Human Action Recognition across Datasets by Foreground-weighted Histogram Decomposition

Waqas Sultani Imran Saleemi
Center for Research in Computer Vision, University of Central Florida
waqassultani@knights.ucf.edu imran@eecs.ucf.edu

Abstract

This paper attempts to address the problem of recognizing human actions while training and testing on distinct datasets, when test videos are neither labeled nor available during training. In this scenario, learning of a joint vocabulary, or domain transfer techniques are not applicable. We first explore reasons for poor classifier performance when tested on novel datasets, and quantify the effect of scene backgrounds on action representations and recognition. Using only the background features and partitioning of gist feature space, we show that the background scenes in recent datasets are quite discriminative and can be used classify an action with reasonable accuracy. We then propose a new process to obtain a measure of confidence in each pixel of the video being a foreground region, using motion, appearance, and saliency together in a 3D MRF based framework. We also propose multiple ways to exploit the foreground confidence: to improve bag-of-words vocabulary, histogram representation of a video, and a novel histogram decomposition based representation and kernel. We used these foreground confidences to recognize actions trained on one data set and test on a different data set. We have performed extensive experiments on several datasets that improve cross dataset recognition accuracy as compared to baseline methods.

1. Introduction

We investigate the problem of human action recognition when training and testing on distinct datasets. Research in recognition strives to develop increasingly generalized methods that are robust to intra-class variability and inter-class ambiguity. Indeed, recent years have seen tremendous strides in improving recognition accuracy [27, 18] on ever larger and complex benchmark datasets [13, 14, 12, 16, 11], comprising actions "in the wild" videos. Unfortunately, the all encompassing, dense, global representations [27, 28] that bring about such improvements often benefit from the inherent characteristics, specific to datasets and classes, that do not necessarily reflect knowledge about the entity to be recognized. This results in increasingly specific models that perform well within datasets but generalize poorly.

The need to mitigate this disconnect has given rise to the application of domain adaptation [1, 5], in recognition of objects [19] and events [6, 30]. Lixin *et al*. [6] employed an adaptive multiple kernel learning approach to minimize the mismatch between distributions from YouTube and consumer videos. Several variations of traditional SVM has been introduced for domain adaptation such as adaptive SVM [30], domain adaptation SVM [1], and domain adaptation machine [5]. However, the major limitation of all these approaches is that they require availability of video labels (or features [2]) from both domains during training.

There is no question that these techniques improve performance across datasets, and are significant in their own right, but it is worth asking whether the same actions in distinct datasets are truly representative of different domains or if their specific characteristics are distracting biases that emanate from data collection criteria and processes. This issue has been raised recently in an interesting work by Torralba and Efros [23] for the problem of image classification and object detection. They have empirically established that most object recognition datasets represent close visual world views and have biases toward specific poses, backgrounds, and locations, etc. In this study, we show that action recognition datasets too are prejudiced towards background scenes – a characteristic that should ideally be inconsequential to human action classes.

Explicit mitigation of dataset bias is possible, e.g., as proposed by Khosla *et al*. in [9] who introduced a data-driven approach where biases of different datasets were combined to classify image in a new dataset. We argue however, that research should focus on video (or image) *representation* instead, so it is invariant to said bias and potentially generalizes better across datasets. The underlying assumption is that the hypothetical, exhaustive set of examples of an action class would be truly representative of our visual world, and treating distinct datasets as training and testing partitions is a step towards realizing such a set. We propose that dataset invariant action representations should attempt to capture features of the actor's motion and appearance along with involved objects, and diminish the effect of scene background and clutter.

Historically, taking a page from image analysis, several *video* interest point detectors were introduced, including space time interest points [12], Dollar interest points [4], and spatiotemporal Hessian detector [29], etc. The obvious idea was to estimate local descriptors only at these important locations and ignore the rest of the video. Representations based on local descriptors estimated at interest points showed promising results on simple datasets such as Weizmann [21] and KTH [20]. Even though these datasets are now considered easier, their generally static, mostly uniform scene backgrounds, coupled with interest point detection, ensured a true action representation, largely devoid of background information.

In recent years, the difficulty in obtaining meaningful locations of interest in contemporary datasets, coupled with the lack of evaluation of action localization, has resulted in a shift in research focus away from interest point detection. Indeed, it has been shown experimentally, that dense sampling of feature descriptors generally outperforms interest point [28] and other detectors (human, foreground, etc.)[10]. Several methods have even been proposed to recognize actions in single images instead of videos [3]. It is then safe to assume that background scene information is a key component of the final representation that allows higher quantitative performance, but in the process 'learns the dataset' rather than the action. We maintain that the goal of action representation schemes and efforts to collect larger datasets should be to increase intra-class generalization for which cross-dataset recognition is a reasonable metric.

While obtaining meaningful interest points, actor contours or silhouettes in modern action datasets is challenging, we nevertheless argue that a truly representative action model that generalizes reasonably well across unseen datasets, would benefit from the same cues that are used for foreground segmentation. Attempting to estimate actor bounding boxes, or binary foreground-background labels is akin to introducing a new problem to solve the first. Therefore, we propose to perform *unsupervised* estimation of pixel-wise real-valued labels for the entire video that can be employed to control influence of different video regions on the final representation.

In this paper, we put forth several methods to exploit these foreground confidences for soft assignment of features within the bag-of-words paradigm. The two main aspects of our proposed methods are: important features should have larger influence in video representation; and regions with a specific level of importance should only be compared with corresponding similarly important parts of other videos. Previously, [25] proposed the soft assignment of features instead of hard quantization for improved accuracy but the features themselves are equally weighted, i.e., their total contribution to the histogram is constant. Context specific histograms were proposed in [22] for image classification, where different words contribute differently to each histogram. However, the context classifiers need to be pre-trained in a supervised manner. Ullah *et al*. [24] segmented a video into different regions and final representation consisted of concatenation of all histograms. Vig *et al*. [26] used saliency to remove features from non-salient regions, however, their method requires thresholding of saliency maps while each of the remaining features contribute equally to the final representation. Instead, in our first representation (weighted bag-of-words), each word contributes to the histogram according to its probability of being foreground. In our second representation (foreground based histogram decomposition), we divide video into arbitrary (potentially *non-contiguous*) regions according to their probability of being foreground, and the final distance between two videos is summation of *weighted* distances between regions that correspond to same quantized probability of being foreground. In this way, the negative effects of using (person, foreground) detectors, such as false positives and false negatives are mitigated. Moreover, the problem being considered in this paper is different from the above mentioned papers. In fact, we are not aware of any previous methods attempting the problem of action recognition across datasets, that do not exploit either labeled or unlabeled videos (or features) from the test dataset.

We believe that the representation of an action should be actor centric, so that a classifier learns the action and not the dataset and hence is able to recognize actions across completely different backgrounds. Moreover, although background and contextual information is useful and should be taken into consideration, its contribution to the final representation should be less than the action itself. We demonstrate that using soft weights instead of binary labels (person or not person), in addition to pixel-wise analysis (instead of bounding boxes) results in significant improvement. The rest of the paper is organized as follows: We propose several measures to quantify the effect of scene and background statistics on action class discriminativity in §2. In §3, we propose methods for obtaining foreground-specific action representations, using motion, appearance, and saliency in a 3D MRF based framework. Experimental setup and results are reported in §4. The paper is concluded in §5.

2. Background Discriminativity in Action Datasets

A recognition dataset should be representative of our surrounding visual world, and therefore diverse as possible. Besides illumination, clutter, etc., the sample actions

STIP Sampling	UCF Sports	UCF Youtube
Foreground only	71.92%	59.80%
Background only	73.97%	55.27%
Dense	75.34%	60.60%

Table 1. Accuracy using STIP in different video regions. There is little decrease in performance even when completely ignoring features on the actor. In UCF Sports, background only features actually perform better than foreground only features.

should vary in terms of actor viewpoint, pose, speed, and articulation. The background should be diverse as well, but *not* discriminative, i.e., it should not aid in recognition of the action class, or it would limit the generalizability of the class model, and consequently result in worse cross-dataset recognition than within dataset. In this section, we quantify the discriminative power of background scenes in a few well known action datasets using two methods. First, we computed motion features on only the background regions to perform recognition within datasets, and second, we measured class-wise confusion within datasets using global scene descriptor.

2.1. Background Motion Features

Computation of background features in a video requires annotation or estimation of regions corresponding to the action. For this analysis, two recent datasets were selected: UCF Sports [17] and UCF Youtube [13], due to availability of manually annotated actor bounding boxes.

Dense space-time interest point descriptor (STIP) [12] was extracted for all videos in both datasets. The features were extracted with a 50% spatiotemporal overlap, using a single spatial and temporal scale. The features with an overlap of more than 50% with the actor bounding boxes were then labeled as foreground, while all remaining features were considered background features. Notice that multi-scale features would not allow such categorization. The train/test process followed the original papers, i.e., leave one actor out classification for UCF Sports, and leave one group out classification for UCF Youtube. The experimental results for both datasets are shown in Table 1. It is evident that even complete removal of foreground words does not have a significant detrimental effect on accuracy. It is reasonable to assume that the action in an arbitrary video can hypothetically be replaced with a different action without a significant change in background feature descriptors. The implication then is that the background alone is almost as discriminative as the action itself. An action model trained on these datasets with dense feature coverage will therefore perform poorly on a novel test set with a different background composition or distribution.

2.2. Global Scene Features

Using a global image descriptor to represent an action video has two inherent disadvantages. First, it would ignore the motion or temporal properties of the video, and secondly, it may not appropriately capture the background scene since a significant part of every image can potentially be the actor. If a scene descriptor, despite these limitations, can be used to reasonably recognize an action, the corresponding classifier is likely to generalize poorly.

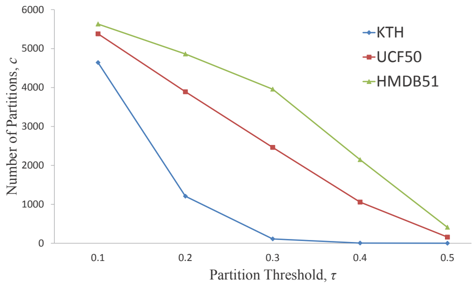


Figure 1. Relative number of partitions c of frames in KTH, UCF50, and HMDB51 datasets at varying minimum inter-partition distances, τ. At the maximum allowed distance (0.5), KTH, UCF50, and HMDB51 have 1, 158, and 411 partitions respectively. Please see text for interpretation.

We used the gist descriptor [15] to quantify discriminativity of background scenes in action datasets. The gist descriptor was computed for every 50th frame of all videos in the KTH [20], UCF50 [16], and HMDB51 [11] datasets. The first experiment we performed was to quantify the relative number of distinct background scenes in each dataset at a fixed level of separation in the feature space. Given n gist descriptors in a dataset, a graph $G = (V, E)$ is constructed, such that $V = \{v_i\}$, $i \in \{1, \ldots, n\}$, is the set of all descriptors, and $E = \{e_{ij}\}$, $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, n\}$, and $e_{ij} = \|v_i - v_j\|_2$ is the Euclidean distance between descriptors i and j. The feature distance matrix, E is then thresholded to obtain U such that $u_{ij} = 1$, if $e_{ij} \leq \tau$, and 0 otherwise. The graph connected component analysis of U then results in a partitioning of the n gist descriptors into c groups.

The number of partitions obtained in this manner are independent of n, and for a specific τ provides a comparison of two datasets in terms of the relative diversity of scenes. The larger the value of c, the more diverse the backgrounds will be. A quantitative comparison of the number of partitions in each of KTH, UCF50, and HMDB51 datasets is shown in Fig. 1. As expected, KTH with largely uniform background, and little camera motion, consistently has the fewest partitions. One would expect that UCF50 and HMDB51 with similarly large number of classes, and complexity, should have similar number of partitions in the gist feature space, at equal inter-partition distances. However HMDB51 has consistently larger values of c for the same τ, as compared to UCF50. This comparison points to the hypothesis that the background scene features would be less helpful for the former dataset. A similar observation was made in [11] when using scene descriptors.

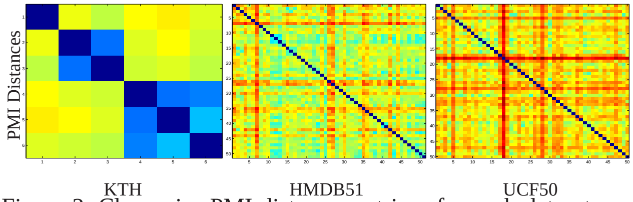


Figure 2. Class-wise PMI distance matrices for each dataset can be considered as the inverse of a confusion matrix. The values are normalized with respect to the maximum of across all 3 datasets, so the same colors correspond to the same absolute value.

k	100	200	300
KTH	5.12	8.25	10.40
HMDB51	7.15	11.07	14.06
UCF50	7.97	11.79	14.38

Table 2. Average discrimination between classes of different datasets, using PMI of gist clusters, for different number of clusters. Discrimination increases (confusion decreases) as number of clusters (background scene codebook size) increases. Even though there are fewer distinct backgrounds in KTH (see Fig. 1), it is the hardest to classify using scene information alone. UCF50 and HMDB51 are comparable, the latter being consistently harder.

The number of partitions c however, does not explicitly reveal the relative importance of background in discriminating action classes from each other. Therefore, we performed another experiment where we clustered the gist descriptors for each dataset into k clusters using K-means. The pointwise mutual information (PMI) between each cluster and an action class, resulting in a $k \times N$ matrix, where N is the number of classes in the dataset. We then computed the class-wise Euclidean distances between all pairs of classes to obtain an $N \times N$ matrix, P (see Fig. 2 for examples). Each element p_{ij} of the matrix represents the discrimination between gist-based representations of classes i and j. The larger the value, the easier it is for gist to classify a test action video. We compute the mean discrimination as the average of matrix P. These values for different values of k are reported in Table 2. It can be noticed that even with fewer action categories, KTH is relatively the hardest to classify using gist (not to be confused with actual accuracy using a classifier like SVM). The relative confusions for HMDB51 and UCF50 are more similar, with the former being consistently harder than the latter.

3. Foreground specific Action Representation

Given our experimental verification of the effect of scene background on action classification, we propose to learn foreground specific action representation to improve recognition on novel test sets. However, the problems of foreground-background segmentation, and human or actor detection are very challenging, and all the more so, in unconstrained videos that make up the more recent action datasets. Since our eventual goal is to recognize actions, rather than segmentation, or actor detection, our proposed framework does not attempt to label each pixel or region as foreground or background. Instead, we estimate the confidence in each pixel being a part of the foreground, and use it directly to obtain the codebook as well as the video representation. This confidence is computed using several cues as explained below.

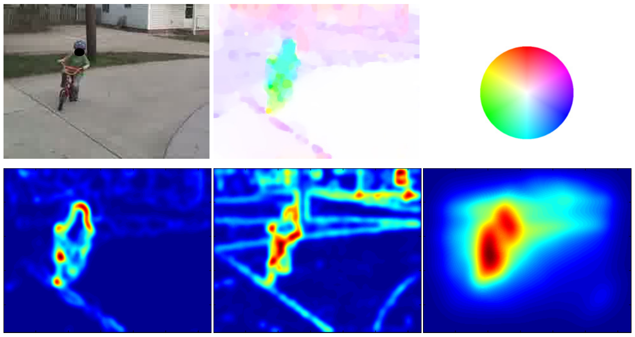


Figure 3. Top: original video frame on the left and optical flow on the right as per the color wheel. Bottom(L-R): optical flow gradient magnitude $f_m(x,y)$; magnitude of color gradient in LAB space $f_c(x,y)$; and saliency $f_s(x,y)$, resp.

3.1. Motion Gradients

Action is mainly characterized by the motion of moving parts. We used this important clue to give high confidence to the locations undergoing articulated motion in a video. However, since most of the realistic datasets involve moving camera, simple optical flow magnitude can be high for background as well. Hence, we used the Frobenius norm of optical flow gradients. The motion gradients based foreground confidence, f_m is then defined as:

$$f_m(x,y) = \sqrt{u_x^2 + u_y^2 + v_x^2 + v_y^2} \ * \ g, \qquad (1)$$

where, u_x, v_x, u_y, and v_y are the horizontal and vertical gradients of optical flow respectively, and g is a 2D Gaussian filter with fixed variance. The main idea behind using optical flow gradients is that it not only helps to remove camera motion but also gives high magnitude around the articulated moving object. A qualitative example of $f_m(x,y)$ for a frame in a moving camera video is shown in Fig. 3.

3.2. Color Gradients

In many videos, the actor has different appearance and color than the background, while the background (such as sky or floor) has relatively uniform color distribution. Therefore, the color gradients can be used as a cue towards estimating the confidence in location of actor and object boundaries, while resulting in low responses for background regions with uniform colors. Specifically, we compute the color gradient based confidence in observing a foreground pixel, f_c, using the Frobenius norm of LAB color space given as:

$$f_c(x,y) = \sqrt{L_x^2 + L_y^2 + a_x^2 + a_y^2 + b_x^2 + b_y^2} \ * \ g, \qquad (2)$$

where (L_x, a_x, b_x) is the horizontal gradient of the color vector at (x, y). A qualitative example of $f_c(x, y)$ is shown in Fig. 3.

3.3. Saliency

We propose to use visual saliency as the third cue to estimate the confidence in observing a foreground pixel. In sports videos (a common type of actions in UCF50, HMDB51 and olympic sports), the player receives most of visual attention, and hence represents the most salient part of the video. A similar observation applies to amateur as well as professional moving camera videos that follow objects with distinct appearance amid relatively homogenous backgrounds. Although, our ultimate goal is to estimate foreground confidences for *videos*, we experimentally observed that, due to large camera motion and noisy optical flow, video or motion based saliency methods do not always result in reasonable outputs. Instead, we used graph based visual saliency [8] to capture the salient regions in each frame individually. We chose this method due its computational efficiency, evident capability in finding salient regions and natural interpretation as decomposition of image into neural network.

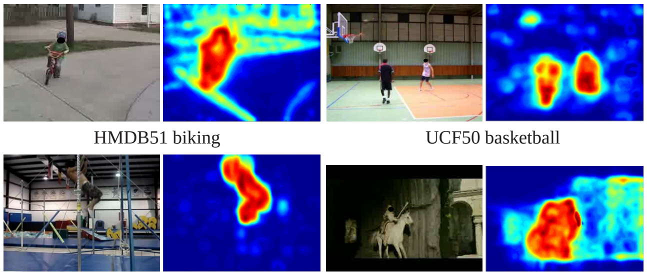
Following [8], we computed contrast, luminance, and four orientation maps corresponding to orientation $\theta = \{0^o, 45^o, 90^o, 135^o\}$ using Gabor filters, all on multiple spatial scales. In the activation step, a fully connected directed graph is built where edge weight between two nodes, corresponding to pixel locations, (i, j) and (p, q) is given as:

$$B_a(i,j,p,q) = \Big|M(i,j) - M(p,q)\Big| \quad exp\left(-\frac{(i-p)^2 + (j-q)^2}{2\varphi^2}\right), \tag{3}$$

where $M(i, j)$ represents the features at (i, j), and φ is a free parameter. Using the graph to define a Markov chain, the stationary distribution of the chain is computed and treated as an activation map, $A(p, q)$. A new graph is then defined on all pixels with the edge weights being:

$$B_n(i,j,p,q) = A(p,q) \; exp\left(-\frac{(i-p)^2 + (j-q)^2}{2\varphi^2}\right). \tag{4}$$

Again, the weights of outbound edges are normalized and the graph is treated as a Markov chain. The equilibrium distribution of the chain is then used as a per pixel saliency measure, $f_s(x, y)$. An example of the final saliency based foreground confidence is shown in Fig. 3.


HMDB51 biking

UCF50 basketball

UCF50 pull up

HMDB51 horse riding

Figure 4. Original video frame, and corresponding confidence of each pixel being the foreground, $f_a(x, y)$, for 4 example videos.

3.4. Coherence of Foreground Confidence

Since saliency and color gradients are computed based on a single frame, they ignore the temporal information as well as coherency. Moreover, color as well as optical flow computation does not explicitly impose spatial coherence constraints. We therefore compute an initial aggregate confidence map, $\hat{f}_a$, as: $log\left(f_m\left(f_c + f_s\right) + 1\right)$. The values of $\hat{f}_a$ are max-normalized for each frame of a video. In order to impose spatiotemporal dependency among neighboring pixels, we use temporal extension of 2D Markov random field [7] similar to [31]. The video is considered as a 3D grid graph, $(\mathcal{V}, \mathcal{E})$, where each node is connected to four spatial and two temporal neighbors. If a labeling ω assigns a weight $\omega_p \in \Omega = [0, 1]$ to a node, $\psi_p \in \mathcal{V}$, then the quality of labeling is given by the following energy function:

$$E(\omega) = \sum_{\psi_p \in \mathcal{V}} D_p(\omega_p) + \sum_{(p,q) \in \mathcal{V}} V(\omega_p - \omega_q). \tag{5}$$

We used quadratic data term defined as $D_p(\omega_p) = (\hat{f}_a(p) - \omega_p)^2$ and truncated quadratic smoothness term given as $V(\omega_p - \omega_q) = min((\omega_p - \omega_q)^2, \kappa)$. During inference, in addition to spatial neighbors, each node receives a message from the temporal neighbors as well. At time t, the message, $m^t_{p \to q}(\omega_q)$, that node p sends to q is given as:

$$\min_{\omega_p} \left(D_p(\omega_p) + V(\omega_p - \omega_q) + \sum_{s \in \mathcal{N}_p \backslash q} m^{t-1}_{s \to p}(\omega_p) \right), \tag{6}$$

where $\mathcal{N}_p \backslash q$ represents the five neighbors of p other than q and the belief vector for node q at time t is,

$$b^t_q(\omega_q) = D_q(\omega_q) + \sum_{s \in N_q} m^t_{s \to q}(\omega_q). \tag{7}$$

The inference starts by sweeping in six directions, and after a fix number of iterations, the weights which give the minimum cost for the belief vector are selected as final, and denoted as f_a. Qualitative examples of the final confidence after 3D MRF is shown in Fig. 4. Note that after imposition of spatiotemporal dependencies the resultant map assigns high weights for the entire region corresponding to the actor and the bicycle and gives low values for the background pixels despite the camera motion.

3.5. Foreground weighted Representation

Given the confidence in each pixel being in a foreground region, we propose to modify the bag-of-words representation of a video in several important ways. The underlying goal is to represent the video so that features corresponding to the actual action, i.e., the foreground, contribute towards the vocabulary as well as the resulting representation, while those on the background have minimal effect when training models or comparing videos. Our proposed ideas are described in the following.

Weighted Codebook: Traditionally, the codebook or vocabulary in a bag-of-words framework is learned using K-means, whereby the set of feature descriptors $X = \{\mathrm{x}_i\}$, obtained from training videos are the data points to be clustered into k clusters, with centers represented as z_j. Given the pixel wise foreground confidence for every frame in every video, we begin by computing the average foreground confidence, $w_i = \sum_{(x,y)\in P_i} f_a(x,y)/|P_i|$, where P_i is the set of pixels in the spatiotemporal volume corresponding to descriptor x_i. We then employ weighted K-means to obtain the codebook, where the goal of clustering is to minimize the following energy function,

$$\underset{C}{argmin} \sum_{j=1}^{K} C(i,j) w_i \left\|\mathrm{x}_i - \mathrm{z}_j\right\|^2, \tag{8}$$

where C is the $|X| \times k$ unknown membership matrix. The resulting vocabulary, $Z = \{\mathrm{z}_j\}$, is a set of points in the feature space that are more similar to descriptors with high confidence of being a foreground, and potentially farther away from descriptors on the background.

Weighted Histogram: In the bag-of-words method, the image or video is represented as a k long vector, $H = [h^1, \ldots, h^K]$, where h^j is the number of times the nearest neighbor of a descriptor in the video is found to be z_j. Given the average foreground confidence, w_i for the descriptor, x_i, we propose to compute the weighted histogram, $\hat{H}$, where $\hat{h}^j$ is the sum of w_i for all descriptors whose nearest codeword is z_j. The weighted histogram therefore is influenced by features with high confidence of being in the foreground regions, while the background features have a minimal effect on the final video representation. The weighted histogram is not to be confused with soft quantization where the weight w_i would have been distributed across bins.

Foreground Confidence based Histogram Decomposition: We notice that despite weighing the influence of features on the histogram, the accumulative effect of background features on different bins of the histogram can sum up to be significant. This is because of the fact that a significant number of pixels in the video, and consequently densely samples descriptors, can have relatively low foreground confidence. In other words, the number of high confidence features contributing to the histogram is far less than those with low confidence of being foreground. This would not be a problem if features with high and low confidences were quantized to different words, but that may not always be the case, especially due to the weighted codebook.

If the foreground and background regions were divided into two distinct classes (binary labeled), it would be straightforward to compute two different histograms for each type of region. However, given that it is desirable to avoid thresholding and binarization of foreground confidence, we propose a novel alternative solution. We begin by categorizing the spatiotemporal regions corresponding to different feature descriptors into R classes. These classes correspond to R equal, non-overlapping, exhaustive partitions of the range of average foreground confidences, w. A set $\hat{\mathbf{H}}$ of R weighted histograms, $\hat{H}_r$, is then computed for all the features in each of the R groups separately. The following kernel function then replaces histogram intersection:

$$\Delta\left(\hat{\mathbf{H}}^i, \hat{\mathbf{H}}^j\right) = \sum_{r=1}^{R} \alpha_r \Theta\left(\hat{H}_r^i, \hat{H}_r^j\right), \tag{9}$$

where Θ is the histogram intersection kernel, and α_r are predefined weights, that increase linearly with $r \in \{1, \ldots, R\}$. As a result, regions of two videos that have approximately the same foreground confidence, are compared only with each other. This process and the effect of our proposed scheme are illustrated in Fig. 5. As can be seen in the figure, the proposed multiple weighted histograms and the weighted average of histogram intersection kernel, show obvious improvement as a representation and measure of similarity between videos, respectively.

4. Experimental Results

The main goal of our experiments is to verify that models trained using features from foreground regions are likely to generalize better and attain higher recognition accuracy than dense sampling, especially when tested on videos from novel, unseen datasets. To this end, we have performed extensive experiments, evaluating the effect of our proposed foreground confidence measure in a weighted bag-of-words framework for cross dataset recognition over three datasets: UCF50, HMDB51, and Olympic sports.

UCF50 has 50 action categories. Since all the videos in UCF50 are taken from Youtube, they are implicitly biased towards a specific type of video shooting, including but not limited to amateur shooting style, cluttered background, and abrupt camera motion. Videos in 51 action categories of HMDB51 are mostly taken from movies and a small number from YouTube and Google Videos. The two datasets have 10 actions with common class labels, namely basketball, biking, pull ups, golf swing, horse riding, punch, fencing, push ups, rock climbing, and walking. In our experiments, we only chose the first 5 classes which are *visually* similar in both datasets. We did not consider other actions because even though they have similar class labels, there are visually very different: almost all the videos of punch in UCF50 correspond to the sport of boxing in a boxing ring, while most of the HMDB51 punch videos are more unconstrained, such as those from fist fights. Similarly the walking action of HMDB51 is quite different from

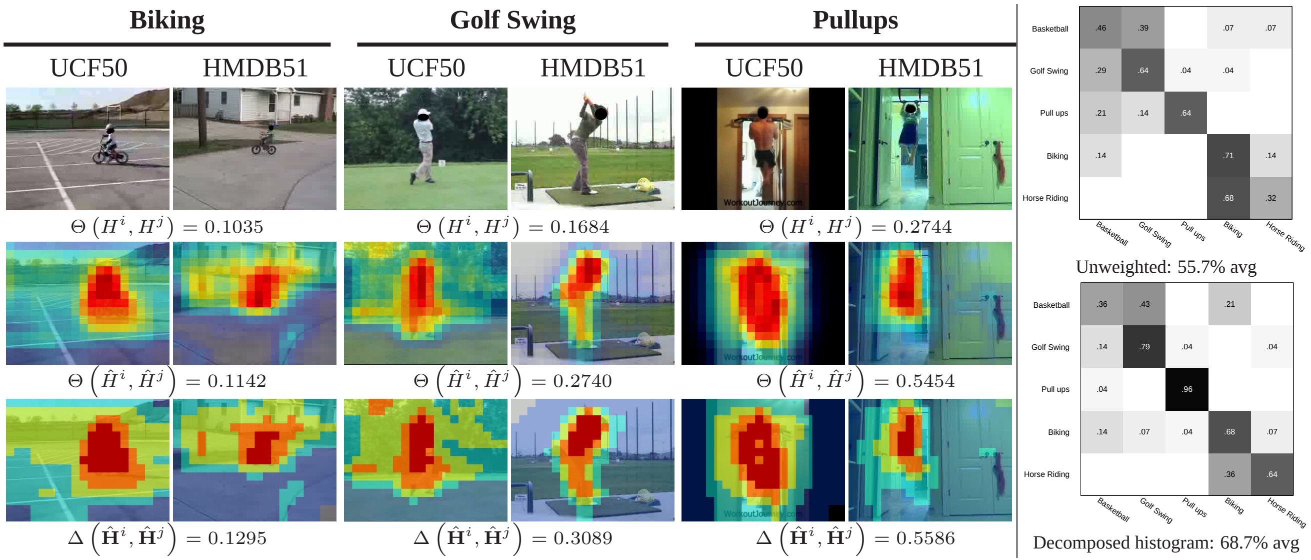


Figure 5. Figures on the left illustrate the effect of weighted histograms, and foreground confidence based decomposed histograms. In each column the top row shows the original image. The middle row shows the relative average foreground confidence or weight w_i of each spatiotemporal cuboid in the video, where shades of red correspond to high values. The third row shows the category or group, $r \in \{1, \ldots, R\}$, out of a total of $R = 5$. Features in each group are compared only with those in corresponding group in other videos. Notice that similarity between same label videos increases with use of weighted histograms, $\hat{H}$, and the proposed kernel Δ for decomposed histograms, $\hat{\mathbf{H}}$. The right column shows confusion tables for unweighted and decomposed weighted histogram classifiers trained on UCF50 and tested on HMDB51. See text for detailed analysis.

the ‘walking with dog’ action of UCF50. We only want to evaluate videos from the same class which are at least visually/semantically similar to a human observer.

For training from and testing on HMDB51, we used the train and test partitions for each of the chosen action as the original setup [11]. In order to have a fair comparison, we selected the same number of training and testing videos from UCF50 by dividing 100 videos of each class into 18 and 7 groups respectively, i.e., an ∼70-30 ratio. We computed dense STIP descriptors over both datasets with cuboid size of 32 by 32 spatially, and 15 frames temporally, with 50% spatial and temporal overlap.

For all datasets, we trained multi-class classifiers. An 3000-words vocabulary was created using K-means algorithm. As in traditional bag-of-words approach, all the STIP features were quantized into 3000-bin histograms for each video, to establish the baseline representation, H. For weighted histograms $\hat{H}$, instead of having equal contribution, each descriptor contributed to the histogram based on its confidence in being the foreground. Finally for foreground based histogram decomposition $\hat{\mathbf{H}}$, we divided the set of all features in each video into $R = 5$ groups based on foreground weights. We used $\alpha_i = \{1, 0.8, 0.6, 0.4, 0.2\}$, for weighted summation of the 5 histogram intersections.

The Olympic dataset consists of 16 human actions, where the videos mostly depict athletes practicing different sports actions. This dataset is also collected from Youtube. Similar to our previous experiment, we used 6 common actions between UCF50 and Olympic Sports (see Table 3 for labels). Due to the small number of training and testing examples in Olympic Sports, we extended the dataset by adding a horizontally flipped version of each video sequence. In order to ensure fairness, we used the same number of videos from UCF50 and added their horizontally flipped counterparts. We ensured that both the original and flipped video pairs are in either training or testing sets but not both.

Our experimental results are reported in Fig. 5(right) and Table 3. It can be seen from the confusion matrices in Fig. 5 corresponding to performance of UCF50-trained classifiers on HMDB51 test videos, compared to the traditional bag-of-words, our proposed representations exhibit consistent improvement in all action classes except basketball. When using baseline classifiers trained on UCF50, 68% of the horse riding examples in HMDB51 are classified as biking. Using the proposed method however, we were able reduce this confusion to 36%. In the complimentary experiment, baseline classifiers trained on HMDB51 categorized 57% of UCF50 biking examples as horse riding, while the proposed method reduced the confusion to 30%. Similarly, reducing dependency on background has significantly improved accuracy for pull ups and golf swing, where the actions are visually similar across the datasets. The drop in basketball accuracy is likely due to variation in actor pose and viewpoint across datasets.

As reported in Table 3, the quantitative results conclusively demonstrate that the proposed framework for estimation of foreground confidence is meaningful, and the con-

Training	Testing	Unweighted	Weighted	Histogram Decomposition	Actions
UCF50	UCF50	70.00	74.20	77.85	Biking, Golf swing, Pull ups, Horse riding, Basketball
UCF50	HMDB51	55.70	60.00	68.70	
HMDB51	HMDB51	65.30	69.30	68.00	
HMDB51	UCF50	63.33	64.00	68.67	
Olympic Sports	Olympic Sports	71.80	73.95	69.79	Basketball, Pole vault, Tennis serve, Diving, Clean & jerk, Throw Discus
UCF50	Olympic Sports	31.25	31.25	33.33	
Olympic Sports	UCF50	16.67	32.29	47.91	

Table 3. Average accuracy of action recognition across different pairs of training and testing datasets. 'Unweighted' is the traditional bag-of-words paradigm, using dense STIP features. The column labeled 'weighted' corresponds to foreground confidence weighted vocabulary and weighted histograms. The column labeled 'histogram decomposition' uses multiple histograms for different range of foreground confidence values, and uses a weighted mean of individual histogram intersections as the kernel. As can be observed, our two proposed representations perform significantly better than the baseline for most experiments.

sistently higher recognition accuracies serve as an empirical verification of our conjecture that the dataset specific background scenes are one of the main causes of deterioration in recognition accuracy across datasets. Moreover, when training and testing on distinct datasets, the histogram decomposition and the newly proposed corresponding similarity measure perform better than even the foreground weighted vocabulary and histograms, for all cross-dataset experiments.

.

5. Conclusion

We have attempted cross dataset action recognition without using labels or features from the test set. In doing so, we have experimentally demonstrated the detrimental effect of background scenes on action recognition dataset. We have also proposed a new process for obtaining per pixel confidence of every video pixel being the foregreound, as well as novel soft assignment, and histogram decomposition schemes for the bag-of-words representation. Our extensive experimental results and discussion validates the proposed ideas and framework.

Acknowledgements: This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20071. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

[1] L. Bruzzone and M. Marconcini. Domain adaptation problems: A dasvm classification technique and a circular validation strategy. *PAMI*, 32(5):770–787, 2010.

[2] L. Cao, Z. Liu, and T. Huang. Cross-dataset action detection. In *CVPR*, 2010.

[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.

[4] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *PETS*, 2005.

[5] L. Duan, D. Xu, and I. W. Tsang. Domain adaptation from multiple sources: A domain-dependent regularization approach. *IEEE Trans. NNLS*, 23(3):504–518, 2012.

[6] L. Duan, D. Xu, I. W. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. In *CVPR*, 2010.

[7] P. Felzenszwalb and D. Huttenlocher. Efficient belief propagation for early vision. In *CVPR*, 2004.

[8] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, 2006.

[9] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba. Undoing the damage of dataset bias. In *ECCV*, 2012.

[10] A. Kläser, M. Marszałek, I. Laptev, and C. Schmid. Will person detection help bag-of-features action recognition? Technical Report RR-7373, INRIA Grenoble - Rhône-Alpes, 2010.

[11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011.

[12] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[13] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild. In *CVPR*, 2009.

[14] N. Niebles, C. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010.

[15] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42:145–175, 2001.

[16] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *MVA*, 24(5):971–981, 2012.

[17] M. Rodriguez, J. Ahmed, and M. Shah. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.

[18] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012.

[19] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

[20] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004.

[21] E. Shechtman, L. Gorelick, M. Blank, M. Irani, and R. Basri. Actions as space-time shapes. *PAMI*, 29(12):2247–2253, 2007.

[22] Y. Su and F. Jurie. Visual word disambiguation by semantic contexts. In *ICCV*, 2011.

[23] A. Torralba and A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

[24] M. M. Ullah, S. N. Parizi, and I. Laptev. Improving bag-of-features action recognition with non-local cues. In *BMVC*, 2010.

[25] J. Van Gemert, C. Veenman, A. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *PAMI*, 32(7):1271–1283, 2010.

[26] E. Vig, M. Dorr, and D. Cox. Space-variant descriptor sampling for action recognition based on saliency and eye movements. In *ECCV*, 2012.

[27] H. Wang, A. Kläser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[28] H. Wang, M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[29] G. Willems, T. Tuytelaars, and L. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*, 2008.

[30] J. Yang, R. Yan, and A. Hauptmann. Cross-domain video concept detection using adaptive svms. In *ICM*, 2007.

[31] Z. Yin and R. Collins. Belief propagation in a 3d spatio-temporal mrf for moving object detection. In *CVPR*, 2007.