Spatio-Temporal Saliency Networks for Dynamic Saliency Prediction

Cagdas Bak, Aysun Kocak, Erkut Erdem, and Aykut Erdem

***Abstract*—Computational saliency models for still images have gained significant popularity in recent years. Saliency prediction from videos, on the other hand, has received relatively little interest from the community. Motivated by this, in this work, we study the use of deep learning for dynamic saliency prediction and propose the so-called spatio-temporal saliency networks. The key to our models is the architecture of two-stream networks where we investigate different fusion mechanisms to integrate spatial and temporal information. We evaluate our models on the DIEM and UCF-Sports datasets and present highly competitive results against the existing state-of-the-art models. We also carry out some experiments on a number of still images from the MIT300 dataset by exploiting the optical flow maps predicted from these images. Our results show that considering inherent motion information in this way can be helpful for static saliency estimation.**

***Index Terms*—dynamic saliency, deep learning**

I. Introduction

As a key part of the human visual system, visual attention mechanisms filter irrelevant visual stimuli in order to focus more on the important parts. Computational models of attention try to mimic this process through the use of machines and algorithms. These models have gained increasing attention lately. The reason behind this growing interest lies in their use in different computer vision and multimedia problems including but not limited to image retrieval [1], visual quality assessment [2], [3] video resizing/summarization [4], [5], action recognition [6], event detection [7] either as a tool for visual feature extraction or as a mechanism for selecting features. These models are also important for generic applications such as advertisement and web design [8] as attention plays a key role in both user interfaces and human-machine interaction.

In the literature, the computational attention models developed so far generally aim to predict where humans fixate their eyes in images [9]. Specifically, they produce the so-called saliency maps from the visual data where a high saliency score at an image location indicates that the point is more likely to be fixated. These models are largely inspired by the hierarchical processing of human visual system [10] and the theoretical studies like Feature Integration Theory [11] or Guided Search Model [12]. They consider low-level image features (color, orientation, contrast, motion, etc.) and/or high-level features (pedestrians, faces, text) while predicting saliency maps. In this process, while low-level features are employed to examine how different an image point from its surroundings, high-level features are employed as it is experimentally shown that humans have a tendency to fixate on certain object classes more than others.

Another current trend in the existing literature is to detect salient objects [13], [14], [15], [16] from the images. These models specifically aim at identifying the most prominent objects in an image that attract attention under free-viewing conditions and then segmenting them out from the background. These computational models for visual attention can be further grouped into two according to their inputs as static and dynamic saliency models. While static models work on images, dynamic models take video sequences as input.

Saliency prediction from videos leads to great challenges when it is compared to carrying out the same task in still images. The reason is that the dynamic saliency frameworks require taking into account both spatial and temporal characteristics of the video sequences. Static saliency models use features like color, intensity and orientation, however dynamic models need to focus more on the moving objects or image parts as it is shown that humans there is a tendency for humans to look at them while viewing. Hence, the preliminary models proposed for dynamic saliency prediction extend the existing saliency models suggested for still images so that they consider extra motion features [17], [18], [19], [20]. However, more recent works approach the same task from a different point of view and propose novel solutions [21], [22], [23].

A. Overview of our approach

Deep learning has been successfully applied to saliency prediction in still images in the last few years, providing state-of-the-art results [24], [25], [26], [27], [28], [29], [30]. The early models utilize pre-trained deep convolutional neural networks (CNNs) proposed to classify images as generic feature extractors and build classifiers on top of those features to classify fixated image regions [27], [24]. Later models, however, approach the problem from an end-to-end perspective and either train networks from scratch or most of the time fine-tune the weights of a pre-trained model [31], [30], [25]. The modifications in the network architectures are usually about integrating multi-scale processing or using different loss functions [32], [30]. It has been investigated that the power of these deep models mainly comes from the property that the features learned by these networks are semantically very rich [29], capturing high-level factors important for saliency detection. Motivated by the success of these works, in this study, we explore the use of two-stream CNNs for saliency

The authors are with the Department of Computer Engineering, Hacettepe University, Ankara, Turkey, TR-06800 (e-mail: cgds77@gmail.com; aysunkocak@cs.hacettepe.edu.tr; erkut@cs.hacettepe.edu.tr; aykut@cs.hacettepe.edu.tr)

prediction from videos. To the best of our knowledge, our work is the first deep model for dynamic saliency, which is trained in an end-to-end manner, that learns to combine spatial and temporal information in an optimal manner within a two-stream network architecture.

B. *Our contributions*

The contributions of our work can be summarized as follows:

1) We study two-stream convolutional neural networks which mimic the visual pathways in the brain and combine networks trained on temporal and spatial information to predict saliency map of a given video frame. Although these network architectures have been previously investigated for some computer vision problems such as video classification [33] and action recognition [34], to our knowledge, we are the first to apply two-stream deep models for saliency prediction from videos in the literature. In particular, in our study, we investigate two different fusion strategies, namely element-wise and convolutional fusion strategies, to integrate spatial and temporal streams.
2) We carry out extensive experiments on DIEM [35] and UCF-Sports [36] datasets and compare our deep spatio-temporal saliency networks against several state-of-the-art dynamic saliency models. Our evaluation demonstrates that the proposed STSConvNet model outperforms these models in nearly all of the evaluation metrics on these datasets.
3) On a number of challenging still images, we also show that our spatio-temporal saliency network can predict the human fixations better than the state-of-the-art deep static saliency models. The key idea that we follow is to extract optical flow from these static images by using a recently proposed method [37] and feed them to our network along with the appearance image.

II. Related Work

In this study, we focus on bottom-up modeling of dynamic saliency. Below, we first summarize the existing dynamic saliency models from the literature and then provide a brief overview of the proposed deep-learning based static saliency models which are related to ours.

A. *Dynamic Saliency*

Early examples of saliency models for dynamic scenes extend the previously proposed static saliency models which process images in a hierarchical manner by additionally considering features related to motion such as optical flow. For instance, in [17], Harel *et al.* propose a graph-theoretic solution to dynamic saliency by representing the extracted feature maps in terms of fully connected graphs and by predicting the final saliency map. In [19], Cui *et al.* extract salient parts of video frames by performing spectral residual analysis on the Fourier spectrum of these frames over the spatial and the temporal domains. In [18], Guo *et al.* propose a similar spectral analysis based formulation. In [38], Sultani and Saleemi extend Harel *et al.* [17]'s model by using additional features such as color and motion gradients and by post-processing the predicted maps via a graphical model based on Markov Random Fields. In [20], Seo and Milanfar employ self similarities of spatio-temporal volumes to predict saliency. [39], Mauthner *et al.* also present a video saliency detection method for using as a prior information for activity recognition algorithms. Instead of using a data driven approach they propose an unsupervised algorithm for estimating salient regions of video sequences.

Following these early works, other researchers rather take different perspectives and devise novel solutions for dynamic saliency. For example, in [21], Hou and Zhang consider rarity of visual features while extracting saliency maps from videos and propose an entropy maximization-based model. In [40], Rahtu *et al.* extract saliency by estimating local contrast between feature distributions. In [41], Ren *et al.* propose a unified model with the temporal saliency being estimated by sparse and low-rank decomposition and the spatial saliency map being extracted by considering local-global contrast information. In [22], Mathe *et al.* devise saliency prediction from videos as a classification task where they integrate several visual cues through learning-based fusion strategies. In another study, Rudoy *et al.* [23] propose another learning based model for dynamic saliency. It differs from Mathe *et al.*'s model [22] in that they take into account a sparse set of gaze locations thorough which they predict conditional gaze transitions over subsequent video frames. Zhou *et al.* [42] oversegment video frames and use low-level features from the extracted regions to estimate regional contrast values. Zhao *et al.* [43] learn a bank of filters for fixations and use it to model saliency in a location-dependent manner. Khatoonabadi *et al.* [44] propose a saliency model that depends on compressibility principle. In a very recent study, Leboran *et al.* [45], propose another dynamic saliency model by using the idea that perceptual relevant information is carried by high-order statistical structures.

B. *Deep Static Saliency*

In recent years, deep neural networks based models provide state-of-the-art results in many computer vision problems such as image classification [46], object detection [47], activity recognition [42], semantic segmentation [48] and video classification [33]. These approaches perform hierarchical feature learning specific to a task, and thus gives results better than the engineered features. Motivated by the success of these models, a number of researchers have recently proposed deep learning based models for saliency prediction from images [24], [25], [26], [27], [29], [30], [49].

Vig *et al.* [27] use an ensemble of CNNs which learns biologically inspired hierarchical features for saliency prediction. Kümmerer *et al.* [24] employ deep features learned through different layers of the AlexNet [50] and learn how to integrate them for predicting saliency maps. Kruthiventi et al. [26] adopt VGGNet [51] for saliency estimation where they introduce a location-biased convolutional layer to model the center-bias, and train the model on SALICON dataset using Euclidean loss. Jetley *et al.* [30] also use the VGGNet architecture but they especially concentrate on investigating different kinds of probability distance measures to define the loss function. Pan *et al.* [25] very recently propose two CNN models having different layer sizes by approaching saliency prediction as a regression task. Li *et al.* [49] employ a fully convolutional neural network within a multi-task learning framework to jointly detect saliency and perform object class segmentation. It is important to note that all these models are proposed for predicting saliency in still images not videos. Bruce *et al.* [29] propose yet another fully convolutional network to predict saliency and they try to understand factors and learned representations when training these type of networks for saliency estimation.

Motivated by the deep static saliency models, in our paper we investigate the use of two-stream CNNs for saliency prediction from videos. In fact, investigating layered formulations is not new for saliency prediction. As discussed earlier, most of the traditional dynamic saliency models are inspired from the hierarchical processing in the low-level human vision [10]. These models, however, employ hand-crafted features to encode appearance and motion contrast to predict where humans look at in dynamic scenes. Since they depend on low-level cues, they often fail to capture semantics of scenes at its full extent, which is evidently important for gaze prediction. More recent models, on the other hand, employ learning-based formulations to integrate these low-level features with the detection maps for faces, persons, and other objects. This additional supervision boosts the prediction accuracies, however, the performance is limited by the discrimination capability of the considered features and the robustness of the employed detectors.

As compared to the previous dynamic saliency models, our deep spatio-temporal saliency networks are trained to predict saliency in an end-to-end manner. This allows us to learn hierarchical features, both low-, mid- and high-level, [52], [53] that are specialized for the gaze prediction task. For instance, while the early layers learn filters that are sensitive to edges or feature contrasts, the filters in the top layers are responsible from capturing more complex, semantic patterns that are important for the task. In our case, our deep two-stream saliency networks learn multiple layers of spatial and temporal representations and ways to combine them to predict saliency.

In particular, in our study we extract temporal information via optical flow between consecutive video frames and investigate different ways to use this additional information in saliency prediction within a deep two-stream spatio-temporal network architecture [34]. These two-stream networks are simple to implement and train, and to our interest, are in line with the hierarchical organization of the human visual system. Specifically, the biological motivation behind these architectures is the so-called two-streams hypothesis [54] which speculate that human visual cortex is comprised of two distinct streams, namely ventral and the dorsal streams, which are respectively specialized to process appearance and motion information.

Here, an alternative deep architecture could be to stack two or more frames together and feeding this input to a deep single-stream CNN, which was investigated in several action recognition networks [55], [56], [57], [58]. In this work, we do not pursue this direction because of two reasons. Firstly, this approach requires learning 3D convolutional filters [55], [56], [57] in order to capture spatio-temporal regularities among input video frames but using 3D filters highly increases the complexity of the models and these 3D convolutional networks are harder to train with limited training data [58] (which is the case for the existing dynamic saliency datasets). Secondly, 3D convolutional filters are mainly used for expressing long-range motion patterns which could be important for recognizing an action since they cannot easily be captured by optical-flow based two-stream models. For dynamic saliency prediction though, we believe that such long-range dependencies are minimally important as human attention shifts continuously, and optical flow information is sufficient to establish the link between motion and saliency.

III. Our Models

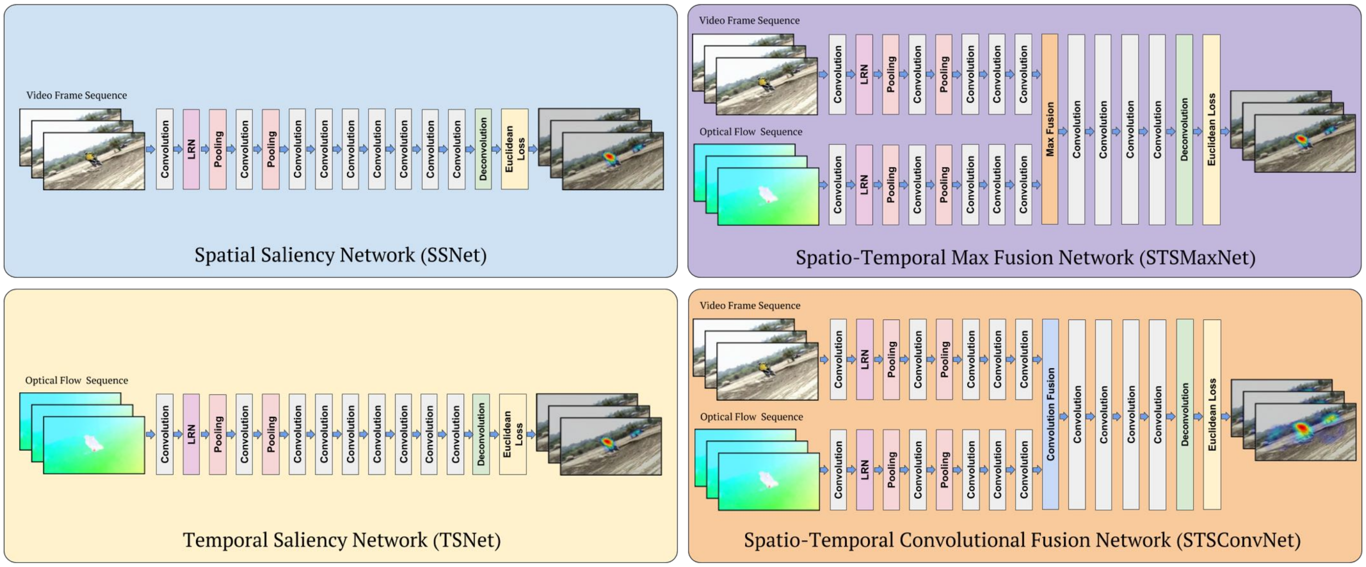
The aim of our study is to investigate the use of deep architectures for predicting saliency from dynamic scenes. Recently, CNNs provided drastically superior performance in many classification and regression tasks in computer vision. While the lower layers of these networks respond to primitive image features such as edges, corners and shared common patterns, the higher layers extract semantic information like object parts, faces or text [29], [53]. As mentioned before, such low and high-level features are shown to be both important and complementary in estimating visual saliency. Towards this end, we examine two baseline single stream networks (spatial and temporal) given in Figure 1(a) and two two-stream networks [34] shown in Figure 1(b), which combine spatial and temporal cues via two different integration mechanisms: element-wise max fusion and convolutional fusion, respectively. We describe these models in detail below.

A. Spatial Saliency Network

For the basic single stream baseline model, we retrain the recently proposed static saliency model in [25] for dynamic saliency prediction by simply ignoring temporal information and using the input video frame alone. Hence, this model does not consider the inter-frame relationships while predicting saliency for a given video. As shown in the top row of Figure 1(a), this CNN resembles the VGG-M model [51] – the main difference being that the final layer is a deconvolution (fractionally strided convolution) layer to up sample to the original image size. Note that it does not use any temporal information and exploits only appearance information to predict saliency in still video frames. We refer to this network architecture as SSNet.

B. Temporal Saliency Network

Saliency prediction from videos is inherently different than estimating saliency from still images in that our attention is highly affected by the local motion contrast of the foreground objects. To understand the contribution of temporal information to the saliency prediction, we develop a second single



(a) Single stream saliency networks (b) Two-stream saliency networks

Fig. 1. (a) The baseline single stream saliency networks. While SSNet utilizes only spatial (appearance) information and accepts still video frames, TSNet exploits only temporal information whose input is given in the form of optical flow images. (b) The proposed two-stream spatio-temporal saliency networks. STSMaxNet performs fusion by using element-wise max fusion, whereas STSConvNet employs convolutional fusion after the fifth convolution layers.

stream baseline. As given in the bottom row of Figure 1(a), this model is just a replica of the spatial stream net but the input is provided in the form of optical flow images, as in [34], computed from two subsequent frames. We refer to this single stream network architecture as TSNet.

C. Spatio-Temporal Saliency Network with Direct Averaging

As a baseline model, we define a network model which integrates the responses of the final layers of the spatial and the temporal saliency networks by using direct averaging. Note that this model does not consider a learning strategy on how to combine these two-stream network and consider each one of the single-stream networks equally reliable. We refer to this two-stream network architecture as STSAvgNet.

D. Spatio-Temporal Saliency Network with Max Fusion

This network model accepts both a video frame and the corresponding optical flow image as inputs and merges together the spatial and temporal single stream networks via element-wise max fusion. That is, given two feature maps $\mathbf{x}^s, \mathbf{x}^t \in \mathbb{R}^{H\times W\times D}$ from the spatial and temporal streams, with W, H, D denoting the width, height and the number of channels (filters), max fusion takes the maximum of these two feature maps at every spatial location i and j, and channel d, as:

$$y^{max}_{i,j,d} = \max\left(x^s_{i,j,d}, x^t_{i,j,d}\right) \ . \tag{1}$$

The use of the $\max$ operation makes the ordering of the channels in a convolutional layer arbitrary. Hence, this fusion strategy assumes arbitrary correspondences between the spatial and temporal streams. That said, this spatio-temporal model seeks filters so that these arbitrary correspondences between feature maps become as meaningful as possible according to the joint loss. After this fusion step, it also uses a deconvolution layer to produce an up-sampled saliency map as the final result as illustrated in the top row of Figure 1(b). We refer to this two-stream network architecture as STSMaxNet.

E. Spatio-Temporal Saliency with Convolutional Fusion

This network model integrates spatial and temporal streams by applying convolutional fusion. That is, the corresponding feature maps $\mathbf{x}^s$ and $\mathbf{x}^t$ respectively from the spatial and temporal streams are stacked together and then combined as follows:

$$\mathbf{y}^{conv} = \begin{bmatrix}\mathbf{x}^s & \mathbf{x}^t\end{bmatrix} * \mathbf{f} + \mathbf{b} \ , \tag{2}$$

where $\mathbf{f} \in \mathbb{R}^{1\times 1\times 2D\times D}$ denotes a bank of 1×1 filters, and $\mathbf{b} \in \mathbb{R}^D$ represents the bias term.

The main advantage of the convolutional fusion over the element-wise max fusion is that the filterbank $\mathbf{f}$ learns the optimal correspondences between the spatial and temporal streams based on the loss function, and reduces the number of channels by a factor of two through the weighted combinations of $\mathbf{x}^s$ and $\mathbf{x}^t$ with weights given by $\mathbf{f}$ at each spatial location. As demonstrated in the bottom row of Figure 1(b), this is followed by a number of convolutions and a final deconvolution layer to produce the saliency map. We refer to this network architecture as STSConvNet.

F. Spatio-Temporal Saliency Network with Direct Fusion

Finally, as another baseline model, we design a single stream network in which the appearance and optical flow images are stacked together and fed to the network as input. This model implements an early fusion strategy at the very beginning of the network architecture and can be seen as a special case of STSConvNet. Here, each layer of the network learns a set of filters that directly acts on the given appearance and motion frames. We refer to this model as STSDirectNet.

IV. Implementation Details

A. Network Architectures

The architecture of our single stream models is the same with that of the deep convolution network proposed in [25]. They take $320 \times 240 \times 3$ pixels images and processes them by the following operations: $C(96,7,3) \rightarrow LRN \rightarrow P \rightarrow C(256,5,2) \rightarrow P \rightarrow C(512,3,1) \rightarrow C(512,5,2) \rightarrow C(512,5,2) \rightarrow C(256,7,3) \rightarrow C(128,11,5) \rightarrow C(32,11,5) \rightarrow C(1,13,6) \rightarrow D$. Here, $C(d,f,p)$ represents a convolutional layer with d filters of size $f \times f$ applied to the input with padding p and stride 1. LRN denotes a local response normalization layer that carries out a kind of lateral inhibition, and P indicates a max pooling layer over 3×3 regions with stride 2. Finally, D is a deconvolution layer with filters of size $8 \times 8 \times 1$ with stride 4 and padding 2 which upscales the final convolution results to the original size. All convolutional layers except the last one are followed by a ReLU layer. Our spatial and temporal stream models in particular differ from each other in their inputs. While the first one processes still images, the next one accepts optical flow images as input.

For the proposed spatio-temporal saliency networks shown in Figure 1(b), we employ element-wise max and convolutional fusion strategies to integrate the spatial and temporal streams. Performing fusion after the fifth convolutional layer gives the best results for both of these fusion strategies. In STSMaxNet, the single stream networks are combined by applying element-wise max operation, which is followed by the same deconvolution layer in the single stream models. On the other hand, STSConvNet performs fusion by stacking the feature maps together and integrating them by a convolution layer $C(512,1,0)$ whose weights are initialized with identity matrices. The remaining layers are the same with those of the single stream models.

B. Data Preprocessing

We employ three publicly available datasets, 1.DIEM (Dynamic Images and Eye Movements) [35], 2. UCF-Sports [36] datasets and 3. MIT 300 dataset [59], which are described in detail in Section V, in our experiments. Since our networks accept inputs of size $320 \times 240 \times 3$ pixels and outputs saliency maps of the same size, all videos and ground truth fixation density maps are rescaled to this size prior to training. We use the publicly available implementation of DeepFlow [60] and we additionally extract optical flow information from the rescaled versions of subsequent video frames. Optical flow images are then generated by stacking horizontal and vertical flow components and the magnitude of the flow together. Some example optical flow images are shown in Figure 2.

C. Data Augmentation

Data augmentation is a widely used approach to reduce the effect of over-fitting and improve generalization of neural networks. For saliency prediction, however, classical techniques such as cropping, horizontal flipping, or RGB jittering are not very suitable since they alter the visual stimuli used in the eye tracking experiments in collecting the fixation data. Having said that, horizontal flipping is used in [25] as a data augmentation strategy although there is no theoretical basis for why this helps to obtain better performance.

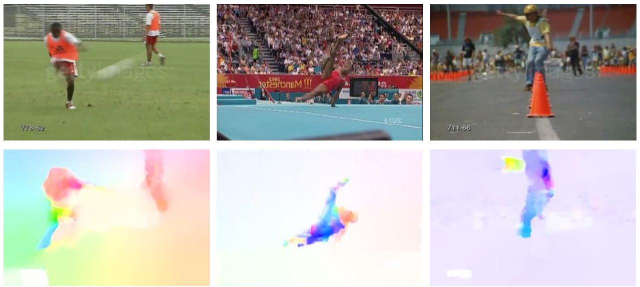


Fig. 2. Sample optical flow images generated for some frames of a video sequence from UCF-Sports dataset.

In our study, we propose to employ a new and empirically grounded data augmentation strategy for specifically training saliency networks. In [61], the authors performed a thorough analysis on how image resolution affects the exploratory behavior of humans through an eye-tracking experiment. Their experiments revealed that humans are quite consistent about where they look on high and low-resolution versions of the same images. Motivated with this observation, we process all video sequences and produce their low-resolution versions by down-sampling them by a factor of 2 and 4, and use these additional images with the fixations obtained from original high-resolution images in training. We note that in reducing the resolution of optical flow images the magnitude should also be rescaled to match with the down-sampling rate. It is worth-mentioning that this new data augmentation strategy can also be used for boosting performances of deep models for static saliency estimation.

D. Training

We employ the weights of the pretrained CNN model in [25] to set the initial weights of our spatial and temporal stream networks. In training the models, we use Caffe framework [62] and employed Stochastic Gradient Descent with Euclidean distance between the predicted saliency map and the ground truth. The networks were trained over 200K iterations where we used a batch size of 2 images, momentum of 0.9 and weight decay of 0.0005, which is reduced by a factor of 0.1 at every 10K iterations. Depending on the network architectures, it takes between 1 day to 3 days to train our models on the DIEM dataset by using a single 2GB GDDR5 NVIDIA GeForce GTX 775M GPU on a desktop PC equipped with 4-core Intel i5 (3.4 GHz) Processor and 16 GB memory. At test time, it takes approximately 2 secs to extract the saliency map of a single video frame.

V. Experimental Results

In the following, we first review the datasets on which we perform our experiments and provide the list of state-of-the-art computational saliency models that we compared against our spatio-temporal saliency networks. We then provide the details of the quantitative evaluation metrics that are used to assess the model performances. Next, we discuss our experimental results.

A. Datasets

To validate the effectiveness of the proposed deep dynamic saliency networks, our first set of experiments is carried out on the DIEM dataset [35]. This dataset is collected for the purpose of investigating where people look at dynamic scenes. It includes 84 high-definition natural videos including movie trailers, advertisements, etc. Each video sequence has eye fixation data collected from approximately 50 different human subjects. In our experiments, we only used the right-eye positions of the human subjects as done in [63].

We perform our second set of experiments on the UCF-Sports [36]. This dataset is collected from broadcast television channels such as the BBC and ESPN which consists of a set of sport actions [36]. The video sequences are collected from wide range of websites. This dataset contains 150 video sequences with 720 × 480 resolution and cover a range of scene and viewpoints. The dataset includes several actions such as diving, golf swing, kicking and lifting, and is used for action recognition. However, recently, additional human gaze annotations were collected in [36]. These fixations are collected over 16 human subjects under task specific and task-independent free viewing conditions.

Lastly, for the experiments on predicting eye fixations on still images, we choose a number of images from the MIT 300 dataset [59]. Selected images are the ones especially depicting an action and including objects that are interpreted as in motion. This dataset has eye fixation data collected from 39 subjects under free-viewing conditions for 3 secs for a total of 300 natural images with longest dimension 1024 pixels and the other dimension varied from 457 to 1024 pixels.

B. The compared computational saliency models

Through our experiments on DIEM and UCF-Sports datasets, we compare our deep network models with eight state-of-the-art dynamic saliency models: GVBS [17], PQFT [18], SR [21], Seo and Milanfar [20], Rudoy *et al.* [23], Fang *et al.* [64], Zhou *et al.* [42], and DWS [45] models. Moreover, we compare our two-stream deep models STSMaxNet and STSConvNet to a certain extent with deep static saliency model DeepSal [25] as the architectures of our TSNet and SSNet models are adapted from this model.

C. Evaluation Measures

We employ Area Under Curve (AUC), shuffled AUC (sAUC) [65], Pearson's Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS) [66], Normalized Cross Correlation (NCC) and χ^2 distance throughout our experiments for performance evaluation. We note that NCC and χ^2 distance are not widely-used measures but we report them as some recent studies [23], [43], [39] employ them in their analysis.

AUC measure considers the saliency maps as a classification map and uses the receiver operator characteristics curve to estimate the effectiveness of the predicted saliency maps in capturing the ground truth eye fixations. An AUC value of 1 indicates a perfect match with the ground truth while the performance of chance is indicated by a value close to 0.5. The AUC measure cannot account for the tendency of human subjects to look at the image center of the screen, i.e. the so-called center bias. Hence, we also report the results of the shuffled version of AUC (sAUC) where the center bias is compensated by selecting the set of fixations used as the false positives from another randomly selected video frame from the dataset instead of the processed frame.

CC treats the predicted saliency and the ground truth human fixation density maps as random variables and measures the strength of a linear relationship between two using a Gaussian kernel density estimator. While a value close to 0 indicates no correlation, a CC value close to +1/-1 demonstrates a good linear relationship. NSS estimates the average normalized saliency score value by examining the responses at the human fixated locations on the predicted saliency map which has been normalized to have zero mean and unit standard deviation. While a NSS value of 0 indicates chance in predicting eye fixtions, a non-negative NSS value, especially that of greater than 1, denotes correspondence between maps above chance.

NCC is a general measure used for assessing image similarity. It treats the ground truth fixation map and the predicted saliency map as images and estimates a score with values close to 1 implying high similarity and negative values indicating low similarity. χ^2 distance considers the saliency maps as a probability distribution map and compares the predicted map with the ground truth human fixation map accordingly. A perfect prediction model needs to provide a distance close to 0 for the χ^2 distance.

D. Experiments on DIEM

In our analysis, we first both qualitatively and quantitatively compare our proposed deep dynamic saliency networks, SSNet, TSNet, STSDirectNet, STSAvgNet, STSMaxNet and STSConvNet, with each other. Following the experimental setup of [63], we split the dataset into a training set containing 64 video sequences and a testing set including the remaining 20 representative videos covering different concepts. Specifically, we use the same set of splits used in [23].

As our STSMaxNet and STSConvNet models integrate spatial and temporal streams by respectively using element-wise and convolutional fusion strategies, we perform an extensive set of initial experiments to determine the optimum layers for the fusion process to take place in these two-stream networks. In particular, we train STSMaxNet and STSConvNet models by considering different fusion layers, and test each one of them on the test set by considering sAUC measure. In Table I, we provide these performance comparisons at different fusion layers. Interestingly, as can be seen from the table, fusing the spatial and temporal streams after the fifth convolution layer achieves the best results for both spatio-temporal networks. In the remainder, we use these settings for our STSMaxNet and STSConvNet models.

In Figure 3, for some sample video frames we provide the outputs of the proposed networks along with the corresponding

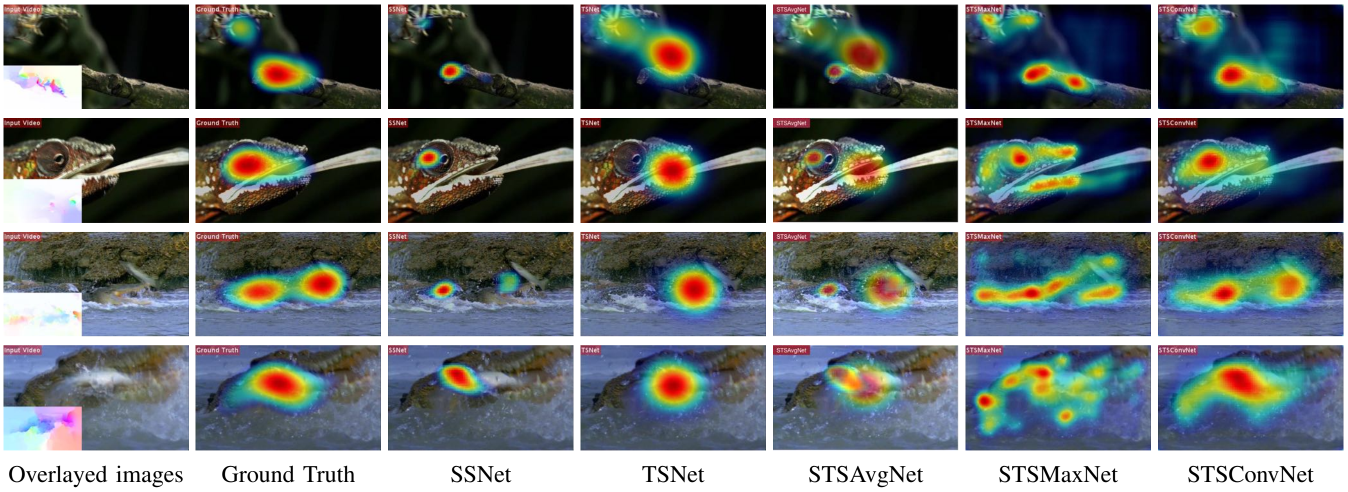


Fig. 3. Qualitative evaluation of the proposed saliency network architectures. For these sample frames from the DIEM dataset, our STSConvNet provides the most accurate prediction as compared to the other network models.

TABLE I
PERFORMANCE COMPARISON OF OUR SPATIO-TEMPORAL SALIENCY NETWORKS AT DIFFERENT FUSION LAYERS ON THE DIEM DATASET. THE REPORTED VALUES ARE SAUC SCORES. BEST PERFORMANCE IS ACHIEVED AFTER THE FIFTH CONVOLUTION LAYER.

Fusion Layers	STSMaxNet	STSConvNet
Conv2	0.52	0.71
Conv3	0.67	0.70
Conv4	0.76	0.83
Conv5	0.81	0.84
Conv6	0.80	0.79
Conv7	0.81	0.79

human fixation density maps (the ground truth). The input frames are given as overlayed images by compositing them with their consecutive frames to show the motion in the scenes. Saliency maps are shown as heatmaps superimposed over the original image for visualization purposes. As can be seen from these results, SSNet, which does employ appearance but not motion information, in general provides less accurate saliency maps and misses the foreground objects or their parts that are in motion. TSNet gives relatively better results, but as shown in the second and the third row, it does not identify all of the salient regions as it focuses more on the moving parts of the images. Directly averaging the saliency maps of these two single stream models, referred to as STSAvgNet, does not produce very satisfactory results either. As expected, STSMaxNet is slightly better since max fusion enforces to learn more effective filters in order to combine the spatial and temporal streams. Overall, STSConvNet is the best performing model. This can be rooted in the application of 1×1 convolutional filters that learn the optimal weights to combine appearance and motion feature maps.

In Table II, we present the quantitative results averaged over all video sequences and frames on the test split of the DIEM dataset. Here, we compare and contrast our single- and two-stream saliency networks with eight existing dynamic saliency methods, namely GVBS [17], SR [21], PQFT [18], Seo and Milanfar [20], Rudoy *et al.* [23][1], Fang *et al.* [64], Zhou *et al.* [42] and DWS [45] models.

TABLE II
PERFORMANCE COMPARISONS ON THE DIEM DATASET.

	AUC	sAUC	CC	NSS	χ^2	NCC
SSNet	0.72	0.69	0.35	1.85	0.48	0.41
TSNet	0.79	0.77	0.41	1.98	0.40	0.43
STSDirectNet	0.71	0.60	0.37	1.53	0.49	0.28
STSAvgNet	0.68	0.62	0.37	1.67	0.49	0.37
STSMaxNet	0.83	0.81	0.46	2.01	0.31	0.45
STSConvNet	0.87	0.84	0.47	2.15	0.29	0.46
STSConvNet*	**0.88**	**0.86**	**0.48**	2.18	**0.28**	**0.47**
GBVS [17]	0.74	0.70	0.47	2.04	0.47	0.38
SR [21]	0.69	0.64	0.30	2.22	0.57	0.40
PQFT [18]	0.71	0.67	0.33	1.77	0.52	0.33
Seo-Milanfar [20]	0.59	0.51	0.10	0.12	0.75	0.28
Rudoy *et al.* [23]	–	0.74	–	–	0.31	–
Fang *et al.* [64]	0.71	0.48	0.21	0.55	0.87	0.40
Zhou *et al.* [42]	0.60	0.52	0.13	0.24	0.71	0.22
DWS [45]	0.81	0.79	0.32	**2.97**	0.40	0.39

Among our deep saliency networks, we empirically find that STSDirectNet provides the worst quantitative results. This is in line with our observation in Table I that delaying the integration of appearance and motion streams to a certain extent leads to more effective learning of mid and low level features. Secondly, we see that SSNet performs considerably lower than Temporal stream network, which demonstrates that motion is more vital for dynamic saliency. STSMaxNet gives results better than those of the single stream models but our STSConvNet model performs even better. It can be argued that STSConvNet learns more effective filters that combine spatial and temporal streams in an optimal manner. In addition, when we employ the data augmentation strategy proposed in the previous section, it further improves the overall performance of STSConvNet. In the remainder, we refer to this model

[1] Since the code provided by the authors is not working correctly, sAUC and χ^2 scores are directly taken from [23].

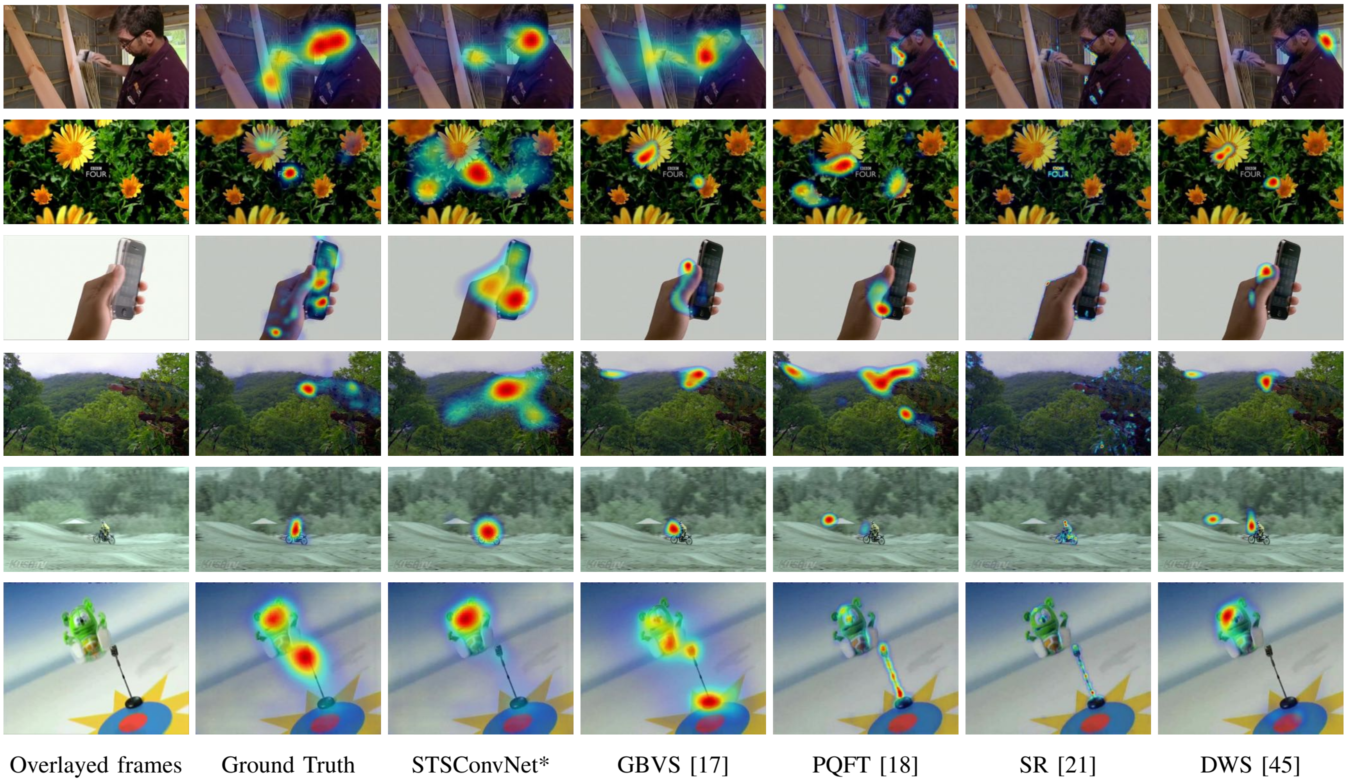


Fig. 4. Qualitative comparison of our STSConvNet* model against some dynamic saliency models on DIEM dataset. Our model clearly produces better results.

with data augmentation as STSConvNet*. When we compare our proposed STSMaxNet, STSConvNet, and STSConvNet* models to the previous dynamic saliency methods, our results demonstrate the advantages of two-stream deep CNNs that they consistently outperform all those approaches, including the very recently proposed DWS model, according to five out of six evaluation measures.

We present some qualitative results in Figure 4 where we again provide the input frames as transparent overlayed images showing the inherent motion. We observe that the proposed STSConvNet* model localizes the salient regions more accurately than the existing models. For example, for the frame given in the first row, none of the compared models correctly capture the fixations over the painting brush. Similarly, for the second and the third frames, only our spatio-temporal saliency network fixates to the text and the cellular phone in the frames, respectively.

E. Experiments on UCF-Sports

Learning-based models might sometimes fail to provide satisfactory results for a test sample due to a shift from the training data domain. To validate generalization ability of our best-performing STSConvNet* model, we perform additional experiments on UCF-Sports dataset. In particular, we do not carry out any training for our model from scratch or fine-tune it on UCF-Sports but rather use the predictions of the model trained only on DIEM dataset.

In Table III, we provide the performance of our model compared to the previous dynamic saliency models which are publicly available on the web. As can be seen, our STSConvNet* model performs better than the state-of-the-art models according to majority of the evaluation measures. It especially outperforms the recently proposed DWS model in terms of all measures. These results suggest that our two-stream network generalizes well beyond the DIEM dataset.

TABLE III
PERFORMANCE COMPARISONS ON THE UCF-SPORTS DATASET.

	AUC	sAUC	CC	NSS	χ^2	NCC
GBVS [17]	0.83	0.52	0.46	1.82	0.54	**0.59**
SR [21]	0.78	0.69	0.26	1.20	0.42	0.52
PQFT [18]	0.69	0.51	0.29	1.15	0.64	0.48
Seo-Milanfar [20]	0.80	0.72	0.31	1.37	0.56	0.36
Fang *et al.* [64]	**0.85**	0.70	0.44	1.95	0.52	0.33
Zhou *et al.* [42]	0.81	0.72	0.36	1.71	0.56	0.37
DWS [45]	0.76	0.70	0.28	2.01	0.40	0.49
STSConvNet*	0.82	**0.75**	**0.48**	**2.13**	**0.39**	0.54

In Figure 5, we present sample qualitative results of Fang *et al.* [64] and DWS model [45] (two recently proposed models) and our STSConvNet model on some video frames. For instance, we observe that for the sample frame given in the first row, our model fixates to both the runner and the crowd. For the second and the third sample frames, the compared models do not accurately localize the weight lifter and the cowboys as salient, respectively. Similarly, the proposed STSConvNet* model predicts the eye fixations better than the competing models for the fourth image containing a guardian walking in a corridor. For the last diving image, STSConvNet* and

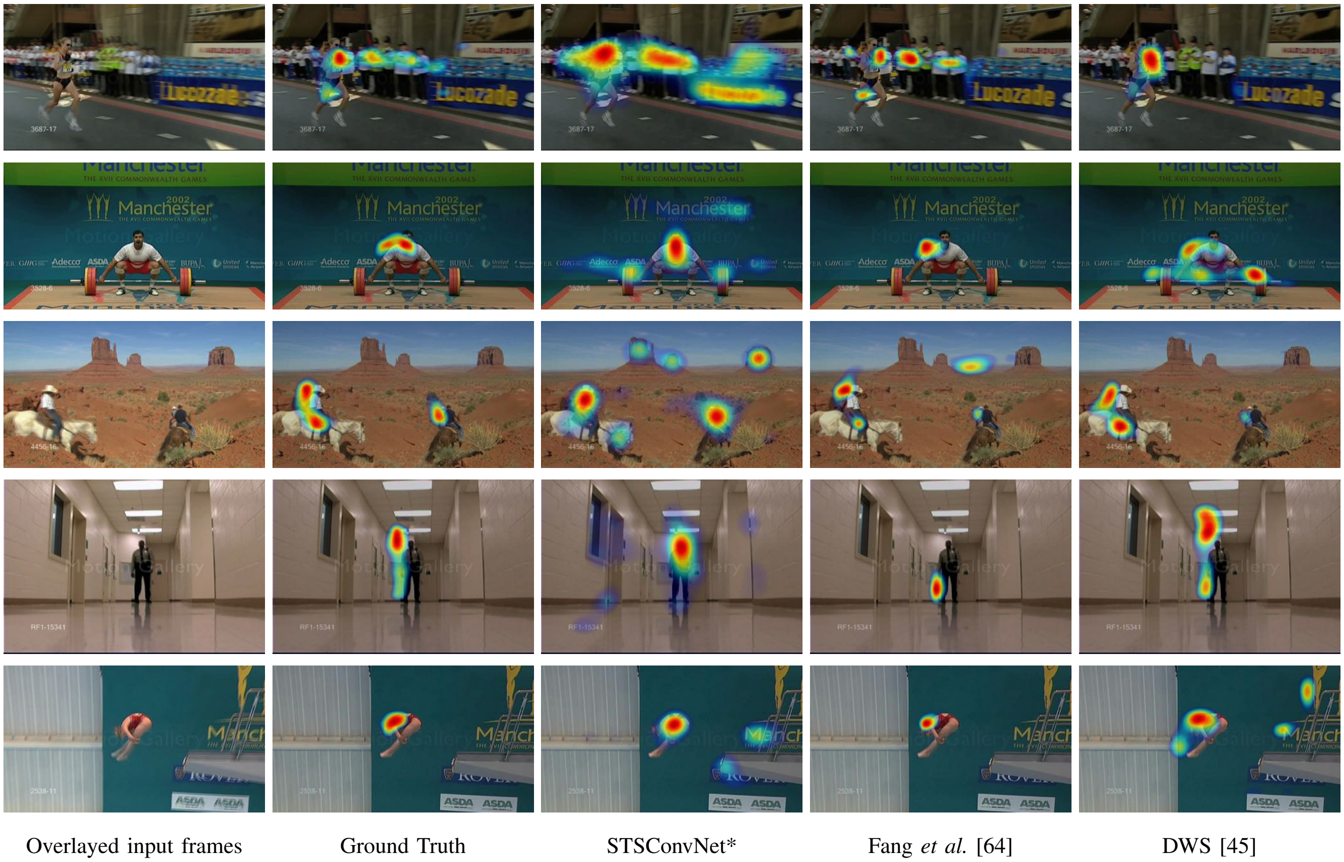


Fig. 5. Qualitative comparison of our STSConvNet* model against some previous dynamic saliency models on UCF-Sports dataset. Our spatio-temporal saliency network outperforms the others.

Fang *et al.* give results fairly close to the ground truth, while DWS output some spurious regions as salient.

F. Experiments on Still Images from MIT300

Deep static saliency networks achieve excellent performances on existing benchmark datasets for static saliency estimation. These models, however, only exploit spatial information captured in still images, but sometimes an image, despite being taken in an instant, might carry plenty of information regarding the inherent motion exist in it. In a recent study by Bylinski *et al.* [53], the authors demonstrate that the areas showing these kind of activities are indeed evidently important for saliency prediction since humans have tendency to look at the objects that they think in motion or that are in interaction with humans or some other objects. Motivated by these observations, in this section, we present the failures or the shortcomings of the current deep static saliency models through some examples, and show how motion information exist in still images can be utilized to fill in the semantic gap exist in the current static saliency models.

Figure 6 presents sample images taken from [53] and which are from the MIT 300 dataset [59] where highly fixated regions (which cover the 95th percentile of the human fixation maps) are highlighted with yellow curves. As can be clearly seen from these examples, the state-of-the-art deep static models generally fail to give high saliency values to regions where an action occurs or which contains objects that are interpreted as in motion. To capture those regions, we employ the deep optical flow prediction model [37] which extracts optical flow from static images. Once we estimate the motion map of a still image, we can exploit this information together with the RGB image as inputs to our spatio-temporal saliency network (STSConvNet) to extract a saliency map. We observe that using these (possibly noisy) motion maps within our framework provides more accurate predictions than the existing deep static saliency models, and even captures the objects of gaze as illustrated in the first two sample images. These experiments reveal that the performances of static saliency networks can be improved by additionally considering motion information inherent in still images.

VI. Conclusion

In this work, we have investigated several deep architectures for predicting saliency from dynamic scenes. Two of these deep models are single-stream convolutional networks respectively trained for processing spatial and temporal information. Our proposed spatio-temporal saliency networks, on the other hand, are built based on two-stream architecture and employ different fusion strategies, namely direct averaging, max fusion and convolutional fusion, to integrate appearance and motion features, and they are all trainable in an end-to-end manner. While training these saliency networks, we additionally employ an effective and well-founded data augmentation method that utilizes low-resolution versions of the video frames and

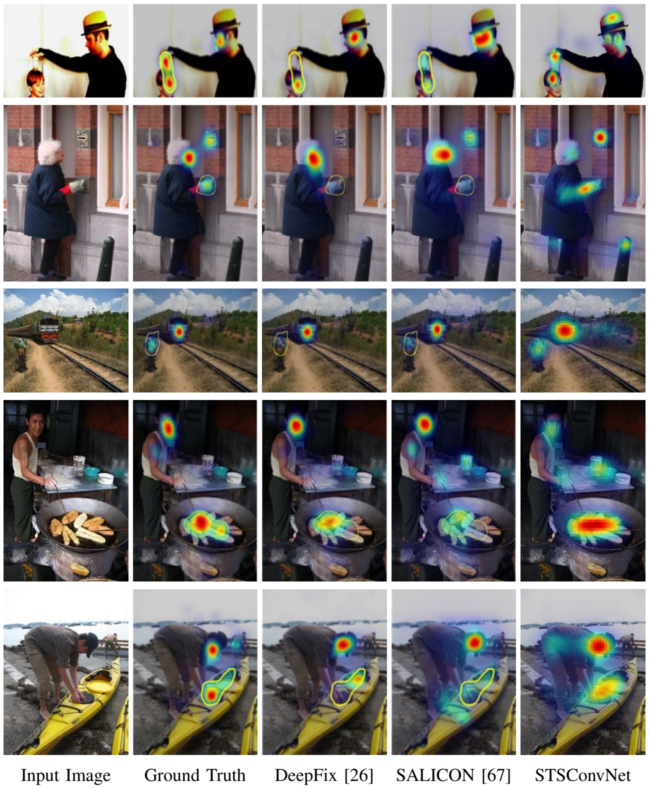


Fig. 6. Some experiments on still images. While the top performing deep static saliency models fail to compute satisfactory results in these images (results taken from [53]), our spatio-temporal saliency network (STSConvNet) can produce better saliency maps using predicted optical flow maps.

the ground truth saliency maps, giving a significant boost in performance. Our experimental results demonstrate that the proposed STSConvNet model achieves superior performance over the state-of-the-art methods on DIEM and UCF-Sports datasets. Lastly, we provide some illustrative example results on a number of challenging still images, which show that static saliency estimation can also benefit from motion information. This is left as an interesting topic for future research.

Acknowledgment

This research was supported in part by TUBITAK Career Development Award 113E497 and Hacettepe BAP FDS-2016-10202.

References

[1] J. Huang, X. Yang, X. Fang, W. Lin, and R. Zhang, "Integrating visual saliency and consistency for re-ranking image search results," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 653–661, 2011.

[2] H. Kim and S. Lee, "Transition of visual attention assessment in stereoscopic images with evaluation of subjective visual quality and discomfort," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2198–2209, 2015.

[3] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, and W. Zhang, "Saliency-guided quality assessment of screen content images," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1098–1110, 2016.

[4] D. Y. Chen and Y. S. Luo, "Preserving motion-tolerant contextual visual saliency for video resizing," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1616–1627, 2013.

[5] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[6] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level cnn: Saliency-aware 3-d cnn with lstm for video action recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 510–514, 2017.

[7] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann, "Devnet: A deep event network for multimedia event detection and evidence recounting," in *CVPR*, 2015, pp. 2568–2577.

[8] C. Shen, X. Huang, and Q. Zhao, "Predicting eye fixations on webpage with an ensemble of early features and high-level representations from deep network," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2084–2093, 2015.

[9] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, 2013.

[10] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cats visual cortex." *The Journal of physiology*, vol. 160, pp. 106–154, Jan 1962.

[11] A. Treisman and G. Gelade, "A feature integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.

[12] J. M. Wolfe, "Guided search 2.0: A revised model of visual search," *Psychonomic Bulletin & Review*, vol. 1, no. 2, pp. 202–238, 1994.

[13] N. Imamoglu, W. Lin, and Y. Fang, "A saliency detection model using low-level features based on wavelet transform," *IEEE Transactions on Multimedia*, vol. 15, no. 1, pp. 96–105, 2013.

[14] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[15] J. G. Yu, G. S. Xia, C. Gao, and A. Samal, "A computational model for object-based visual saliency: Spreading attention along gestalt cues," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 273–286, 2016.

[16] J. Lei, B. Wang, Y. Fang, W. Lin, P. L. Callet, N. Ling, and C. Hou, "A universal framework for salient object detection," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1783–1795, 2016.

[17] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *NIPS*, 2006, pp. 545–552.

[18] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *CVPR*, 2008, pp. 1–8.

[19] X. Cui, Q. Liu, and D. Metaxas, "Temporal spectral residual: fast motion saliency detection," in *ACM MM*, 2009, pp. 617–620.

[20] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vision*, vol. 9, no. 12, pp. 15–15, 2009.

[21] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *CVPR*, 2007, pp. 1–8.

[22] S. Mathe and C. Sminchisescu, "Dynamic eye movement datasets and learnt saliency models for visual action recognition," in *ECCV*, 2012, pp. 842–856.

[23] D. Rudoy, D. Goldman, E. Shechtman, and L. Zelnik-Manor, "Learning video saliency from human gaze using candidate selection," in *CVPR*, 2013, pp. 1147–1154.

[24] M. Kummerer, L. Theis, and M. Bethge, "Deep gaze I: Boosting saliency prediction with feature maps trained on imagenet," in *ICLR Workshop*, 2015.

[25] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *CVPR*, 2016.

[26] S. S. S. Kruthiventi, K. Ayush, and R. V. Babu, "Deepfix: A fully convolutional neural network for predicting human eye fixations," arXiv preprint arXiv:1510.02927, 2015.

[27] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *CVPR*, 2014, pp. 2798–2805.

[28] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *CVPR*, 2015, pp. 1265–1274.

[29] N. D. B. Bruce, C. Catton, and S. Janjic, "A deeper look at saliency: Feature contrast, semantics, and beyond," in *CVPR*, 2016.

[30] S. Jetley, N. Murray, and E. Vig, "End-to-end saliency mapping via probability distribution prediction," in *CVPR*, 2016.

[31] S. S. S. Kruthiventi, V. Gudisa, J. H. Dholakiya, and R. Venkatesh Babu, "Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation," in *CVPR*, 2016.

[32] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *ICCV*, 2015, pp. 262–270.

[33] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014, pp. 1725–1732.

[34] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *NIPS*, 2014, pp. 568–576.

[35] 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*, vol. 3, no. 1, pp. 5–24, 2011.

[36] 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*, vol. 37, 2015.

[37] J. Walker, A. Gupta, and M. Hebert, “Dense optical flow prediction from a static image,” in *ICCV*, 2015, pp. 2443–2451.

[38] W. Sultani and I. Saleemi, “Human action recognition across datasets by foreground-weighted histogram decomposition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 764–771.

[39] T. Mauthner, H. Possegger, G. Waltner, and H. Bischof, “Encoding based saliency detection for videos and images,” in *CVPR*, 2015, pp. 2494–2502.

[40] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, “Segmenting salient objects from images and videos,” in *European Conference on Computer Vision*. Springer, 2010, pp. 366–379.

[41] Z. Ren, L.-T. Chia, and D. Rajan, “Video saliency detection with robust temporal alignment and local-global spatial contrast,” in *ICMR*, 2012, pp. 47:1–47:8.

[42] F. Zhou, S. B. Kang, and M. F. Cohen, “Time-mapping using space-time saliency,” in *CVPR*, 2014.

[43] J. Zhao, C. Siagian, and L. Itti, “Fixation bank: Learning to reweight fixation candidates,” in *CVPR*, 2015, pp. 3174–3182.

[44] S. Hossein Khatoonabadi, N. Vasconcelos, I. V. Bajic, and Y. Shan, “How many bits does it take for a stimulus to be salient?” in *CVPR*, 2015, pp. 5501–5510.

[45] V. Leboran, A. Garcia-Diaz, X. Fdez-Vidal, and X. Pardo, “Dynamic whitening saliency,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2016.

[46] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016.

[47] R. Girshick, “Fast R-CNN,” in *ICCV*, 2015, pp. 1440–1448.

[48] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *CVPR*, 2015, pp. 3431–3440.

[49] X. Li, L. Zhao, L. Wei, M. H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, “Deepsaliency: Multi-task deep neural network model for salient object detection,” *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3919–3930, 2016.

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *NIPS*, 2012, pp. 1097–1105.

[51] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *ICLR*, 2015.

[52] R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva, “Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence,” *Scientific Reports*, 2016.

[53] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand, “Where should saliency models look next?” in *European Conference on Computer Vision*. Springer, 2016, pp. 809–824.

[54] M. A. Goodale and A. D. Milner, “Separate visual pathways for perception and action,” *Trends in Neurosciences*, vol. 15, no. 1, pp. 20–25, 1992.

[55] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *ICCV*, 2015, pp. 4489–4497.

[56] 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*, vol. 35, no. 1, pp. 221–231, 2013.

[57] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, “Convolutional learning of spatio-temporal features,” in *ECCV*, 2010, pp. 140–153.

[58] L. Sun, K. Jia, D. Y. Yeung, and B. E. Shi, “Human action recognition using factorized spatio-temporal convolutional networks,” in *ICCV*, 2015, pp. 4597–4605.

[59] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, “Mit saliency benchmark.”

[60] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, “Deepflow: Large displacement optical flow with deep matching,” in *ICCV*, 2013, pp. 1385–1392.

[61] T. Judd, F. Durand, and A. Torralba, “Fixations on low-resolution images,” *Journal of Vision*, vol. 11, no. 4, pp. 14–14, 2011.

[62] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” arXiv preprint arXiv:1408.5093, 2014.

[63] A. Borji, D. N. Sihite, and L. Itti, “Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study,” *IEEE Trans. Image Processing*, vol. 22, no. 1, pp. 55–69, 2013.

[64] Y. Fang, Z. Wang, W. Lin, and Z. Fang, “Video saliency incorporating spatiotemporal cues and uncertainty weighting,” *IEEE Trans. Image Processsing*, vol. 23, no. 9, pp. 3910–3921, 2014.

[65] L. Zhang, M. H. Tong, T. K. M. anf H. Shan, and G. W. Cottrell, “SUN: A Bayesian framework for saliency using natural statistics,” *Journal of Vision*, vol. 8, no. 7, pp. 1–20, 2008.

[66] R. J. Peters, A. Iyer, L. Itti, and C. Koch, “Components of bottom-up gaze allocation in natural images,” *Vision Research*, vol. 45, no. 8, pp. 2397–2416, 2005.

[67] M. Jiang, S. Huang, J. Duan, and Q. Zhao, “Salicon: Saliency in context,” in *CVPR*, 2015, pp. 1072–1080.

Cagdas Bak received his B.Sc. degree in Computer Engineering from Hacettepe University, Ankara, Turkey in 2013. He got his Master of Science degree from the same deparment in 2016. His current research interests include image and video processing, visual saliency and deep learning.

Aysun Kocak received her B.Sc. degree in Computer Engineering from Hacettepe University, Ankara, Turkey in 2012. She is currently pursuing her Ph.D. studies in the same department. Her current research interests include image and video processing, scanpath estimation, visual saliency and deep learning.

Erkut Erdem has received his B.Sc. and M.Sc. degrees respectively in 2001 and 2003 from the Department of Computer Engineering, Middle East Technical University. After completing his Ph.D. work at the Middle East Technical University in 2008, he continued his post-doctoral research studies at Télécom ParisTech, Ecole Nationale Supérieure des Télécommunications between 2009 and 2010. He is an Assistant Professor at the Department of Computer Engineering, Hacettepe University since 2014. While his research interests in general concern computer vision and machine learning, he is conducting research activities specifically lie in image editing and smoothing, visual saliency prediction and language and vision.

Aykut Erdem has received his B.Sc. and M.Sc. degrees in Computer Engineering in 2001 and 2003 from Middle East Technical University (METU), Ankara, Turkey. Upon receiving his Ph.D. degree in 2008, he worked as a post-doctoral researcher in the Computer Science Department of CaFoscari University of Venice, Italy from 2008-2010. In 2010, he joined Hacettepe University, Ankara, Turkey, where he is now an Assistant Professor at the Department of Computer Engineering. His research interests include computer vision and machine learning, currently focused on image matting, summarization of videos and large image collections, and integrating language and vision.