Pattern Recognition 47 (2014) 3343–3361

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A survey on still image based human action recognition

Guodong Guo*, Alice Lai

Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, United States

ARTICLE INFO

Article history:
Received 23 July 2013
Received in revised form
16 March 2014
Accepted 13 April 2014
Available online 9 May 2014

Keywords:
Action recognition
Still image based
Various cues
Databases
Survey
Evaluation

ABSTRACT

Recently still image-based human action recognition has become an active research topic in computer vision and pattern recognition. It focuses on identifying a person's action or behavior from a single image. Unlike the traditional action recognition approaches where videos or image sequences are used, a still image contains no temporal information for action characterization. Thus the prevailing spatio-temporal features for video-based action analysis are not appropriate for still image-based action recognition. It is more challenging to perform still image-based action recognition than the video-based one, given the limited source of information as well as the cluttered background for images collected from the Internet. On the other hand, a large number of still images exist over the Internet. Therefore it is demanding to develop robust and efficient methods for still image-based action recognition to understand the web images better for image retrieval or search. Based on the emerging research in recent years, it is time to review the existing approaches to still image-based action recognition and inspire more efforts to advance the field of research. We present a detailed overview of the state-of-the-art methods for still image-based action recognition, and categorize and describe various high-level cues and low-level features for action analysis in still images. All related databases are introduced with details. Finally, we give our views and thoughts for future research.

1. Introduction

Recognizing human motion and action has been an active research topic in Computer Vision for more than two decades. This can also be indicated by a series of survey papers in the literature. Earlier review papers focused on human motion analysis and discussed human action recognition as a part, such as the surveys by Cedras and Shah [1], Aggarwal and Cai [2], and Gavrila [3]. Later on, the survey paper by Kruger et al. [4] classified human action recognition approaches based on the complexity of features to represent human actions and considered potential applications to robotics. The survey paper by Turaga et al. [5] covered human activity recognition with a categorization based on the complexity of activities and recognition methodologies. In Poppe's survey [6], various challenges in action recognition were addressed and novelties of different approaches were discussed. In Ji and Liu's survey [7], the concentration was on view-invariant representation for action recognition. They discussed related issues such as human detection, view-invariant pose representation and estimation, and behavior understanding. Finally, the most recent survey was given by Aggarwal and Ryoo in [8], who performed a comprehensive review of recognizing action, activity, gesture, human–object interaction, and group activities. It discussed the limitations of many existing approaches and listed various databases for evaluations. The real-time applications were also mentioned.

Although motion-based/video-based human action recognition is still an active research topic in computer vision and pattern recognition, recent studies have started to explore action recognition in *still images*. As shown in Fig. 1, many action categories can be depicted unambiguously in single images (without motion or video signal), and these actions can be understood well based on human perception. This evidence supports the development of computational algorithms for automated action analysis and recognition in still images. Considering the large number of single images distributed over the Internet, it is valuable to analyze human behaviors in those images. Actually, it has become an active research topic very recently [9].

An analogy to human (body-based) action recognition is facial expression recognition [10,11], sometimes called the facial behavior understanding. In facial expression analysis, either single face images or face videos can be used. Different from action recognition, the studies of facial expressions using single images or videos are almost in parallel. The number of publications using either single images or videos is probably comparable in facial expression recognition. However, in action recognition, a large number of publications are video-based. Only very recently, researchers have begun to focus on still image-based action understanding.

* Corresponding author.
E-mail address: guodong.guo@mail.wvu.edu (G. Guo).

http://dx.doi.org/10.1016/j.patcog.2014.04.018
0031-3203/

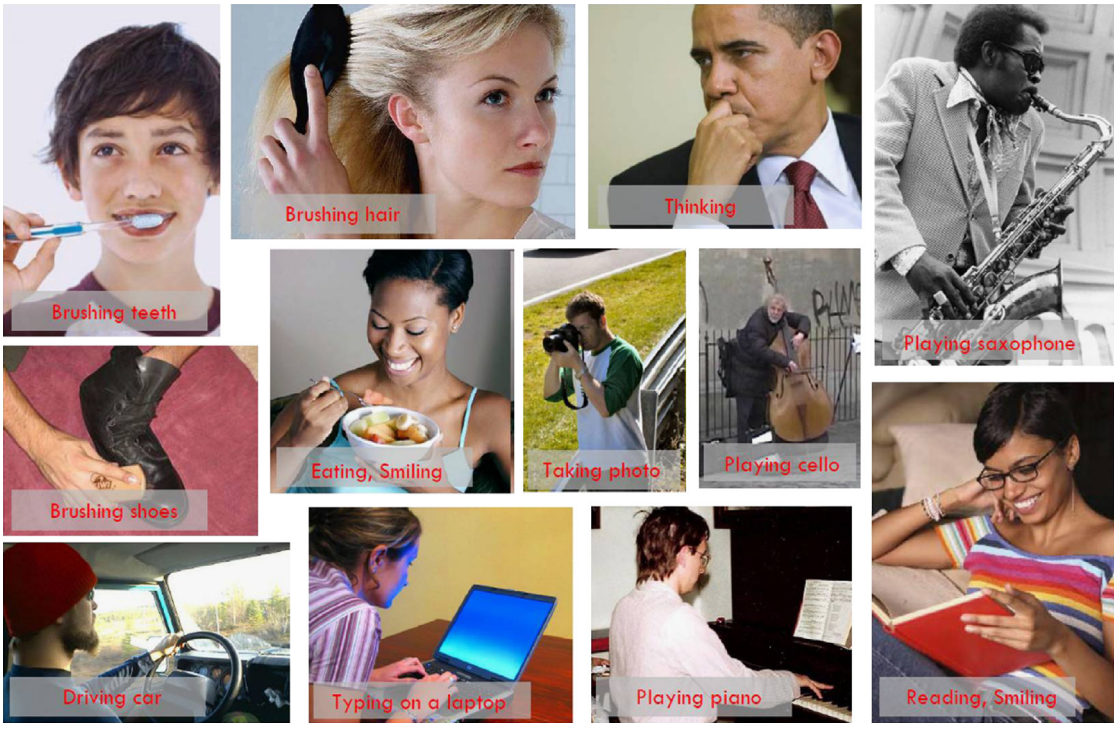

Fig. 1. Some examples of still image-based action recognition. Only single images are sufficient to tell the corresponding actions, originally shown in [9].

Compared to the traditional video-based action recognition, still image-based action recognition has some special properties. For example, there is no motion in a still image, and thus many spatiotemporal features and methods that were developed for traditional video-based action recognition are not applicable to still images. And also, it is not trivial to segment the humans from the background in still images [12–14], since there is no motion cue to utilize and the scene can be very cluttered. Thus there are new challenges in solving the problem of still image-based action recognition.

Still image-based action recognition has quite a few useful applications: (1) image annotation. A huge amount of still images are distributed over the Internet, and new images are being acquired more and more. Automated action recognition in still images can help to annotate "verbs" (for actions) on Internet images (such as the examples shown in Fig. 1). (2) Action or behavior based image retrieval. Similar to off-line image annotation, the automated action recognition can also help search and retrieve online images based on action queries. (3) Video frame reduction in video-based action recognition. When still image based action recognition can achieve a high accuracy for some categories of actions, the long video sequences for those actions can be reduced to a small number of single frames for action representation, and thus significantly lower the redundant information without satisfying the action recognition accuracy. (4) Human computer interaction (HCI). Similar to the traditional video-based action recognition, still image based action recognition can also be useful for HCI, especially for actions that do not require a long time period to execute the whole process, e.g., touching, thinking, and smiling.

Although there are useful applications in practice, the study of still image-based action recognition has a very short history, compared to the video-based action recognition research. Starting at about the year of 2006, it appears to have some research papers on still image-based action recognition. Following 2006, only a very small number of papers related to action recognition based on single images appeared, since not many researchers have realized that it is an interesting topic. More papers were published recently. To show the trend of publications on this topic, the annual and the accumulative number of published papers are drawn in a yearly basis and shown in Fig. 2. The criterion of selecting papers on still image based action recognition is based on whether still images are used as the test data for action recognition. The training examples can be purely still images or single image frames extracted from video sequences. The time period to collect the related publications for this survey starts from the year that the earliest paper was published in 2006 until 2013. From Fig. 2, we can see that there are more publications in very recent years, such as 2011 and 2012, with about 10 papers each year, while much less numbers before 2011. The research on this topic has become more active since 2011. We also noticed that there were less numbers of publications after 2011. One reason is that there might be a "bottleneck" to improve the recognition accuracies significantly. New ideas and approaches are needed to address this challenging problem.

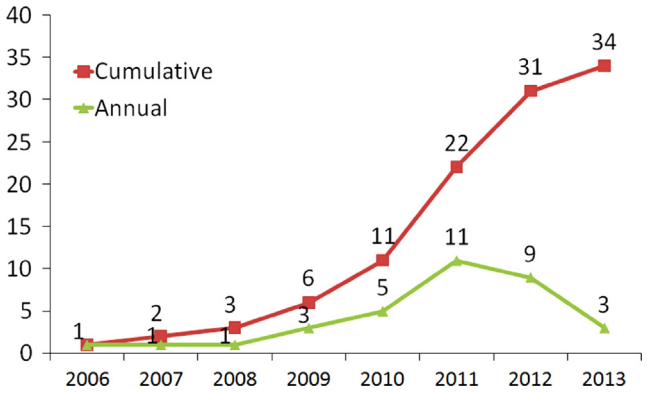

Fig. 2. Graphic display of the number of publications annually and accumulatively on still image-based action recognition. Based on the slope of the piecewise curve, there are more publications in 2011 and 2012.

We believe that it is now a good time to look back on the research works and summarize the progresses in still image-based action recognition. The purpose of this survey is hopefully to raise the attention from the image and visual computing community, encourage more deep thinking about the problem, review the related methods and available databases, inspire new ideas, and eventually advance the field of research.

The survey is organized as follows. We review various high-level cues for still image-based action recognition together with the low-level features to represent the cues and action analysis in Section 2. We present various methods that have been applied to action learning in Section 3. The databases that have been assembled and used for action recognition are introduced in Section 4. The action recognition performance on the most popular databases is presented in Section 5, such that the readers may have some basic idea of the recognition accuracies and results. Some other research topics related to still image-based action recognition are described in Section 6. Then we discuss some future research directions in Section 7. Finally, we draw conclusions.

2. Various cues for action representation

In still image-based action recognition, there is no temporal information available, and thus the traditional spatiotemporal features [6] cannot be applied anymore. Further, in traditional video-based action recognition, the low-level features extracted from space-time volume can be used directly for action recognition, e.g., the spatiotemporal interest point (STIP) based features [15]. However, in still image-based action recognition, usually the low-level features extracted directly from the whole image cannot work well. Thus previous works seldom use the whole image or scene only for low-level feature extraction and action recognition.

Since only the spatial information is available in single images with cluttered background, researchers have pursued different high-level cues in still images in order to characterize actions better than using low-level features in the whole image.

The high-level cues can be characterized through various low-level features. Then different high-level cues can be combined to recognize the actions in still images. In this section, we will present the high-level cues and low-level features that have been used for still image-based action recognition.

2.1. High-level cues

The most popular high-level cues for still image-based action recognition include the human body, body parts, action-related objects, human object interaction, and the whole scene or context. These cues can characterize human actions from different aspects. A summarization of these cues is given in Table 1. From the table, one can see that some approaches employed more cues, while some others used less. The table tells clearly which cues were used in which paper.

We will introduce the high-level cues first, and then the low-level features.

2.1.1. Human body

Human body is an important cue for still image-based action recognition. Most of the existing approaches use the human body cue for action representation. The human body can be detected automatically [12,14] in images or manually labeled [9]. Usually the bounding box of the human is used to indicate the location of the person and determine the image region for human body information extraction. For example, Li et al. [34,37] defined a so-called exemplarlet which is the manually selected and segmented minimum bounding box which contains enough visual information to identify the human body for action analysis in a still image. The resulted exemplarlets show that visual information from human body is dominant in action image analysis. There are also approaches using other kinds of information from the body, e.g., contour [16], and poses [18].

There are also approaches that extract features in areas within or surrounding the human bounding boxes, e.g., Delaitre et al. [26]. They defined a person setting in each image with 1.5 × the size of the human bounding box, and resized each cropped region into a new size such that the larger dimension is 300 pixels. These regions are then represented using some low-level features.

Some methods for action recognition rely heavily on the human body contour information from the image, rather than just the bounding box. For example, the approach by Wang et al. [16], which is probably the earliest work on still image-based action recognition, exploited the overall coarse shape of human body in the image represented by a collection of edge points obtained via the Canny edge detector [47], as the features to cluster and label images into different actions.

In addition to body shape and bounding boxes, the human body pose is also useful to extract the cue from body images. Thurau and Hlavac [18] used human body poses based on extracting a set of Non-negative Matrix Factorization (NMF) [48]

Table 1
High-level cues used in existing approaches for still image-based action recognition.

Approach	Human body	Body parts	Objects	h–o Interact.	Scene
Wang et al. [16]	√				
Li and Fei-Fei [17]	√		√		√
Thurau and Hlavac [18]	√				√
Ikizler et al. [19]	√				
Ikizler-Cinbis et al. [20]	√				
Gupta et al.[21]	√		√		√
Yao and Fei-Fei [22]	√				
Desai et al. [23]	√		√	√	
Yang et al. [24]		√			
Yao and Fei-Fei [25]	√	√	√	√	
Delaitre et al. [26]	√	√			√
Shapovalova et al. [27]	√		√	√	√
Maji et al. [28]		√	√	√	
Yao et al. [29]	√				√
Yao et al. [30]	√	√	√	√	√
Raja et al. [31]		√			
Komiusz and Mikolajczyk [32]	√				
Komiusz and Mikolajczyk [33]	√				
Li et al. [34]	√				
Yao et al. [35]	√	√	√		
Delaitre et al. [36]		√	√	√	
Li and Ma [37]	√				
Prest et al. [13]	√		√	√	√
Yao and Fei-Fei [38]	√	√	√	√	√
Sharma et al. [39]	√				
Yao and Fei-Fei [40]		√			
Desai and Ramanan [12]		√	√	√	
Kumar et al. [41]		√			
Sener et al. [42]	√	√	√		√
Zheng et al. [43]		√			√
Ikizler-Cinbis et al. [44]	√				
Le et al. [45]			√		
Khan et al. [14]	√				
Sharma et al. [46]		√	√		√

Fig. 3. Pose rectangle extraction, originally shown in [19]. Two example images and their extracted poses: region-based, edge-based, and probability-based parsing. The edge and region features can be used, and two deformable models [49,19] were constructed using the Conditional Random Field (CRF). The edge-based deformable model estimates the initial body part positions. Then a region model (parse) that represents an image for each body part is created using this position estimate. Information obtained from the part histograms becomes the basis of the region-based deformable model. These two parses are used as the basis to extract silhouettes by thresholding over the probability maps.

bases. Ikizler et al. [19] also used the body poses extracted from images by the method in [49], which uses edge and region features and constructs two deformable models using the Conditional Random Field (CRF). This is illustrated in Fig. 3, showing example images and their corresponding poses. The initial parsing of the poses was used as the starting point for body silhouettes extraction by thresholding of the probability maps.

On the other hand, some critical patches can be found within the bounding box of human body. For instance, Yao et al. [29] used the random forest method [50] with some variations (e.g., using the support vector machine (SVM) classifier at each node) to search the useful, discriminative patches from the human body region for action recognition. A saliency map also keeps the critical patch information [39].

Semantic features such as the attributes can also be used to describe the actions in images with the human body. In [35], the authors took a global representation of the attributes as in [51], and used binary (yes/no) classifiers to learn each attribute for action analysis. The attributes are defined as linguistically related descriptions of human actions. Most of the attributes used in [35] were related to verbs in human language. For instance, the attributes of "riding" and "sitting (on a bike seat)" can be used to describe the action of "riding a bike." And also, one attribute may be shared by different actions. For instance, "riding" can be shared by "riding a bike" and "riding a horse." However, there are different attributes between any two actions so that the actions can be differentiated using the attributes.

2.1.2. Body parts

Rather than the whole human body, body parts can be more related to action execution. When performing different actions, e.g., throwing a ball or using a computer, the body parts, e.g., the arms, can be in different locations or with different poses. Based on this, the cue of body parts may be used for action characterization.

Delaitre et al. [26] combined the results from a body part detector in the score level fusion with other features, e.g., those based on the spatial pyramid bag-of-features. They demonstrated that the fusion with body parts can improve the action recognition accuracy in their experiments on three datasets: their own Willow Dataset, the Sports Dataset [21] and the PPMI dataset [22]. In their approach, body part detectors need to be trained and the detection results might not be good in some cases.

Poselet [52] is usually extracted from body parts, and can capture the salient body poses specific to certain actions. The use of poselet is suitable to analyze human body parts for action recognition in still images, e.g., the studies in Maji et al. in 2011 [28] and Zheng et al. in 2012 [43].

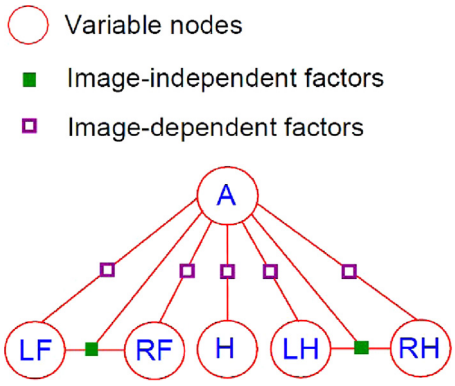

Fig. 4. The graphical model, originally shown in [31]. The model of a person contains six variable nodes encoding the positions of five body parts and the action label. The considered parts $p=\{H, RH, LH, RF, LF\}$ correspond to head, right-hand, left-hand, right-foot and left-foot, respectively, as well as K action classes A. The links between the nodes encode action dependent constraints on the relative position of body parts and their appearance.

A graph is usually a good model to represent the connections and the relations between different body parts. Yang et al. [24] chose to use a coarse exemplar-based poselet representation, which is an action-specific variant of the original poselet [52]. For each image, let L be the pose of the person, denoted as $L=(l_0, l_1 \ldots l_{K-1})$, where $K=4$ corresponding to four body parts: the upper-body, legs, left-arm, and right-arm. The configuration of the kth part l_k contains the locations and the indices of the chosen poselets for the kth body part. Each body part may have more than 20 poselets, which may be resulted from different actions with different body poses.

Raja et al. [31] considered a graphical model containing six nodes, encoding positions of five body parts and the action label. See Fig. 4 for an illustration of their work. Body parts $p=\{H, RH, LH, RF, LF\}$ correspond to head, right-hand, left-hand, right-foot and left-foot, respectively, as well as the action class label node A. The links between the nodes encode action-dependent constraints on the relative positions of body parts and their appearance. Using the body part detection results, relations between positions of hands and feet in images are also modeled and interpreted as being proportional to the joint probability of the right- and the left-part location for a given action.

Yao and Fei-Fei [40] proposed a 2.5D graph for an action image, consisting of a set of nodes that are key-points in the human body, as well as a set of edges that depict spatial relationships between

the nodes. Each key-point is represented by view-independent 3D positions and local 2D appearance features. The 3D position of these points is obtained by using first a pictorial structure [53] to estimate their positions in 2D, and then using the method [54] with additional constraints [55] on the relative lengths of the body parts to effectively enforce all the kinematic constraints associated with the joints, except for joint angle limits. It was used to recover the depth information. This depth recovery approach is simple and effective. Finally, the similarity between two action images can then be measured by matching their corresponding 2.5D graphs.

2.1.3. Objects

When performing actions by humans, there are objects related to the actions. This can be observed in many still images of human actions. Thus it is natural to consider the related objects for human action characterization. Different actions might be related to different objects. By knowing the related objects, it can help to recognize the corresponding actions. For example, a horse (with a human) is possibly related to the action of "riding a horse," or a phone (with a person) could be related to the action of "phoning."

Researchers have realized the importance of using object information to help action recognition in still images. Prest et al. [13] used the results from objectness [56] to calculate the probability of a certain patch being an object. The objectness method can find multiple candidates of objects that can be related to actions regardless of the actual classes of the objects, such as bike, horse and phone.

Sener et al. [42] extracted several candidate object regions and used them in a Multiple Instance Learning (MIL) framework [57]. They sampled 100 windows from each image-based on the candidates from objectness measure [56] where feature vectors can be extracted from each of the 100 windows. The features can be computed by the bag-of-words approach with clustering. They grouped the 100 windows into 10 clusters, and used the cluster centers as the representation of candidate object regions. The object candidates act as the corresponding instances inside a 'bag' (which is the image here) to apply the multiple instance learning method for action recognition.

Yao et al. [35] used a part model composed of objects and human poses. See Fig. 5 for an illustration. The related objects are either manipulated by persons (e.g., a bike is rid by a person) or related to the scene context of the action (e.g., the grass in the scene of "horse riding in grassland"). They used the ImageNet [58] dataset with provided bounding boxes to train an object detector by using the Deformable Parts Model [59]. The DPM method is a discriminatively trained part model using a latent SVM method, used extensively for object detection.

In [45], Le et al. decomposed the input images into groups of recognized objects. Then they used a language model to describe all possible actions in the configurations of objects.

While most approaches using object detectors to determine the occurrence of individual objects for action recognition [25,36,30,38], some works integrate objects with the scene as context, e.g., Zheng et al. [43] learned context-based classifiers, which uses image content from both the foreground and the background as the context.

2.1.4. Human–object interaction

In addition to the co-occurrence of humans and objects and modeling of them separately, the *interaction* between humans and objects is also useful for action recognition in still images, for instance, the relative position between a person and the action-related object (e.g., a book for reading), and the relative angle between the person and the object (e.g., the person is above the

Fig. 5. The attributes are linguistically related descriptions of human actions. The parts are composed of objects and human poses. Attributes and parts are used as action bases to model actions in still images, originally shown in [35].

bike when he/she is riding a bike), the relative size of the person and the object (e.g., a phone (in calling) is much smaller than a horse (in riding) in the two different actions), etc. The configuration of humans and objects in executing actions has been pursued by several researchers.

Desai et al. [23] used the contextual information for action recognition, which was derived from the object layout obtained by their discriminative models [60]. Fig. 6 shows a visualization of a spatial histogram feature d_{ij} from multi-class object layout [60]. They considered the location of the center window j with respect to a coordinate frame defined by window i, labeled as the thick outlined box in Fig. 6. The dashed and dotted rectangles represent regions over which the center of window j are binned. The relative location of j can be either far or near with respect to i. For near windows, they considered the above, on top, below, and symmetric next-to bins as shown in the figure. This scheme makes d_{ij} a six-dimensional, sparse binary vector. The score associated with a particular action in an image is obtained by searching over all possible configurations of objects that are consistent with the learned configuration model. Similarly, Shapovalova et al. [27] integrate this sparse spatial interaction feature d_{ij} as well in their action model.

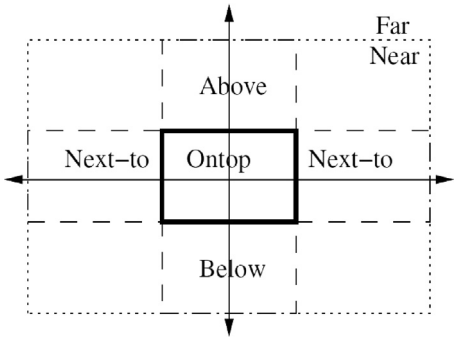

Fig. 6. A visualization of the spatial histogram feature d_{ij} based on the multi-class object layout, originally shown in [60]. The relative location of j can be either far or near w.r.t i. For near windows, the above, on top, below, and symmetric next-to bins were considered. Thus the resulted d_{ij} is a six dimensional, sparse binary vector.

Maji et al. [28] learned a mixture model of the relative spatial locations between the person's bounding box and the object's bounding box in still images, as shown in Fig. 7. For each object type, they fit a two component mixture model of the predicted bounding box to model various relative locations between the person and the object.

Yao et al. [25] presented a graphical modeling of the Human–Object Interaction (HOI), as shown in Fig. 8. They modeled the spatial relationship between the object and the human body parts as well as the dependence of the object with its corresponding image evidence. In their later methods [30,38], the model was extended to deal with any number of objects, instead of being limited to the interactions between one person and one object.

Prest et al. [13] proposed to model human object interaction with four different spatial relations, in order to obtain object candidates from a large number of windows delivered by the objectness measure [56]. The relations include the following: (1) the relative scale difference between the object and the human; (2) the Euclidean distance between the object and the human; (3) the overlapped area of the object and the human (normalized by the area of human); and (4) the relative location between the object and the human. Given the relations between human and object in each image, the differences of these recurring relations between every two images are used to select the

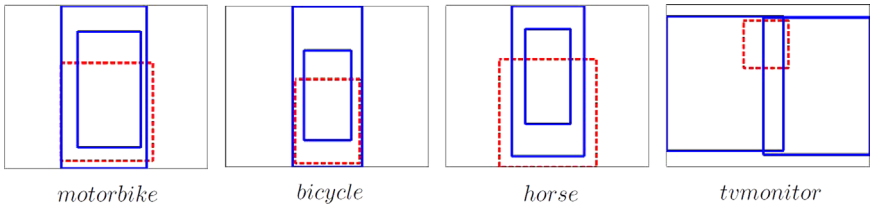

Fig. 7. A spatial model of object person interaction. Each one shows the modes of the bounding boxes of the person (blue) relative to the bounding box of the object (red), originally shown in [28]. For motorbike, bicycle and horse categories, the two modes capture front and side views of the object w.r.t. the person, while for the TV monitor, it captures the fact that TV monitors are often at the left or right corner of the person's bounding box. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

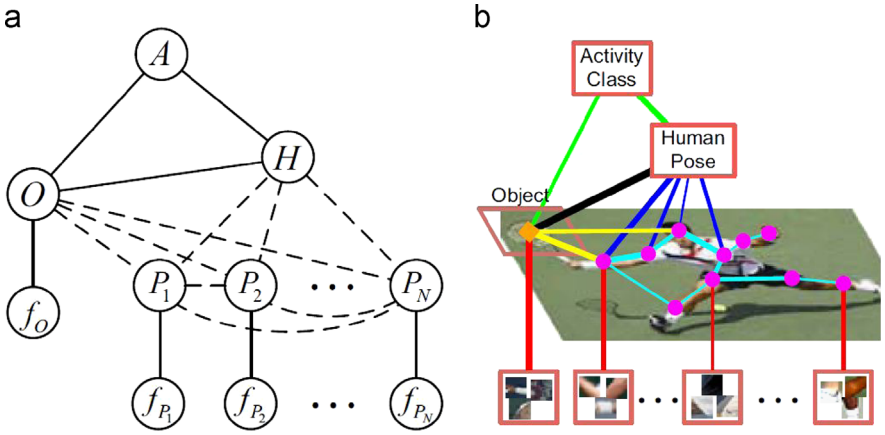

Fig. 8. Illustration of the graphical modeling of human and object interaction. (a) A specific graphical modeling proposed in [25]. The edges represented by dashed lines indicate that their connectivity will be obtained by structure learning. "A" denotes an HOI activity class, "H" the human pose class, "P" a body part, and "O" the object. f_O and f_P are image appearance information of O and P respectively. (b) Using the model in an image example of "human playing tennis." Different types of potentials are denoted by lines with different colors, originally shown in [25]. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

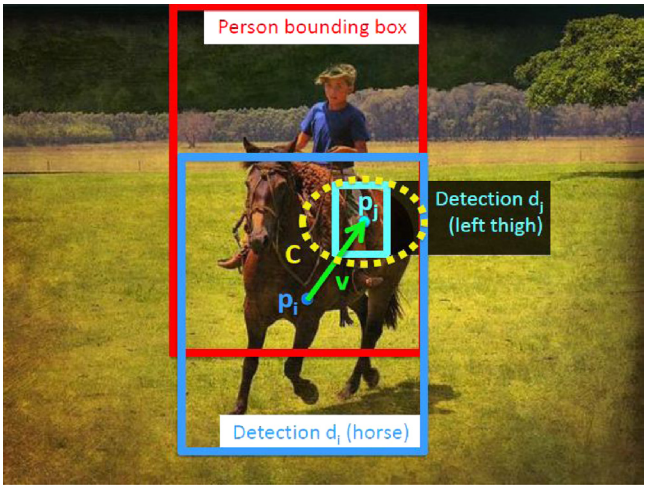

Fig. 9. Representing person–object interactions by pairs of body part (cyan) and object (blue) detectors. To get a strong interaction response, the pair of detectors (here visualized at positions p_i and p_j) must fire in a particular relative 3D scale-space displacement (given by the vector v) with a scale-space displacement uncertainty (deformation cost) given by diagonal 3×3 covariance matrix C (the spatial part of C is visualized as a yellow dotted ellipse). This image representation is defined by the max-pooling of interaction responses over the whole image, solved efficiently by the distance transform, originally shown in [36]. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

optimal object candidates to fit the global configuration for each action class.

Delaitre et al. [36] defined human object interactions based on the pairs of detectors (d_i, d_j) as well as the spatial and scale relations between them. See Fig. 9 for an illustration. Each pair of detectors constitutes a two-node tree where the position and the scale of the leaf are related to the root node by scale-space offset and spatial deformation cost. These object and body part interaction pairs were used for learning a new mid-level feature for action recognition.

2.1.5. Context or scene

The background in an image usually refers to the image region with the foreground human and/or object removed. It may be taken as the context or scene of an executed action. In some cases, the whole image could be considered as the context or scene for action analysis, especially when the foreground (e.g., human and object) occupies a relatively small area in the still image. In reality, some actions are performed in specific scenes, e.g., swimming in water, and driving on the road. So extracting information from the action context or the whole scene can be helpful for still image-based action analysis and recognition [43].

Delaitre et al. [26] studied the efficiency of using whole image, only area of human bounding box, or the combination of these two cues for action recognition, based on the bag-of-features approach. It was shown that the integration of human bounding boxes with the spatial pyramids of background gives an improved performance.

Li and Fei-Fei [17] introduced a method to recognize actions, based on only the occurrences of action-specific scene and objects using spatial and appearance information. The scene-related parameters were learned independently, and integrated for action analysis. Gupta et al. [21] encoded the scene for action image analysis. Their Bayesian model consists of four types of nodes, corresponding to the scene/event (S), scene objects (SO), manipulable objects (MO), and human (H). See Fig. 10 for an illustration of the approach.

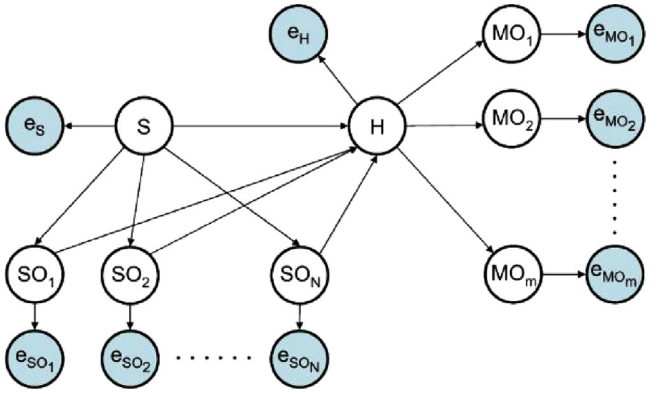

Fig. 10. The Bayesian model. The model consists of four types of nodes, corresponding to scene/event (S), scene objects (SO), manipulable objects (MO), and human (H). The observed (evidence) and hidden nodes are shown in blue and white, originally shown in [21]. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Some other approaches used the background information based on whole image classification. The extracted feature or resulted confidence measure from the background may be fused with other results at either the feature- or decision-level. For example, Prest et al. [13] extracted features from the whole image with GIST [61] for action image analysis. Sener et al. [42] used multiple features extracted from the whole image as the scene for action recognition.

Although the action context or scene is useful for action recognition, the image background may have negative effects on action analysis, especially when the background is too noisy and cluttered. Furthermore, different actions might be performed in the same or similar scene in which the context or scene may not provide helpful information to separate those actions.

2.1.6. A summary

We have presented various high-level cues used for still image based action recognition. The related papers and the applied cues in each paper are listed in Table 1. Since the study on still image based action recognition is still in the early stage, it is difficult to say which cues are more useful than others. Depending on how to "encode" the cues and the experimental databases, so far there is no cue which can always outperform others significantly. In practice, it is easier to use the cue of scene, especially using the whole image for action recognition (i.e., no separation between the foreground and background). In contrast, it is relatively difficult to use human body or object cues, since it is still challenging to detect the human body or objects automatically with a very high accuracy. That is why some public databases provide the manually annotated human body bounding boxes, assuming that they are available for action recognition. Finally, it is even harder to extract the cues from body parts or human object interactions, since it is very challenging to detect human body parts or the interactions with a good performance, although these cues could provide more detailed information for action analysis in still images. For example, the action of eating may involve the hand and the mouth, while taking photo may involve hand and eye (body parts), and a camera (i.e., interaction with object).

2.2. Low-level features

In previous subsection, we introduced various high-level cues for action analysis in still images. These high-level cues are usually characterized by using the low-level features. Different low-level features have been attempted in previous approaches. We will

Table 2
Low-level Features used in still image-based action recognition. See text for details.

Approach	DSIFT	HOG	SC	GIST	Other features
Wang et al. [16]			√		
Li and Fei-Fei [17]	√				
Thurau and Hlavac [18]		√			
Ikizler et al. [19]					Circular HOR
Ikizler-Cinbis et al. [20]		√			
Gupta et al. [21]		√	√	√	Color histogram, edge distance
Yao and Fei-Fei [22]	√				
Desai et al. [23]		√			
Yang et al. [24]		√			
Yao and Fei-Fei [25]			√		
Delaitre et al. [26]	√	√			
Shapovalova et al. [27]	√	√			
Maji et al. [28]		√			
Yao et al. [29]	√				
Yao et al. [30]		√			
Raja et al. [31]		√			
Komiusz and Mikolajczyk [32]	√				
Komiusz and Mikolajczyk [33]	√				
Li et al. [34]	√	√			
Yao et al. [35]	√	√			
Delaitre et al. [36]		√			
Li and Ma [37]	√	√		√	
Prest et al. [13]				√	SURF, color histogram
Yao and Fei-Fei [38]	√	√			
Sharma et al. [39]	√				
Yao and Fei-Fei [40]	√	√			
Desai and Ramanan [12]		√			
Kumar et al. [41]		√			
Sener et al. [42]	√	√			
Zheng et al. [43]	√	√			
Ikizler-Cinbis et al. [44]		√			
Le et al. [45]	√	√			RGB-SIFT, opponent SIFT
Khan et al. [14]	√				SIFT variants, RGB, HUE, Opp-angle, HS, C, CN
Sharma et al. [46]	√				

present the low-level features for still image-based action recognition in this subsection.

Typical low-level features include the dense sampling of scale invariant feature transform (DSIFT), histogram of oriented gradient (HOG), shape context (SC), GIST, or some other features. We summarize the low-level features in Table 2. One can see clearly which low-level features were used in which paper, and how many low-level features were used in each paper. From the table, we can also observe that the SIFT and HOG features have been used in most existing approaches to action recognition in still images.

2.2.1. *Scale-invariant feature transform (SIFT)*

A dense sampling of the gray scale images is often carried out to extract low-level features for action analysis, using the Scale-Invariant Feature Transform (SIFT) [62] method. It can be denoted by DSIFT. Local features can be described by the DSIFT descriptor in image regions or patches. The original SIFT algorithm was proposed by Lowe in 1999 [62], which can detect the interest point locations too. The SIFT feature has been applied in many problems, including object recognition, robotic mapping and navigation, image stitching, 3D modeling, gesture recognition, and video tracking. A 128 dimensional feature vector resulted in using the SIFT descriptor. In dense sampling of the SIFT feature, or DSIFT, a regular grid is used to "assign" interest point locations for feature extraction. Given the DSIFT features extracted from many image patches, a clustering is usually executed to obtain a limited number of "keywords" or "codebook", and the histogram can be computed and used as the feature for each image.

Many action recognition approaches have used the DSIFT based feature, including Li and Fei-Fei [17], Delaitre et al. [26], Yao and Fei-Fei [22], Shapovalova et al. [27], Yao et al. [29,35,38,40], Komiusz and Mikolajczyk [32,33], Li et al. [34,37], Delaitre et al. [36], Sharma et al. [39], Sener et al. [42], and Zheng et al. [43]. The DSIFT feature can be computed from the whole image or certain regions, such as human bounding box or detected object area. The computed DSIFT feature can be used as the input for direct classification of actions, or for high-level cue representation. In [14], Khan et al. also examined several color SIFT descriptors using different color channels.

2.2.2. *Histogram of oriented gradients (HOGs)*

Another often-used low-level feature descriptor is the Histogram of Oriented Gradients (HOGs) [63]. The HOG feature was originally proposed by Dalal and Triggs in 2005 for pedestrian detection [63]. Since then, the HOG feature has been used to solve many other computer vision problems, e.g., object detection, and human detection. The HOG feature counts occurrences of discretized gradient orientations within a local image patch, similar to the edge orientation histogram, SIFT, and shape context (SC).

The HOG feature was used frequently for still image-based action recognition, e.g., Thurau and Hlavac [18], Gupta et al. [21],

Desai et al. [23,12], Shapovalova et al. [27], Raja et al. [31], Delaitre et al. [36], Yao and Fei-Fei [38], Li and Ma [37], and Sener et al. [42].

2.2.3. Shape context (SC)

Shape context (SC) was proposed by Belongie and Malik in 2000 [64] for shape feature extraction for object matching. The SC can also be used for action recognition in still images. It can help to detect and segment the human contour. Some approaches have used the SC feature, such as Wang et al. [16], Gupta et al. [21], and Yao and Fei-Fei [25]. The usage of the SC feature is crucial for high-level cue representation of human body silhouettes for action recognition.

2.2.4. Spatial envelop or GIST

Spatial envelop or called GIST was proposed by Oliva and Torralba in 2001 [61]. A set of holistic, spatial properties of the scene can be computed by the GIST method. The GIST is an abstract representation of the scene that spontaneously activates memory representations of scene categories (a city, a mountain, etc.), as pointed out early in 1979 by Friedman [65]. The GIST feature is mainly used to integrate scene or background information. It has been used in Gupta et al. [21], Prest et al. [13], and Li and Ma [37] for action recognition in still images.

2.2.5. Other low-level features

In addition to the frequently used features mentioned above, there are also other low-level features for action recognition. The Speeded Up Robust Features (SURF) [66] feature is a scale- and rotation-invariant interest point detector and descriptor, proposed by Bay et al. in 2006 [66], partly inspired by the SIFT [62]. It has good performance in tasks such as object recognition or 3D reconstruction [66]. The SURF operator is based on computing the sums of 2D Haar wavelet responses using the integral images, and thus it can be computed efficiently. For action recognition, Prest et al. [13] showed the use of SURF to extract features from the candidate bounding boxes of action-related objects.

Circular Histogram of Oriented Rectangles (CHORs) were derived from the Histogram of Oriented Rectangles [67], and were used by Ikizler et al. [19] to represent the extracted human silhouettes for action recognition. Rectangular regions were searched over human silhouettes using convolution of a rectangular filter with different orientations and scales. Fig. 11 shows the CHOR-based pose representation for action recognition.

Besides using the HOG descriptor, Gupta et al. [21] also classified each image patch as belonging to one of the N candidates of scene object classes, using an AdaBoost classifier with color histogram (eight bins in each color channel), and histograms of edge distance map values within the neighborhood as the classifier inputs. The color histogram was also adopted in Prest et al. [13] to describe the appearance of candidate objects.

In [14], Khan et al. evaluated the performance of several pure color descriptors: RGB descriptor (RGB), C descriptor (C), Hue-Saturation descriptor (HS), Robust Hue descriptor (HUE), Opponent Derivative descriptor (OPP), Color Name (CN). Details of these color features can be found in [14]. They used the term "pure" to emphasize that these descriptors do not code any shape information from the local patch. Since no other previous methods emphasized the usage and effects of color in action images, they set up several experiments to exploit how to incorporate various color features.

3. Action learning

Given various image representations, either high-level cues or low-level features, the next step is to learn the actions from training examples. The learned models or classifiers can then be used to recognize actions from the unseen, test images.

Different learning methods have been proposed by researchers. We categorize the action learning methods into different categories, such as general models, discriminative learning, learning mid-level features, fusing multiple features, extracting spatial saliency, conditional random field, and pose matching. We will introduce various action learning methods under the seven categories in this section.

3.1. Generative models

Generative models usually learn the statistical distributions for action classes, which can randomly generate the observable data.

Li and Fei-Fei [17] proposed a generative model for event (action) recognition, incorporating the appearance and spatial information of the scene and the object. They took humans as a special kind of objects. Fig. 12 illustrates the graphical model representation of their approach. The parameters of the generative model include ψ, ρ, π, λ, θ, β in a hierarchical structure, as shown in Fig. 12. Given the event E, the scene and object images are assumed to be independent of each other, therefore the scene-related and object-related parameters can be learnt separately. Each object is represented by the most possible patches given the object, and the scene class label can be obtained based on the maximum likelihood estimation of the image features given the scene class.

As shown in Fig. 12, for scene recognition, each patch X only encodes appearance information. For object recognition, two types of information are obtained for each patch: the appearance information A, and the layout/geometry related information G. ψ is a multinomial parameter that governs the distribution of S

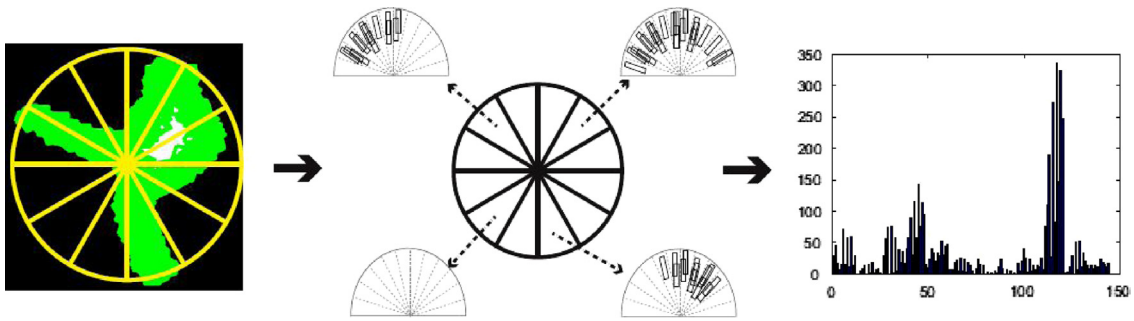

Fig. 11. Pose representation using Circular Histogram of Oriented Rectangle (CHOR), originally shown in [19]. The histogram of extracted rectangular regions is computed based on their orientations. For still images, the histogramming is over the spatial circular grids with circular HORs (CHORs), as opposed to the original $N \times N$ grid. This is mainly because it is difficult to know the explicit height of the human figure. Using circular grid helps to capture the angular positions of the parts more reliably. The center of the highest probabilistic region of the parse is used as the center of the circular grid. The bins of this circular histogram are 30° apart, making 12 bins in total.

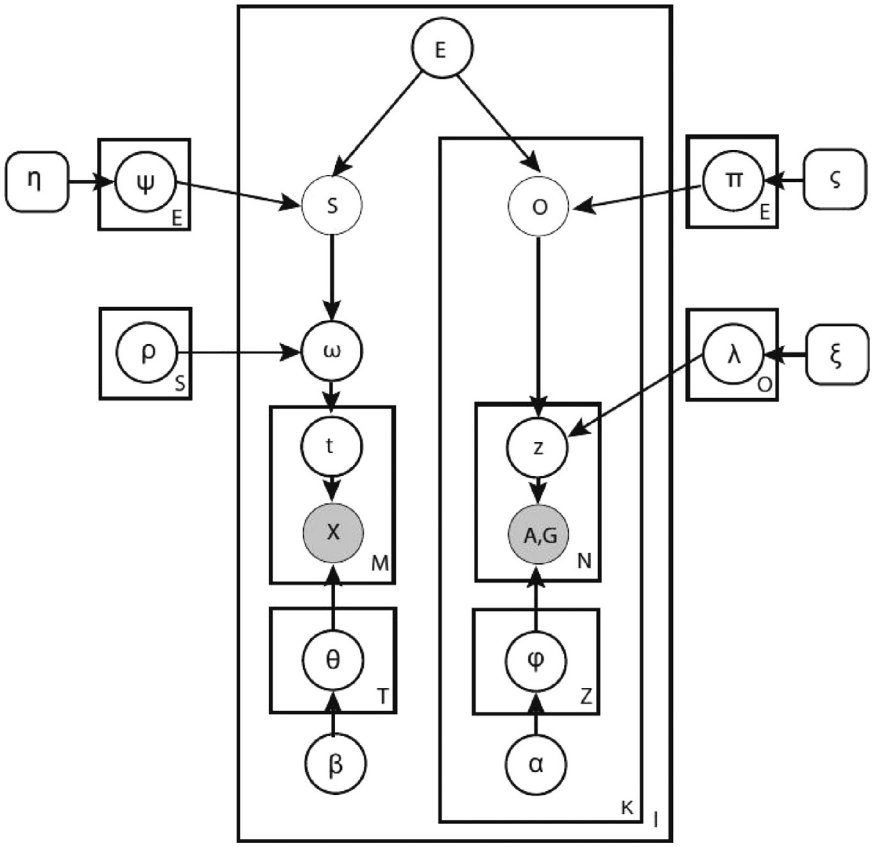

Fig. 12. Graphical model, originally shown in [17]. E, S, and O represent the event, scene, and object labels, respectively. An open node is a latent variable whereas a darkened node is observed during training. The lighter gray nodes (event, scene and object labels) are only observed during training whereas the darker gray nodes (image patches) are observed in both training and testing. The parameters either govern the distributions of certain topics or are Dirichlet priors to others.

given E. Given S, the mixing parameters ω govern the distribution of scene patch topics. Elements of ω sum to 1 as the multinomial parameter of the latent topics t, and t is a discrete variable indicating which latent topic this patch comes from. θ is the multinomial parameter for discrete variable X. A multinomial parameter π governs the distribution of O given E. z is a discrete variable indicating which latent topic this patch will come from, whereas λ is the multinomial parameter for z. φ is a multinomial parameter for discrete variable A or G, and η, β, ς, ξ, ρ, α all act as the Dirichlet priors for the parameters they point to.

However, an action is not defined by isolated scene or objects. Instead, the images usually present human as the action center, interacting with object and scene to some degree. A Bayesian model was used in Gupta et al. [21]. See Fig. 10 for an illustration. A fully supervised approach was taken to train the Bayesian model for image interpretation. The parameters were learned for individual likelihood functions and the conditional probabilities which model the interactions between the object and the scene. To learn parameters of individual likelihood functions, they trained individual detectors separately. Given the evidential variables or the observations, the goal is to estimate scene variable, scene objects, human, manipulable objects, using the loopy belief propagation algorithm for inference over the graphical model.

3.2. Discriminative learning

Discriminative learning is appropriate for distinguishing different action classes, without turning to learning the complex generative models.

Ikizler et al. [19] used the circular Histogram of Oriented Rectangles (CHORs) features extracted from human silhouettes of still images, and applied the Linear Discriminant Analysis (LDA). Then the one-vs-all SVM classifiers were trained for action classification.

Komiusz and Mikolajczyk [32,33] used the Kernel Discriminant Analysis (KDA) [68] classifier with χ^2 kernel, and linear SVM [69] for action recognition. Their feature descriptors contain spatial location information and the visual words are optimized with a smoothing factor for soft assignment. Delaitre et al. [26] applied the SVM too for action classification. They showed different performance when using different spatial pyramid levels, vocabulary sizes and classifier kernels.

Sharma et al. [46] proposed a model using a collection of discriminative templates with associated scale space locations. Image matching is a process of partially 'reconstructing' the important regions from the discriminative templates.

3.3. Learning mid-level features

Different from low-level features such as the SIFT and HOG some middle level features can be learned from the action images. Most of them are based on the extracted low-level features.

Yao et al. [22] proposed a mid-level feature named grouplet, using an AND/OR [70] structure on low-level features, which are computed from the SIFT [62] codebook over the dense grid. The feature unit, denoted by (A, x, σ), indicates that a codeword of visual appearance A is observed in the neighborhood of location x (relative to a reference point). The spatial extent of A in the neighborhood of x is expressed as a 2D Gaussian distribution $N(x,\sigma)$. Each feature unit captures a specific appearance, location, and spatial extent information of an image patch. Then given images with class labels, an expectation maximization (EM)

algorithm was used to estimate parameters, indicating the importance of each grouplet for each class, and the importance can be directly used for classification.

Delaitre et al. [36] built a new person–object interaction feature based on spatial co-occurrences of individual body parts and objects. See Fig. 9 for an illustration. Each pair of detectors constitutes a two-node tree where the position and the scale of the leaf nodes are related to the root by scale-space offset and a spatial deformation cost. Given a set of M discriminative interactions for each action class, and a scale-space pyramid with D cells, each image can be represented by concatenating M features from each of the K classes into a $M \times K \times D$-dimensional vector. A non-linear SVM with RBF kernel was used for action classification.

Maji et al. [28] learned 1200 action-specific poselets. Based on the assumption that if a pose is discriminative, there will be many examples of that poselet from the same action class, they measured the discriminativeness by the number of within class examples of the seed windows in the top k nearest examples to the poselet. The representative poselets for each action class are trained using HOG-based features and the SVM classifiers. Similar to poselet, Desai and Ramanan [12] generated new features called phrase-lets by clustering configurations of pose and nearby objects, following a supervised learning framework to learn parts and relations [71,72].

Yao et al. [35] defined action bases, consisting of action attributes as the verbs that describe the properties of human actions, as well as parts of actions which are objects and poselets [52] closely related to the actions. They modeled the attributes and parts jointly by learning a set of sparse action bases that are shown to carry much semantic meaning. See Fig. 5 for an illustration. A vector of the normalized confidence scores was obtained from the object and poselet classifiers or detectors. Then, the attributes and parts of an action image can be reconstructed from sparse coefficients with respect to the learned bases. The reconstruction coefficients of these bases are used to represent the image, and are fed into the SVM classifiers.

3.4. Multiple features fusion

Multiple features can be extracted to help improve the action recognition accuracy, in feature level (e.g., histogram concatenation) or score level. The assumption in fusion-based approaches is that multiple features may complement each other and a combination of them may characterize the actions better than each single feature. Thus the fusions of multiple features are expected to improve the action recognition accuracies. Some fusion methods were carefully investigated in [14] for different features, incorporating shape and color information.

Shapovalova et al. [27] performed a fusion in feature level, representing images with concatenated histograms to model pose, scene and object interactions. A human pose model H_P resulted from the concatenation of appearance and shape representations. The global scene of the image is represented by a histogram H_{BG}, which is a concatenation of histograms of spatial pyramid levels. Spatial human–object interactions are combined by two interaction models: (i) a local interaction model, which is a SIFT based BoW histogram calculated over the local neighborhood around the bounding box, and (ii) a global interaction model (see Fig. 6). Then these fused features are classified using an intersection-kernel based SVM classifier.

Prest et al. [13] extracted three descriptors: human–object interaction, whole image, and human pose cues. They used a separate RBF kernel for each descriptor and computed a linear combination of them. A multi-kernel one-vs-all SVM classifier was learned.

Sener et al. [42] used multiple features, such as representative object regions and detected faces extracted from the whole image or image patches. The SVM classifiers are then used. The score level fusion is used to combine each confidence measure to make the final decision of action labels.

In [14], Khan et al. exploited the efficiency of color in still image recognition. They examined various color descriptors for representing action images. Then they tried different fusions of shape and color features for both action classification and detection.

Zheng et al. [43] combined poselet and context based classifier confidence for action analysis, emphasizing the equal importance of both foreground and background in action images.

3.5. Spatial saliency

Sharma et al. [39] defined image saliency as a mapping $s : G \rightarrow R$, where G is a spatial pyramid like uniform grid [73] of image, $c \in G$ is a region of the image, and $s(c)$ gives the saliency of the region. They proposed a model consisting of three components: (i) the separating hyperplane w, (ii) the image saliency maps s^i, and (iii) a generic saliency map $\bar{s}$ to regularize the image saliency maps. The saliency map of an image maximizes the classification score while penalizing the deviation from the generic saliency map. See Fig. 13 for some illustrations. The images were represented by concatenation of cells of bag-of-features, weighted by the image saliency maps. Using a latent SVM, they optimized iteratively the hyperplane vector w while keeping the saliency maps of the positive images fixed, and the saliency map while keeping w fixed.

The random forests with discriminative decision tree nodes were used in Yao et al. [29], to find the most discriminative image patches for action recognition. The discriminative patches can be considered as another kind of saliency extraction.

3.6. Conditional random field

Yao and Fei-Fei [25] used a conditional random field (CRF) model for action analysis. As illustrated in Fig. 14, the inference procedure is the following: to detect the tennis racket in the image, the likelihood of the image is maximized, given the models learned for tennis-forehand. This is achieved by finding a best configuration of human body parts and the object (tennis racket) in the image, which is denoted as $\max_{O,H} \Psi - (A_k, O, H, I)$ in the figure where A indicates action, H is the pose, O is the object, and I is the image. In order to estimate the human pose, they computed $\max_{O,H} \Psi - (A_k, O, H, I)$ for each activity class and find the class A^* that corresponds to the maximum likelihood score. This score can be used to measure the confidence of activity classification as well as human pose estimation.

Later, Yao and Fei-Fei [38,30] extended their model by introducing a set of atomic poses. They learned an overall relationship among different activities, objects, and human poses, rather than modeling the human–object interactions for each activity separately as in [25]. Instead of limiting to one human and one object interaction [25], the extended model can deal with the human interactions with any number of objects. The new model [38,30] incorporates a discriminative action classification component and uses the state-of-the-art object and body part detectors, which further improves the recognition accuracy.

3.7. Pose matching

Some approaches to action recognition are mainly based on matching human body poses. The matching scheme is especially to exploit body shape and pose information. From a sketch of human body poses, it was assumed that there is a great similarity among intra-class poses and the matching of poses can recognize

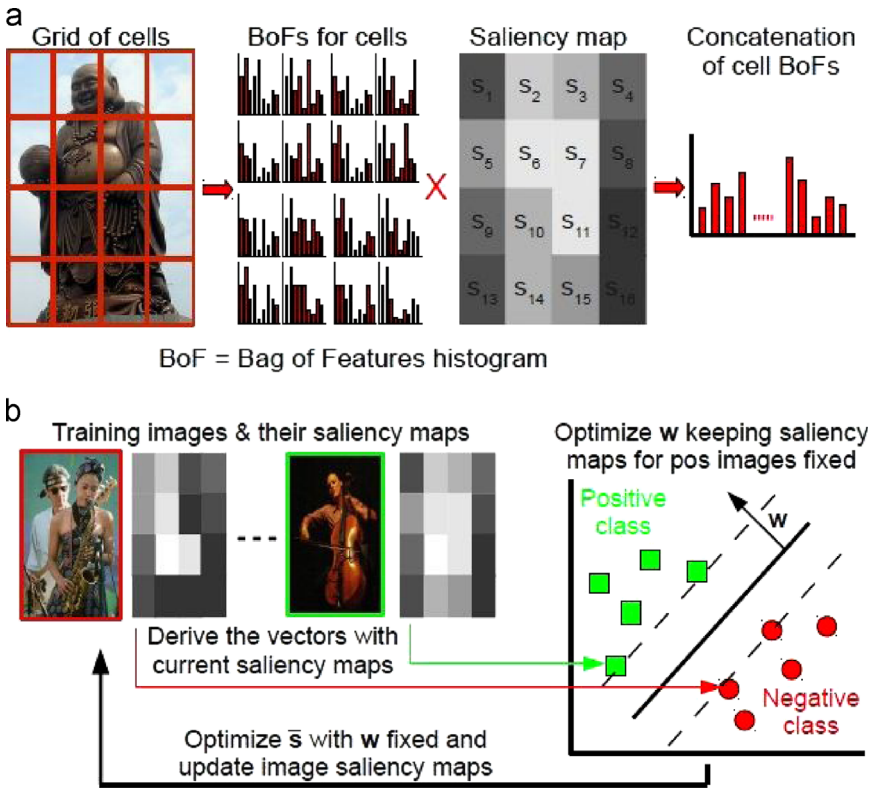

Fig. 13. Saliency representation (a) and training with saliency maps (b), originally shown in [39]. The images are represented by concatenation of cell bag-of-features weighted by the image saliency maps. A block coordinate descent algorithm is used to learn the model (Section 2.4 from [39]). Using a latent SVM, they optimized iteratively the hyperplane vector w while keeping the saliency maps of the positive images fixed, and optimized the saliency while keeping w fixed.

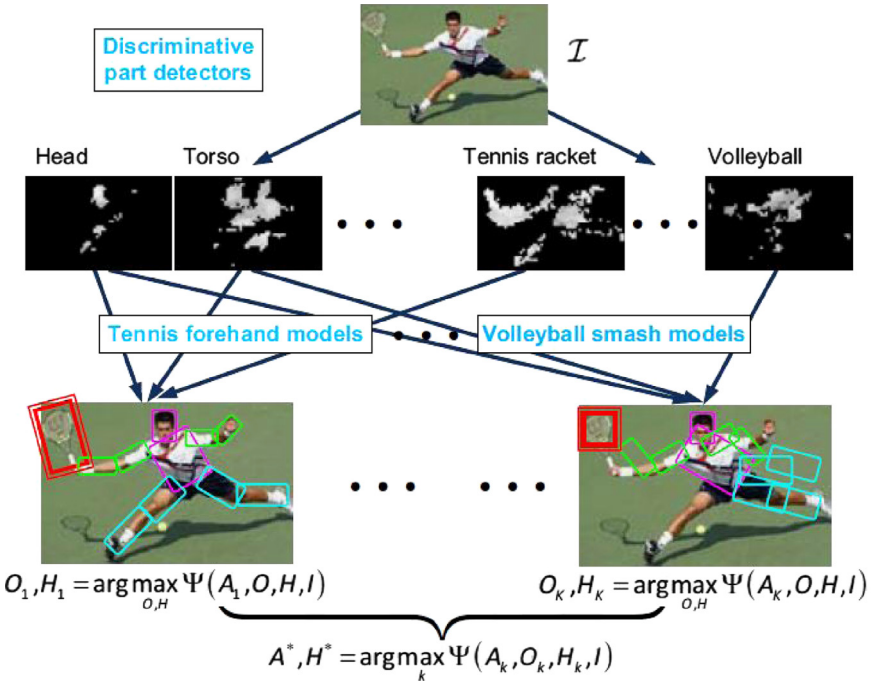

Fig. 14. The framework of the inference method, originally shown in [25]. Given an input image I, the inference results are the following: (1) object detection results O_k (e.g., O_1 is the tennis racket detection result); (2) human pose estimation result H^*; (3) activity classification result A^*.

actions. To deal with variations of intra-class poses, multiple poses can be used to represent a certain class [40].

Wang et al. [16] performed spectral clustering [74] using the sparse matrix of specific distances between pairs of images. For n images, an n-by-n affinity matrix W is constructed, where W_{ij} is the distance between images i and j. Given these manually label-assigned clusters of actions, prototypes of these clusters can be used to classify new images according to these labels. For each new test image, they matched it to one of the clusters using the k nearest neighbor classification.

Thurau and Hlavac [18] defined pose primitives in each image as the unique cues for a specific action. They used a standard agglomerative clustering method [75] on all training pose primitives, action recognition is then achieved on a single recognized pose primitive, or a sequence of recognized poses, with a histogram comparison inspired by [76].

Yao and Fei-Fei [40] used a 2.5D exemplar-based action classification approach, where a set of representative images was selected for each action class. The selected images cover large within-class variations and carry discriminative information with the 2.5D exemplars. They constructed the 2.5D graph for each training action image, consisting of a set of nodes that are keypoints of the human body, as well as a set of edges that are spatial relationships between nodes. To formally define the dominate images of human actions, they defined the coverage set of an image I, $Cov(I)$, belonging to the same class as I, with a larger similarity value than all images in other classes. The problem is essentially a minimum dominating set problem [77], and can be solved by using an improved reverse heuristic algorithm [78].

4. Databases

There are many public datasets available to validate different methods for action recognition in still images. We present the widely used datasets and categorize them into different categories. The statistics of all datasets are shown in Table 3 with other information as well, e.g., the source of the datasets, and which papers conducted experiments on each dataset.

4.1. Sports related action databases

In collecting action databases, action images in sports are the earliest and of the most popular usage for recognition, probably due to the relatively small human pose variations within the same actions in sports activity, and the distinctiveness and uniqueness of specific sports actions in single images.

4.1.1. The sports dataset

The sports dataset first used in [21] has six actions: tennis-forehand, tennis-serve, volleyball-smash, cricket-defensive shot, cricket-bowling and croquet-shot. The authors illustrated that the changes between actions are mainly on poses rather than the object or scene. Each action class contains 20–30 training images and testing images. The classes were selected so that there are significant confusions due to the same scene and similar poses. For example, the poses during volleyball-smash and tennis-serve are quite similar and the scenes in tennis-forehand and tennis-serve are exactly the same. This is a widely used dataset for human action recognition involving objects. Experimental results could be found in [21,23,26,25,30,13,38] using this dataset. The accuracies have rose from 78% [21] to 87% [30], as shown in Fig. 15.

4.1.2. Skating dataset

Wang et al. assembled three datasets in 2006 [16]. The first dataset is a collection of images from six videos of different figure skaters. These videos were automatically filtered. Frames with complicated backgrounds (consisting of a large number of edges) were removed, resulting in a simplified set of 1400 images. The figure skating clusters were given the following 10 labels: face close-up picture, skates with arms down, skates with one arm out, skater leans to his right, skates with both arms out, skates on one leg, sit spin leg to left of the image, sit spin leg to right, camel spin leg to left, and camel spin leg to right.

4.1.3. Baseball dataset

Wang et al. [16] collected the baseball clusters consisting of 4500 images, which were collected by querying the captions of sports news photos for professional sports team names. These datasets are significantly more challenging, containing substantial background clutter, and a wide range of content. The baseball clusters have seven labels: face close-up picture, right-handed pitcher throws, right-handed pitcher cocks his arm to throw, runner slides into base, team celebrates, batter swings, batter finished swinging.

4.1.4. Basketball dataset

The basketball clusters were also collected by Wang et al. [16] with 8500 images. There are eight labels: a player goes for a lay-up above the defenders, a player goes for a lay-up against a defender, a player goes for a jumpshot while another one tries to block, a player goes for a lay-up leaning to his right, a player drives past another, a player has his shot blocked, a player leaps by his defender for a shot, and a player posts up.

4.1.5. Sports events dataset

Li et al. compiled a dataset [17], containing eight sports event categories collected from the Internet: bocce, croquet, polo,

Table 3
Databases assembled for still image-based action recognition.

Dataset	# of images	# of classes	Source	Used in papers
The Sports Dataset [21]	300 in total	6	Internet	[21,26,23,25,30,38,13]
*Skating dataset [16]	1400 in total	10	Videos	[16,19]
*Baseball dataset [16]	4500 in total	7	Sports news	[16]
*Basketball dataset [16]	8500 in total	8	Sports news	[16]
*Sports Events [17]	137 to 250 per class	8	Internet	[17]
Pascal VOC 2010 [9]	50 to 100 per class	9	Internet	[35,28,29,27,36,32,33,13,43,14,12]
Pascal VOC 2011 [9]	200 or more per class	10	Internet	[40,12,41,43]
Pascal VOC 2012 [9]	400 or more per class	10	Internet	
Stanford 40 Actions [35]	180 to 300 per class	40	Internet	[35,14,46,42]
Willow Dataset [26]	968 in total	7	Internet	[26,36,39,43,14,46]
89 Action Dataset [45]	2038 in total	89	Pascal 2012 trainval set	[45]
*Action Images by Ikizler [19]	467 in total	6	Google Images, Flickr, BBC Motion database, etc.	[19]
*Retrieved Web Images [20]	2458 in total	5	Internet	[20,24]
*Action Images by Li [34]	400 per class	6	Internet and Pascal VOC 2010	[34,37]
TBH [13]	341 in total	3	Google Images and the IAPR TC-12 dataset [79]	[13]
Weizmann dataset [18]	–	10	Videos	[18]
KTH dataset [31]	789 in total	6	Videos	[31]
PPMI [22]	300 per class	7	Internet	[22,26,39,38,29,40]

*Dataset names are given by us.

rowing, snowboarding, badminton, sailing, and 5 rock climbing. The number of images in each category varies from 137 (bocce) to 250 (rowing). They had also obtained a thorough ground truth annotation for every image in the dataset. This annotation provides information for event class, background scene class(es), most discernable object classes, and detailed segmentation of each objects. For each event class in their experiments, 70 randomly selected images were used for training and 60 for testing.

4.2. Daily activity databases

Datasets in this category contain common activities performed by humans in daily lives, which have less controversy than other categories of databases.

4.2.1. Pascal VOC action datasets

Pascal VOC competition includes still image-based action recognition starting from 2010. There are nine actions: phoning, playing a musical instrument, reading, riding a bicycle or motorcycle, riding a horse, running, taking a photograph, using a computer, or walking. Only subset of people are annotated (bounding box of the human + action). All people in dataset are labelled with exactly one action class. Results reported on VOC 2010 could be found in [35,28,29,27,36,32,33,13,43,14,12] and visualized in Fig. 17.

Later in 2011, this dataset was extended about five times larger in size, and one more action called jumping was added to the original 2010 dataset. There is a minimum of around 200 people per action category. Actions are not mutually exclusive, which means that there could be one person with more than one action labels in the same image. Besides these changes, training and test images belonging to 'other' action class were collected in the dataset, increasing the difficulty in action analysis.

In 2012, the dataset was expanded again: about 90% increase in size over VOC 2011. There is a minimum of around 400 people per action category. A single point located somewhere in the human body was also annotated for each image.

4.2.2. Stanford 40 actions dataset

Yao et al. [35] collected a challenging, large scale dataset, called Stanford 40 Actions, containing 40 diverse daily human actions, such as brushing teeth, cleaning the floor, reading book, and throwing a frisbee. All images were obtained from Google, Bing, and Flickr. 180–300 images were collected for each class. There are 9352 images in total. They provided bounding boxes for the humans who are doing one of the 40 actions in each image. The authors randomly selected 100 images in each class for training, and the remaining for testing.

4.2.3. Willow dataset

Delaitre et al. [26] collected the willow action dataset from original consumer photographs, depicting seven common human actions: interacting with computers, photographing, playing a musical instrument, riding bike, riding horse, running and walking. Images for the riding bike action were taken from the Pascal 2007 VOC Challenge and the remaining images were collected from Flickr by querying on keywords such as running people or playing piano, resulting in a total of 968 photographs with at least 108 images for each class. They split the dataset into two parts: 70 images per class for training and the remaining for testing. Each image was manually annotated with bounding boxes indicating the locations of people.

4.2.4. The 89 action dataset

Le et al. [45] assembled a dataset from 11,500 images of the PASCAL 2012 VOC trainval set [9], selecting all those images representing a human action, resulting in 2038 images. They manually annotated these images with a verb to obtain the label of the human action (verb–object). The dataset was annotated with 19 objects and 36 verbs, which are combined to form 89 actions. Similar to the training vs. validation split used in the PASCAL competition, their human action dataset consists of 1104 images for training and 934 images for validation.

4.2.5. Action images by Ikizler

Ikizler et al. [19] built a dataset from various sources like Google Image Search, Flickr, and BBC Motion database. This dataset consists of 467 images and includes six different actions: running, walking, catching, throwing, crouching and kicking.

4.2.6. Retrieved web images

Ikizler-Cinbis et al. [20] retrieved images from the web. Part of their works were done on refining the initial results of keyword queries to an image search engine. In the end, there are 2458 images in total in the dataset, containing 384 running, 307 walking, 313 sitting, 162 playing golf, and 561 dancing images.

4.2.7. Action images by Li

Li et al. [34,37] collected about 2400 images for six action queries: phoning, playing guitar, riding bike, riding horse, running and shooting. Most of the images were collected from Google Image, Bing and Flickr, and others are from PASCAL VOC 2010 [9]. Each action class contains about 400 images.

4.2.8. TBH dataset

Prest et al. [13] introduced an action dataset called TBH. It was built from Google Images and the IAPR TC-12 dataset, containing three actions: playing trumpet, riding bike,and wearing hat. Split 100 positive images into training (60) and testing (40) for playing trumpet class. For the actions of riding bike and wearing hat, images from the IAPR TC-12 dataset were used. The dataset contains 117 images for riding bike (70 training, 47 testing) and 124 images for wearing hat (74 training, 50 testing). Images were only annotated with the action class labels.

4.3. Frames from action videos

Still images may also be extracted from some action videos. The extracted image frames usually have a relative static or cleaner background.

4.3.1. Weizmann dataset

Thurau and Hlavac [18] used still images extracted from the popular Weizmann action videos [80]. The dataset contains 10 different actions: bend, jack, side, skip, run, pjump, jump, walk, wave1, and wave2, performed by nine subjects.

4.3.2. KTH dataset

Raja et al. [31] executed human pose estimation and action recognition in still image frames extracted from the KTH dataset [81]. The dataset contains images of six classes: boxing, handclapping, hand-waving, jogging, running and walking. The training and test sets were separated by the identities, containing 461 and 328 cropped images, respectively.

4.4. Music instruments action dataset

Yao and Fei-Fei [22] assembled a dataset called the People-playing-musical-instruments (PPMI). The PPMI consists of seven different musical instruments: bassoon, erhu, flute, French horn, guitar, saxophone, and violin. Each class includes 150 PPMI+ images (humans playing instruments) and 150 PPMI− images (humans holding the instruments without playing).

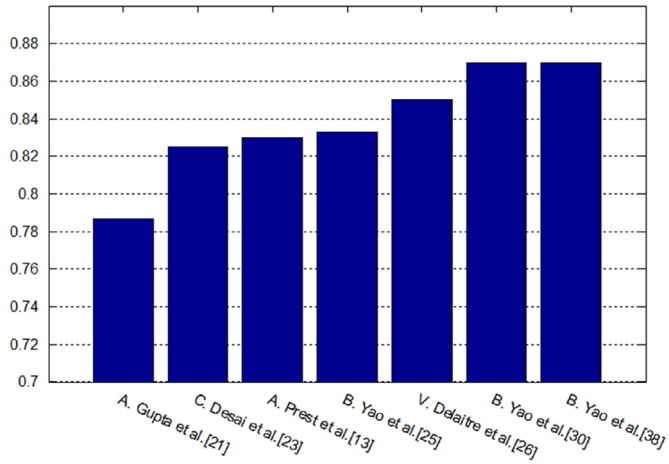

Fig. 15. Performance comparison of different methods on the Sports dataset [21].

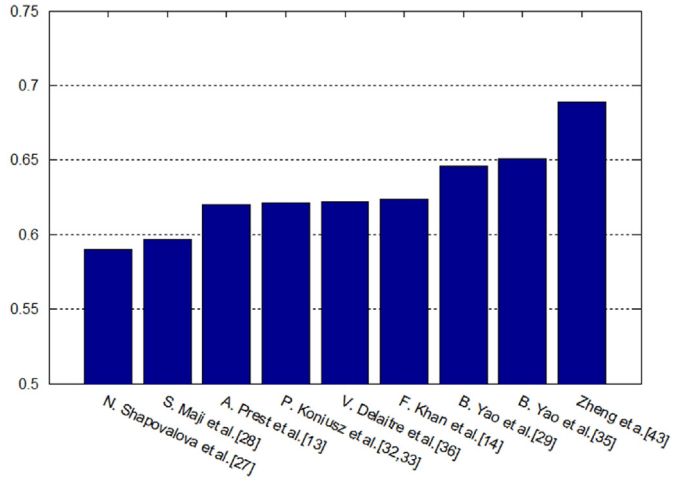

Fig. 16. Performance comparison of different methods on Pascal VOC 2010 Action Dataset [9].

5. Action recognition performance

To understand the current status of still image-based action recognition, we present the action recognition accuracies obtained in previous approaches. We select to present the reported recognition results on the most popular action databases, which can be found from Table 3. They are the Sports dataset [21] and the Pascal VOC 2010 Action Dataset [9]. The recognition accuracies of different methods on these two databases are shown in Figs. 15–17. VOC datasets set up a new measurement of mean Average Precision (mAP), which is a calculation of the area under the precision–recall curve. This is different from the traditional accuracy measure, somehow becoming a standard for performance evaluation in action recognition.

We can observe that the recognition accuracies are in the range from 78% to 87% on the Sports Dataset, and from about 59% to 68% on Pascal VOC 2010 Action Dataset. The accuracies are improving with the research advances, but the improvements are not big, e.g., less than 10%. This indicates that new approaches are still demanding to make more significant progresses. Hope our review of the existing works can inspire new thoughts and efforts.

6. Relation to other research topics

Still image-based action recognition is not an isolated topic. It is closely related to some other research problems, such as video based action recognition, object recognition, scene recognition, image retrieval and pose estimation. See Fig. 18 for a visualization of the relations. The figure illustrates that object recognition can help still image-based action recognition, while action recognition in still images can help image retrieval and video-based action recognition. Further, still image-based action recognition and scene recognition can help each other, i.e., a mutual help relation. This mutual help relation also holds for pose estimation and action recognition in still images. More details about these relations are presented in the following subsections.

The progress in those related topics may help to enhance action recognition in still images. On the other hand, action recognition

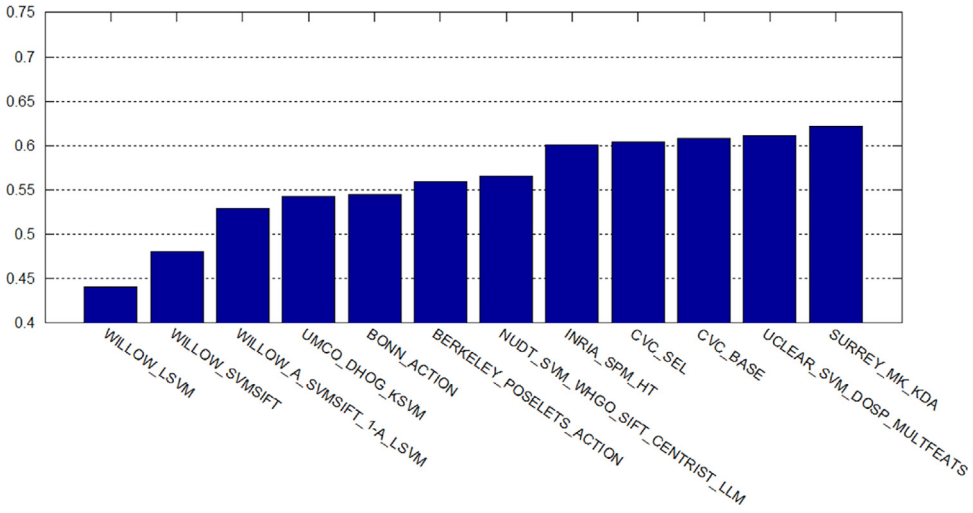

Fig. 17. Competition results from different research groups on Pascal VOC 2010 Action Dataset. Details of the research groups can be found from [9], while most of these results do not have formal papers published.

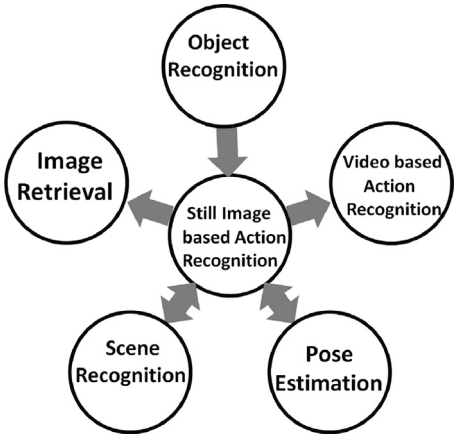

Fig. 18. Other topics related to still image-based action recognition. "A ⇒ B" means that task A can help B.

can help other research problems by providing refined semantic meaning in images, e.g., what the persons are doing in the still images to be annotated.

6.1. Video based action recognition

Ignoring the motion cue, video based action recognition could be done by analyzing individual frames, i.e., through performing image-based action recognition. Thurau and Hlavac [18] got training data from video frames and their method was applied to video-based action recognition on the KTH dataset [81]. In their approach, actions were represented by histograms of pose primitives, without using motion information from video sequences. Their work indicates that still image based action recognition can be applicable to video-based action recognition. Another work is by Ikizler-Cinbis et al. [20,44], where the single images collected from the Internet were used to learn image representations of actions, and then the learned classifiers were used for action recognition in videos.

Therefore, the developed techniques on image-based action recognition could be useful for video-based action recognition [24,21]. Based on this, the number of frames could be reduced significantly in a video sequence for the purpose of action recognition. An interesting question is, what kinds of actions are appropriate to use still image-based techniques, and what actions need to use spatiotemporal features in videos for recognition. By addressing this question, we can understand the applicability of still image based methods to video-based action recognition deeply and specifically.

6.2. Object recognition

In performing various actions, humans often interact with objects. For example, "playing cello" has a cello object involved and "brushing hair" has a comb object. So it is not difficult to understand that recognizing the corresponding object will help a lot for still image-based action recognition. Further, the detected and recognized object can help model the interaction between human and object, which is an important component for action recognition in still images [13].

Object detection and recognition are active research topics in computer vision, however, in the context of action recognition, probably there is a difference from the traditional object detection and recognition [62,59]. In still image-based action recognition, typically the human (body or face, see some examples in Fig. 1) is the main focus of attention, which occupies a larger part in still images than the involved object, while in traditional object recognition, usually the objects are the main focus and the images do not necessarily contain humans. In this sense, it might be more challenging to detect and recognize objects in still images for the task of action recognition.

6.3. Pose estimation

Pose estimation is an important problem in computer vision by itself [53]. The estimated body pose can be useful for still image-based action recognition [28,31,13,12]. It is intuitive that human actions are often associated with articulated body poses. For instance, the actions of "taking photos" and "playing piano" have very different body poses. Thus the progress and performance improvement in pose estimation can be very beneficial to still image-based action recognition.

On the other hand, the potential action categories could be used to constrain the search space in pose estimation from still images. For example, the body poses of "eating" and "typing" are different. When an action is recognized as "eating," the corresponding body pose should be different from the pose of "typing." The information of action categories can be utilized to reduce the space of possible candidate poses. It is even better by performing joint pose estimation and action recognition [31], since the two tasks can be beneficial to each other.

6.4. Scene recognition

Usually scene recognition can provide the contextual information for action recognition, i.e., where the action takes place. For example, the office scene may be related to actions such as reading, writing, and using computer. So scene understanding can provide useful cues for action recognition in still images.

Sharma et al. [39] performed action recognition in a scene dataset [73], which contains 15 scene categories, e.g., beach and office. The recognized actions can help scene understanding. For instance, driving is often in a traffic scene rather than office. Therefore scene understanding and action recognition can be mutually helpful in image analysis.

6.5. Image retrieval

In Li et al. [34], action recognition was conducted on the web images retrieved by text-based query. Typically, still image based action recognition can be used to annotate the action related keywords, e.g., eating, driving, playing music instrument, for a given image. So we can say that still image based action recognition can benefit image retrieval especially on images distributed over the Internet, which may lack text annotations from the users.

7. Some thoughts of future research

Based on our overview in previous sections, one can see that significant progresses have been made in still image-based action recognition. However, the field of research is still in its early stage. Deeper studies of methodologies are expected to make a breakthrough in the area. Here we present our own views and thoughts, and hope these can inspire new research efforts.

(1) How many action classes can be collected for still image-based action recognition? In Section 4 and Table 3, we introduced a list of databases for action recognition in still images. Most of the databases contain about 10 action classes or less, while two databases have 40 or up to 89 action classes. So the question is, how many action classes can be collected? What is the maximum number of action classes in reality? We believe that 89 is not the maximum number of action classes. As a multi-class classification

problem, the number of classes does matter in evaluating different methods. Suppose we know the maximum number of action classes, denoted by M, then a unique benchmark dataset might be built, and all future developed methods can use the same database for validation and comparisons. Some related issues include the following: among the M total classes, which action classes are the most difficult to separate from others? Does there exist any "easy" or "hard" actions to recognize, among the M classes in total?

In determining the total number of action classes, the vocabulary of verbs in the large lexical database called WordNet (http://wordnet.princeton.edu/) could provide some guidance on identifying the potential action categories, with consideration of the availability of corresponding images from the Internet.

(2) How many cues can be found for still image-based action recognition? In Section 2.1, we presented various high-level cues for action analysis in still images, including human body, body parts, object, human object interaction, and the scene or context. A question is, can we find some new cues to enhance action recognition? If yes, what are those new cues? How to represent them? By addressing these questions, one may develop new approaches to improve the action recognition performance.

(3) What features are appropriate for image-based action recognition? In Section 2.2, we presented various low-level features for high-level cues representation and action recognition. Almost all those features were originally developed for other computer vision problems. A question may be asked: can we have some "special" features that are unique for action representation in still images? There are some existing approaches to learn mid-level features, as we introduced in Section 3.3. However, new features are expected to target the action patterns specifically, and develop a better representation than the current features.

(4) How to combine action recognition with other research problems? In traditional video-based action recognition, the problem is often taken as an independent one. While in still image-based action recognition, it usually needs the cues about human, objects, human body pose, human object interaction, and so on. The research progresses in related areas, as we discussed in Sections 2.1 and 6, will definitely help to improve still image-based action recognition. For instance, the human bounding boxes are currently provided in many action databases based on manual labeling. It will be nice if an automated human detection can achieve a high accuracy, e.g., above 95%, in action images. On the other hand, action recognition in still images might help to enrich solving other problems. For example, automated action analysis in still images can "tag" online images with the performed actions for image search or retrieval. This will make the action recognition more interesting, and may bring more attention to researchers in image or multimedia retrieval community.

(5) How to solve the occlusion problem in action recognition? As we presented in Section 2, many high-level cues and low-level features are usually needed for action analysis. Sometimes, occlusions could be serious. For instance, the human body may occlude objects (fully or partially) or be occluded partially by the objects, or human body parts may self-occlude each other. The occlusions may cause it difficult to extract the related high-level cues or low-level features. There are some approaches, e.g., [12], using a visibility flag to indicate a particular body part being occluded or not. However, further efforts are needed to deal with the occlusions, making the computational approach insensitive to full or partial occlusions.

(6) How to do action learning in the small sample case? As we discussed in Section 4, some of the existing databases have more image examples in each class, while some others have less. The Pascal VOC databases are continuously increasing the number of image examples for each class in each year. So some questions can be raised: How many examples are needed to learn the action classes? Is there a minimum number of training examples for each action class? Do we have enough training examples to represent all possible variations for each action class? If the number of training examples is too small, compared to all possible variations, we have the problem of learning in the small sample case. How to develop robust methods to learn actions with small samples?

In order to study these problems and get statistically meaningful results, researchers need to perform a comprehensive, empirical investigation of the action recognition performance with respect to the number of training samples, using sufficiently large databases.

(7) Which actions are appropriate to use videos and to use still images? There are many action databases, either videos or still images. If we look at the specific actions in these two types of action databases, we can find that some actions appear in both videos and still images, such as walking and running. There are also actions that appear in still image databases but not in videos, or vice versa. We may ask a question: Which actions are appropriate to use videos for representation, and which are proper to use still images? If an action can be characterized completely by a still image, probably there is no need to capture and store in a video. Even stored in a video, one may use a small number of image frames for analysis of that action which is appropriate to use still images. On the other hand, if some actions need to use video data for a better representation, we may not expect a good performance for those actions using still images, and we know why. As a result, researchers will not waste time to develop new algorithms to improve the performance for those actions appeared in still images.

(8) Are discriminative methods good enough to separate action classes? Or are generative models better for certain actions? For action learning, there are both general and discriminative approaches. We may ask questions: (1) Which learning method is better for action recognition on standard databases: discriminative or generative? (2) Does discriminative learning perform better than the generative for some actions, but worse for some others? Based on these studies, one may find the appropriate learning methods for action recognition in still images.

To address the related questions, researchers may perform a comprehensive evaluation of representative methods for discriminative and generative learning of actions, using a commonly adopted database. The focus is to study the difference of recognition performance for each specific action.

8. Conclusions

We have conducted a comprehensive survey of existing approaches on still image-based action recognition. We have introduced different approaches based on a categorization of them with high-level cues and low-level features. Different action learning methods have been discussed too. Various action databases are grouped and summarized with specific details. We have also presented some research topics that are related to action analysis in still images, and given some thoughts for future research. As a relatively new area, the research on still image-based action recognition is in the early stage. The recognition accuracies are not high based on examining the results on several often-used databases. Hope our survey can motivate some new research efforts to advance the field of research on human action analysis using still images.

Conflict of Interest

None declared.

Acknowledgments

The authors thank the anonymous reviewers for their helpful suggestions and comments to improve the paper.

References

[1] C. Cedras, M. Shah, Motion-based recognition a survey, Image Vis. Comput. 13 (1995) 129–155.
[2] J. Aggarwal, Q. Cai, Human motion analysis: a review, Comput. Vis. Image Underst. 73 (1999) 428–440.
[3] D. Gavrila, The visual analysis of human movement: a survey, Comput. Vis. Image Underst. 73 (1999) 82–98.
[4] V. Kruger, D. Kragic, A. Ude, C. Geib, The meaning of action: a review on action recognition and mapping, Adv. Robot. 21 (2007) 1473–1501.
[5] P. Turaga, R. Chellappa, V. Subrahmanian, O. Udrea, Machine recognition of human activities: a survey, IEEE Trans. Circuits Syst. Video Technol. 18 (2008) 1473–1488.
[6] R. Poppe, A survey on vision-based human action recognition, Image Vis. Comput. 28 (2010) 976–990.
[7] X. Ji, H. Liu, Advances in view-invariant human motion analysis: a review, IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 40 (2010) 13–24.
[8] J. Aggarwal, M. Ryoo, Human activity analysis: a review, ACM Comput. Surv. 43 (2011) 16.
[9] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The Pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2010) 303–338.
[10] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recognit. 36 (2003) 259–275.
[11] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 39–58.
[12] C. Desai, D. Ramanan, Detecting actions, poses, and objects with relational phraselets, in: European Conference on Computer Vision Workshops and Demonstrations, Springer-Verlag, 2012, pp. 158–172.
[13] A. Prest, C. Schmid, V. Ferrari, Weakly supervised learning of interactions between humans and objects, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 601–614.
[14] F.S. Khan, M.A. Rao, J. van de Weijer, A.D. Bagdanov, A. Lopez, M. Felsberg, Coloring action recognititon in still images, Int. J. Comput. Vis. (2013) 1–17.
[15] I. Laptev, On space-time interest points, Int. J. Comput. Vis. 64 (2005) 107–123.
[16] Y. Wang, H. Jiang, M. Drew, Z.-N. Li, G. Mori, Unsupervised discovery of action classes, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2006, pp. 1654–1661.
[17] L.-J. Li, L. Fei-Fei, What, where and who? classifying events by scene and object recognition, in: IEEE Conference on Computer Vision, IEEE, 2007, pp. 1–8.
[18] C. Thurau, V. Hlavac, Pose primitive based human action recognition in videos or still images, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
[19] N. Ikizler, R. Cinbis, S. Pehlivan, P. Duygulu, Recognizing actions from still images, in: International Conference on Pattern Recognition, IEEE, 2008, pp. 1–4.
[20] N. Ikizler-Cinbis, R. G. Cinbis, S. Sclaroff, Learning actions from the Web, in: IEEE International Conference on Computer Vision, IEEE, 2009, pp. 995–1002.
[21] A. Gupta, A. Kembhavi, L. Davis, Observing human-object interactions: using spatial and functional compatibility for recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 1775–1789.
[22] B. Yao, L. Fei-Fei, Grouplet: A structured image representation for recognizing human and object interactions, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, San Francisco, USA, 2010, pp. 9–16.
[23] C. Desai, D. Ramanan, C. Fowlkes, Discriminative models for static human-object interactions, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2010, pp. 9–16.
[24] W. Yang, Y. Wang, G. Mori, Recognizing human actions from still images with latent poses, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 2030–2037.
[25] B. Yao, L. Fei-Fei, Modeling mutual context of object and human pose in human-object interaction activities, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 17–24.
[26] V. Delaitre, I. Laptev, J. Sivic, Recognizing human actions in still images: a study of bag-of-features and part-based representations, in: Proceedings of the British Machine Vision Conference, BMVA, 2010, p. 7.
[27] N. Shapovalova, W. Gong, M. Pedersoli, F. X. Roca, J. Gonzalez, On importance of interactions and context in human action recognition, in: Pattern Recognition and Image Analysis: 5th Iberian Conference, vol. 6669, Springer, 2011, p. 58.
[28] S. Maji, L. Bourdev, J. Malik, Action recognition from a distributed representation of pose and appearance, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 3177–3184.
[29] B. Yao, A. Khosla, L. Fei-Fei, Combining randomization and discrimination for fine-grained image categorization, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 1577–1584.
[30] B. Yao, A. Khosla, L. Fei-Fei, Classifying actions and measuring action similarity by modeling the mutual context of objects and human poses, in: International Conference on Machine Learning, Omnipress, Bellevue, USA, 2011, p. D3.
[31] K. Raja, I. Laptev, P. Pérez, L. Oisel, Joint pose estimation and action recognition in image graphs, in: IEEE International Conference on Image Processing, IEEE, 2011, pp. 25–28.
[32] P. Koniusz, K. Mikolajczyk, Soft assignment of visual words as linear coordinate coding and optimisation of its reconstruction error, in: IEEE International Conference on Image Processing, IEEE, 2011, pp. 2413–2416.
[33] P. Koniusz, K. Mikolajczyk, Spatial coordinate coding to reduce histogram representations, dominant angle and colour pyramid match, in: IEEE International Conference on Image Processing, IEEE, 2011, pp. 661–664.
[34] P. Li, J. Ma, S. Gao, Actions in still web images: visualization, detection and retrieval, in: Web-Age Information Management: 12th International Conference, vol. 6897, Springer, 2011, p. 302.
[35] B. Yao, X. Jiang, A. Khosla, A. Lin, L. Guibas, L. Fei-Fei, Human action recognition by learning bases of action attributes and parts, in: IEEE International Conference on Computer Vision, IEEE, 2011, pp. 1331–1338.
[36] V. Delaitre, J. Sivic, I. Laptev, Learning person-object interactions for action recognition in still images, in: Advances in Neural Information Processing Systems (NIPS), MIT Press, 2011.
[37] P. Li, J. Ma, What is happening in a still picture?, in: IEEE Asian Conference on Pattern Recognition, IEEE, 2011, pp. 32–36.
[38] B. Yao, L. Fei-Fei, Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses, IEEE Trans. Pattern Anal. Mach. Intell. 99 (2012) 1691–1703.
[39] G. Sharma, F. Jurie, C. Schmid, Discriminative spatial saliency for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3506–3513.
[40] B. Yao, L. Fei-Fei, Action recognition with exemplar based 2.5 d graph matching, in: European Conference on Computer Vision Workshops and Demonstrations, Springer-Verlag, 2012, pp. 173–186.
[41] M.P. Kumar, B. Packer, D. Koller, Modeling latent variable uncertainty for loss-based learning, in: Proceedings of International Conference on Machine Learning, Omnipress, New York, NY, USA, 2012, pp. 465–472.
[42] F. Sener, C. Bas, N. Ikizler-Cinbis, On recognizing actions in still images via multiple features, in: European Conference on Computer Vision Workshops and Demonstrations, Springer, 2012, pp. 263–272.
[43] Y. Zheng, Y.-J. Zhang, X. Li, B.-D. Liu, Action recognition in still images using a combination of human pose and context information, in: IEEE International Conference on Image Processing, IEEE, 2012, pp. 785–788.
[44] N. Ikizler-Cinbis, S. Sclaroff, Web-based classifiers for human action recognition, IEEE Trans. Multimed. 14 (2012) 1031–1045.
[45] D.T. Le, R. Bernardi, J. Uijlings, Exploiting language models to recognize unseen actions, in: Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, ACM, 2013, pp. 231–238.
[46] G. Sharma, F. Jurie, C. Schmid, Expanded Parts Model for Human Attribute and Action Recognition in Still Images, in: IEEE Conference on Computer Vision Pattern Recognition, IEEE, Portland, OR, United States, 2013.
[47] J. Canny, A computational approach to edge detection, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-8 (1986) 679–698.
[48] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in: Advances in Neural Information Processing Systems (NIPS), MIT Press, 2000, pp. 556–562.
[49] D. Ramanan, Learning to parse images of articulated bodies, in: Advances in Neural Information Processing Systems (NIPS), MIT press, 2006, pp. 1129–1136.
[50] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[51] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1778–1785.
[52] L. Bourdev, J. Malik, Poselets: Body part detectors trained using 3d human pose annotations, in: IEEE International Conference on Computer Vision, IEEE, 2009, pp. 1365–1372.
[53] B. Sapp, A. Toshev, B. Taskar, Cascaded models for articulated pose estimation, in: European Conference on Computer Vision, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 406–420.
[54] C.J. Taylor, Reconstruction of articulated objects from point correspondences in a single uncalibrated image, Comput. Vis. Image Underst. 80 (2000) 349–363.
[55] H.-J. Lee, Z. Chen, Determination of 3d human body postures from a single view, Comput. Vis. Graph. Image Process. 30 (1985) 148–168.
[56] B. Alexe, T. Deselaers, V. Ferrari, What is an object?, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 73–80.
[57] Y. Chen, J. Bi, J. Wang, Miles: multiple-instance learning via embedded instance selection, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 1931–1947.
[58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.
[59] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 1627–1645.
[60] C. Desai, D. Ramanan, C. Fowlkes, Discriminative models for multi-class object layout, in: IEEE International Conference on Computer Vision, IEEE, 2009, pp. 229–236.
[61] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (2001) 145–175.

[62] D. Lowe, Object recognition from local scale-invariant features, in: IEEE International Conference on Computer Vision, vol. 2, IEEE, 1999, pp. 1150–1157.
[63] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2005, pp. 886–893.
[64] S. Belongie, J. Malik, Matching with shape contexts, in: IEEE Workshop on Content-based Access of Image and Video Libraries, IEEE, 2000, pp. 20–26.
[65] A. Friedman, Framing pictures: the role of knowledge in automatized encoding and memory for gist, J. Exp. Psychol.: General 108 (1979) 316–355.
[66] H. Bay, T. Tuytelaars, L. V. Gool, Surf: Speeded up robust features, in: European Conference on Computer Vision, Springer, Berlin, Heidelberg, 2006, pp. 404–417.
[67] N. Ikizler, P. Duygulu, Human action recognition using distribution of oriented rectangular patches, in: Workshop on Human Motion, Springer, 2007, pp. 271–284.
[68] M. Tahir, J. Kittler, K. Mikolajczyk, F. Yan, K. van de Sande, T. Gevers, Visual category recognition using spectral regression and kernel discriminant analysis, in: IEEE International Conference on Computer Vision Workshops, IEEE, 2009, pp. 178–185.
[69] B. Scholkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT press, 2001.
[70] Y. Chen, L. Zhu, C. Lin, A. L. Yuille, H. Zhang, Rapid inference on a novel and/or graph for object detection, segmentation and parsing, in: Advances in Neural Information Processing Systems (NIPS), MIT Press, 2007, pp. 289–296.
[71] Y. Yang, D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 1385–1392.
[72] M. Kumar, A. Zisserman, P. Torr, Efficient discriminative learning of parts-based models, in: IEEE International Conference on Computer Vision, IEEE, 2009, pp. 552–559.
[73] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, IEEE, 2006, pp. 2169–2178.
[74] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 888–905.
[75] J.H. Ward, Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. 58 (1963) 236–244.
[76] R. Hamid, A. Johnson, S. Batta, A. Bobick, C. Isbell, G. Coleman, Detection and explanation of anomalous activities: representing activities as bags of event n-grams, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2005, pp. 1031–1038.
[77] S.T. Hedetniemi, R.C. Laskar, Bibliography on domination in graphs and some basic definitions of domination parameters, Discret. Math. 86 (1991) 257–277.
[78] B. Yao, H. Ai, S. Lao, Building a compact relevant sample coverage for relevance feedback in content-based image retrieval, in: European Conference on Computer Vision, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 697–710.
[79] M. Grubinger, P. Clough, H. Müller, T. Deselaers, The IAPR TC-12 benchmark: a new evaluation resource for visual information systems, in: International Workshop OntoImage, ESSLLI, 2006, pp. 13–23.
[80] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: IEEE International Conference on Computer Vision, IEEE, 2005, pp. 1395–1402.
[81] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: IEEE International Conference on Pattern Recognition, vol. 3, IEEE, 2004, pp. 32–36.

Guodong Guo received his B.E. degree in Automation from Tsinghua University, Beijing, China, in 1991, the Ph.D. degree in Pattern Recognition and Intelligent Control from Chinese Academy of Sciences, in 1998, and the Ph.D. degree in computer science from the University of Wisconsin–Madison, in 2006. He is currently an Assistant Professor in the Lane Department of Computer Science and Electrical Engineering, West Virginia University. In the past, he visited and worked in several places, including INRIA, Sophia Antipolis, France, Ritsumeikan University, Japan, Microsoft Research, China, and North Carolina Central University. He won the North Carolina State Award for Excellence in Innovation, in 2008, and Outstanding Researcher (2013-2014) and New Researcher of the Year (2010–2011) at CEMR, WVU. He was selected as the "People's Hero of the Week" by the Broadband and Social Justice Blog under the Minority Media and Telecommunications Council (MMTC) in June 2013 for his work on estimating Body Mass Index (BMI) from image data. His research areas include computer vision, machine learning, and multimedia. He has authored a book, Face, Expression, and Iris Recognition Using Learning-based Approaches (2008), co-edited a book, Support Vector Machines Applications (2014), published over 60 technical papers in face, iris, expression, and gender recognition, age estimation, multimedia information retrieval, and image analysis, and filed three patents on iris and texture image analysis.

Alice Lai is a Master degree student in the Lane Department of Computer Science and Electrical Engineering at West Virginia University. She received her Bachelor's degree in Software Engineering from South China University of Technology, Guangzhou, China, in 2011. Her current research focuses on the problem of action recognition in still images. Her research interests include computer vision, pattern recognition, and machine learning.