Object, Scene and Actions: Combining Multiple Features for Human Action Recognition

Nazli Ikizler-Cinbis and Stan Sclaroff

Department of Computer Science, Boston University

Abstract. In many cases, human actions can be identified not only by the singular observation of the human body in motion, but also properties of the surrounding scene and the related objects. In this paper, we look into this problem and propose an approach for human action recognition that integrates multiple feature channels from several entities such as objects, scenes and people. We formulate the problem in a multiple instance learning (MIL) framework, based on multiple feature channels. By using a discriminative approach, we join multiple feature channels embedded to the MIL space. Our experiments over the large YouTube dataset show that scene and object information can be used to complement person features for human action recognition.

1 Introduction

Action recognition "in the wild" is often a very difficult problem for computer vision. When the camera is non-stationary, and the background is fairly complicated, it is often difficult to infer the foreground features and the complex dynamics that are related to an action. Moreover, motion blur, serious occlusions and low resolution present additional challenges that cause the extracted features to be largely noisy.

Under such challenges, we argue that the features of the scene and/or moving objects can be used to complement features extracted from people in the video. The intuition behind this is straightforward: the presence (or absence) of particular objects or scene properties can often be used to infer the possible subset of actions that can take place. For example, if there is a pool within the scene, then "diving" becomes a possible action. On the contrary, if there is no pool, but a basketball court, then the probability of the "diving" action reduces. In this work, our aim is to capture such relationships between objects, scenes and actions.

Our approach starts with extracting a large set of features for describing both the shape and the motion information in the videos. All the features are extracted densely, allowing spatial and temporal overlap, and we operate over tracks when the temporal continuity is available. We do not use any explict object detectors, but treat each moving region as an object candidate. In the end, the videos are represented with multiple feature vectors acquired from different feature channels.

We are particularly interested in human action classification in the real-world, i.e. in unconstrained video sources like YouTube. In this problem, the videos are weakly annotated; there is a class label for each video sequence, however we do not know where or when in the video sequence the action occurs. Moreover, there may be more

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part I, LNCS 6311, pp. 494–507, 2010.

than one person or moving object in the video, and only a subset of the detected regions are involved in the action. Our aim is to be able to train our action models in the presence of such diverse conditions.

For this purpose, we formulate our problem within a multiple instance learning (MIL) framework, where the training set is ambiguous and the training labels are associated with bags of instances, rather than single instances as in a fully supervised system. The obvious advantage of using such an approach is to tolerate the large amount of irrelevant instances or false detections in the input videos.

In order to accomodate multiple heterogenous feature types, we define an agglomerative multiple instance learning framework, where each video is represented with multiple bags and each bag corresponds to a different feature channel. The MIL positivity constraint on a bag is therefore extended over multiple bags, i.e., at least one bag is required to contain one positive instance for the particular action. We then formulate a discriminative learning strategy with globally weighted or unweighted combinations of these multiple bags. We test our approach over the extensive YouTube dataset provided by Liu et al [22], and the results demonstrate that the proposed framework effectively combines different and noisy feature channels for accurate human action recognition.

2 Related Work

Human action recognition has been a very active research topic over the recent years. This makes the comprehensive listing of the related literature impossible, while Forsyth et al. [10] presents an extensive review of the subject. Some of the recent works include [8,29,18,20,16,13]. In most of the earlier works, the focus is on simpler scenarios, where the background was stable and the foreground human figure is easy to extract [4,18]. However, this scenario is hardly realistic; videos from the real world are fairly complicated, especially when taken in uncontrolled environments. Some recent approaches try to deal with such complex scenarios [19,25,20,16].

Joint modeling of object and action interactions has been a recent topic of interest. Moore et al.'s work [26] is one of the earliest attempts to consider actions and objects together. They use belief networks for modeling object and hand movements extracted from static camera sequences. Gupta et al. [12] try to improve the localization of both objects and actions by using a graphical Bayesian model. Marszalek, et al. [24] use movie scripts as automatic supervision for scene and action recognition in movies. Han et al [13] use context and higher level bags-of-detection descriptors for action recognition. They assume that the objects related to each action are known beforehand and corresponding object detectors are available. In our case, we do not rely on explicit object detectors and try to discover related objects in an unsupervised manner. We consider each moving region as a candidate object region and we utilize shape and motion descriptors for all candidate regions. While doing this, we have no explicit knowledge about their class membership.

Multiple Instance Learning (MIL) paradigm has been explored in quite a number of studies, both in machine learning([2]) and computer vision([23,31,5]). Computer vision problems are in fact very suitable application domains for MIL algorithms, because of the high-level of ambiguity in the domain. There are also some recent works

which use multiple instance learning for tracking and human actions. Babenko et al. [3] introduce an online MIL algorithm for target tracking. Ali and Shah use multiple-instance embedding [1] to facilitate classification with their kinematic mode features. Hu et.al [14] utilize a simulated annealing based MIL algorithm for finding the exact location of the actions over HOG features.

YouTube videos have been a focus of interest recently, due to its popularity being a widespread source that contains various challenging videos. Niebles, et al. [27] present a method for detecting moving people from such videos. Ikizler-Cinbis, et al. [17] use web images to facilitate action recognition in uncontrolled videos. Tran, et al. [30] work on YouTube Badminton videos. Recently, Liu, et al. [22] collected a large action dataset, and presented a method based on PageRank algorithm to prune the large number of space-time interest points in these videos.

3 Features

Features constitute the basic building blocks of our algorithm. Here, the idea is to extract as many meaningful and informative features as possible, both at the high and low-level. These features are extracted densely, in the sense that there can be spatial or temporal overlap between them. We will rely on the learning algorithm to extract useful patterns that associate each action with the combination of these different sets of features. There are three sets of features, namely "person-centric","object-centric" and "scene-centric" features. All these feature channels are depicted in Fig. 1. In this section, we first describe the video pre-processing steps and then go into the details of the feature extraction procedure.

3.1 Stabilizing Videos

When there is camera motion and the background is not static, the optical flow of the foreground objects is not easy to estimate, amidst the noisy flow field. Therefore, before extraction of features, especially the motion-related ones, the videos should be stabilized. We use a dominant motion compensation procedure for this purpose.

In order to estimate the foreground flow field, we make use of a homography-based motion compensation approach, similar to [21]. Assuming that the background is relatively dominant in the scene, we can estimate the background flow by calculating the homography between consecutive frames. For this purpose, we first extract Harris corner features from each frame. By establishing feature correspondences between frames, we estimate the homography using RANSAC. Once the homography between consecutive frames is estimated, it can be used for calculating both the background flow vector per pixel $\mathbf{m}_b(x, y)$ and as a prior to a block-based optical flow algorithm for computing the overall flow $\mathbf{m}_o(x, y)$. Then, the foreground flow at each pixel (x,y) is calculated as

$$\mathbf{m}_f(x, y) = (\mathbf{m}_o(x, y) - \mathbf{m}_b(x, y)). \tag{1}$$

The noisy motion flow fields are mostly stabilized by this procedure. Example resultant flows can be seen in Fig 2.

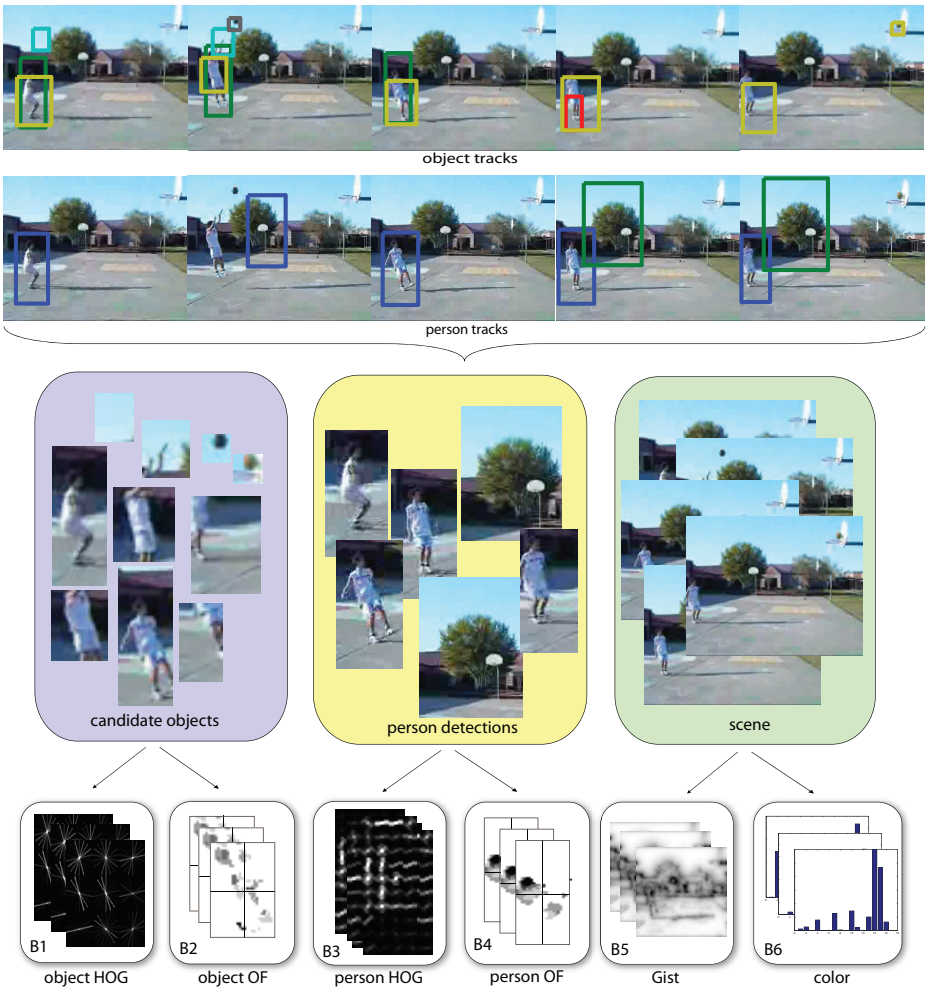


Fig. 1. There are three main feature channels, namely person-centric, object-centric and scene features. Once the videos are stabilized, we extract candidate person and object tracks. An example track from a basketball sequence is shown above (for details of track extraction, see the text). From each track, we extract multiple features. Each of the feature channels may contain noisy detections, as well as the true detections. There can be multiple people and multiple objects within the video. Since we do not have explicit supervision on which feature or detection region may be relevant, each feature channel is defined as a MIL bag. We then combine these feature channels using two different approaches.

Fig. 2. The videos include an extensive amount of camera motion, thus, uneven flow fields. We use a homography based approach to estimate the flow of the foreground objects, following [21]. (a) shows the original frames from four videos. (b) shows the flow estimate (in green) without the motion stabilization. (c) shows the foreground flow estimate obtained following stabilization. As seen, this estimate gives us the ability to concentrate on the moving foreground objects.

3.2 Person-Centric Features

To extract person-centric features, first, one should have a rough estimate of the location(s) of the person(s) and corresponding person tracks in the video. This is tough, especially when the background clutter is dominant.

We approach the problem by using a "tracking-by-detection" method and use Felzenswalb et al.'s human detector [9]. This person detector has shown to perform quite well in detecting people of various poses; however, due to motion blur and pose variations, it is not able to locate the person in every frame. In order to compensate for this, we use mean-shift tracking [6] to fill the gaps in which the person detector did not fire, by using the person detection bounding box to initiate the tracker in each case. We initiate a separate track for each individual person detection and discard the short tracks (with ≤ 5 frames) as being noise.

Figure 1 shows some example tracks. Here, we define a track as the series of bounding boxes associated with the detected regions over the video. While the final tracks are not perfect and some of the tracks may still be irrelevant, they provide fairly usable person localizations. From each detected track, we extract two types of features: person-centric motion and shape features.

Person-centric motion features: Optical flow has been shown to be a useful feature for describing human actions [8]. We use the intersection of the estimated foreground flow (computed by stabilizing the video as in Sec.3.1) and the person tracks to locate the regions of the optical flow map that belong to a person. We describe the flow in each detection bounding box with a spatial histogram, by dividing the optical flow field equally into 2×2 spatial regions, and represent each spatial bin with four major flow orientations. In order to accommodate for the noise in the optical flow, we use a windowing scheme over the tracks and extract histograms from every snippet of six frames.

Each subwindow is considered as an instance in the MIL setting (see Section 4). The final descriptor of each instance has $4 \times 4 \times 6 = 96$ feature dimensions.

Person-centric shape features: We expect the shape feature to be complimentary to motion features, especially when the motion field of the video has excessive noise. In order to account for this shape information, we use the Histogram of Oriented Gradients (HOG) [7]. We downsize each bounding box region to $[64 \times 32]$ pixels and extract HOGs using 8 pixel cell size and 8 pixel cell step. We use eight orientation bins and use a temporal window of five frames over the tracks to accumulate the temporal pattern. The final descriptor has $8 \times 8 \times 4 \times 5 = 1280$ feature dimensions.

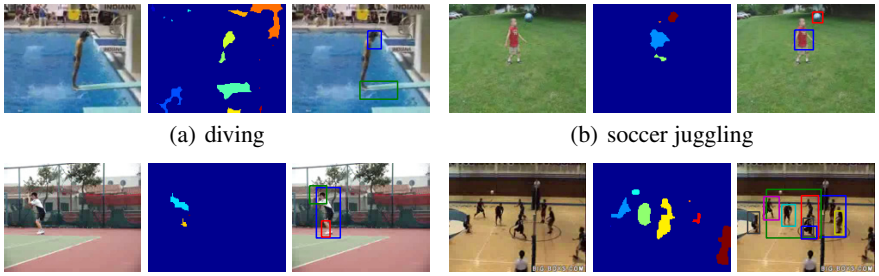
3.3 Object-Centric Features

Actions may involve certain objects. For example, for a throwing action, the presence of a ball and/or the shape of its trajectory are important discriminative cues. With this intuition, we find candidate object regions and extract object-centric features from these regions. We do not use any explicit object detector for this purpose. We consider any moving region that has sufficient temporal and spatial coherence as a "*candidate object*". Once a candidate object region is detected, we try to find the associated tracks and the corresponding features from each track.

Extracting candidate object tracks: We extract candidate object regions as follows: First, we estimate the foreground motion as described in Sec 3.1. Then, we find the consistent regions among the estimated foreground motion fields. We do this by looking at temporal, spatial and appearance consistency of the detections in sequential frames. More formally, given a video frame and its estimated foreground flow, we first find the connected components of the flow field, yielding possible object regions. We follow an agglomerative clustering approach to group each of these regions based on their appearance and spatial coherence within the video. In this agglomerative clustering, the similarity between regions is computed by using the χ^2 distance of color histograms and Euclidean distance of their midpoint coordinates. During this clustering, we allow a temporal gap of up to 10 frames. In the end, the small clusters (i.e., less than 5 frames) are considered to be noise and discarded. This clustering serves as an initial preprocessing step to remove noisy and discontinuous detections.

After this initial step, we form tracks for each remaining region. We follow a greedy approach for generating tracks: we start from the region with the largest area and we track that region using mean-shift tracking [6], forward and backward in time. For the remaining regions, we check if there is already a track that overlaps with that region to a certain degree (30%). The degree of overlap is calculated as the ratio of the intersection over the union of the corresponding bounding boxes. If there is no previous overlapping track, we create a new track. Example outputs of this procedure are shown in Fig. 3.

Object-centric motion features: Once the candidate object tracks are found, we extract the motion features from each track. We describe the flow region in each detection bounding box by dividing it into 2×2 spatial partitions, and representing each spatial partition with histograms of the four major flow orientations. We extract these motion histograms over a snippet of 5 frames from each track. We apply a windowing scheme



(a) diving (b) soccer juggling

(c) tennis swing (d) volleyball

Fig. 3. Objects are extracted by grouping the optical flow regions and tracking them over time. Short tracks are eliminated, so that we are left with object regions that have a considerable amount of motion throughout the sequence.

over each object track to extract all the instances that will be input to the MIL algorithm. The final descriptor has $4 \times 4 \times 5 = 80$ feature dimensions.

Object-centric shape features: This feature channel includes shape information for the moving objects in the sequence. With this feature, we aim to capture the immediate context information for the action by defining the shape of nearby objects, such as bicycles, horses or smaller objects, like rackets, balls, etc. We define the shape of object regions by using HOGs. Since the size and scale of the object bounding boxes are not constant and we cannot define a single width/height ratio, we extract HOGs from the spatial grids. That is, instead of using a regular cell and block size, we assume that the object region is divided into an equal number of spatial blocks. We use 3×3 spatial bins and represent each spatial bin with nine gradient orientations. We then normalize each descriptor with respect to the object size. The final descriptor size for each object bounding box has $9 \times 9 = 81$ feature dimensions.

3.4 Scene Features

Apart from the person and object related features, the overall properties of the scene can give us related contextual information about the action taking place. For example, if we see a basketball hoop and a court, the probability of observing people playing basketball is higher. In order to exploit these properties, we extract shape and color features.

Scene shape features: To describe the scene structure, we extracted Gist [28] features from five frames selected randomly from each video. We use the original parameter settings provided in [28] and the final Gist descriptor has 512 feature dimensions.

Scene color features: Color features can be complementary to the shape information for the scene. For example, the presence of a "blue rectangular region" (i.e., a swimming pool) may be helpful in identifying the "diving" action. We extract color features respecting the coarser spatial layout of the scene. For this purpose, we divide each scene horizontally into three equal regions and extract color histograms from each region. For the color histogram, we discretize the RGB colorspace into 16 bins. The final descriptor has $16 \times 3 = 48$ dimensions. We do this for three randomly selected frames.

4 Combining Features - A Multiple MIL Approach

In our problem, we are given a set of videos with labels that tell us the presence of an action class in each video. However, we do not know the exact spatio-temporal location of the specified action in each video, nor do we know what related objects or scene information will contribute to the identification of that class. There may be many object and/or person tracks extracted from each video. Some of these tracks may be relevant to the action, e.g., the track of a basketball or a jumping person, whereas some of the tracks may be irrelevant or caused by noise.

This scenario suggests the particular suitability of "multiple instance learning"(MIL) [2]. In MIL, the given class label is associated with bags (rather than instances as in the case of fully supervised learning), where each bag consists of one or more instances. A bag is labelled as positive if at least one instance x_{ij} in the bag is known to be positive. A bag is labelled as negative if all the instances in that bag are known to be negative. Individual labels of the instances are unknown. Since the labels are given to bags rather than instances, the learning procedure operates over the bags.

In our case, a bag contains all the instances extracted from a video sequence for a particular feature channel. For example, for the Gist feature, the bag would contain five instances, one Gist feature vector per each of the randomly selected frames from the video. For the person-centric motion feature, there would be several feature vectors x_{ij} extracted by employing the windowing procedure over each detected person track and each of these feature vectors is considered to be an instance inside the bag of the corresponding feature channel.

Formally, for each video i, we have one bag $\mathbf{B}_i^f$ per feature channel $f \in \{1, \ldots, F\}$. Each $\mathbf{B}_i^f$ contains multiple instances x_{ij} such that $\mathbf{B}_i^f = \{x_{ij}^f \ : \ j = \{1, \ldots n_i^f\}\}$. Here, n_i^f is the number of instances of that feature type in video i. Each bag has an associated label $Y_i \in A$, where $A = \{a_1, ..., a_M\}$ is the possible set of M actions .

In order to represent these bags in the MIL framework, we first embed the original feature space x, to the instance domain $\mathbf{m}(B)$, via the instance embedding framework of [5]. In [5], each bag is represented by its similarity to each of the instances in the dataset. In our case, this is infeasible, given the large size of the dataset and number of instances per bag. Therefore, we cluster the data using k-means to find potential target concept instances $c_l^f \in C^f$. We do this for each action class separately, setting k to a constant value (we use $k = 50$). The total size of C^f becomes $N = k \times M$ for each feature channel. The similarity between bag $\mathbf{B}_i$ and concept c_l^f is defined as

$$s(c_l^f, \mathbf{B}_i^f) = \max_j \exp\left(-\frac{D(x_{ij}, c_l^f)}{\sigma}\right), \tag{2}$$

where $D(x_{ij}, c_l^f)$ measures the distance between a concept instance c_l^f and a bag instance x_{ij}. In our case, since all the features are histogram-based, we can use the χ^2 distance $D(x_{ij}, c_l^f) = \chi^2(x_{ij}, c_l^f) = \frac{1}{2}\sum_d \frac{(x_{ij}(d) - c_l^f(d))^2}{x_{ij}(d) + c_l^f(d)}$, where d is a feature dimension of the instance feature vector. For the bandwidth parameter σ, we use the standard deviation of each feature embedding.

Each bag can then be represented in terms of its similarities to each of these target concepts and this mapped representation $\mathbf{m}(B_i^f)$ can be written as

$$\mathbf{m}(B_i^f) = [s(c_1^f, B_i), s(c_2^f, B_i), \ldots, s(c_N^f, B_i)]^T. \tag{3}$$

We convert the instances from each feature channel to their MIL representation separately. Subsequently, we need a way to combine these different feature channels. For this purpose, we propose two combination techniques. The first technique concatenates all feature channels and treats the problem as a classification problem over the joint set of features. More formally, in our first method, we represent each bag B_i with its concatenated embeddings over F feature channels, such that

$$\hat{\mathbf{m}}(B_i) = [\mathbf{m}(B_i^1)\mathbf{m}(B_i^2)\ldots\mathbf{m}(B_i^F)]. \tag{4}$$

We use an L2-regularized linear SVM for the classification over these concatenated bag representations $\hat{\mathbf{m}}(B)$. In this way, the positivity constraint of the MIL framework is extended over multiple feature channels. If a B_i^f is empty for a particular f, we simply assign the corresponding $\mathbf{m}(B_i^f)$ to zero.

In the above formulation, each feature channel is treated equally. However, there may be certain cases where a particular feature channel is more informative than the other feature channels for a specific action . Likewise, some of the feature channels may contain redundant information for specific actions. In this case, we may be interested in learning global weights for individual feature channels.

This observation motivates the formulation of our second method, which employs a joint formulation for learning the global weights for feature channels. This global weighting is analogous to learning the kernel weights in multiple kernel learning (MKL) [11]. In MKL, the task is to select informative kernels, whereas here we try to select informative feature channels. We formulate the optimization as follows:

$$\begin{aligned} \min_{w,\alpha,b} \quad & \sum_f (w^f)^T w^f + \beta \alpha^T \alpha + \gamma \sum_i L\left(y_i, \sum_f \alpha_f (w^f)^T m^f + b\right), \\ \text{s.t.} \quad & \alpha \geq 0 \end{aligned} \tag{5}$$

In this formulation w^f is the weight vector for individual features in fth feature channel, α_f is the global weight of the whole feature channel, L is the loss function (we use Hinge loss). β and γ are the regularization parameters and b is the bias term. Here, each α_f defines a global combination weight for the instances of feature channel f. The first term in this objective function in Eq. 5 stands for the regularization of the individual feature weights on each channel, whereas the second term corresponds to the regularization on global feature channel weights. If a feature channel tends to be very noisy, this global weighting scheme can help in deemphasizing that feature channel completely by assigning the corresponding α_f to a small value.

The objective function in Eq. 5 becomes convex when the α or w vector is fixed. Therefore, we follow an iterative alternating optimization approach in the primal space, which is a coordinate descent method. In this iterative approach, we first fix α and solve for w and b, such that

$$\min_{w,b} \sum_f (w^f)^T w^f + C \sum_i L\left(y_i, \sum_f \sum_{z \in G_f} (w_z^f)^T (\alpha_f m_z^f) + b\right). \quad (6)$$

Here, G_f represents the group of features for feature type f. Once Eq. 6 is optimized with respect to w and b, we then fix the w and b, and optimize α such that,

$$\min_{\alpha} \beta \alpha^T \alpha + C \sum_i L\left(y_i, \sum_f \alpha_f ((w^f)^T m^f) + b\right). \quad (7)$$

Note that in this formulation, both steps minimize the same objective, so convergence is guaranteed. In our experiments, we observe that convergence to a local minimum is achieved in ≈ 10 iterations.

5 Experiments

In order to test our approach, we use the YouTube dataset collected by Liu et al [22]. This is a very large dataset that consists of 1168 videos in total. This is a particularly suitable dataset for studying the effects of object and scene properties of actions, since there are actions involving specific objects (like basketball) and scenes (like diving). The dataset contains videos of 11 actions; these are basketball shooting, biking/cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. It is a quite challenging dataset with lots of camera movement, cluttered backgrounds, different viewing directions and varying illumination conditions. Videos for each category of action are divided into 25 related subsets, and leave-one-out cross validation is applied over these subsets, following the same evaluation methodology of [22].

5.1 Evaluation

Table 1 summarizes the overall quantitative results. The classification results here are normalized with respect to the number of videos for each action type.

We first evaluate the performance of the individual feature channels. The first six rows of Table 1 include the individual classification accuracies for each of the feature channels represented in the embedded MIL domain. Note that, in our setup, the object tracks are allowed to overlap with or include person tracks, so the object tracks may sometimes include person track information as well. The results show that, even using the single feature channel Gist gives 53.20% average classification accuracy in this dataset, whereas the simple color histogram feature is able to perform with 49.28% average accuracy. These numbers are noticably high, compared to the chance level in this dataset, which is 9.09%. This observation shows that using scene features can provide a great deal of useful information about the possible action, especially in this dataset. For example, for the diving action, the simple color features achieve 86% accuracy, whereas for volleyball spiking, the Gist features give 81%. These results suggest that when the person is less visible, the scene features can be used for reducing the set of possible actions considered. On the other hand, where the person is more visible (e.g. videos of

Table 1. Overall performance evaluation of individual feature channels and their combination. perOF/objOF: person/object optical flow features, perHOG/objHOG: person/object HOG features. p+s: person and scene features, p+o : person and object features and o+s: object and scene feature combinations. The best results for each action class are shown in bold. We see that action recognition benefits from both object and scene features in most of the action types.

% correct classification using single feature channels												
	b_shoot	bike	dive	golf	h_ride	s_ juggle	swing	t_swing	t_jump	v_spike	walk	**Avg**
perOF	20.20	44.83	51.0	69.0	45.0	44.0	36. 0	32. 0	64. 0	29. 0	29.27	42.72
perHOG	28.28	57.93	56.0	40.0	51.0	36.0	43. 0	45. 0	34. 0	49. 0	39.84	43.64
objOF	14.14	45.52	24.0	36.0	51.0	20.0	42. 0	14. 0	59. 0	25. 0	33.33	33.09
objHOG	21.21	44.14	62. 0	55. 0	38. 0	22. 0	42. 0	44. 0	42. 0	45. 0	21.95	39.75
gist	38.38	60.69	69. 0	61. 0	66. 0	9. 0	42. 0	61. 0	54. 0	81. 0	43.09	53.20
color	33.33	44.83	86. 0	65. 0	43. 0	22. 0	27. 0	47. 0	57. 0	73. 0	43.90	49.28
% correct classification using combinations of channels												
p+s	44.44	70.34	92. 0	87. 0	63. 0	35. 0	56. 0	75. 0	84. 0	84. 0	56.91	67.97
p+o	40.40	70.34	84. 0	91. 0	63. 0	**54. 0**	63. 0	60. 0	84. 0	78. 0	50.41	67.11
o+s	47.47	73.79	91. 0	90. 0	73. 0	35. 0	64. 0	75. 0	83. 0	**89. 0**	56.10	70.67
% correct classification using all feature channels												
p+o+s	48.48	**75.17**	95. 0	**95. 0**	**73. 0**	53. 0	**66. 0**	77. 0	**93. 0**	85. 0	66.67	**75.21**
w[p+o+s]	43.43	**75.17**	**96. 0**	94. 0	72. 0	47. 0	65. 0	74. 0	**93. 0**	85. 0	67.48	73.83
Liu [22]	**53.0**	73.0	81.0	86.0	72.0	**54.0**	57.0	**80.0**	79.0	73.3	**75.0**	71.2

trampoline jumping, golf, juggling actions), the optical flow of the person detections seems to be the most informative feature.

Second, we look at the joint performance of these feature channels. In Table 1, "p+o" refers to combination of person-centric and object-centric features together, i.e. the first four feature channels, ignoring the scene dimension. The rows "p+s" and "o+s" correspond to combination of the "person and scene" and the "object and scene" feature channels, respectively. Looking at these feature combinations, we see that, in most of the cases, using them in combination improves classification accuracy significantly. For example, for the trampoline jumping action, the maximum response from the individual feature channels is 64% for person optical flow features, whereas, accuracy increases to 84% if person features are considered in combination with object or scene features.

Third, we look at the overall combination results. As described in Sec 4, we have combined all feature channels using two different methods; the first method uses L2-regularized linear SVM over the concatenated embedded feature channels (represented as p+o+s in Table 1), the second method learns global combination weights over each feature bag (w[p+o+s]). In 9 out of 11 actions, using all the features together yields higher classification accuracy. These results demonstrate that all feature channels are informative and complementary to each other, and that each of them introduces some amount of useful information for the identification of the actions in this dataset. We see that both of the proposed combination techniques introduce an improvement over the best reported results in this dataset [22], while the average improvement is higher without the global weights (4% and 2.6% respectively). We believe that this difference is due to the the high amount of noise in some of the feature channels. The excessive noise may cause the weighting scheme to underestimate the exact weights of that feature channel during training.

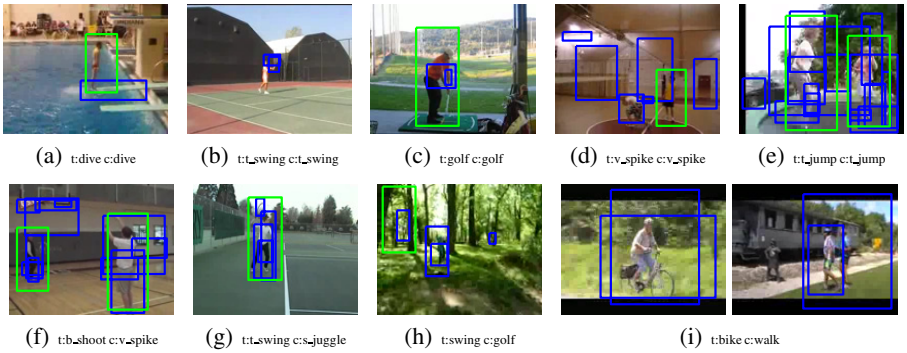


Fig. 4. Example classification results. The person regions are shown in green and the candidate objects are shown in blue. The subcaptions shows the true class label and output classification label, respectively. See text for details.

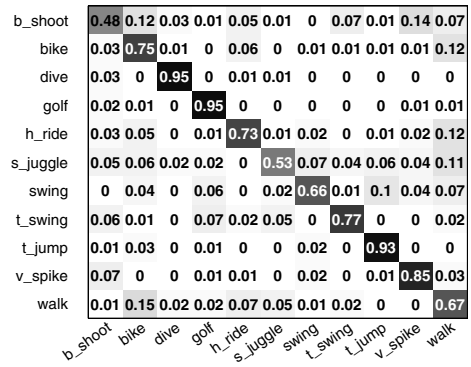


Fig. 5. Overall confusion matrix of [p+o+s]. The average accuracy is 75.21%.

Example classification results are shown in Fig. 4. The first row shows the correct classifications. For the diving action in Fig. 4(a), both the person and the related object (diving board) are detected and their features are complementary to the scene features. In Fig. 4(b), the person detection has failed, but the tennis racket is found and helps the identification of the tennis swing action. In Fig 4(e), the candidate objects are noisy, but the person tracks seem reliable. Example misclassifications are shown in the second row of Fig. 4. The failure cases are mostly caused by the multiple people, noisy detections, and/or multiple actions. For instance, in Fig. 4(f) there are multiple people and many candidate objects. In Fig. 4(i), although there is a biking person in the first half of the video, in the second part there is a walking detection.

Figure 5 shows the confusion matrix of our approach. Most of the confusion occurs between walking and biking actions. This is mostly due to the higher frequency of close-up recording in these actions. When there is extensive close-up in the video, the motion stabilization procedure fails to estimate the homography of the scene correctly, because of the increased ratio of the moving regions. Basketball shooting and volleyball

actions are also confused in some cases; this is largely because most of the time, the basketball and volleyball sports use very similar courts.

In a typical video, our moving region grouping procedure results in 20-30 object tracks on average in the YouTube dataset. While the complete annotation of these tracks is infeasible, we estimate that approximately five tracks on average are relevant in each video. The experiments indicate that our framework can succeed, even under such challenging conditions.

6 Conclusion

In this paper, we present an approach for combining features of the people, objects and scene for better recognition of actions. The videos available for training our approach are only weakly annotated; we do not know where or when in the video the action occurs, nor do we know which objects or scene features will contribute to the identification of that action. To discover these automatically during training, we use a MIL-based framework.

Our results show that, scene and object properties can indeed be used as complementary to person features for the correct identification of actions. This is especially true when the person is seen from a far distance and the distinct features of the human body are not fully visible. In that case, the moving regions nearby or the overall scene gist can give an idea about what the person/people is up to in that scene.

Action recognition in YouTube videos is an especially good application domain for our method. The low resolution and the unstable camera conditions can make a single feature channel unreliable on its own. In that case, the recognition of actions is likely to benefit from multiple feature channels, as we demonstrate in this work.

We use three main types of features and ignore the temporal and spatial relationships of these features. Our framework can be extended to handle more feature channels, like space-time interest points [29] and can benefit from more complex video object segmentation methods like [15]. Future work includes exploring these techniques and more feature channels, together with their spatio-temporal relationships.

Acknowledgments. This material is based upon work supported in part by the U.S. National Science Foundation under Grant No. 0713168.

References

1. Ali, S., Shah, M.: Human action recognition in videos using kinematic features and multiple instance learning. IEEE TPAMI 32(2) (2010)
2. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: NIPS, pp. 561–568. MIT Press, Cambridge (2003)
3. Babenko, B., Yang, M.-H., Belongie, S.: Visual Tracking with Online Multiple Instance Learning. In: CVPR (2009)
4. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV (2005)
5. Chen, Y., Bi, J., Wang, J.Z.: Miles: Multiple-instance learning via embedded instance selection. IEEE TPAMI 28(12), 1931–1947 (2006)
6. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE TPAMI 25(5), 564–575 (2003)

7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. 886–893 (2005)
8. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: ICCV '03, pp. 726–733 (2003)
9. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
10. Forsyth, D.A., Arikan, O., Ikemoto, L., O'Brien, J., Ramanan, D.: Computational studies of human motion: part 1, tracking and motion synthesis. Found. Trends. Comput. Graph. Vis. 1(2-3), 77–254 (2005)
11. Gehler, P.V., Nowozin, S.: On feature combination for multiclass object classification. In: ICCV (2009)
12. Gupta, A., Davis, L.S.: Objects in action:an approach for combining action understanding and object perception. In: CVPR (2007)
13. Han, D., Bo, L., Sminchisescu, C.: Selection and context for action recognition. In: ICCV (2009)
14. Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., Huang, T.S.: Action detection in complex scenes with spatial and temporal ambiguities. In: ICCV (2009)
15. Huang, Y., Liu, Q., Metaxas, D.N.: Video object segmentation by hypergraph cut. In: CVPR (2009)
16. Ikizler, N., Forsyth, D.: Searching for complex human activities with no visual examples. IJCV 80(3) (2008)
17. Ikizler-Cinbis, N., Cinbis, R.G., Sclaroff, S.: Learning actions from the web. In: ICCV (2009)
18. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: ICCV (2007)
19. Ke, Y., Sukthankar, R., Hebert, M.: Event detection in crowded videos. In: ICCV (2007)
20. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR (2008)
21. Liu, F., Gleicher, M.: Learning color and locality cues for moving object detection and segmentation. In: CVPR (2009)
22. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos "in the wild". In: CVPR (2009)
23. Maron, O., Ratan, A.L.: Multiple-instance learning for natural scene classification. In: ICML, pp. 341–349 (1998)
24. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR (2009)
25. Mikolajczyk, K., Uemura, H.: Action recognition with motion-appearance vocabulary forest. In: CVPR (2008)
26. Moore, D.J., Essa, I., Hayes, M.H.: Exploiting human actions and object context for recognition tasks. In: ICCV (1999)
27. Niebles, J.C., Han, B., Ferencz, A., Fei-Fei, L.: Extracting moving people from internet videos. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 527–540. Springer, Heidelberg (2008)
28. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV 42(3), 142–175 (2001)
29. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: ICPR (2004)
30. Tran, D., Sorokin, A.: Human activity recognition with metric learning. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 548–561. Springer, Heidelberg (2008)
31. Vijayanarasimhan, S., Grauman, K.: Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization. In: CVPR (2008)