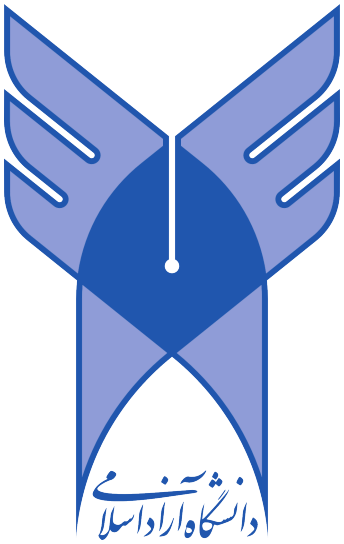

Visual Attention Detection in Video Sequences Using Spatiotemporal Cues report

Paper's Authors

Yun Zhai, Mubarak Shah

Advisor

Dr. Maryam Abedi

Student

Mohammad Shahpouri

October

2022

Contents

List of Figures

List of Equations

1 Proposed framework

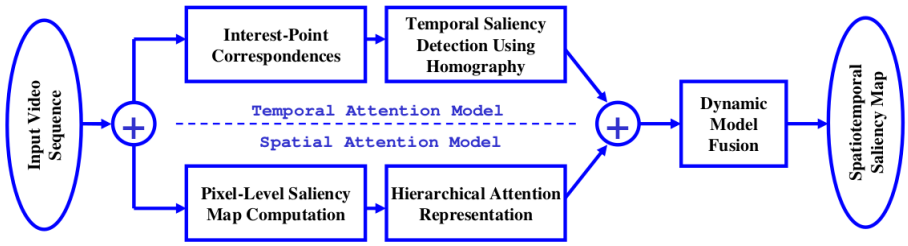

Figure 1. Work flow of the proposed spatiotemporal attention detection framework. It consists of two components, temporal attention model and spatial attention model. These two models are combined using a dynamic fusion technique to produce the overall spatiotemporal saliency maps.

2 Temporal attention model

To find the interest points, the Scale Invariant Feature Transformation (SIFT [1]) is applied. To model the motion contrast between the target point and other points the temporal saliency value $SalT(\mathbf{p}_i)$ of point $\mathbf{p}_i$ is calculated by:

$$SalT(\mathbf{p}_i) = \sum_{j=1}^{n} DistT(\mathbf{p}_i, \mathbf{p}_j) \tag{1}$$

where n is the total number of correspondences. $DistT(\mathbf{p}_i, \mathbf{p}_j)$ is some distance function between $\mathbf{p}_i$ and $\mathbf{p}_j$. The planar transformations are modelled using homography. The interesting point $\mathbf{p} = [x, y, 1]^T$ and its correspondence $\mathbf{p}' = [x', y', 1]^T$ can be associated by:

$$\begin{bmatrix} \hat{x}' \\ \hat{y}' \\ \hat{w}' \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ a_7 & a_8 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2}$$

Here, $\hat{p}' = [\hat{x}', \hat{y}', \hat{w}']^T$ is the projection of $\mathbf{p}$ in the form of homogeneous coordinates. Parameters $\{a_i, i = 1, \ldots, 8\}$ capture the transformation between two matching planes, and they can be estimated by providing at least four pairs of correspondences. Projection error of $\hat{\mathbf{p}}'$ and $\mathbf{p}'$ is computed as:

$$\epsilon(\mathbf{p}_i, \mathbf{H}) = \|\hat{\mathbf{p}}'_i - \mathbf{p}'_i\| \tag{3}$$

where $\mathbf{H}$ represents the transformation matrix.

Homography is insufficient to model all the correspondences in the imagery. Therefore, RANSAC algorithm are applied on the point correspondences to estimate multiple homographies that model different motion segments in the scene.

For each homography H_m estimated by RANSAC, a list of points $\mathbf{L}_m = \{\mathbf{p}_1^m, \ldots, \mathbf{p}_{n_m}^m\}$ are considered as its inliers, where n_m is the number of inliers for $\mathbf{H}_m$. Now, it is possible to redefine 1 as:

$$DistT(\mathbf{p}_i, \mathbf{p}_j) = \epsilon(\mathbf{p}_i, \mathbf{H}_m) \tag{4}$$

where $\mathbf{p}_j \in \mathbf{L}_m$, The temporal saliency value of each point $\mathbf{p}$ is then computed as:

$$SalT(\mathbf{p}) = \sum_{j=1}^{m} f_j \times \epsilon(\mathbf{p}, \mathbf{H}_j) \tag{5}$$

where f_j is the size of the inlier set of $\mathbf{H}_j$. Sometimes, relatively larger moving objects/regions may contribute less trajectories, while smaller regions but with richer texture provide more trajectories. To avoid this problem, they incorporate the spanning area information of the moving regions. The spanning area of a homography $\mathbf{H}_m$ is computed as:

$$\alpha_m = (max(x_i^m) - min(x_i^m)) \times (max(y_i^m) - min(y_i^m)) \tag{6}$$

where $\forall \mathbf{p}_i^m \in \mathbf{L}_m$, and α_i is normalized with respect to the image size, such that $\alpha_i \in [0, 1]$. If $max(x_i^m) = min(x_i^m)$ or $max(y_i^m) = min(y_i^m)$, to avoid zero values of α_m, the corresponding term in Equation 6 is replaced with a non-zero constant number (it is 0.1). The temporal saliency value of a target point $\mathbf{p}$ is finally computed as:

$$SalT(\mathbf{p}) = \sum_{j=1}^{M} \alpha_j \times \epsilon(\mathbf{p}, \mathbf{H}_j) \tag{7}$$

where M is the total number of homographies in the scene.

3 Spatial attention model

To compute the spatial saliency maps of images, image color histograms are ustilized. The color contrast between image pixels build the saliency map of an image. The saliency value of a pixel I_k in an image I is defined as:

$$SalS(I_k) = \sum_{\forall I_i \in I} \|I_k - I_i\| \tag{8}$$

where the value of I_i is in the range of $[0, 255]$, and $\|\cdot\|$ represent the distance metric between color values. This equation is expanded to have the following form:

$$SalS(I_k) = \|I_k - I_1\| + \|I_k - I_2\| + \cdots + \|I_k - I_N\| \tag{9}$$

where N is the total number of pixels in the image. Let $I_k = a_m$, and Equation 9 is restructured as:

$$\begin{aligned} SalS(I_k) &= \|a_m - a_0\| + \cdots + \|a_m - a_1\| + \cdots + \cdots \\ SalS(a_m) &= \sum_{n=0}^{255} f_n \|a_m - a_n\| \end{aligned} \tag{10}$$

where f_n is the frequency of pixel value a_n in the image. The frequencies are expressed in the form of histograms. $\|a_m - a_n\|$ is the color distance map D. Now, The saliency value for a pixel I_k is computed as:

$$SalS(I_k) = SalS(a_m) = \sum_{n=0}^{255} f_n D(m, n) \tag{11}$$

Now, To compute the attended regions based on attended point, computed previously, Given an attended point $\mathbf{c}$, a rectangular box centered at $\mathbf{c}$ with the unit dimensions is created as the seed region $\mathbf{B_c}$. The seed region is then iteratively expanded by moving its sides outward by analyzing the energy around its sides. The attended region expansion algorithm is described as follows:

1. For each side $i \in 1, 2, 3, 4$ of region $\mathbf{B}$ with length l_i, two energy terms $E(s_i)$ and $E(s_i')$ are computed for both its inner and outer sides s_i and s_i', respectively. The potential for expanding side i outward is defined as follows:

$$EP(i) = \frac{E(s_i)E(s_i')}{l_i^2} \tag{12}$$

   where l_i^2 is normalization term.

2. Expand the region by moving side i outward with a unit length if $EP(i) > Th$, where Th is the stopping criteria for the expansion. In the experiment, the unit length is 1 pixel.

3. Repeat steps 1 and 2 until no more side of $\mathbf{B}$ can be further expanded, i.e., all the corresponding expansion potentials are below the defined threshold.

4 Dynamic model fusion

The spatiotemporal saliency map of an image I in the video sequence is constructed as:

$$Sal(I) = \kappa_T \times SalT(I) + \kappa_S \times SalS(I) \tag{13}$$

where κ_T and κ_S are the dynamic weights for the temporal and spatial attention models, respectively. The weights κ_T and κ_S are then defined as:

$$\kappa_T = \frac{PVarT}{PVarT + Const}, \quad \kappa_S = \frac{Const}{PVarT + Const} \tag{14}$$

where $PVarT = max(SalT(I)) - median(SalT(I))$ and $Const$ is constant number.

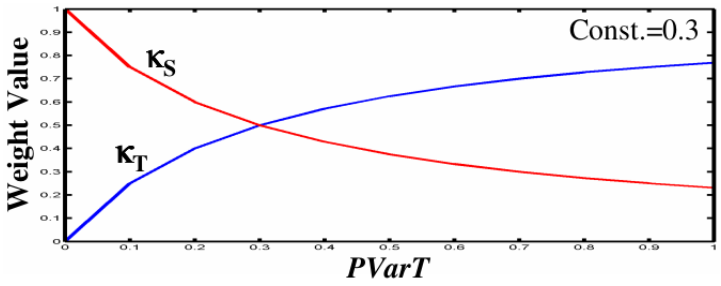

Figure 2. Plots of the dynamic weights, κ_T and κ_S , with respect to the pseudo-variance $PVarT$ of the temporal saliency map, where $Const = 0.3$. As it is clear in the figure, the fusion weight of the temporal attention model increases with the pseudo-variance of the temporal saliency values.

5 Experimental results

TRECVID 2005 BBC rushes data [2] is used to evaluate the proposed spatiotemporal attention model. Testing Set 1 contains nine video sequences, each of which has one object moving in the scene, such as moving car or flying airplane. The second testing set, Testing Set 2, contains video sequences without prominent motions. The videos mainly focus on the static scene settings or with uniform global motions, i.e., there is no motion contrast in the scene. In this case, the spatial attention model should be dominant over the temporal model.

5.1 Quantitative result

System Performance Evaluation			
Data Set	Good	Acceptable	Failed
Testing Set 1 (Moving Objects)	0.82	0.16	0.02
Testing Set 2 (Attended Points)	0.70	0.12	0.18
Testing Set 2 (Attended Regions)	0.80	0.12	0.08

Figure 3. System performance evaluation for three categories, Testing Set 1 with moving objects, Testing Set 2: attended point detection and Testing Set 2: attended region detection.

5.2 Qualitative result

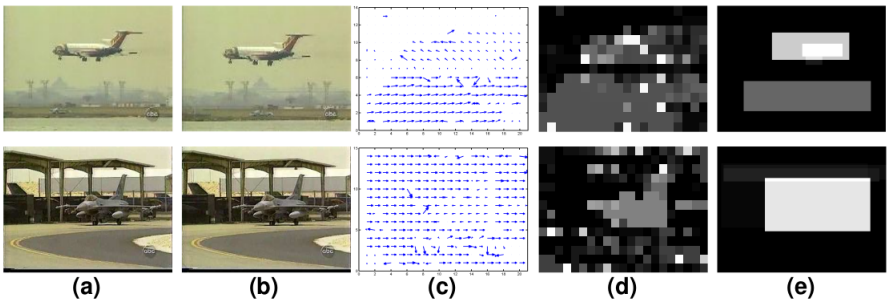

Figure 4. Comparison between the optical-flow based saliency computation and our proposed spatiotemporal attention detection method. For each sequence, two consecutive images are given in (a) and (b). Figure (c) shows the dense optical flow field using SSD block matching. Figure (d) shows the saliency map computed based on the contrast between the optical flow vectors. Figure (e) shows the saliency map generated by the proposed method. In the results, the performance of the optical-flow based method is greatly degraded by the noise around the moving objects.

References

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, Nov 2004. 1

[2] "TRECVID BBC Rushes Data," *TREC Video Retrieval Evaluation Forum*, 2005. NIST. 4