tracked coordinates from the feature tracker. The further the actual coordinates of a feature deviate from the prediction, the more salient the feature is. The dissimilarity metric can be regarded as the saliency metric and hence a preliminary saliency map can be constructed.

The suppressor post-processes this preliminary saliency map. The features are subject to a threshold process and those whose dissimilarity values are less than the threshold are disregarded. The neighboring pixels of each feature are then searched. The feature only remains salient and advances into the final saliency map if there are sufficient number of other features surviving the threshold process within the neighborhood, otherwise the feature is disqualified. Under this suppression scheme, only clusters of salient features remain and outliers are removed. This has the biological implication of mimicking excitatory and inhibitory behavior of receptive field of human retinal cell [1], where center cells are excited if surrounding cells are of the same type and inhibited otherwise. Moreover, the suppressor has the effect of disentangling noise from genuine salient movement by relying on multiple saliency responses.

4. PERFORMANCE EVALUATION

Experiments are conducted to demonstrate the viability and robustness of the proposed spatiotemporal framework using the model described in Section 3. In all the experiments, the maximum number of features are limited to 1000. The suppressor threshold is 2. Each feature must have at least 1 neighboring threshold feature within a circle of radius three pixels in order to be qualified as salient feature. The Kalman filter noise covariance matrices \( Q \) and \( R \) are both set to be identity matrices, i.e. covariance noises of 1, which implies that position, speed and acceleration variables are uncorrelated. Video demo can be found at (http://cas.ee.ic.ac.uk/people/yi100/demo.zip).

4.1. Object movement saliency responses

The car experiment illustrated in Figure 2 shows how the movement of the car triggers the spatiotemporal saliency responses under different circumstances. For presentation purpose, the centers of the circles are the feature coordinates that give rise to the saliency responses and the radius of the circles are determined by the dissimilarity values.

- In frame \( F_1 \) and \( F_2 \), the car moves at constant velocity, therefore no saliency is detected by the model.
- The car starts accelerating in \( F_3 \). This sudden acceleration is unpredictable and leads to large saliency responses.
- The car starts turning in \( F_4 \). Again, this is highly unpredictable and causes large saliency responses.
- The car carries on turning and moving forward in \( F_5 \), which is less surprising and generates smaller saliency responses.

It can be concluded that the proposed spatiotemporal framework detects motion pop-up phenomena similar to humans.

4.2. Complex scene saliency analysis

This experiment demonstrates the robustness of the spatiotemporal saliency model dealing with a complex scene. In Figure 3, massive number of fish flock towards one spot where food is supplied. In the region around that spot, fish movements are especially vigourous and irregular, hence these points are the most unpredictable in the scene.

Figure 3 (a) shows spatiotemporal saliency detection results with only the first two frames of the video. The saliency
model suggests a majority of features as salient due to the lack of sufficient temporal information. Hence, there are salient responses all over the place.

Figure 3 (b) shows saliency detection result at a later time. Although there are only two consecutive frames available to the saliency model, it has implicitly picked up earlier temporal information through parameter correction of the Kalman filters. Evidently in Figure 3 (b), the saliency model indeed emphasizes the region around the feeding spot with large saliency responses and smaller responses in other regions.

The particle filter adapted slower to the motions of the features due to the loss of diversity among the particles as a result of resampling. Techniques to tackle that such as Markov Chain Monte Carlo (MCMC) would increase the computational cost. Future work will involve further investigation of employing particle filters as predictors under the current framework with more complicated motion models.

Moreover, the proposed framework has the flexibility to be steered to specific motion patterns depending on the requirements. Under a linear motion assumption, we suggest the Kalman filter as the predictor. Future research will target other motion models that give rise to saliency under different circumstances.

6. REFERENCES


5. CONCLUSION

This paper presents an unsupervised spatiotemporal saliency framework that combines low-level processing and statistical inference. The core idea of the paper is the definition of the spatiotemporal saliency as the unpredictability in the motion of the feature. Experimental results show that the model using this framework has the capability of distinguishing between predictable and unpredictable motions hence spatiotemporal saliency without the involvement of object detection and segmentation even in a complex scene.

An alternative predictor to the Kalman filter can be the particle filter. Given the saliency definition of linearity in motion, initial investigation shows that the particle filters generate inferior saliency results compared to the Kalman filter.

Fig. 2. Spatiotemporal saliency results: In F1 and F2, the car moves at constant velocity; The car starts accelerating in F3; The car starts turning in F4 and turning more in F5.

Fig. 3. Complex scene: massive number of fish moving towards fish feed. (a) saliency result from the very first two frames only; (b) saliency result at a later time.