# Action Recognition with Exemplar Based 2.5D Graph Matching

Bangpeng Yao Li Fei-Fei

Department of Computer Science, Stanford University
{bangpeng,feifeili}@cs.stanford.edu

**Abstract.** This paper deals with recognizing human actions in still images. We make two key contributions. (1) We propose a novel, 2.5D representation of action images that considers both view-independent pose information and rich appearance information. A 2.5D graph of an action image consists of a set of nodes that are key-points of the human body, as well as a set of edges that are spatial relationships between the nodes. Each key-point is represented by view-independent 3D positions and local 2D appearance features. The similarity between two action images can then be measured by matching their corresponding 2.5D graphs. (2) We use an exemplar based action classification approach, where a set of representative images are selected for each action class. The selected images cover large within-action variations and carry discriminative information compared with the other classes. This exemplar based representation of action classes further makes our approach robust to pose variations and occlusions. We test our method on two publicly available datasets and show that it achieves very promising performance.

## 1 Introduction

Humans can effortlessly recognize many human actions from still images, such as "playing violin" and "riding a bike". In recent years, much effort has been made in computer vision [1–8] with the goal of making this process automatic. Automatic recognition of human actions in still images has many potential applications, such as image search and personal album management.

Considering the close relationship between actions and human poses, in this paper, we aim to develop a robust action recognition approach by modeling human poses. The idea of using human poses for action recognition has been studied in some previous work which either detect local pose features [4, 6] or model the spatial configuration between human body parts and objects [2, 3, 7, 8]. However, while such approaches sound promising, the winning method [9] in the recent PASCAL challenge [10] simply treats action recognition as an image classification problem, without explicitly modeling human poses.

The challenges in modeling human poses for action recognition are illustrated in Fig.2. On the one hand, because of the variations of camera angles, the same human pose can correspond to very different body parts configurations on the 2D

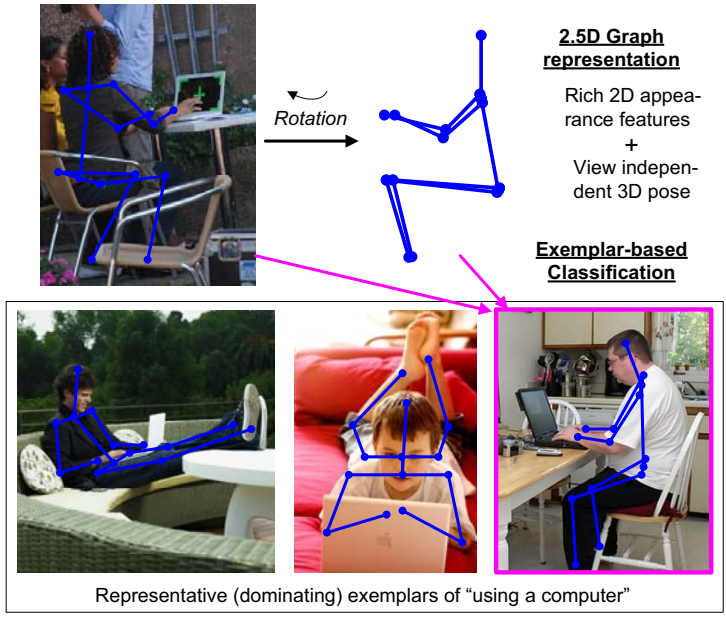



**Fig. 1.** An overview of our action recognition algorithm. We represent an action image as a 2.5D graph consisting of view-independent 3D pose and 2D appearance features. In recognition, the 2.5D graph is matched with a set of exemplar graphs for each action class, allowing more robust handling of within-action variations.

image plane, which poses challenges in reliable measurement of action similarities. On the other hand, human poses in the same action can change drastically, while very similar human poses might correspond to many different human actions, and therefore it is difficult to build a single pose model to distinguish one action from all the others.

In this paper, we propose a novel action recognition approach (Fig.1) to address the above two challenges. Specifically, we make two key contributions:

- **2.5D graph for action image representation.** We propose a 2.5D graph representation for action images. The nodes of the graph are key-points of the human body represented by view-independent 3D positions and rich 2D appearance features. The edges are relative distances between the key-points. Estimating the similarity between two action images then becomes matching their corresponding graphs.
- **Exemplar-based action classification.** Considering that a single pose model is not enough to distinguish one action from all the others, we propose an exemplar based approach for action classification. For each action class, we select a minimum set of "dominating images" that are able to cover all within-class pose variations and capture all between-class distinctions.

The rest of this paper is organized as follows. Related work is discussed in Sec.2. The 2.5D graph representation of action images and exemplar-based

(a) (b)

**Fig. 2. (a)** The same action might contain very large pose variations. **(b)** Due to different camera angles, even the same human pose looks differently in 2D images.

action recognition algorithm are elaborated in Sec.3 and Sec.4, respectively. Experiments are represented in Sec.5. We conclude our paper in Sec.6.

## 2 Related Work

Human poses have been used for action recognition in existing literatures. Both global silhouette [2] and local pose units [4, 6] have been adopted for distinguishing different human actions. In [3, 7], action recognition is treated as a human-object interaction problem, where spatial relationships between different body parts and objects are modeled. The interactions are also represented as a set of bases of action attribute-object-pose in [5]. While most of such approaches rely on annotations of human poses, a weakly-supervised method was proposed to model human-object interactions in [8]. All those methods, however, model human poses in 2D only, and therefore are difficult to deal with the within-class pose variations caused by camera angle changes, as shown in Fig.2(b).

There has been some work for view-independent action recognition, mostly dealing with videos. [11] renders Mocap data from multiple viewpoints, which is time and storage consuming. [12] projects 2D features to a 3D visual hull. Manifold based warping methods are adopted in [13]. View-invariant feature descriptors have also been proposed [14, 15]. Most of such methods rely on temporal information, and therefore are not suitable to our problem.

In this work, we aim at view-independent action recognition from single images. We extract key-points of the human body [16] and then convert the 2D key-points to 3D positions without any supervision [17, 18]. The 3D positions of key-points allow us to rotate human skeletons from different views to the same view-point (Fig.1), hence making view-independent matching possible. Inspired by [19], where it shows that the combined pose and appearance features help improve action recognition performance, our 2.5D action graph is constructed by combining the view-independent 3D human skeletons and 2D appearance features [20, 21]. 2.5D graph representations have been used in computer vision systems before [22–24]. While most of these papers focus on modeling scene layers or rigid objects such as human faces, our method is designed for recognizing articulated objects such as human bodies.

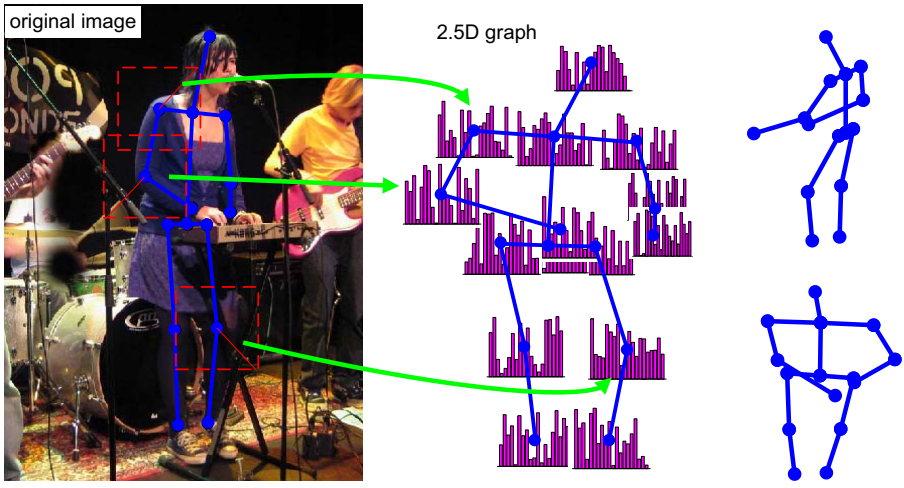



**Fig. 3. (a)** Illustration of an image and its corresponding 2.5D action graph. The histograms represent appearance features extracted from the corresponding image regions. **(b)** The human body skeleton from the other views.

While the majority of work in computer vision are model based, exemplar based methods have also been applied in object recognition [25–27] and video classification [28]. Different from most previous work where all training samples are treated as candidate exemplars, our method aims at selecting a compact set of images for each action class that are able to cover the within-class pose variation and capture all between-class distinctions. We show that the problem is essentially a minimum dominating set problem [29], and can be solved by using an improved reverse heuristic algorithm [30].

## 3 A 2.5D Graph of Human Poses and Appearances

### 3.1 The 2.5D Graph Representation

The term, *2.5D graph*, is borrowed from stereoscopic vision [31]. It refers to the outcome of reconstructing 3D information from 2D but the appearance cues are still 2D. A graphical illustration of our 2.5D representation of action images are shown in Fig.3. It combines view-independent 3D configuration of human skeletons and 2D appearance features.

A 2.5D graph $\mathcal{G}^{\mathcal{I}}$ representing an action image $\mathcal{I}$ consists of $V$ nodes connected by $E$ edges. The nodes correspond to a set of key points of the human body, as shown in Fig.3. A node $v$ is represented by the 3D position of this node $\mathbf{l}_v^{\mathcal{I}}$ and 2D appearance features $\mathbf{f}_v^{\mathcal{I}}$ extracted in a local image region surrounding this point. An edge $e$ is a three-dimensional vector $\Delta\mathbf{l}_e^{\mathcal{I}} = \mathbf{l}_v^{\mathcal{I}} - \mathbf{l}_{v'}^{\mathcal{I}}$, where node $v$ and node $v'$ are connected by $e$. Note that our model allows the human body to

rotate in 3D (as shown in Fig.3(b)), which will result in different 3D positions of key-points and hence edge vectors. Also, because some key points might be outside of the boundary of the image, we introduce an auxiliary variable $h_v^{\mathcal{I}}$ for each $v$, and a $h_e^{\mathcal{I}}$ for each $e$. $h_v^{\mathcal{I}} = 1$ if key-point $v$ is within the boundary of image $\mathcal{I}$, otherwise $h_v^{\mathcal{I}} = 0$. Similarly, $h_e^{\mathcal{I}} = 1$ if and only if both two points connected by $e$ are within the image boundary.

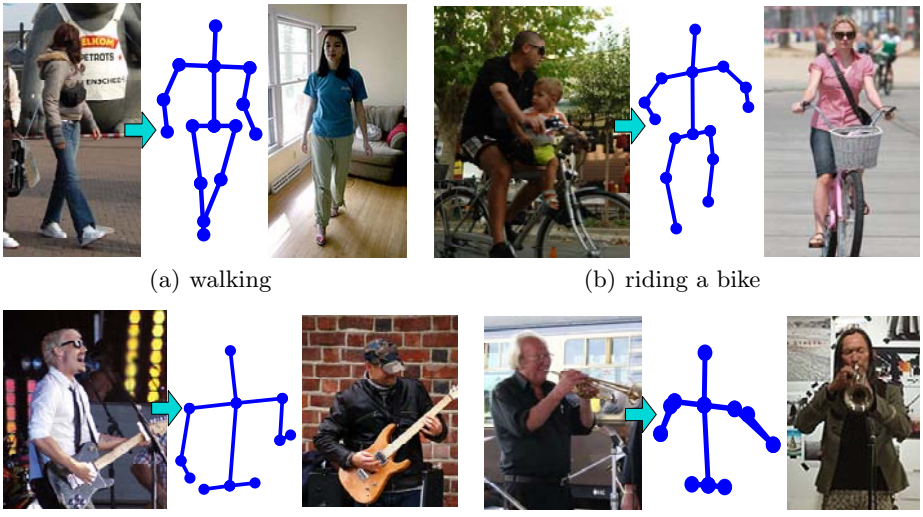
**Implementation details.** We consider 15 key-points of human bodies: top head, left-middle-right shoulders and hips, left-right elbows, wrists, knees, and ankles. Given an image, the 3D position of these points are obtained by first using pictorial structure [16] to estimate their positions in 2D, and then using the method in [17] with additional constraints [32] to recover the depth information. The key-point locations are then normalized such that the center of the torso is at $(0, 0, 0)$, and the height of the torso (distance between middle shoulder and middle hip) is 100 pixels. Although human pose estimation itself is challenging and the 3D points we obtain are not perfect, our approach can still achieve very good action recognition performance, even comparing with the setting that uses ground-truth key-point locations. We will show this in Sec.5.

The detailed process of using pictorial structure to estimate 2D key-points locations is as follows. Following the standard settings in [10], we assume that there is a bounding box surrounding each person whose action is to be recognized. As in [33], the image is normalized by extending the bounding box to contain $1.5\times$ the original size of the bounding box, and cropping and resizing it such that the large image dimension is 300 pixels. To deal with the situation that the legs are outside of the image boundary, we train a full human detector and an upper body detector excluding the key-points below hips. Given a normalized image, if the calibrated response score obtained from the full body detector is larger than 0.8 times of the score obtained from the upper body detector, we regard that the full human body is visible, otherwise upper body only. Because of the provided bounding boxes of the humans, the detection results are very reliable in almost all the images. Based on whether full body or only upper body is visible, we use the appropriate pictorial structure [16] model to estimate the location of the key points, considering or ignoring the key-points below hips. In our experiments (Sec.5), we re-train a pictorial structure model on each dataset, where the body part detectors are obtained using the deformable part models [34].

The appearance feature $\mathbf{f}_v^{\mathcal{I}}$ is a two-level spatial pyramid [21] of SIFT [20] features with locality-constrained linear coding [35] in a $60 \times 60$ image region centered at point $v$ of image $\mathcal{I}$. We consider two image sizes, one is the normalized image of which the larger dimension is 300 pixels, the other is the image where the length of the torso is 100 pixels. We use a 512 codebook size for SIFT features, and therefore the dimensionality for $\mathbf{f}_v^{\mathcal{I}}$ is 2560. If the point $i$ is outside of the image boundary, then all values of $\mathbf{f}_v^{\mathcal{I}}$ are set to 0.

### 3.2 Measuring Similarity of 2.5D Graphs

To use the 2.5D graph constructed in Sec.3.1 for action recognition (details in Sec.4), we need to match a graph $\mathcal{G}^{\mathcal{I}}$ to a "template graph" $\mathcal{G}^{\mathcal{M}}$ and compute



(a) walking (b) riding a bike

(c) playing guitar (d) playing trumpet

**Fig. 4.** The 3D representation of human body key-points allows us to rotate one image to the same view-point of the other image, and thus achieve view-independent similarity matching. In each subfigure, from left to right: human in profile view, its pose in frontal view, and the other human with the same action in the frontal view.

their similarity. As described in Sec.3.1, the graph $\mathcal{G}^{\mathcal{I}}$ is denoted by $\{\mathbf{f}_v^{\mathcal{I}}, h_v^{\mathcal{I}}, v = 1, \cdots, V; \Delta\mathbf{l}_e^{\mathcal{I}}, h_e^{\mathcal{I}}, e = 1, \cdots, E\}$. The template graph $\mathcal{G}^{\mathcal{M}}$ is denoted as $\{\mathbf{f}_v^{\mathcal{M}}, h_v^{\mathcal{M}}, \mathbf{w}_v^{\mathcal{M}}, v = 1, \cdots, V; \Delta\mathbf{l}_e^{\mathcal{M}}, h_e^{\mathcal{M}}, \mathbf{w}_e^{\mathcal{M}}, e = 1, \cdots, E\}$, where $\mathbf{w}_v^{\mathcal{M}}$ and $\mathbf{w}_e^{\mathcal{M}}$ are the feature weights for the corresponding node and edge. How to obtain the weights will be described in Sec.4.

When matching the similarity between $\mathcal{G}^{\mathcal{I}}$ and $\mathcal{G}^{\mathcal{M}}$, we deal with the 2D appearance features (nodes) and 3D pose features (edges) separately. The similarity between the appearance features if node $v$ is simply the weighted histogram intersection between $\mathbf{f}_v^{\mathcal{I}}$ and $\mathbf{f}_v^{\mathcal{M}}$, denoted as $\mathbf{w}_v^{\mathcal{M}} \cdot I\left(\mathbf{f}_v^{\mathcal{I}}, \mathbf{f}_v^{\mathcal{M}}\right)$. For the pose features, as shown in Fig.4, the 3D representation allows us to rotate the 3D key-point locations $\{\mathbf{l}_v^{\mathcal{I}}\}_{v=1}^V$ to the same view-point of $\{\mathbf{l}_v^{\mathcal{M}}\}_{v=1}^V$, and then match the view-independent similarity score.

Let $\mathbf{L}^{\mathcal{I}}$ and $\mathbf{L}^{\mathcal{M}}$ be $V \times 3$ matrices of the 3D positions of the key-points in $\mathcal{I}$ and $\mathcal{M}$. We want to find a $3 \times 3$ rotation matrix $\mathbf{R}^*$ that rotates $\mathbf{L}^{\mathcal{I}}$ to the same view of $\mathbf{L}^{\mathcal{M}}$, i.e.

$$\mathbf{R}^* = \arg\min_R \|\mathbf{L}^{\mathcal{M}} - \mathbf{R}\mathbf{L}^{\mathcal{I}}\|^2 \tag{1}$$

We use a least-square method [36] to find $\mathbf{R}^*$. Let $\mathbf{U}\mathbf{D}\mathbf{V}^T$ a singular decomposition of $\mathbf{L}^{\mathcal{M}^T}\mathbf{L}^{\mathcal{I}}$, and define $\mathbf{S} = \mathbf{I}$ if $\det(\mathbf{L}^{\mathcal{M}^T}\mathbf{L}^{\mathcal{I}}) \geq 0$, otherwise $\mathbf{S} = \mathrm{diag}(1, \cdots, 1, -1)$. Then we have $\mathbf{R}^* = \mathbf{U}\mathbf{S}\mathbf{V}^T$. Fig.4 gives some example results of rotating an image to similar view-points of the other images.

Combining the similarity values obtained from appearance and pose features, the similarity between $\mathcal{G}^{\mathcal{I}}$ and $\mathcal{G}^{\mathcal{M}}$ is

$$\mathcal{S}\left(\mathcal{G}^{\mathcal{I}}, \mathcal{G}^{\mathcal{M}}\right) = \exp\left\{\sum_{v} h_v^{\mathcal{I}} h_v^{\mathcal{M}} \cdot \mathbf{w}_v^{\mathcal{M}} \cdot I\left(\mathbf{f}_v^{\mathcal{I}}, \mathbf{f}_v^{\mathcal{M}}\right) + \sum_{e} h_e^{\mathcal{I}} h_e^{\mathcal{M}} \cdot \mathbf{w}_e^{\mathcal{M}} \cdot \left(\mathbf{R}^* \Delta \mathbf{l}_e^{\mathcal{I}} - \Delta \mathbf{l}_e^{\mathcal{M}}\right)\right\} \tag{2}$$

$\mathcal{S}(\cdot,\cdot)$ is not symmetric, i.e. in most situations $\mathcal{S}\left(\mathcal{G}^{\mathcal{I}}, \mathcal{G}^{\mathcal{M}}\right) \neq \mathcal{S}\left(\mathcal{G}^{\mathcal{M}}, \mathcal{G}^{\mathcal{I}}\right)$.

## 4 Exemplar-Based Action Recognition

### 4.1 Dominating Sets of Action Classes

We adopt an exemplar-based approach for action recognition. Exemplar-based approaches allow using multiple exemplars to represent an action class, enabling more flexibility in overcoming the challenge of large within-action pose variations (Fig.2(b)). Rather than matching a testing image with all the training images as in most previous exemplar-based systems, for each action class, we select a small set of representative training images that are able to cover all pose variations of this action while maximizing the distinction between this action and all the others. Selecting such images is equivalent to the *minimum dominating set problem* [29, 30] in graph theory, and therefore we call those images *dominating images*, denoted as $Dom(k)$ for class $k$.

To formally define the dominating images of human actions, we first define the *coverage set* of an image $\mathcal{I}$, $Cov(\mathcal{I})$. The images in $Cov(\mathcal{I})$ belong to the same class as $\mathcal{I}$, and each image has a larger similarity value with $\mathcal{I}$ than all the images of different classes. Mathematically speaking, assume we have a set of training images $\{\mathcal{I}_1, \cdots, \mathcal{I}_N\}$, where each $\mathcal{I}_i$ is associated with an action class label $y_i \in \{1, \cdots, K\}$. The coverage set of $\mathcal{I}$ is defined as

$$Cov\left(\mathcal{I}\right) = \left\{ \mathcal{I}_i \mid \mathcal{S}\left(\mathcal{G}^{\mathcal{I}_i}, \mathcal{G}^{\mathcal{I}}\right) > T + \eta,\ T = \max_{\forall j, y_j \neq y} \mathcal{S}\left(\mathcal{G}^{\mathcal{I}_j}, \mathcal{G}^{\mathcal{I}}\right) \right\}, \tag{3}$$

where $T$ is the maximum similarity between $\mathcal{I}$ and images of the other classes. $\eta > 0$ controls the margin of the similarity difference. As shown in Fig.5, $Cov\left(\mathcal{I}\right)$ defines a set of images where the 3D pose configurations and visual appearances are similar to $\mathcal{I}$. For an action class $k$, the dominating image set $Dom(k)$ are a minimum set of images such that the joint of their coverage sets contain all the images of class $k$, i.e.

$$\forall\, \mathcal{I}_i \text{ where } y_i = k,\ \exists\, \mathcal{I}_j \in Dom(k) \text{ such that } \mathcal{S}\left(\mathcal{G}^{\mathcal{I}_i}, \mathcal{G}^{\mathcal{I}_j}\right) > T_j + \eta \tag{4}$$

If there exist another $\widetilde{Dom}(k)$ satisfies the above condition, $|Dom(k)| \leq |\widetilde{Dom}(k)|$, where $|Dom(k)|$ is the number of images in $Dom(k)$.

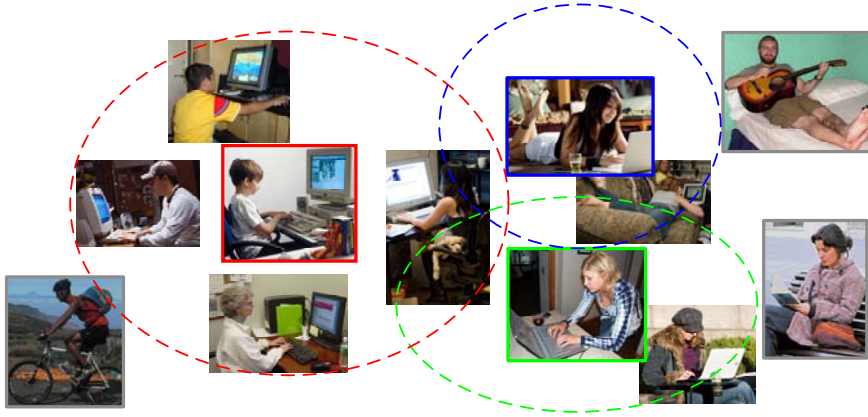


**Fig. 5.** Illustration of the dominating images of "using a computer". The images surrounded by red, blue, and green rectangles are dominating images. Dotted ellipses representing the corresponding coverage sets. The images surrounded by gray are images of the other actions, which are used to define the boundary of the coverage sets.

### 4.2 Obtaining Minimum Dominating Sets for Each Action

Our method of obtaining the minimum dominating sets consists of two steps. Firstly, we learn image-specific feature weights $\mathbf{W}^{\mathcal{I}} = \left\{\mathbf{w}_v^{\mathcal{I}}, v = 1, \cdots, V; \mathbf{w}_e^{\mathcal{I}}, e = 1, \cdots, E\right\}$ for each image $\mathcal{I}$ to maximize $|Cov\left(\mathcal{I}\right)|$. Then we use an improved reverse heuristic method [30] to find the images that belong to $Dom(k)$ for each class $k$. We elaborate on the two steps separately.

For each image $\mathcal{I}$, $\mathbf{W}^{\mathcal{I}}$ maximizes the distinction between $\mathcal{I}$ and images of the other action classes. Finding a globally optimal $\mathbf{W}^{\mathcal{I}}$, however, is not a convex problem, because which images belong to $Cov\left(\mathcal{I}\right)$ is uncertain. We therefore resort to a suboptimal solution which aims at separating within-class similarities from between-class similarities. We compute the histogram intersections of appearance features and distances of the key-point 3D positions between $\mathcal{I}$ and each image $\mathcal{I}_i$. This results to a feature vector

$$\left[h_v^{\mathcal{I}_i} h_v^{\mathcal{I}} \cdot I\left(\mathbf{f}_v^{\mathcal{I}_i}, \mathbf{f}_v^{\mathcal{I}}\right), v = 1, \cdots, N; h_e^{\mathcal{I}_i} h_e^{\mathcal{I}} \cdot \mathbf{R}^* \Delta \mathbf{l}_e^{\mathcal{I}_i} - \Delta \mathbf{l}_e^{\mathcal{I}}, e = 1, \cdots, E\right]. \quad (5)$$

If $\mathcal{I}_i$ and $\mathcal{I}$ belong to the same class, this vector is regarded as a positive sample, otherwise negative. We then train a binary SVM classifier to discriminate positive samples from negative samples. The obtained SVM feature weights are $\mathbf{W}^{\mathcal{I}}$.

Based on $\mathbf{W}^{\mathcal{I}}$ learned for each image that belong to class $k$, i.e. $y = k$, we can compute their coverage sets (Eq.3) and then find $Dom(k)$. But finding the minimum dominating set is also a NP-hard problem. We use the improved reverse heuristic (IRH) method [30], which selects the samples in $Dom(k)$ iteratively for each $k$. The heuristic rule is, on the one hand, the images have large coverage sets are more likely to be selected; on the other hand, the images that are covered by many other ones are less likely to be selected. In order to incorporate the latter

- For each class $k \in \{1, \cdots, K\}$, denote all the images of this class as $Im(k)$.
- Initialize $Dom(k) = \emptyset$.
  1. Compute $Cov(\mathcal{I})$ and $Reach(\mathcal{I})$ for each $\mathcal{I} \in Im(k)$;
  2. Find $\mathcal{I}^* \in Im(k)$ that maximizes $Cov(\mathcal{I}) - \lambda \cdot Reach(\mathcal{I})$;
  3. Add $\mathcal{I}^*$ to $Dom(k)$, and remove all $\mathcal{I} \in Cover(\mathcal{I}^*)$ from $Im(k)$;
  4. If $Im(k) \neq \emptyset$, return to step 1.

**Fig. 6.** The improved reverse heuristic method for selecting dominating images for each action class.

heuristic rule, we define the reachability of an image $\mathcal{I}$,

$$Reach\left(\mathcal{I}\right) = \left\{\mathcal{I}_i \mid \mathcal{S}\left(\mathcal{G}^{\mathcal{I}}, \mathcal{G}^{\mathcal{I}_i}\right) > T_i + \eta,\ y_i = y\right\} \tag{6}$$

Based on the coverage set and reachability set of each image, the IRH method are shown in Fig.6.

### 4.3 Action Recognition Using the Dominating Sets

To recognize the human action in a test image $\mathcal{I}'$, we construct a 2.5D graph for this image and match it with the dominating images in all the action classes. The action class that correspond to the largest normalized similarity is the recognition result, i.e.

$$k' = \arg\max_k \mathcal{S}(\mathcal{I}', k), \text{ where } \mathcal{S}(\mathcal{I}', k) = \arg \max_{\mathcal{I}_i \in Dom(k)} \frac{\mathcal{S}(\mathcal{I}', \mathcal{I}_i)}{T_i} \tag{7}$$

## 5 Experiments

We carry out experiments on two publicly available datasets: the people playing musical instrument (PPMI) dataset [37] and the PASCAL VOC 2011 action classification dataset [10]. In all the experiments described below, all training processes are conducted on only training images, including human pose estimation, etc. Please refer to Sec.3 and Sec.4 for implementation details of our approach. On both datasets, we use mean Average Precision (mAP) for performance evaluation.

### 5.1 Results on the PPMI Dataset

The PPMI dataset [37] is a collection of images of people interacting with twelve different musical instruments: bassoon, cello, clarinet, erhu, flute, French horn, guitar, harp, recorder, saxophone, trumpet, and violin. It is a 24-class classification problem. For each instrument, there are images of people playing the instrument, as well as images of people holding the instrument but not playing. We use the normalized images on this dataset. For each class, there are 100 images for training and 100 images for testing.

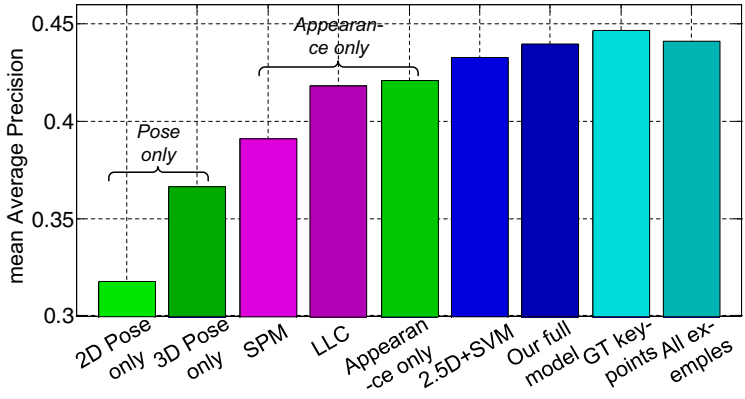



**Fig. 7.** Comparison of different methods on the PPMI dataset. The performances are evaluated by mean Average Precision. Magenta colors indicate existing methods. Green, blue, and cyan colors indicate our method or control experiments.

We compare our approach with a number of control settings and some state-of-the-art classification systems described below.

- *Bag-of-Words (BOW) baseline:* Extract SIFT features [20] and use bag-of-words for classification. The codebook size for SIFT features is 1024.
- *Locality-constrained linear (LLC) coding + spatial pyramid:* Image features are multi-scale, multi-resolution color-SIFT [38] features with locality-constrained linear coding [35]. The features are max-pooled on a three-level image pyramid [21] with linear SVM for classification. This is the best result reported in the website of the dataset.
- *Control - 3D pose only:* 2D image appearances are not used for image representation. Everything else is the same as our method. This is equivalent to setting $I\left(\mathbf{f}_v^{\mathcal{I}}, \mathbf{f}_v^{\mathcal{M}}\right)$ to $\mathbf{0}$ in Eq.2.
- *Control - 2D pose only:* Using only the original 2D locations for recognition, without rotating 3D key-point positions when matching two images.
- *Control - 2D appearance only:* The location of 3D key-points are not used for image representation. Everything else is the same as our method. This is equivalent to setting $\left(\mathbf{R}^* \Delta \mathbf{l}_e^{\mathcal{I}} - \Delta \mathbf{l}_e^{\mathcal{M}}\right)$ to 0 in Eq.2.
- *Control - 2.5D graph + SVM:* Using the 2.5D graph for image representation, and train a multiclass classifier based on 2D appearances and 3D poses.
- *Control - using ground-truth key-points:* Instead of using pictorial structure to estimate the 2D key-point locations. We use ground-truth positions of key-points.
- *Control - using all training images as exemplars:* Instead of selecting dominating images for each action, we match a testing image to all training images for classification.

The mAP of different methods are shown in Fig.7. Our method outperforms the existing methods by achieving a 43.9% mAP, even comparing with LLC,

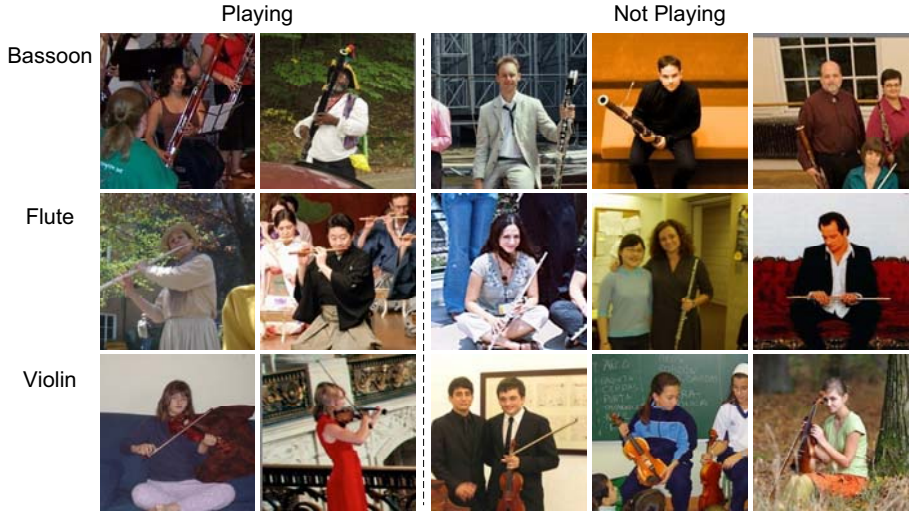



**Fig. 8.** Examples of dominating images selected from the PPMI training set.

which is the current best result on this dataset. Because the images of people playing some musical instruments are very similar (e.g. playing saxophone and playing bassoon, as shown in Fig.8.), using human pose only cannot achieve very good performance on this dataset. But 3D poses achieve much better results than 2D poses. Using the local appearance features extracted based on the key-point positions, our appearance feature performs comparable with LLC. The full 2.5D graph representation, which combines the 3D position information and 2D appearance information, outperforms both methods that use any one of them. This shows that our method effectively captures the complementary information between poses and appearances. Our full model also performs better than training a multiclass SVM classifier on the 2.5D graph features, demonstrating the effectiveness of the exemplar-based classification.

In Fig.7, our method is only 0.7% worse than the approach that uses ground-truth key-point locations to construct the 2.5D graphs. This shows that although our 2.5D graphs are constructed based on imperfect key-point locations (using the criteria in [39], our key-point detection accuracy is 65.7%), it can still achieve satisfactory recognition performance. Finally, our method performs comparable with the approach that uses all training images as exemplars. But our classification is much faster because we only need to match each testing image with 3.6 images (the average number of selected dominating images) per class, as compared with matching 100 images in the "all-exemplars" setting.

Fig.8 shows the dominating images selected from some action classes. On the classes of people playing the instrument, human poses are very similar in each class. Therefore the dominating images mainly capture with-class appearance variations. On the classes of people holding the instruments but not playing, the variations in both human pose and image appearance are captured.

### 5.2 Results on the PASCAL Dataset

The PASCAL 2011 action dataset contains around 8,000 images of ten actions: "jumping", "phoning", "playing instrument", "reading", "riding bike", "riding horse", "running", "taking photo", "using computer", and "walking". The dataset also contains images that do not belong to any of the ten actions. All images are downloaded from flickr, and represent very large variations in both human pose and appearance.

We compare our approach with a number of methods that achieve good performance on the challenge [10]. The results are shown in Table 1. We observe that our method performs the best on three out of the ten classes, especially on the classes of "jumping" and "playing instrument" which contain large human pose variations, and obtains the highest mean average precision over all the classes. On the classes of "riding a horse" and "riding a bike", our method does not perform as good as ATTR_PART, which explicitly detects objects such as horses and bikes in the images and relies on independent dataset to train the object detectors. Table 1 also shows that using pose features only, our method achieves better performance thatn POSELETS, demonstrating the effectiveness of our view-independent 3D pose representation.

**Table 1.** Results on the PASCAL 2011 action dataset. The numbers are percentage of mean average precision. The best results are marked by bold fonts.

| Action | HOBJ_ DSAL | CON- TEXT | RF_ SVM | POSE_ LETS | ATTR_ PART | Our Method | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | Pose | App. | Full |
| jumping | 71.6 | 65.9 | 66.0 | 59.5 | 66.7 | 64.6 | 68.9 | **72.4** |
| phoning | **50.7** | 41.5 | 41.0 | 31.3 | 41.1 | 41.2 | 44.5 | 48.3 |
| playing instrument | 77.5 | 57.4 | 60.0 | 45.6 | 60.8 | 68.3 | 72.9 | **77.7** |
| reading | 37.8 | 34.7 | 41.5 | 27.8 | 42.2 | 36.0 | 39.2 | **43.2** |
| riding bike | 86.5 | 88.8 | 90.0 | 84.4 | **90.5** | 81.4 | 86.6 | 89.0 |
| riding horse | 89.5 | 90.2 | 92.1 | 88.3 | **92.2** | 80.4 | 87.1 | 90.0 |
| running | 83.8 | **87.9** | 86.6 | 77.6 | 86.2 | 79.4 | 83.0 | 86.8 |
| taking photo | 25.1 | 25.7 | 28.8 | **31.0** | 28.8 | 21.6 | 25.1 | 27.9 |
| using computer | 58.9 | 54.5 | 62.0 | 47.4 | **63.5** | 51.5 | 56.9 | 60.5 |
| walking | 59.2 | 59.5 | **65.9** | 57.6 | 64.2 | 52.8 | 59.7 | 62.1 |
| mean | 64.1 | 60.6 | 63.4 | 55.1 | 63.6 | 57.7 | 62.4 | **65.8** |

## 6 Conclusion

In this paper, we propose a 2.5D graph for action image representation. The 2.5D graph integrates 3D view-independent pose features and 2D appearance features. An exemplar-based approach is used for action recognition, where a small set of images that are able to cover the large with-action pose variations are used as

the exemplars for each class. One direction of future research is to study how the alignment of 3D positions can provide better usage of 2D appearance features.

## Acknowledgement

This research is partially supported by an ONR MURI grant, the DARPA CSSG program, an NSF CAREER grant (IIS-0845230), a research sponsorship from Intel to L.F-F., and the SAP Stanford Graduate Fellowship and Microsoft Research PhD Fellowship to B.Y.

## References

1. Ikizler, N., Cinbis, R.G., Pehlivan, S., Duygulu, P.: Recognizing actions from still images. In: ICPR. (2008)
2. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE T. Pattern Anal. Mach. Intell. **31** (2009) 1775–1789
3. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR. (2010)
4. Yang, W., Wang, Y., Mori, G.: Recognizing human actions from still images with latent poses. In: CVPR. (2010)
5. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L.J., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: ICCV. (2011)
6. Maji, S., Bourdev, L., Malik, J.: Action recognition from a distributed representation of pose and appearance. In: CVPR. (2011)
7. Delaitre, V., Sivic, J., Laptev, I.: Learning person-object interactions for action recognition in still images. In: NIPS. (2011)
8. Prest, A., Schmid, C., Ferrari, V.: Weakly supervised learning of interactions between humans and objects. IEEE T. Pattern Anal. Mach. Intell. **34** (2012) 601–614
9. Yao, B., Khosla, A., Fei-Fei, L.: Combining randomization and discrimination for fine-grained image categorization. In: CVPR. (2011)
10. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results (2011)
11. Natarajan, P., Nevatia, R.: View and scale invariant action recognition using multiview shape-flow methods. In: CVPR. (2008)
12. Yan, P., Khan, S.M., Shah, M.: Learning 4D action feature models for arbitaray view action recognition. In: CVPR. (2008)
13. Gong, D., Medioni, G.: Dynamic manifold warping for view invariant action recognition. In: ICCV. (2011)
14. Weinland, D., Ozuysal, M., Fua, P.: Making action recognition robust to occlusions and viewpoint changes. In: ECCV. (2010)
15. Junejo, I.N., Dexter, E., Laptev, I., Perez, P.: View-independent action recognition from temporal self-similarities. IEEE T. Pattern Anal. Mach. Intell. **33** (2011) 172–185
16. Sapp, B., Toshev, A., Taskar, B.: Cascade models for articulated pose estimation. In: ECCV. (2010)

17. Taylor, C.J.: Reconstruction of articulated objects from point correspondences in a single uncalibrated image. Volume 80. (2000) 349–363
18. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3D human pose annotations. In: ICCV. (2009)
19. Yao, A., Gall, J., Fanelli, G., van Gool, L.: Does human action recognition benefit from pose estimation? In: BMVC. (2011)
20. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60** (2004) 91–110
21. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006)
22. Szeliski, R., Anandan, P., Baker, S.: From 2D images to 2.5D sprites: A layered approach to modeling 3D scenes. In: MMCS. (1999)
23. Duan, Y., Qin, H.: 2.5D active contour for surface reconstruction. In: VMV. (2003)
24. Zafeiriou, S., Petrou, M.: 2.5d elastic graph matching. Comput. Vis. Image Und. **115** (2011) 1062–1072
25. Sung, K.K., Poggio, T.: Example-based learning for view-based human face detection. IEEE T. Pattern Anal. Mach. Intell. **20** (1998) 39–51
26. Frome, A., Singer, Y., Sha, F., Malik, J.: Learning globally-consistent local distance functions for shape-based image retrieval and classification. In: ICCV. (2007)
27. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-SVMs for object detection and beyond. In: ICCV. (2011)
28. Willems, G., Becker, J.H., Tuytelaars, T., van Gool, L.: Exemplar-based action recognition in video. In: BMVC. (2009)
29. Hedetniemi, S.T., Laskar, R.C.: Bibliography on domination in graphs and some basic definitions of domination parameters. Discrete Math. **86** (1990) 257–277
30. Yao, B., Ai, H., Lao, S.: Building a compact relevant sample coverage for relevance feedback in content-based image retrieval. In: ECCV. (2008)
31. Read, J.C.A., Phillipson, G.P., Serrano-Pedraza, I., Milner, A.D., Parker, A.J.: Stereoscopic vision in the absence of the lateral occipital cortex. PLoS ONE **5** (2010)
32. Lee, H.J., Chen, Z.: Determination of human body posture from a single view. Comp. Vision, Graphcs, and Image Proc. **30** (1985) 148–168
33. Delaitre, V., Laptev, I., Sivic, J.: Recognizing human actions in still images: a study of bag-of-features and part-based representations. In: BMVC. (2010)
34. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE T. Pattern Anal. Mach. Intell. **32** (2010) 1627–1645
35. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Learning locality-constrained linear coding for image classification. In: CVPR. (2010)
36. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE T. Pattern Anal. Mach. Intell. **13** (1991) 376–380
37. Yao, B., Fei-Fei, L.: Grouplet: A structured image representation for recognizing human and object interactions. In: CVPR. (2010)
38. Burghouts, G.J., Geusebroek, J.M.: Performance evaluation of local colour invariants. Comput. Vis. Image Und. **113** (2009) 48–62
39. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR. (2008)