UCF-STAR: A Large Scale Still Image Dataset for Understanding Human Actions

Marjaneh Safaei, Pooyan Balouchian, Hassan Foroosh
Department of Computer Science
University of Central Florida (UCF)
Orlando, FL 32816-2362
{marjaneh.safaei, pooyan}@knights.ucf.edu, Hassan.Foroosh@ucf.edu

Abstract

Action recognition in still images poses a great challenge due to (i) fewer available training data, (ii) absence of temporal information. To address the first challenge, we introduce a dataset for STill image Action Recognition (*STAR*), containing over $1M$ images across 50 different human body-motion action categories. *UCF-STAR* is the largest dataset in the literature for action recognition in still images. The key characteristics of *UCF-STAR* include (1) focusing on human body-motion rather than relatively static human-object interaction categories, (2) collecting images from the *wild* to benefit from a varied set of action representations, (3) appending multiple human-annotated labels per image rather than just the action label, and (4) inclusion of rich, structured and multi-modal set of metadata for each image. This departs from existing datasets, which typically provide single annotation in a smaller number of images and categories, with no metadata. *UCF-STAR* exposes the intrinsic difficulty of action recognition through its realistic scene and action complexity. To benchmark and demonstrate the benefits of *UCF-STAR* as a large-scale dataset, and to show the role of "latent" motion information in recognizing human actions in still images, we present a novel approach relying on predicting temporal information, yielding higher accuracy on 5 widely-used datasets.

Introduction

We introduce a new large-scale multi-modal dataset, *UCF-STAR*, to advance the current still image-based action recognition research, and to promote future research opportunities through its additional rich and structured metadata. *UCF-STAR* contains 1,038,622 annotated still images, collected from the *wild*, more than 40 times the size of the largest previous action image dataset; i.e. BU-101 (Ma et al. 2017). Images are annotated with multiple labels, including the action, and are accompanied with rich textual metadata. Excerpts from *UCF-STAR* dataset are shown in Fig.1.

Broadly, human actions fall into two categories: (1) human body-motion actions, and (2) static actions. Actions can be either person-centric, or group activity. *UCF-STAR* is focused on still images of person-centric body-motion actions. Hence, the choice of keywords to crawl the *Web* to collect images was driven accordingly. For instance, *swinging tennis racket* took precedence over simply *tennis*, as the term *swinging racket* suggests body-motion compared to *tennis*. We collected 50 different action categories, using *Bing's Cognitive Services API*, allowing 250 transactions per second, where a transaction refers to a successful Bing API call request. *UCF-STAR* includes a rich set of metadata including human visibility, number of humans, human-object interaction, caption, tags, bounding boxes, action labels, and other metadata explained in section *UCF-STAR Construction*.

Human body-motion categories. While in existing datasets many action classes, e.g. *photography*, are relatively static and dependent on human-object interaction, our emphasis is on human body-motion actions, i.e. actions dependent on body motion. Thus, *UCF-STAR* is constructed by collecting images of actions with body motion, and is benchmarked in section *Action Recognition Method* by proposing a new method, tackling missing motion in still images.

Person-centric action categories. Our focus is on the actions performed by people, treated as individual agents. There can be multiple people in a scene, however each one serves as an individual agent. Here, the annotations are constructed to refer to the main agent.

The realistic complexity of *UCF-STAR* exposes the inherent difficulty of human body-motion action recognition, overlooked by many well-known datasets. We perform comparative benchmarking of well-known methods on Stanford-40 (Yao et al. 2011), Willow (Delaitre, Laptev, and Sivic 2010), WIDER (Xiong et al. 2015), BU-101 (Ma et al. 2017) and *UCF-STAR*. Results confirm the more challenging nature of *UCF-STAR* compared to other datasets.

While in videos one can readily infer motion (Sun, Junejo, and Foroosh 2011; Ashraf, Sun, and Foroosh 2014; Sun, Tappen, and Foroosh 2014; Sun et al. 2015), such information is missing in a single image. Thus, action recognition poses a bigger challenge in still images, due to the absence of temporal information (Zhao, Ma, and Chen 2017; Safaei and Foroosh 2018; Zhang et al. 2016; Oquab et al. 2014; Gkioxari, Girshick, and Malik 2015; Khan et al. 2013). The issue is exacerbated when there is no contextual information, e.g. interaction with a recognizable object. To address

Figure 1: Excerpts from *UCF-STAR*: (a) Examples depicting body-motion actions; (b) Examples of associated metadata and labels, e.g. bounding boxes, action class, captions, tags, number of humans, human visibility and human-object interaction.

this gap and hence boost the accuracy, we propose a new method of modeling the "latent" temporal information in a still image, and use it as prior knowledge in a two-stream deep network (Chéron, Laptev, and Schmid 2015; Donahue et al. 2015; Feichtenhofer, Pinz, and Zisserman 2016; Girdhar et al. 2017; Gkioxari, Girshick, and Malik 2015).

Related Work

Still image based action recognition datasets. Most popular action classification datasets, such as KTH (Schuldt et al. 2012), Weizmann (Blank et al. 2005), Hollywood-2 (Marszałek, Laptev, and Schmid 2009), HMDB (Jhuang et al. 2011), UCF101 (Soomro et al. 2012) consist of short clips, manually trimmed to capture a single action in a video clip. They serve a valuable purpose, but address a different need than what *UCF-STAR* has to offer.

Action images in sports (Gupta, Kembhavi, and Davis 2009; Li and Li 2007) are among the earliest datasets introduced for research. Daily activity datasets (Yao and Fei-Fei 2012; Le, Bernardi, and Uijlings 2013) contain common human activities in daily life. The latest version of Pascal VOC (Maji, Bourdev, and Malik 2011) competition includes ten categories of still image actions, with only a subset of people annotated (bounding box + action). People in the dataset are labeled with exactly one action class. There is a minimum of around 400 people per action category. In contrast, *UCF-STAR* is focused on human body-motion actions rather than relatively static actions such as reading, using computer, etc.

Datasets by (Delaitre, Laptev, and Sivic 2010; Ikizler et al. 2008; Ikizler-Cinbis, Cinbis, and Sclaroff 2009; Li, Ma, and Gao 2011; Prest, Schmid, and Ferrari 2012; Yao and Fei-Fei 2010) contain 968, 467, 2,458, 2,400, 341 and 2,100 images, respectively. Images were collected from different sources like Google Image search, Flickr and PASCAL VOC 2010 to build 3 to 7 action categories. The main differences with *UCF-STAR* dataset are the small number of action classes and the small number of overall images. Furthermore, classes contain actions with less human body motion i.e. playing/holding instruments and wearing hat, which are not the primary focus in *UCF-STAR*.

Thurau and Hlavac (Thurau and Hlavác 2008) and Raja et al. (Raja et al. 2011) extracted frames from popular action videos to build 10 and 6 action classes, respectively. The images are usually depicting relatively static actions with clean background. Yao et al. (Yao et al. 2011) introduced Stanford 40, containing 40 daily human actions in 9,352 images, obtained from Google, Bing, and Flickr. Le et al. (Le, Bernardi, and Uijlings 2013) assembled a dataset from the PASCAL 2012 VOC trainval set by selecting a subset of 2,038 images with human actions, over 89 action classes.

Given the demanding nature of deep learning methods for training data, there is a need for a larger dataset with larger number of images in each class. This motivated the construction of *UCF-STAR* with a large number of action classes; i.e. 50, and a large number of images per class; i.e. an average of 20,366. Finally, *UCF-STAR* provides not only multiple labels for each image, but also a rich set of metadata further explained in section *UCF-STAR Construction*. Table 1 compares *UCF-STAR* against some well-known image datasets for action recognition providing statistics on each dataset.

Methods for action recognition in still images. Body parts and pose-based approaches are challenging due to the limited number of poses they can detect and the fact that many different human actions share almost the same poses. Moreover, the work in (Prest, Schmid, and Ferrari 2012; Yao et al. 2011; Gkioxari, Girshick, and Malik 2015) rely on the presence and detection of objects as additional contextual information, posing a challenge when the action involves only a human with no object interaction.

Still image action recognition has recently benefited from the outstanding performance of CNN models (Gao, Xiong, and Grauman 2018; Gkioxari, Girshick, and Malik 2015; Safaei and Foroosh 2019; Hoai 2014; Oquab et al. 2014; Rahman and Wang 2016). The tradeoff is the need for millions of parameters and dependency on huge training sets. *UCF-STAR* will thus play an invaluable role in future research. To benchmark *UCF-STAR*, we explore the idea of predicting the "latent" human body motion, outperforming the state of the art (see section *Action Recognition Method*).

UCF-STAR Construction

Construction of *UCF-STAR* was a five step process involving (1) action category selection, (2) semantic grounding, (3) collecting images from the *wild*, (4) image annotation, and (5) enhancing dataset size. Below, we provide the details.

Action Category Selection

We followed two principles in selecting action categories. First, only actions involving significant human body-motion

Table 1: Comparison of *UCF-STAR* dataset with other still image action recognition datasets.

Dataset	#classes	#images	Labels			Caption	Source	Tag	Bounding box
			number of humans?	human-object interaction?	human visibility?				
Stanford-40 (Yao et al.)	40	9,532	No	No	No	No	No	No	Yes
Willow (Delaitre et al.)	7	911	No	No	No	No	No	No	Yes
Pascal VOC 2010 (Maji et al.)	9	50 to 100 per class	No	No	No	No	No	No	Yes
Pascal VOC 2011 (Maji et al.)	10	200 or more per class	No	No	No	No	No	No	Yes
Pascal VOC 2012 (Maji et al.)	10	200 or more per class	No	No	No	No	No	No	Yes
PPMI (Yao and Fei-Fei)	7	2,100	No	No	No	No	No	No	No
89 Action Dataset (Le et al.)	89	2,038	No	No	No	No	No	No	No
BU-101 (Ma et al.)	100	23,782	No	No	No	No	No	No	No
Action Images by Ikizler (Ikizler et al.)	6	467	No	No	No	**Yes**	No	No	No
Sport Dataset (Gupta et al.)	6	300	No	No	No	**Yes**	No	No	No
UCF-STAR	**50**	**1,038,622**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

were selected. Second, key poses providing clear visual signatures were considered for each action; i.e. *tennis swing* or *tennis serve* taking precedence over *playing tennis*. Needless to emphasize on the generic nature of the term *playing tennis* compared to *tennis swing/serve*.

Semantic Grounding

To search for images, we applied semantic grounding to find synonymous terms for actions, leading to more accurate data retrieval. This was done by querying *WordNet* synsets and collecting synonymous terms. This step was performed to not only expand the search space, but also help retrieve more relevant images, minimizing search misses of conventional keyword-only searches. To retrieve human-centric images, we appended keywords like "human", "person", "woman" or "man" to form *n-grams*; i.e. human + <action>, person + <action>, man + <action> and woman + <action>. This leads to a significant reduction in noise that would otherwise include images with no visible human body.

Image Collection

To crawl the *Web* for images, we took advantage of *Bing's Cognitive Services API* as proposed in (Balouchian, Safaei, and Foroosh 2019), supporting 250 transactions per second, where a transaction is defined as a successful Bing API call request. This API provides support for an array of filters including *face-only*, *include body parts*, etc., as part of its *Image Search API*. We flagged each *n-gram* with the relevant filters, such as *face-only* and *include body parts*. These flags help the final search results require less manual effort, and reduce noise. Using this approach, we collected 29,037 images, which we refer to as the *strongly labeled* dataset. Even though a dataset of 29,037 images would be considered as the largest action image dataset in the literature, we further enhanced the dataset size using the approach explained next.

The images in each class have large variations in background, appearance and pose. To further enhance *UCF-STAR*, we also collected a rich set of *metadata* for each image. Bing Image Search API returns *insightsToken* that can be used to submit a second query for collecting a rich set of metadata on each image, including: 1) *BRQ* which is the best representative query that is defined as a term that best describes the image, 2) *Caption*, which provides textual information that may contain entities and links to other related entities, 3) *Collections* providing a list of related images, 4) *PagesIncluding* providing a list of webpages that include the image, 5) *RecognizedEntities* representing a list of entities (people) that were recognized in the image, 6) *RelatedSearches* offering a list of related searches made by others, 7) *SimilarImages* providing a list of images that are visually similar, and 8) *Tags* providing characteristics of the type of content found in the image. For example, if the image is of a person, the tags may indicate gender or type of clothes.

Image Annotation Process

We used Amazon Mechanical Turk (AMT) for annotating images, with questions designed to capture (1) the observed human action, (2) number of humans in the image, if any, (3) whether or not there exists at least one whole human body in the image, and (4) whether or not a human-object interaction is present. To reduce noise, each image was annotated by three independent AMT workers, and a label was considered as ground truth if confirmed by majority. Our AMT workers flagged 57.7% of the images as correct; i.e. 57.7% of images matched the weak labels they initiated from, resulting in 16,756 noise-reduced strongly-labeled images.

Enhancing Dataset Size

Our dataset, at this step of the process, included only the noise-reduced images labeled by AMT workers. To enhance the size of our dataset, we took advantage of Bing's feature available in its *Image Search API* that enables queries for visually similar images. Taking advantage of this feature, our system re-crawled the *Web* and collected 1,315,714 images. Next we removed the duplicates using *fdupes*, resulting in 1,038,622 unique labeled images. We further split the 1,038,622 images into mutually exclusive 664,718 training, 166,180 validation, and 207,724 test images.

Dataset Statistics

A key characteristic of *UCF-STAR* is its human body-motion based action classes. The rich set of metadata offered by *UCF-STAR* would also enable the community to benefit from its multi-modal nature, benchmarking methods relying on both image and text modalities. Figure 2 shows *UCF-STAR*'s distribution of action classes as well as the annotations thereof. Even though the number of images per class is different (averaging at 20,366), the large-scale nature of the dataset, however, would enable us to easily sub-sample the dataset to avoid the *class imbalance* problem.

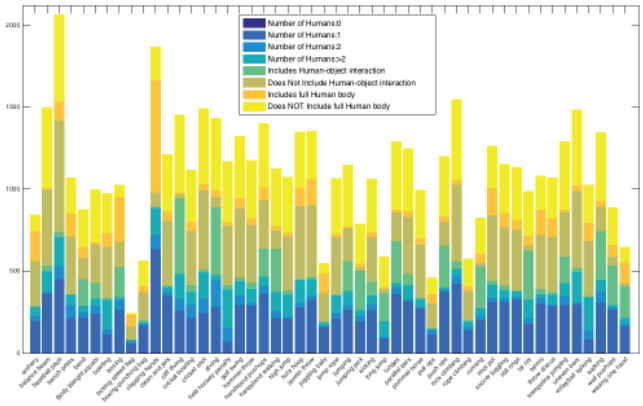

Figure 2: Distribution of UCF-STAR's annotations per class.

Our Action Recognition Method

Popular datasets such as Standford-40, PASCAL VOC and Willow have been widely used by recent methods in still image action recognition. However, the small number of action classes, limited number of images in each class, and the distinctive nature of action categories may present an exaggerated picture of the state of the art. Difficulties arise when the number of classes are large, human-object interaction is not a determining factor in recognition, actions are only subtly different in poses, and background scenes are not informative. *UCF-STAR* has all these aspects aplenty. To prove this, we developed a new action recognition method inspired by recent motion prediction approaches, and compared the results with recent state-of-the-art still image action recognition methods on both *UCF-STAR* and other existing popular datasets mentioned above.

Unlike previous efforts, our method learns and takes advantage of the "latent" temporal information in still images, rather than relying solely on spatial information. The key idea is to transfer the temporal information learned from video frames into still images to aid action recognition. To achieve this, we developed a two-stream spatiotemporal network (TSSTN), similar to networks used in the video literature, and decomposed still image action recognition into spatial and predicted temporal streams as described below.

Temporal stream network. Our goal is to derive a new image representation, named *dynamic-skeleton map*, by learning motion from video frames and then transferring it to still images. Therefore, a dynamic-skeleton map serves to model the missing temporal information. Dynamic-skeleton represents motions of human body pixels in a predefined time window, hallucinating the human body motion.

The concept of dynamic image; i.e. modeling video evolution for action recognition, is inspired by (Fernando et al. 2015; Bilen et al. 2016). While they propose methods to capture video-wide temporal information for action recognition, we generate a dynamic-skeleton map for every frame in a video to serve as the temporal label for that frame. We then use these labeled frames to learn a model for predicting the dynamic-skeleton maps of still images.

Generating dynamic-skeletons for video frames is essentially done as a ranking process (Fernando et al. 2017), where the parameters of the linear ranking function are used to encode pixel evolution. To learn such a ranking machine, we use the supervised learning proposed in (Yu and Kim 2012). In ranking algorithms, a training set represents an ordering of data. Let $V = [v_{t_1}, v_{t_2}, ..., v_{t_n}]$ represent a sequence of frames, where the frame order also dictates the evolution of the frame appearances. We focus on the relative orderings of the frames, i.e. $v_{t+1} > v_t$ if v_{t+1} succeeds v_t.

A linear Rank-SVM represents a pairwise linear ranking machine that learns a linear mapping of the form $\Psi(V; M) = M^T V$ (Yu and Kim 2012; Fernando et al. 2017). We envision the order of the sequence in the training set V as $v_{t_n} < ... < v_{t_2} < v_{t_1}$. The ranking score of v_t is derived by $\Psi(v_t; m) = m^T v_t$ and satisfies the pairwise constraints ($v_{t+1} < v_t$), while avoiding over-fitting. Consequently, we aim to learn a parametric vector $m \in M$ such that it satisfies all constraints.

$$\Psi(v_{t_i}; m) = m^T v_{t_i} > \Psi(v_{t_j}; m) = m^T v_{t_j} \quad (1)$$
$$\forall v_{t_i}, v_{t_j}, v_{t_i} > v_{t_j}$$

The problem of learning the optimal linear kernel for V reduces to solving the following optimization problem (Yu and Kim 2012):

$$\arg\min_M \frac{1}{2} \| M \|^2 + W \sum_{\forall v_{t_i}, v_{t_j}, v_{t_i} \geq v_{t_j}} \epsilon_{ij} \quad (2)$$
$$s.t. \quad M^T (v_{t_i} - v_{t_j}) \geq 1 - \epsilon_{ij}, \quad \epsilon_{ij} \geq 0,$$

where ϵ_{ij} are slack variables and W represents a regularization parameter. Solving this optimization problem leads to learning a vector of parameters M. As the parameters of M define the order of frames in V, they encode the evolution of pixels. Therefore, we used vector M learned on sequence of skeletons as the dynamic-skeleton map, d_s for short. Therefore, ds_i represents the dynamic skeleton associated with the i_{th} frame in training sequence of $V = [v_{t_1}, v_{t_2}, ..., v_{t_i}]$. ds_i represents a compact temporal representation of all skeleton frames from t_1 to t_i.

To make use of the learned dynamic-skeleton models in still image action recognition, we need to transfer the learned model from videos to still images, i.e. predict the d_s for a still image. As described next, we show that this can be done through a process similar to semantic segmentation (Badrinarayanan, Kendall, and Cipolla 2015), using a deep convolutional encoder-decoder architecture.

Every prior dynamic-skeleton, generated from video frames is first quantized into C clusters by k-means. The problem is then treated in a manner similar to semantic segmentation, where each pixel in the image is classified as a particular cluster of the dynamic-skeleton map. The output is generated as softmax probabilities over the clusters for each pixel. The loss is summed over all pixels in a mini-batch. This implicitly assumes a uniform probability mass function (pmf) for the segmentation classes, which is prone to noise. We, therefore, developed a custom loss function 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 pretrained network using

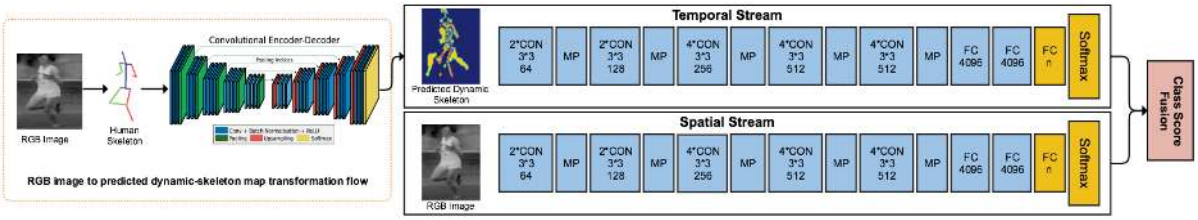

Figure 3: Two-stream still image action recognition network, using predicted dynamic-skeleton map as input to temporal stream.

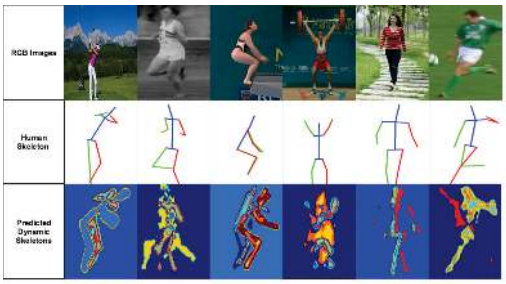

Figure 4: Mapping examples from RGB to skeleton, and then to predicted dynamic-skeleton domain.

the custom loss function. Let I represent the image and Y be the ground truth; i.e. d_s labels, represented as quantized clusters. Then the loss function $L(I, Y)$ is:

$$\hat{L}(I, Y) = -\sum_{i=1}^{M \times N} \sum_{r=1}^{C} \omega_r P_{i,(r)}, \tag{3}$$

where ω_r are some weight factors, and

$$P_{i,(r)} = \mathbb{1}(Y_i = (r)) \log F_{i,(r)}(I) \tag{4}$$

is the pmf in descending order of values; i.e. $P_{i,(1)} \geq P_{i,(2)} \geq ... \geq P_{i,(C)}$. We set $k = 3$ and assumed $\omega_r = \frac{1}{K}$ for $P_{i,(1)}, \ldots, P_{i,(K)}$, and $\omega_r = 0$ otherwise. The $F_{\mathrm{i,r}}(I)$ represents the probability that the i_{th} pixel belong to cluster r, and $\mathbb{1}(Y_{\mathrm{i}} = r)$ is an indicator function.

A predicted d_s represents a compact temporal representation of a a still image, as a hallucinated sequence of human skeleton images. These temporal patterns can then be used as auxiliary information for action recognition in still images. Fig. 4 depicts examples of predicted dynamic-skeleton maps for some still images. Predicted d_ss are used as the input to the temporal stream in our two-stream action recognition network depicted in Fig. 3.

Spatial stream network. The spatial stream network operates on individual RGB images, performing action recognition from still images. The static appearance by itself is a useful cue, and improves the final classification score.Since the spatial stream is essentially an image classification network, we can build upon the recent advances in large-scale image recognition methods.

We build TSSTN using a CNN architecture similar to (Simonyan and Zisserman 2014). Our temporal and spatial streams are trained to selectively focus on the corresponding features, respectively. Each stream is formed by 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 $s \times s$. The input to our CNN 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. We change the last *FC* layer, used smaller learning rates for layers that are being fine-tuned to further promote our goal to inject the predicted motion prior into still images.

Final classification score. Stacked generalization is a method of using a high-level model to combine lower-level models to achieve greater accuracy. Stacking with Multi-response regression (MRR) uses linear regression to perform classification (Ting and Witten 1999). If the original classification problem has I classes, it is converted into I separate regression problems.

Given a data set $\mathcal{D} = \{(y_n, x_n), n = 1, ..., N\}$, where y_n is the class value and x_n is a vector representing the attribute values of the n-th instances, randomly split the data into J almost equal parts, the linear regression for class c is simply:

$$\mathcal{R}_c(x) = \sum_{k}^{K} \alpha_{kc} \mathcal{P}_{kc}(x) \tag{5}$$

let $\mathcal{P}_{kc}(x)$ denote the probability of the c_{th} output class obtained by the k_{th} model for an instance x. Next, we choose the linear regression coefficients $\{\alpha_{kc}\}$ to minimize

$$\sum_{d} \sum_{(y_n, x_n) \in \mathcal{D}_j} (y_n - \sum_{k} \alpha_{kc} \mathcal{P}_{kc}^{-j}(x_n))^2 \tag{6}$$

A least-square algorithm under non-negativity constraint is then employed to derive the linear regression for each action class. Finally, to classify a new instance x, $\mathcal{R}_c(x)$ for all C classes is computed and the instance x is assigned to that class c with the greatest value:

$$\mathcal{R}_c(x) > \mathcal{R}_{\acute{c}}(x) \quad for \quad all \quad c \neq \acute{c} \tag{7}$$

Since diversity is relatively high among the two classifiers before fusion, simply averaging fusion produces poor results compared to the MLR fusion.

Experiments

In this section, we experimentally analyze the key features of *UCF-STAR* and the challenges it introduces.

Datasets and Metrics

UCF-STAR. As shown in Fig. 2, all classes are of sufficient and roughly equal size, therefore there are no issues of unbalanced classes. Our resulting benchmark consists of a total of 664,718 training, 166,180 validation and 207,724 test examples on 50 classes.

Other Datasets. We fully compare *UCF-STAR* with existing image datasets in terms of their challenges. ***Stanford-40*** (Yao et al. 2011) contains 40 classes and 9,532 images. We split this dataset into 2 categories with 11 and 29 actions, respectively[1]. ***Willow*** (Delaitre, Laptev, and Sivic 2010) contains 911 images split into 7 action categories. We divided the 7 action categories into two main groups, Body-Motion and Non-Body-Motion actions. Interacting with computer, Photographing, Playing music are considered as Non-Body-Motion actions, since they are relatively static and highly dependent on human-object interactions. ***WIDER*** (Xiong et al. 2015) dataset includes 14 human attribute labels and 30 event class labels containing 13,789 images. We considered 6 actions as the Body Motion category; i.e. running, basketball, football, soccer, skiing, hockey. The ***BU$_{101}$*** (Ma et al. 2017) consists of $23.8K$ images that correspond to the 101 action classes in the UCF101 video dataset. We used the train and test splits provided by the original authors for all datasets.

Metrics. For evaluation, we compute the average precision per class and report the average over all classes.

Dynamic-skeleton prediction and analysis

In order to learn the "latent" motion prior from video frames, we extracted over 36,000 frames from *UCF-101* (Soomro, Zamir, and Shah 2012), *UCF-Sport* (Rodriguez, Ahmed, and Shah 2008), *WEIZMANN* (Blank et al. 2005), *KTH* (Schuldt, Laptev, and Caputo 2004) video datasets categorized in 50 different action classes. These extracted frames, after the sampling process, were post-processed to eliminate frames with no clearly-visible human subject. The extracted frames were also augmented by flipping images, resulting in 57,600 frames. We labeled these frames with the video action they were sampled from. Next, RGB frames were converted to human body skeleton representation using Stacked Hourglass Networks (Newell, Yang, and Deng 2016). Skeleton is a lower dimensional shape description of an object. Consequently, the domain of inferred motion, based on human skeletons, helps to get rid of irrelevant information, and mitigate over-fitting.

Since the video data are available for the 57,600 extracted frames, we generate d_s labels for them as described earlier in section *Action Recognition Method* (Temporal Stream Network). Generated d_ss, using the Rank-SVM algorithm, form our labels for training a model to predict d_s for a still image with no video data available. As discussed in section *Action Recognition Method*, we devise a pixel-wise semantic segmentation encoder-decoder architecture for d_s prediction.

[1]Body Motion categories: climbing, jumping, cleaning floor, riding bike, riding horse, rowing boat, running, walking dog, shooting arrow, throwing frisby and waving hands.

Table 2: mAP of predicted d_s for different epochs.

Model	50 Epochs	200 Epochs	500 Epochs
Training from scratch	43%	54%	65%
Fine-tuning using VGG model	48%	64%	76%
Fine-tuning with custom loss function	55%	72%	**83%**

Table 2 shows prediction accuracy of d_s by modifying the loss function as described in section *Action Recognition Method*, compared with using default softmax loss layer.

Table 3: Action classification performance for *UCF-STAR*.

Method	mAP(%)
Object Bank (Li et al. 2010)	26.7
LLC (Wang et al. 2010)	31.5
R*CNN (Gkioxari, Girshick, and Malik 2015)	65.3
im2flow (Gao, Xiong, and Grauman 2018)	70.9
Temporal Stream-Trained from scratch	61.3
Temporal Stream-Fine-tuned all layers	68.3
Temporal Stream-Fine-tuned 7 top layers	**86.2**
Spatial Stream-Fine-tuned 7 top layers	**26.3**
TSSTN	**91.9**

Table 4: mAP(%) results on Stanford-40.

Method	Body-Motion	Non-Body-Motion	All
Gkioxari et al. (Gkioxari, Girshick, and Malik 2015)	93.87	89.73	90.9
Khan et al. (Khan et al. 2014a)	56.92	51.51	53
Khan et al. (Khan et al. 2013)	53.51	51.28	51.9
Yan et al. (Yan, Smith, and Zhang 2017)	92.26	87.07	88.5
Zhao et al. (Zhao, Ma, and You)	-	-	83.4
Zhao et al. (Zhao, Ma, and Chen 2017)	-	-	54.5
Zhao et al. (Zhao, Ma, and Chen 2016)	-	-	80.6
Zhou et al. (Zhou et al. 2014)	-	-	55.3
Sharma et al. (Sharma, Jurie, and Schmid 2017)	-	-	72.3
Khan et al. (Khan et al. 2015)	-	-	75.4
Gao et al. (Gao, Xiong, and Grauman 2018)	0	0	74.9
Ours-TSSTN	**97.8**	80.2	86.3

Comparison to the state of the art Tables 3-6 show action recognition performance of the proposed TSSTN method on *UCF-STAR*, as well as 4 other standard image datasets. TSSTN obtains state-of-the-art performance on Stanford-40, Willow, WIDER and BU-101, outperforming well-established baselines. However, table 3 shows that models in (Li et al. 2010; Wang et al. 2010; Gkioxari, Girshick, and Malik 2015; Gao, Xiong, and Grauman 2018) obtain relatively low performance on *UCF-STAR*. We attribute this to 1) the human body-motion characteristics of *UCF-STAR*, and 2) existence of visually similar poses performing different actions. Therefore, rich temporal prediction models may be needed to succeed at *UCF-STAR*, posing a new challenge for visual action recognition.

A very key observation is that unlike conventional action classification methods, TSSTN treats actions with similar poses, e.g. *running* vs. *walking*, differently. This is due to presence of temporal information in dynamic-skeletons carrying information on the pixels evolution, which would otherwise be missing. Our promising performance on body-motion categories in tables 4-6 shows the impact of the temporal prediction models in our action recognition method.

Table 5: mAP(%) results on the Willow dataset.

Method	Bike-ride	Horse-ride	Run	Walk	Overall (Body-Motion)
Delaitre et al. (Delaitre, Laptev, and Sivic 2010)	82.43	69.60	44.53	54.18	62.7
Delaitre et al. (Delaitre, Sivic, and Laptev 2011)	90.39	75.03	59.73	57.64	70.7
Sharma et al. (Sharma, Jurie, and Schmid 2012)	87.8	84.2	56.1	56.5	71.1
Sharma et al. (Sharma, Jurie, and Schmid 2013)	91.0	87.6	55.0	59.2	73.2
Khan et al. (Khan et al. 2014b)	87.2	77.2	63.7	60.6	72.2
Khan et al. (Khan et al. 2014a)	93.8	87.9	67.2	63.3	78.05
Liang et al. (Liang et al. 2014)	98.17	92.72	46.16	58.88	74.0
Zhao et al. (Zhao, Ma, and Chen 2017)	93.0	86.2	65.7	72.6	79.3
Khan et al. (Khan et al. 2013)	90.3	84.3	64.7	64.6	76.0
Ours-TSSTN	80.6	89.8	84.6	83.8	**84.6**

Table 6: Left: mAP (%) results on WIDER. - Right: mAP (%) results BU_{101} by categories.

Method	mAP (%)	
RCNN	80.0	
R*CNN	80.5	
DHC	81.3	
ResNet-SRN	86.2	
VeSPA	82.4	
Ours-TSSTN	Body-Motion	Non-Body-Motion
	90.3	71.7

Categories	mAP (%)
Human-Object	59.6
Body-Motion	93.8
Human-Human	68.9
Playing-Instrument	67.0
Sport	74.7

Conclusion

This paper introduces *UCF-STAR*, the largest annotated still image dataset for action recognition, having over $1M$ images annotated with multi-modal set of metadata. In addition, we propose TSSTN, a two stream spatiotemporal network that outperforms the current state of the art on standard benchmarks to serve as a baseline. TSSTN proves that predicting the "latent" temporal information in still images improves action recognition performance. Moreover, *UCF-STAR* highlights the need for developing new action recognition approaches based on predicting temporal information in still images.

References

Ashraf, N.; Sun, C.; and Foroosh, H. 2014. View invariant action recognition using projective depth. *CVIU* 41–52.

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2015. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*.

Balouchian, P.; Safaei, M.; and Foroosh, H. 2019. LUCFER: A large-scale context-sensitive image dataset for deep learning of visual emotions. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1645–1654.

Bilen, H.; Fernando, B.; Gavves, E.; Vedaldi, A.; and Gould, S. 2016. Dynamic image networks for action recognition. In *Proc. IEEE CVPR*.

Blank, M.; Gorelick, L.; Shechtman, E.; Irani, M.; and Basri, R. 2005. Actions as space-time shapes. In *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*.

Chéron, G.; Laptev, I.; and Schmid, C. 2015. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE international conference on computer vision*.

Delaitre, V.; Laptev, I.; and Sivic, J. 2010. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC 2010*.

Delaitre, V.; Sivic, J.; and Laptev, I. 2011. Learning person-object interactions for action recognition in still images. In *Proc. NIPS*.

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. IEEE CVPR*.

Feichtenhofer, C.; Pinz, A.; and Zisserman, A. 2016. Convolutional two-stream network fusion for video action recognition. In *Proc. IEEE CVPR*.

Fernando, B.; Gavves, E.; Oramas, J. M.; Ghodrati, A.; and Tuytelaars, T. 2015. Modeling video evolution for action recognition. In *Proc. IEEE CVPR*.

Fernando, B.; Gavves, E.; Oramas, J.; Ghodrati, A.; and Tuytelaars, T. 2017. Rank pooling for action recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*.

Gao, R.; Xiong, B.; and Grauman, K. 2018. Im2flow: Motion hallucination from static images for action recognition. In *Proc. IEEE CVPR*.

Girdhar, R.; Ramanan, D.; Gupta, A.; Sivic, J.; and Russell, B. 2017. Actionvlad: Learning spatio-temporal aggregation for action classification. In *Proc. IEEE CVPR*.

Gkioxari, G.; Girshick, R.; and Malik, J. 2015. Contextual action recognition with R*CNN. In *Proc. ICCV*.

Gupta, A.; Kembhavi, A.; and Davis, L. S. 2009. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*.

Hoai, M. 2014. Regularized max pooling for image categorization. In *Proc. BMVC*.

Ikizler, N.; Cinbis, R. G.; Pehlivan, S.; and Duygulu, P. 2008. Recognizing actions from still images. In *2008 19th International Conference on Pattern Recognition*.

Ikizler-Cinbis, N.; Cinbis, R. G.; and Sclaroff, S. 2009. Learning actions from the web. In *2009 IEEE 12th International Conference on Computer Vision*.

Jhuang, H.; Garrote, H.; Poggio, E.; Serre, T.; and Hmdb, T. 2011. A large video database for human motion recognition. In *Proc. of IEEE International Conference on Computer Vision*.

Khan, F. S.; Anwer, R. M.; van de Weijer, J.; Bagdanov, A. D.; Lopez, A. M.; and Felsberg, M. 2013. Coloring action recognition in still images. *Int. Journal of Computer Vision (IJCV)*.

Khan, F. S.; van de Weijer, J.; Anwer, R. M.; Felsberg, M.; and Gatta, C. 2014a. Semantic pyramids for gender and action recognition. *IEEE Trans. on Image Processing*.

Khan, F. S.; Van De Weijer, J.; Bagdanov, A. D.; and Felsberg, M. 2014b. Scale coding bag-of-words for action recognition. In *Int. Conf. on Pattern Recognition (ICPR)*.

Khan, F. S.; Xu, J.; Weijer, J.; Bagdanov, A. D.; Anwer, R. M.; and Lopez, A. M. 2015. Recognizing actions through action-specific person detection. *IEEE Trans. Image Proc.*

Le, D. T.; Bernardi, R.; and Uijlings, J. 2013. Exploiting language models to recognize unseen actions. In *ACM conference on International conference on multimedia retrieval*.

Li, L.-J., and Li, F.-F. 2007. What, where and who? classifying events by scene and object recognition. In *Iccv*.

Li, L.-J.; Su, H.; Fei-Fei, L.; and Xing, E. P. 2010. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Proc. NIPS*.

Li, P.; Ma, J.; and Gao, S. 2011. Actions in still web images: visualization, detection and retrieval. In *International Conference on Web-Age Information Management*.

Liang, Z.; Wang, X.; Huang, R.; and Lin, L. 2014. An expressive deep model for human action parsing from a single image. In *Multimedia and Expo (ICME), 2014*.

Ma, S.; Bargal, S. A.; Zhang, J.; Sigal, L.; and Sclaroff, S. 2017. Do less and achieve more: Training cnns for action recognition utilizing action images from the web. *Pattern Recognition*.

Maji, S.; Bourdev, L.; and Malik, J. 2011. Action recognition from a distributed representation of pose and appearance. In *Proc. IEEE CVPR*.

Marszałek, M.; Laptev, I.; and Schmid, C. 2009. Actions in context. In *Proc. CVPR*.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *Proc. of ECCV*.

Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *Proc. IEEE CVPR*.

Prest, A.; Schmid, C.; and Ferrari, V. 2012. Weakly supervised learning of interactions between humans and objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence*.

Rahman, M. A., and Wang, Y. 2016. Learning neural networks with ranking-based losses for action retrieval. In *Proc. Conf. on Computer and Robot Vision (CRV)*.

Raja, K.; Laptev, I.; Pérez, P.; and Oisel, L. 2011. Joint pose estimation and action recognition in image graphs. In *2011 18th IEEE International Conference on Image Processing*.

Rodriguez, M. D.; Ahmed, J.; and Shah, M. 2008. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Proc. IEEE CVPR*.

Safaei, M., and Foroosh, H. 2018. A zero-shot architecture for action recognition in still images. In *IEEE Int. Conf. on Image Processing (ICIP)*, 460–464.

Safaei, M., and Foroosh, H. 2019. Still image action recognition by predicting spatial-temporal pixel evolution. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 111–120.

Schuldt, C.; Laptev, I.; and Caputo, B. 2004. Recognizing human actions: a local svm approach. In *Proc. ICPR*, 32–36.

Sharma, G.; Jurie, F.; and Schmid, C. 2012. Discriminative spatial saliency for image classification. In *Proc. IEEE CVPR*.

Sharma, G.; Jurie, F.; and Schmid, C. 2013. Expanded parts model for human attribute and action recognition in still images. In *Proc. IEEE CVPR*.

Sharma, G.; Jurie, F.; and Schmid, C. 2017. Expanded parts model for semantic description of humans in still images. *IEEE Trans. Pattern Analysis and Machine Intelligence*.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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

Sun, C.; Junejo, I.; Tappen, M.; and Foroosh, H. 2015. Exploring sparseness and self-similarity for action recognition. *IEEE Trans. Image Processing* 2488–2501.

Sun, C.; Junejo, I.; and Foroosh, H. 2011. Action recognition using rank-1 approximation of joint self-similarity volume. In *Proc. of IEEE Int. Conf. on Computer Vision*.

Sun, C.; Tappen, M.; and Foroosh, H. 2014. Feature-independent action spotting without human localization, segmentation or frame-wise tracking. In *Proc. IEEE CVPR*.

Thurau, C., and Hlavác, V. 2008. Pose primitive based human action recognition in videos or still images. In *Proc. IEEE CVPR*.

Ting, K. M., and Witten, I. H. 1999. Issues in stacked generalization. *Journal of artificial intelligence research* 10:271–289.

Wang, J.; Yang, J.; Yu, K.; Lv, F.; Huang, T.; and Gong, Y. 2010. Locality-constrained linear coding for image classification. In *IEEE Conf. on CVPR*.

Xiong, Y.; Zhu, K.; Lin, D.; and Tang, X. 2015. Recognize complex events from static images by fusing deep channels.

Yan, S.; Smith, J. S.; and Zhang, B. 2017. Action recognition from still images based on deep vlad spatial pyramids. *Signal Processing: Image Communication*.

Yao, B., and Fei-Fei, L. 2010. Grouplet: A structured image representation for recognizing human and object interactions. In *Proc. IEEE CVPR*.

Yao, B., and Fei-Fei, L. 2012. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE Trans. Pattern Analysis and Machine Intelligence*.

Yao, B.; Jiang, X.; Khosla, A.; Lin, A. L.; Guibas, L.; and Fei-Fei, L. 2011. Human action recognition by learning bases of action attributes and parts. In *Proc. ICCV*.

Yu, H., and Kim, S. 2012. Svm tutorial—classification, regression and ranking. In *Handbook of Natural computing*.

Zhang, Y.; Cheng, L.; Wu, J.; Cai, J.; Do, M. N.; and Lu, J. 2016. Action recognition in still images with minimum annotation efforts. *IEEE Trans. Image Processing*.

Zhao, Z.; Ma, H.; and Chen, X. 2016. Semantic parts based top-down pyramid for action recognition. *Pattern Recognition Letters*.

Zhao, Z.; Ma, H.; and Chen, X. 2017. Generalized symmetric pair model for action classification in still images. *Pattern Recognition*.

Zhao, Z.; Ma, H.; and You, S. Single image action recognition using semantic body part actions. *CoRR*.

Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; and Oliva, A. 2014. Learning deep features for scene recognition using places database. In *Proc. NIPS*.