Still Image Action Recognition by Predicting Spatial-Temporal Pixel Evolution

Marjaneh Safaei
*University Of Central Florida
Department of Computer Science
marjaneh.safaei@knights.ucf.edu

Hassan Foroosh
University Of Central Florida
Department of Computer Science
foroosh@cs.ucf.edu

Abstract

Both spatial and temporal patterns provide crucial information for recognizing human actions. However, lack of temporal information in still images is a major obstacle in single-image action recognition. In this paper, (i) We introduce a novel image representation domain, Ranked Saliency Map and Predicted Optical Flow *or $Rank_{SM-POF}$ for short. This domain captures both actor appearance and the predicted future movement patterns of the actor. This is accomplished through capturing the temporal ordering of each pixel by training a linear ranking machine on the predicted tensor of spatial-temporal representation of images. (ii) We employ a transfer learning approach to propose a new spatial-temporal Convolutional Neural Network, named STCNN for the task of single image action classification, by fine-tuning a CNN which is pre-trained specifically for appearance based classification. (iii) Finally, extensive experiments on six benchmarks clearly demonstrate that appearance and predicted motion are complementary sources of information and using both leads to significant performance improvement in single image action recognition, hence outperforming state-of-the-art methods.*

1. Introduction

Considering the tremendous growth in the number of images on the Web, it is of paramount importance to automate the analysis of human behavior in still images. There has been a remarkable progress on human action recognition in video data. In contrast, action recognition in still images remains more challenging and less attended by researchers.

Still image-based action recognition has numerous useful applications, including (1) image annotation. Automated action recognition in still images can help to annotate verbs (for actions) on Internet images; (2) Action or behavior based image retrieval; (3) Video frame reduction in video-based action recognition; (4) Human-computer interaction (HCI).

Figure 1. Still images depicting actions recognizable by humans.

Action recognition in still images is particularly challenging due to the absence of motion information. The problem becomes exacerbated when there is no contextual information, *e.g.* when no objects (other than the human) are present in the image. Interestingly, humans are less challenged to perform this task compared to machines. Human brain is able to not only recognize what is present in the image but also predict what action might take place next, by guessing the most probable future motions for each pixel. Therefore, predicting the future motion plays a very important role in human action recognition, especially when the action relies mainly on human body motions, and not the human-object interactions. Fig.1 depicts a few actions that can be recognized well by humans.

While recent methods on still image action recognition typically represent actions by spatial features, in this paper we argue for the importance of a spatial-temporal representation for a single image, derived from both static and pre-

[*] This work was supported in part by the National Science Foundation under grant IIS-1212948

978-1-7281-1975-5/19/$31.00 ©2019 IEEE
DOI 10.1109/WACV.2019.00019

dicted dynamic cues, suitable for action recognition. Motion is a missing information in an image. However, it is a valuable cue for action recognition. Therefore, recognizing actions depends not only on the spatially-salient pixels but also their motion information. Here, we also argue that the predicted motion of each pixel depends on how spatially salient that pixel is, or its relative spatial position with respect to other salient pixels. Therefore, it is reasonable to assume that actions have a characteristic evolution from spatial to motion.

In this paper, our main contributions are threefold: (1) Domain Mapping: This step is based on Rank-SVM, projecting images onto a new domain, that we refer to as $Rank_{SM-POF}$. We propose to use the predicted optical flow in a static image as a means of compensating for the missing temporal information, while using the saliency map to represent the spatial information about the location and the shape of the predicted significant regions of the image. In this step, the *Saliency Map* and the *Predicted Optical Flow* are employed to covert the raw still image to a novel image representations that capture both spatially salient parts of the actor as well as the predicted movement patterns of the actor by learning the functional parameters of the linear ranking functions. (2) Action classification in single images through *STCNN*: We propose a deep CNN that is trained to classify human actions based on both spatial and predicted temporal features, i.e. using the images in the $Rank_{SM-POF}$ as input. We take a network trained on a different domain for a different source task, and adapt it for the proposed domain $Rank_{SM-POF}$ and the new target task, which is action classification in single images. We do a supervised domain adaptation via fine-tuning the pre-trained network on the new domain. (3) We created a new large dataset of *2M* still images, UCFSI-101, by sampling random video frames from the UCF-101 dataset [34]. Our experiments on UCFSI-101, Willow [3], Stanford-40 [39], WIDER [37], BU_{101} [21] datasets, as well as our newly collected still images from the wild demonstrate that the proposed domain mapping method is extremely effective in capturing action attributes in still images. As a result, *STCNN* outperforms the state-of-the-art methods by a significant margin. Figures 2 and 3 depict the architecture of the proposed approach as well as the domain mapping stage in detail, respectively.

2. Related Work

In the context of still image action recognition, a big gap is the lack of temporal information in the image, and thus the traditional spatiotemporal features [25] cannot be directly extracted. Therefore, unlike conventional action recognition methods using videos, the low-level features extracted from space-time volume, e.g. Space-Time Interest Point (STIP) based features [17], cannot be applied.

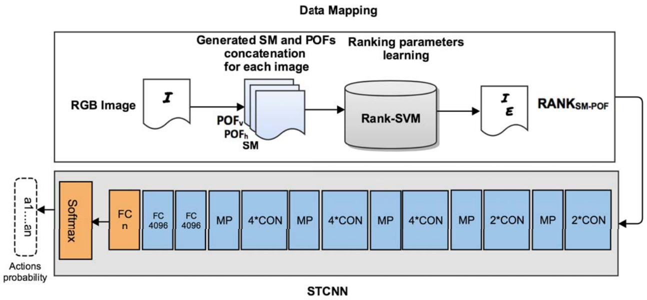

Figure 2. Schematic overview of the proposed framework. Domain Mapping: RGB images are mapped to the new proposed domain $Rank_{SM-POF}$. CNN action classifier: Deep features from the activation of the fully connected layers of a STCNN model are used as input to a softmax layer, modified to fit our domain, to determine the final classification.

Human action recognition in still images has gained increasing attention in recent years [40, 23, 10, 13] due to its challenging nature and its importance in applications such as image search and retrieval, image annotation, video summarization, and human-computer interaction (HCI), to name a few. Recent methods [23, 10, 13, 42] on single image action recognition typically represent actions only by spatial features. Sharma et al. [30] proposed a model using a collection of discriminative templates with associated scale space locations. Liang et al. [19] proposed a hybrid algorithm using the DPM (Deformable Part Model) to detect the human body parts or objects and then train a DBN (Deep Belief Network) to recognize actions using the locations of the body parts and the objects. Herein, we propose a novel spatiotemporal domain for still image action recognition, where the missing temporal dimension is predicted.

A comprehensive survey was performed by [11] on this problem, where existing action recognition methods are categorized based on low-level features (e.g SIFT [20] and HOG [2]), and high-level cues, such as (1) attributes [1, 39], (2) body parts and pose [3, 44, 22, 27], which are challenging due to the limited number of poses they can detect and also the fact that many different human actions share almost the same poses, and (3) human-object interaction [26, 39, 38, 4, 19], which rely on the presence and detection of objects, as additional contextual information. This poses a challenge when the action merely involves a human with no object interaction. Hence we advocate that the salient parts of the body and their predicted motion can play crucial roles for action recognition.

Action classification in still images has recently benefited from CNN models [10, 12, 23], which offer an outstanding performance. The tradeoff, however, is that training CNNs requires millions of parameters and often a huge number of annotated images. This poses a challenge when fewer training data are available. CNNs are high-capacity classifiers with very large numbers of parameters that must be learned from training examples. Action recognition in still images suffers from lack of annotated images for a wide range of action classes. Recent works on single image action recognition focus on a limited number of action classes and primarily on human-object interactions.

What sets our work apart from existing efforts, in this context, is that we propose to use not only spatial cues but also predict temporal patterns for recognizing human actions in still images. We define *spatial* as the salient human body pixels that describe the properties of human actions, while the term *temporal* represents the predicted optical flow for pixels in still images. In our method the characterization of the underlying human action does not require any bounding box annotations. In contrast, approaches adopted by related works require additional input, e.g. Delaitre et al. [3] extract features in areas within or surrounding the human bounding boxes. Some other approaches employ Bag of Words (BOW)-based image representations [13, 14, 29].

It is worth mentioning that most of the large image datasets such as Caltech [7], Pascal VOC [6], and ImageNet [5] have been created for the purpose of object classification tasks, but not for action classification. To overcome this issue, we cast the problem as a supervised domain-adaptation problem. This would allow us to build a ‘spatial-temporal’ CNN action classifier by fine tuning an existing object-classifying CNN model on the proposed new domain. The fine-tuning process would require a smaller training set, as long as the transfer learning method is effective.

3. Domain Mapping

Which parts of the human body are likely to move? What are their overall shapes? And in which direction they may move? Answering these questions may be viewed as a non-semantic form of action prediction when dealing with static images. For identifying an action not all pixels carry the same importance. Some pixels capture less meaningful information or even carry misleading information, while others carry more discriminative information. Here we describe the domain mapping stage that projects data from the source domain of static RGB images onto the $Rank_{SM-POF}$, which only focuses on pixels that have crucial role in human action recognition. Our representation reduces a single image in a manner that emphasizes regions that are “appearance-salient” as well as pixels in those regions that are predicted to be “motion-salient”, and thus more likely to represent an action. This leads to a significant reduction in noise by discarding pixels that are not contributing to the actor’s appearance or their predicted body motion.

The *Domain Mapping* stage itself consists of two steps. First, we create an intermediate third-order tensor, $\mathcal{Q} \in \mathbb{R}^{P \times F \times 3}$, for each action class $a_i \in A$. Second, tensor $\mathcal{Q}$ is projected onto a new domain $Rank_{SM-POF}$, providing a compact representation of the spatial-temporal attributes using an ordinal regression that characterizes the extent of contribution of a pixel for recognizing an action. Below is the description of these two steps and the motivation/intuition behind each.

3.1. Forming tensor $\mathcal{Q}$

The third-order tensor $\mathcal{Q} \in \mathbb{R}^{P \times F \times 3}$ is a feature space tensor consisting of spatial and predicted temporal information related to the underlying action for all training images in action class a_i. Each column in tensor $\mathcal{Q}$ corresponds to an image I. Therefore, P denotes the number of pixels in I and F denotes the number of training samples in action class a_i. We propose to use the Predicted Optical Flow (POF) in a static image as a means of compensating for the missing temporal information, while using the Saliency Map (SM) to represent the spatial information about the predicted significant regions of the image. In tensor $\mathcal{Q}$, the first channel represents the Sailency Map (SM) of images, and the second and the third channels represent the Predicted Horizontal Optical Flow (POF_h) and the Predicted Vertical Optical Flow (POF_v) for each pixel in images, respectively.

The first channel of $\mathcal{Q}$, represents the static saliency map of the image using a bottom-up approach [28], where each pixel indicates the statistical likelihood of saliency of a feature matrix given its surrounding feature matrices. We set all the values below some threshold τ in the SM channel to zero, to ensure a better localization and representation of the shape of the salient regions of the image. The threshold τ was selected automatically using Otsu’s method [24].

For the predicted *temporal* information, we used a CNN model similar to the one proposed by Walker et al. [35] to predict a dense optical flow. This optical flow map represents how and where each pixel in the input static image is predicted to move. For this purpose, the optical flow vectors are first quantized into 40 clusters by *k*-means. The problem is then treated in a manner similar to semantic segmentation, where each region in the image is classified as a particular cluster of the optical flow. These clusters are then used to predict the motion direction for each pixel. A softmax loss layer at the output is then used for computing gradients. We generate the output as softmax probabilities over the optical flow vectors for each pixel. The softmax loss is spatial, summing over all the individual region losses. This leads to an $M \times N \times C$ softmax layer, where M, N and C represent the number of rows, columns and clusters, respectively. Let I represent the image and Y be the ground truth optical flow labels represented as quantized clusters. The spatial loss function $L(I, Y)$ as defined by Walker et al. is

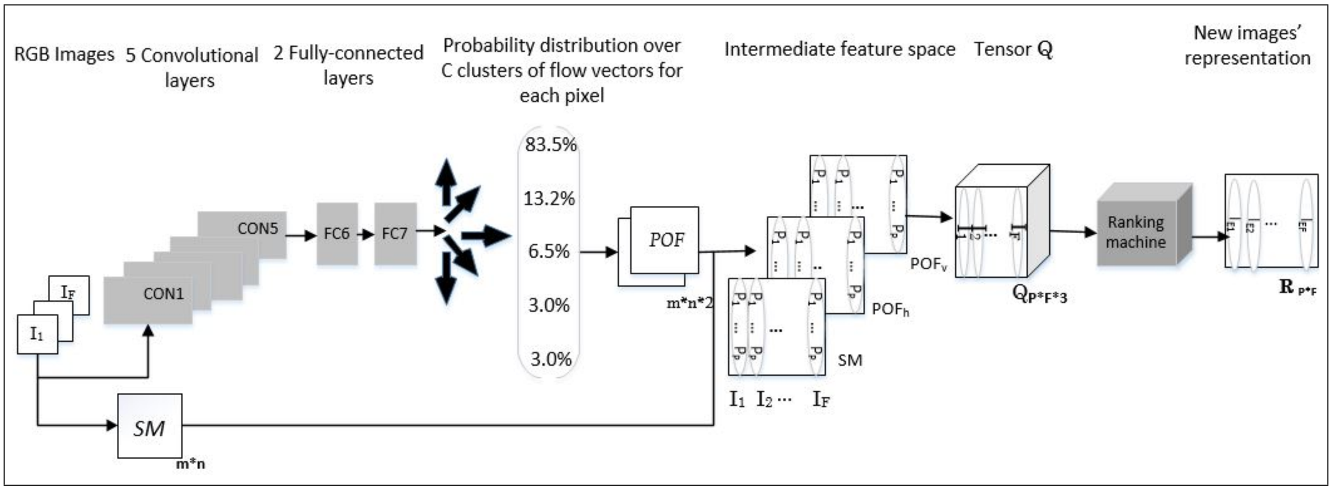

Figure 3. Domain Mapping process. Step 1: Converting RGB images to the intermediate feature space tensor Q by concatenating their Saliency Map (SM) and Predicted Optical Flows (POF). Step 2: Mapping tensor Q onto matrix R, representing I_Es, new representations through Rank-SVM algorithm.

given by:

$$L(I, Y) = -\sum_{i=1}^{M\times N}\sum_{r=1}^{C} \mathbb{1}(Y_i = r)\log F_{i,r}(I), \qquad (1)$$

where $F_{i,r}(I)$ represents the probability that the i_{th} pixel will move according to cluster *r*, and $\mathbb{1}(Y_i = r)$ is an indicator function. A problem with this loss function is that it implicitly assumes a uniform probability mass function (pmf) for the motion clusters, which is very unlikely and prone to noise. Therefore, we modified Eq. (1) in order to minimize the noise by taking into account only the K most-likely clusters, i.e. the K clusters with the highest probability and optimized the pre-trained CNN [35] using our custom loss function in Eq. (2). This amounts to replacing the second summation in Eq. (1) with an order statistic filter as follows:

$$\hat{L}(I, Y) = -\sum_{i=1}^{M\times N}\sum_{r=1}^{C} \omega_r P_{i,(r)}, \qquad (2)$$

where ω_r are some weight factors, and

$$P_{i,(r)} = \mathbb{1}(Y_i = (r))\log F_{i,(r)}(I) \qquad (3)$$

is the pmf in descending order of values, i.e. $P_{i,(1)} \geq P_{i,(2)} \geq \ldots \geq P_{i,(C)}$. We set $K = 10$ and assumed $\omega_r = \frac{1}{K}$ for $P_{i,(1)}, \ldots, P_{i,(K)}$, and $\omega_r = 0$, otherwise. This is equivalent to averaging over the probabilities of the K most-likely clusters. The two components of the predicted optical flow in this manner are then used as the second and third channels in tensor $\mathcal{Q}$, i.e. POF$_h$ and POF$_v$.

The SM, POF$_h$ and POF$_v$ channels for training images in action class a_i, are then normalized in the unit interval and concatenated to form the tensor $\mathcal{Q}$ for each action a_i. We denote each column in $\mathcal{Q}$ as c such that $\forall c_{i,j} | i \in F, j \in [1, 2, 3]$. For instance, $c_{i,1}, c_{i,2}$ and $c_{i,3}$ represent the SM, POF$_h$ and POF$_v$ features of the I_i image, respectively.

This may be viewed as a noisy or approximate prediction of the spatial-temporal attributes of the action. The inaccuracies are of course due to the nature of the prediction process of spatial-temporal attributes in a single image, which is a lot less accurate than estimating them when video data is available. Therefore, our next goal is to regress a more compact and better representation of the spatial-temporal information in images, described in the next section.

3.2. Mapping $\mathcal{Q}$ to $Rank_{SM-POF}$ domain

Here, we propose a new spatial-temporal image representation using an ordinal regression to show the extent of contribution of a pixel for recognizing an action. We also argue that the predicted motion direction of each pixel depends on how spatially-salient that pixel is. The key to the success of this task is to extract discriminative spatial-temporal features to efficiently predict/model pixel-wise evolutions. We further borrow inspiration from [8] to learn pixels' evolution from *spatial* to *temporal*, which we show is an important cue for classifying actions. This is, essentially, a ranking process, where the parameters of the linear ranking functions encode the pixel evolution from spatial salient to predicted motion in a principled way. To learn such ranking machines, we use supervised learning to rank. We propose to use the parameters of the ranking machine as the new spatial-temporal image representation to characterize the extent of significance of spatial-temporal attributes of each pixel in recognizing an action.

Let $\mathbf{V} = [v_{t_1}, v_{t_2}, v_{t_3}]^T$ represent a column in 3-channel tensor $\mathcal{Q}$, such that $v_{t_1}{=}c_{j,1}, v_{t_2}{=}c_{j,2}$ and $v_{t_3}{=}c_{j,3}$. Rows in the vector V are the j_{th} columns in tensor $\mathcal{Q}$, representing SM, POF_h and POF_v for the j_{th} image in action class a_i.

A linear Rank-SVM is basically a pairwise linear ranking machine that learns a linear function of the form $\Psi(\mathbf{V};\mathbf{E}) = \mathbf{E}^T\mathbf{V}$. We assume that the order of the sequence in V is: $v_{t_1}{<}v_{t_2}{<}v_{t_3}$, i.e. SM is always the first channel, followed by the second channel POF_h, and finally the third channel POF_v in tensor $\mathcal{Q}$. The problem of learning the optimal linear kernel for V reduces to solving the following convex optimization problem [8]:

$$\arg\min_{\mathbf{E}} \frac{1}{2}\| \mathbf{E} \|^2 + \lambda \sum_{\forall v_{t_i}, v_{t_j}, v_{t_i} \geq v_{t_j}} \epsilon_{ij} \tag{4}$$

$$s.t. \quad \mathbf{E}^T(v_{t_i} - v_{t_j}) \geq 1 - \epsilon_{ij}, \quad \epsilon_{ij} \geq 0, \tag{5}$$

where ϵ_{ij} are slack variables and λ is a regularization parameter. By solving the above optimization problem, we learn a vector of parameters E that satisfies the order constraint. Therefore, vector E represents the pixels evolution from appearance to predicted motion. In other words, how “motion-salient” a pixel is, as well as how the direction of its predicted motion depends on its spatial saliency. We use the vector of parameters E as our new image representation containing both spatial and temporal information. By projecting images to $Rank_{SM-POF}$, pixel values in the projected domain would best represent its significance in discriminating different actions. In section 5.2, we describe how we benefit from this new domain to classify actions in still images. The new image representation I_E, in $Rank_{SM-POF}$, represents the space-time saliency of pixels as well as their direction of motion. Fig. 4 depicts a transformation flow from RGB to I_E for a single image.

Since for the similar actions within the same action phases, the SM, POF_h and POF_v features are similar, their feature vectors $\mathbf{V}$ are also similar. Therefore, it is expected that the corresponding learned vectors E for them would be similar. In Section 4, we describe how we benefit from this new domain to classify actions in still images.

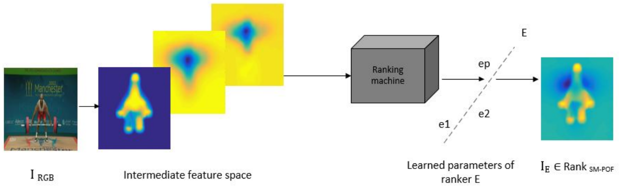

Figure 4. Transformation flow from RGB image to intermediate feature space $\mathcal{Q}$ and finally onto Rank$_{\text{SM-POF}}$ domain.

4. STCNN Architecture

We build our model on top of a CNN, which learns to selectively focus on spatial-temporal features and the pixels evolution from spatial to temporal space. Rather than having two independent spatial-CNN and motion-CNN to capture appearance and motion features separately [32], our proposed method presents a joint unified spatial-temporal CNN, named *STCNN* trained on $Rank_{SM-POF}$ to capture the spatial-temporal combined features and classify actions accordingly. Our STCNN is similar to the architecture proposed in [33]. This network is formed from sixteen successive convolutional layers followed by three fully connected layers. We denote the convolutional layers as CON(*k*,*s*), indicating that there are *k* kernels, of size $s{\times}s$. We also denote the max-pooling layer as MP. The input to our STCNN is a fixed-size 224×224 image. The convolution stride is fixed to 1 pixel. Max-pooling is performed over a 2×2 pixel window, with stride 2. Finally, FC(*n*) denotes a fully connected layer with *n* neurons. Our network architecture can be described as: 2*CON(64,3) → MP → 2*CON(128,3) → MP → 4*CON(256,3) → MP → 4*CON(512,3) → MP → 4*CON(512,3) → MP → FC(4096) → FC(4096) → FC(n).

The model in [33] was trained on Imagenet [5]. In order to accomplish transfer-learning by adapting it to the $Rank_{SM-POF}$, we changed the *FC8* layer in order to adapt it to our target domain classes. In all our experiments, we kept some of the earlier layers fixed, and fine-tuned some higher-level portions of the network, and finally trained the new classifier layer *FC8* on the target dataset from scratch. Even though it appears that we can afford to train a network from scratch when the target dataset is large enough, in practice it is quite often still beneficial to initialize the weights from a pre-trained model. In this case, we have enough data and confidence to fine-tune through the entire network. We use a smaller learning rate for layer weights that are being fine-tuned, in comparison to the weights for the new linear classifier that computes the class scores. Our goal was to learn a mapping between the $Rank_{SM-POF}$ and the action class.

We benefit from CNN's outstanding classification capability, through supervised domain adaptation by fine-tuning a network specifically pre-trained for appearance based classification on RGB images, to train the STCNN for appearance-motion (action) based classification on $Rank_{SM-POF}$ domain. We further evaluate the proposed *STCNN* in section 5.2.

5. Experiments

In this section, we demonstrate the effectiveness of our proposed domain-mapping approach along with the proposed STCNN network for single image action recognition. We ran experiments using our proposed architecture, STCNN, on *UCFSI-101*, *Willow*, *Stanford-40*, *WIDER*, *BU*$_{101}$, and finally our newly created dataset from the *wild*, details of which are described below.

5.1. Datasets

We created a large annotated ***still-image*** dataset ***UCFSI-101***[1], by extracting over *2M* frames randomly from the original UCF-101 video dataset [34]. UCF-101 has 13,320 videos from 101 action categories that spread across 5 broad groups, that is (1) Human-Object Interaction, (2) Body-Motion, (3) Human-Human interaction, (4) Playing Instruments and (5) Sports. Not all frames are useful for image-based action recognition. Therefore, after the sampling process, we eliminated frames with no human subjects clearly visible. We collected 1,585,071 frames to serve as our training set and labeled them based on the video labels they belong to, as well as 617,321 frames to help form our test set. Our train/test frames are not extracted from the same videos. In *UCFSI-101*, visual variance of extracted frames from the same video depends on action categories. Action classes with fast body motion and different key action phases, e.g. Diving, have high visual variance compared to actions that are relatively stationary.

While most related works perform their experiments on datasets with low number of classes, focused on human-object interactions, our method mainly focuses on the human body motion with 100 classes. Considering that our method is based on the patterns associated with the human motion and also the overall shape and location of the most salient parts in the human body, we naturally expect to get better results on Body-Motion categories as opposed to other categories that are highly dependent on detecting the presence of a specific object in the scene.

The ***Willow*** action dataset contains 911 images split into seven action categories: Interacting with computer, Photographing, Playing music, Riding bike, Riding horse, Running and Walking. We used the train and test splits provided by the original authors. We also used standard data augmentation, i.e. randomly mirrored images to avoid spatial biases (such as humans always centered in the image). We further divided the 7 action categories into two main groups, Body-Motion and Non-Body-Motion actions. The first three actions, pointed out in the beginning of this paragraph are considered as Non-Body-Motion actions since they are highly dependent on human-object interactions, and the rest are considered as Body-Motion actions.

The ***Stanford-40 dataset*** is a large and challenging dataset. It consists of 40 actions and 9,532 images. In each category, 100 images are used for training and the others are used for test. We divided this dataset to 2 categories: (1) Body-motion and (2) Non-body motion, with 11 and 29 actions respectively. Climbing, Jumping, Cleaning-floor, Riding-bike, Riding-horse, Rowing-boat, Running, Walking-dog, Shooting-arrow, Throwing-frisby and Waving-hands are considered as Body-motion and the rest as Non-body motion human action classes.

The ***WIDER*** attribute dataset includes 14 human attribute labels and 30 event class labels containing 13,789 images with 57,524 person bounding boxes. We then considered 6 actions under the Body Motion category; i.e. running, basketball, football, soccer, skiing and hockey.

The ***BU***$_{101}$ consists of $23.8K$ action images that correspond to the 101 action classes in the UCF101 video dataset. Action categories are divided into five types: Human-Object Interaction, Body-Motion, Human-Human Interaction, Playing Musical Instruments, Sports. For each action class, images are downloaded from the Web using corresponding key phrases, and then manually remove irrelevant images or drawings and cartoons. We used the train and test splits provided by the original authors for all datasets.

Finally, we formed a dataset that plays a key role in our experiments using ***Google and iStockPhoto*** images [2], representing 10 action categories including Diving, Golf-swing, Kicking, Lifting, Riding-horse, Running, Skateboarding, Swing-bench, Swing-side and Walking. This dataset contains 1000 images, with a 700 and 300 train/test split, representing 10 action categories, with the human body located in the center of the image. All these images are further used for testing out the models trained in STCNN. The reason why this dataset was formed is to support the claim that STCNN is capable of performing well even when given images, belonging to the mentioned categories, are downloaded from the Web.

5.2. Evaluation

After mapping the collected still images from RGB onto I_E, i.e. $I_E \in Rank_{SM-POF}$, we convert the 1-D vectors I_E to 3-D RGB representations using the integer value of each pixel as an index to a colormap. We then trained our deep STCNN network on training sets associated with the target domain. The experiments were run on a single GeForce GTX Titan GPU with 15GB of memory. To evaluate the performance of our domain-mapping, the approach mentioned in section 4 was applied on UCFSI-101 dataset.

As shown in table 1 and 2, our deep STCNN outperforms the non-domain-mapped baseline, where CNN is trained over the RGB images, SM and POF of images in *UCFSI-101*, separately. Experiments on using RGB images rather than saliency map demonstrate that focusing on the most salient pixels rather than all the pixels in an image improves the results by reducing noise, introduced otherwise by incorporating all pixels. The intuition behind using saliency map helps detect the most salient regions in the image, whereas the predicted optical flow helps identify which pixels in the salient regions are predicted to have motion and in which direction.

[1] UCFSI-101, still images dataset, will be made publicly available.

[2] Google and iStockPhoto images will be made publicly available.

Table 1. Results on UCFSI-101 RGB images, Saliency Map (SM) and the Predicted Optical Flow(POF) and also results on intermediate feature space tensor Q (before Rank-SVM).

Model	Top-1 Accuracy	Top-5 Accuracy
Train from scratch, no domain mapping(RGB)	14.8%	22.1%
Train from scratch, no domain mapping(SM)	18.9%	27.8%
Train from scratch, no domain mapping(POF)	14.2%	23.7%
Train from scratch, feature space Q	21.2%	35.0%
Fine-tune all layers, feature space Q	41.8%	55.3%
Fine-tune top 5 layers, feature space Q	**63.7%**	**70.8%**

We also tested our method on both tensor $\mathcal{Q}$, which is an intermediate image representation before applying Rank-SVM as well as I_Es, which represents the new image representation after applying Rank-SVM, explained in 3.1 and 3.2, respectively. Results in tables 1-3 show that using the parameters of the ranking machine as a *spatial-temporal* compact representation of the image improves the action classification results. Moreover, applying Rank-SVM on spatial-temporal tensors indicates that actions have a characteristic evolution from spatial to predicted temporal patterns, improving single image action classification by a large margin. Our transfer learning method also yields more promising results compared to the case where a CNN is trained from scratch on $Rank_{SM-POF}$ domain. Since we expect that CNNs learn more generic features on the bottom layers of the network, and more convoluted dataset-specific features near the top layers of the network, we considered two different approaches for our transfer learning experiments: *Fine-tuning all layers.* Under this approach, we retrained all network parameters, including all convolutional layers on the bottom of the network; *Fine-tuning the last five layers.* Rather than only re-training the final classifier layer from scratch, we performed fine-tuning on the last five layers. Moreover, as shown in table 2, we also compared the proposed STCNN with the commonly used late fusion strategy [32]. Here, we implemented two spatial (SM) and tempral (POF) streams using deep CNNs separately. Next, we combined the softmax scores of each stream via late fusion in order to fully validate the merits of STCNN, our joint *spatial-temporal* network.

We further broke down our performance by 5 broad groups of classes present in the UCFSI-101 dataset. We computed the average precision of every class and then computed the mean average precision over classes in each group. In all experiments, as shown in tables 1-3, fine-tuning the last 5 layers of the network increased the performance on all action categories. Also among all groups, as we expected, we obtained the best results on the *Body-Motion* group, for which actions heavily depend on predicted body motion and also the most salient part of the body rather than human-object interaction.

We then removed the last fully-connected layer from our network, which was pre-trained on the generated UCFSI-101 and trained a linear softmax classifier for other datasets.

Table 2. Results on UCFSI-101 with domain mapping (after Rank-SVM).

Model	Top-1 Accuracy	Top-5 Accuracy
Two-Stream (Late fusion)	19.7%	24.3%
STCNN- Train from scratch	30.3%	38.9%
STCNN- Fine-tune all layers	46.6%	58.7%
STCNN- Fine-tune top 5 layers	**67.9%**	**79.1%**

Table 3. mAP (%) results on UCFSI-101 by category groups.

	Without domain mapping	**Intermediate feature space tensor Q**			**After domain mapping** $Rank_{SM-POF}$		
STCNN	From Scratch	From Scratch	Fine-Tune all layers	Fine-Tune 5 layers	From Scratch	Fine-Tune all layers	Fine-Tune 5 layers
Human-Object	20.1	17.0	35.2	56.3	23.1	36.0	57.0
Body-Motion	11.2	42.1	**58.2**	**84.0**	49.2	**60.0**	**89.1**
Human-Human	10.3	35.1	42.2	58.0	0.42	44.0	60.3
Playing-Instrument	20.0	18.2	30.1	60.0	21.3	35.0	65.3
Sport	17.3	23.1	42.1	65.4	30.0	50.0	69.0
All group	16.0	24.4	42.1	**65.3**	32.0	48.4	**70.2**

Table 4. mAP (%) results on Willow dataset.

Method	**Non-Body-Motion**	**Body-Motion**
Delaitre et al. [3]	55.6	62.7
Delaitre et al. [4]	55.4	70.7
Sharma et al. [29]	59.0	71.1
Sharma et al. [30]	60.1	73.2
Khan et al. [15]	62.4	72.2
Khan et al. [14]	64.1	78.05
Liang et al. [19]	89.0	74.0
Zhao et al. [42]	67.8	79.3
Khan et al. [13]	62.2	76.0
STCNN- on intermediate feature space tensor Q	62.9	69.1
STCNN- on $Rank_{SM-POF}$ domain	64.5	**78.7**

Table 5. mAP (%) results on Stanford-40.

Method	**Body-Motion**	**Non-Body-Motion**	**All**
Gkioxari et al. [9]	93.87	89.73	90.9
Khan et al. [14]	56.92	51.51	53
Khan et al. [13]	53.51	51.28	51.9
Zhao et al. [43]	-	-	83.4
Zhao et al.[42]	-	-	54.5
Zhao et al.[41]	-	-	80.6
Zhou et al. [45]	-	-	55.3
Sharma et al. [31]	-	-	72.3
Khan et al. [16]	-	-	75.4
STCNN	**94.3**	73.1	81.76

As shown in tables 4-9, STCNN achieved promising results compared to state-of-the art approaches on all datasets. Considering that we focus on the human body salient pixels and their motion direction, we achieved the best performance on the Body-Motion group on both Stanford-40 and Willow datasets. While [14] constructs the spatial pyramids on the full-body, upper-body and face bounding boxes, temporal information is completely ignored in their method. As shown in table 5, [14] performs poor on Stanford-40 compared to Willow dataset. The large number of action categories in Stanford-40 makes this dataset particularly challenging. There are many common human body poses among different action classes which makes action recognition harder when temporal cues are ignored. As shown in tables 4 and 5, STCNN maintains its performance on both datasets.

In tables 6 and 7, we evaluate STCNN on the Google/iStockPhoto dataset collected from the wild to

Table 6. mAP (%) on Google/iStockPhoto dataset using STCNN.

Model	Dive	Golf	Kick	Lift	Walk	Skate	Swing-side	Swing-bench	Horse-ride	Run	All
Trained from scratch	67.0	65.2	48.0	63.3	49.4	43.1	48.0	39.3	57.4	52.2	53.0
Fine-tuned all layers	75.1	72.0	55.3	74.0	57.2	55.4	59.1	47.0	62.2	61.3	62.0
Fine-tuned 5 top layers	88.3	85 .0	76.3	87.2	72.0	74.0	73.3	69.1	86.0	84.2	79.1
Model learned from UCFSI-101	82.2	86.1	67.3	85.1	66.0	65.2	73.3	64.0	81.2	82.4	75.0

Table 7. Comparison of action classification on Google/iStockPhoto dataset.

Method	mAP(%)
Object Bank [18]	28.1
LLC [36]	33.5
R*CNN [9]	73.3
STCNN	**75.0**

Table 8. mAP (%) results on WIDER.

Method	mAP(%)	
RCNN	80.0	
R*CNN	80.5	
DHC	81.3	
ResNet-SRN	86.2	
VeSPA	82.4	
STCNN	Body-Motion	Non-Body-Motion
	86.8	59.5

Table 9. STCNN results on BU_{101} dataset.

mAP (%)				
Human-Object	Body-Motion	Human-Human	Playing-Instrument	Sport
61.1	84.4	58.7	71.3	74.8

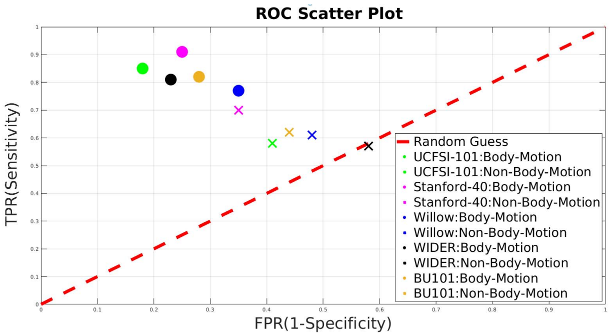

Figure 5. Data points illustrate the classifying ability of STCNN for body-motion and non-body-motion action categories over different datasets. As depicted, STCNN yields points in the upper left corner, representing higher sensitivity (less false negatives) and higher specificity (less false positives), for actions having considerable human body motion.

prove the beneficial effect of our approach on an unseen set of still images. We compared STCNN against three baselines in table 7. Results show that knowing the most salient parts in the human body and also information on how and in which direction the salient parts are likely to move in the future, can extremely help to solve action recognition problem in still images.

As shown in table 8, experiments on WIDER dataset also yield a promising results, marginally outperforming state-of-the-art. In [21], BU_{101} is used to study the utility of filtered web action images for video-based action recognition using CNNs. However, we ran experiments on BU_{101} to perform action recognition in still images by leveraging the pre-trained model on UCFSI-101. Table 9 reports STCNN results on BU_{101} per class category.

Presence of particular objects [4, 3] as well as their spatial and scale relations with humans can characterize the action in the images, however, many human action images with no specific object involved, pose a serious challenge for current human-object interaction methods. The proposed methods in [29, 19, 30] have mainly relied only on body parts and poses. The extensive manual labeling of human bounding boxes and having enough number of body pose and parts detectors, and the fact that different actions might have very similar poses can be very challenging. As shown in all mentioned experiments focusing on pixels' evolution within a generated spatia-temproal image representation improves action classification in still images by a significant margin, especially on Body-motion group.

6. Conclusion

We propose a novel domain mapping technique to capture pixels evolution from appearance to predicted motion in order to obtain a unique compact spatial-temporal representation for human action in a single image. We show that actions have a characteristic evolution from spatial to motion over time. We then employ a transfer learning approach to propose STCNN, a spatial-temporal CNN, for the task of single image action classification, by fine-tuning a CNN which is pre-trained specifically for appearance-based classification. We benefit from the proposed domain mapping approach for single image action recognition - a problem that suffers from lack of temporal information. We carried out a comprehensive testing and evaluation of all components of our proposed architecture. Results on multiple benchmarks demonstrate that appearance and predicted motion are complementary sources of information, hence using both leads to significant performance improvement in single image action recognition.

References

[1] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1543–1550. IEEE, 2011.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[3] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC 2010-21st British Machine Vision Conference*, 2010.

[4] V. Delaitre, J. Sivic, and I. Laptev. Learning person-object interactions for action recognition in still images. In *Advances in neural information processing systems*, pages 1503–1511, 2011.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[7] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[8] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars. Modeling video evolution for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5378–5387, 2015.

[9] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1080–1088, 2015.

[10] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. R-cnns for pose estimation and action detection. *arXiv preprint arXiv:1406.5212*, 2014.

[11] G. Guo and A. Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10):3343–3361, 2014.

[12] M. Hoai12. Regularized max pooling for image categorization. 2014.

[13] F. S. Khan, R. M. Anwer, J. Van De Weijer, A. D. Bagdanov, A. M. Lopez, and M. Felsberg. Coloring action recognition in still images. *International journal of computer vision*, 105(3):205–221, 2013.

[14] F. S. Khan, J. van de Weijer, R. M. Anwer, M. Felsberg, and C. Gatta. Semantic pyramids for gender and action recognition. *IEEE transactions on image processing*, 23(8):3633–3645, 2014.

[15] F. S. Khan, J. Van De Weijer, A. D. Bagdanov, and M. Felsberg. Scale coding bag-of-words for action recognition. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 1514–1519. IEEE, 2014.

[16] F. S. Khan, J. Xu, J. Van De Weijer, A. D. Bagdanov, R. M. Anwer, and A. M. Lopez. Recognizing actions through action-specific person detection. *IEEE transactions on image processing*, 24(11):4422–4432, 2015.

[17] I. Laptev. On space-time interest points. *International journal of computer vision*, 64(2-3):107–123, 2005.

[18] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems*, pages 1378–1386, 2010.

[19] Z. Liang, X. Wang, R. Huang, and L. Lin. An expressive deep model for human action parsing from a single image. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2014.

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[21] S. Ma, S. A. Bargal, J. Zhang, L. Sigal, and S. Sclaroff. Do less and achieve more: Training cnns for action recognition utilizing action images from the web. *Pattern Recognition*, 68:334–345, 2017.

[22] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. 2011.

[23] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.

[24] N. Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[25] R. Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.

[26] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):601–614, 2012.

[27] T. Qi, Y. Xu, Y. Quan, Y. Wang, and H. Ling. Image-based action recognition using hint-enhanced deep neural networks. *Neurocomputing*, 267:475–488, 2017.

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

[29] G. Sharma, F. Jurie, and C. Schmid. Discriminative spatial saliency for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3506–3513. IEEE, 2012.

[30] G. Sharma, F. Jurie, and C. Schmid. Expanded parts model for human attribute and action recognition in still images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–659, 2013.

[31] G. Sharma, F. Jurie, and C. Schmid. Expanded parts model for semantic description of humans in still images. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):87–101, 2017.

[32] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances*

in neural information processing systems, pages 568–576, 2014.

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.

[34] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[35] J. Walker, A. Gupta, and M. Hebert. Dense optical flow prediction from a static image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2443–2451, 2015.

[36] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360–3367. IEEE, 2010.

[37] Y. Xiong, K. Zhu, D. Lin, and X. Tang. Recognize complex events from static images by fusing deep channels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1600–1609, 2015.

[38] B. Yao and L. Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1691–1703, 2012.

[39] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1331–1338. IEEE, 2011.

[40] Y. Zhang, L. Cheng, J. Wu, J. Cai, M. N. Do, and J. Lu. Action recognition in still images with minimum annotation efforts. *IEEE Transactions on Image Processing*, 25(11):5479–5490, 2016.

[41] Z. Zhao, H. Ma, and X. Chen. Semantic parts based top-down pyramid for action recognition. *Pattern Recognition Letters*, 84:134–141, 2016.

[42] Z. Zhao, H. Ma, and X. Chen. Generalized symmetric pair model for action classification in still images. *Pattern Recognition*, 64:347–360, 2017.

[43] Z. Zhao, H. Ma, and S. You. Single image action recognition using semantic body part actions. In *2017 IEEE International Conference on Computer Vision (ICCV), Venice*, pages 3411–3419, 2017.

[44] Y. Zheng, Y.-J. Zhang, X. Li, and B.-D. Liu. Action recognition in still images using a combination of human pose and context information. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 785–788. IEEE, 2012.

[45] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.