RECOGNITION BY DETECTION: PERCEIVING HUMAN MOTION THROUGH PART-CONFIGURED FEATURE MAPS

*Lei Wang, Jun Wu, Zhimin Zhou, Yuncai Liu, Xu Zhao**

Shanghai Jiao Tong University
Institute of Image Processing and Pattern Recognition
{wltongxing, zhugetian, 115236, whomliu, zhaoxu}@sjtu.edu.cn

ABSTRACT

Visually perceiving human motion at semantic level is an important however challenging problem in multimedia area. In this work, we propose a novel approach to map the low-level responses from visual detection to semantically sensitive description to human actions. The feature map is triggered by the output of deformable part model detection, in which the critical information about body parts configuration is contained implicitly under the specific human actions. We map the filter responses of the detectors to an effective feature description, which encodes the position and appearance information of the root and every body parts simultaneously. Statistically, the obtained feature map captures the significance of relative configuration of body parts, therefore is robust to the false detections occurred in the individual part detectors. We conduct comprehensive experiments and the results show that the method generates discriminative action features and achieves remarkable performance in most of the cases.

1. INTRODUCTION

Perceiving human motion at semantic level from visual media like image and video is an important problem in multimedia area. The principle target of this mission is to recognize human action and furthermore achieve a through understanding to the events or environments centered on human activities. The application widely ranges from surveillance, content based retrieval, social activity analysis to human-computer interface, monitoring of patients and so forth. This problem however is very challenging due to the huge variations of human motion and imaging conditions. More specifically, it's a critical yet hard step to find a discriminative low-level description, via which one can efficiently bridge the gap between visual observation and semantical level understanding about human motion.

On feature formation for action recognition, most of the existing methods focus on extracting classification beneficial information by describing the body as a whole or picking some visual patches randomly from the whole body regions [1]. This kind of description is too rough to get the detailed pose information about how the body parts are configured for the specific action types, which in fact play essential role in recognizing human motions. Most recently, a few works [2, 3, 4] introduce pose as a cue for action classification and achieve inspiring performance. A critical shortcoming of these methods however is that generally they need the intermediate results of pose estimation and lack tolerance to the wrongly estimated pose configuration. In this paper, we propose a novel approach to map the low-level responses from visual detection of human body to semantically sensitive description to human actions. Our strategy is different from most of the previous methods and builds a recognition by detection procedure, which integrates the two parts: body detection and action recognition, seamlessly into a complete feature map framework.

The feature map is triggered by the output of mixtures of Deformable Part Model (DPM) detection, which was proposed by Felzenszwalb *et al.* [5] and validated very effective in object detection. From DPM, one can get the responses from the root filter and each of the part filters in multiple scales. The model captures not only the properties of the whole body, but also the local appearance properties of the individual body parts, so the critical information about body parts configuration under the specific human actions is contained implicitly in the filters' response. Actually the filter responses encode the position and appearance information of the root and every body parts simultaneously. We model the responses around locations of detection as a feature map. The obtained feature map statistically captures the significance of relative configuration of body parts, therefore is robust to the false detections occurred in the individual part detectors. In doing so, a discriminative description about human motion is naturally built from the body detection results, which always is the prerequisite of human motion analysis.

We conduct comprehensive experiments on three datasets: YouTube, HAT and Willow. The action types involved in these date sets are diverse and challenging to recognize. The experimental results show that our feature map generates dis-

This work is supported partially by NSFC 61273285, 61375019 and China 973 program (2011CB302203).

criminative action descriptions and achieves remarkable performance in most of the cases while there is space for improvement in some tough cases.

To our best knowledge, the proposed method is the first attempt that introduces the responses from object part detectors to serve for semantical recognition. This recognition-by-detection framework potentially is capable of providing a compact and efficient access to solve people detection, pose estimation and action recognition in an integrated framework.

2. FEATURE FORMATION

We use the detector designed by Felzenszwalb *et al.* [5] as the starting point of the feature construction. This detector is based on mixtures of deformable part models. For a single star-structured model of mixture models, there are several part filters and a root filter. Each part of a person model captures local appearance properties, and the root captures properties of the whole body. We model the responses of part and root filters around locations of detection as a descriptor.

2.1. Detection

The details of the detection algorithm (mixtures of deformable part models) are fully presented in [5]. We give a brief description about the model in the following.

Let a mixture model have $\boldsymbol{m}$ components, and each component have one root filter and $\boldsymbol{n}$ part filters. Let H be a pyramid of a variation of Histogram of Oriented Gradients (HOG) features [5], and $p = (x, y, l)$ defines a position (x, y) in the l-th level of H. An object hypothesis $z = (p_0, ..., p_n)$ specifies the location of filters of a model component in H. p_0 is the position of root filter, and p_i is the position of i-th part filter. The score of z on model component c is given by scores of each filter minus a deformation cost, plus a bias:

$$\text{score}^c(p_0, ..., p_n) = \sum_{i=0}^{n} F_i' \cdot \phi(H, p_i) - \sum_{i=1}^{n} d_i \cdot \phi_d(dx_i, dy_i) + b, \quad (1)$$

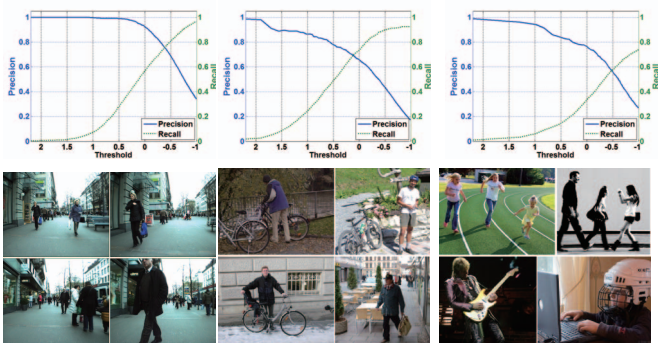
where $\phi(H, p_i)$ is the feature vector at position p of H, F_0' is the root filter vector, F_i' is the i-th part filter vector, $\phi_d(dx_i, dy_i)$ is the deformation feature vector, d_i is the coefficient of deformation cost for the i-th part, and b is the bias.

The overall score for each root location is computed according to the best possible placement of the parts, which define a full object hypothesis and the root locations define the detections

$$\text{score}^c(p_0) = \max_{p_1,...,p_n} \text{score}^c(p_0, ..., p_n). \quad (2)$$

To detect objects using a mixture model, the score at a root location is defined by the highest score across all the components,

$$\text{score}(p_0) = \max_c \text{score}^c(p_0). \quad (3)$$



(a) (b) (c)

Fig. 1. Precision and recall on three datasets (a) ETH pedestrian dataset, (b) INRIA person dataset and (c) Willow action dataset. Precisions and recalls under different thresholds are shown in the top row and the corresponding sample images are shown in the bottom row.

Let T be a detection threshold, the set of detections D is defined by all the root locations with scores over T,

$$D = \{p_0 | \text{score}(p_0) > T\}. \quad (4)$$

To eliminate repeated detections, non-maximum suppression [5] is applied on the detection set D. The detections in D are sorted by score, and the highest scoring ones are selected first. The following detections with bounding boxes that are covered by at least 50% by a previous one are skipped.

A detection is counted to be correct if the overlap (intersection over union) of detected and groundtruth bounding box is greater than a threshold,

$$\frac{\text{area}(D) \cap \text{area}(G)}{\text{area}(D) \cup \text{area}(G)} > t, \quad (5)$$

where area(D) is the area covered by detected bounding box and area(G) is for the ground truth. t is the threshold which determines the detection accuracy. We set t to 0.5 according to the PASCAL [6] criterion.

We use the model trained on dataset of PASCAL VOC 2010 [6]. To analyze the effects of threshold on detection's precision and recall, we test the detector on three data sets: ETH pedestrian [7], INRIA person [8] and Willow action [9]. The results are shown in Fig. 1, where the precisions and recalls under different thresholds are shown in the top row and the sample images are shown in the bottom row. With the same threshold, we get higher precisions and recalls on datasets of ETH pedestrian and INRIA person than on Willow action dataset. When the detector is applied to more challenging action datasets, the threshold should be lower.

2.2. Feature Map for Recognition

The configurations of body parts contain discriminative information for action recognition. Therefore we model the body

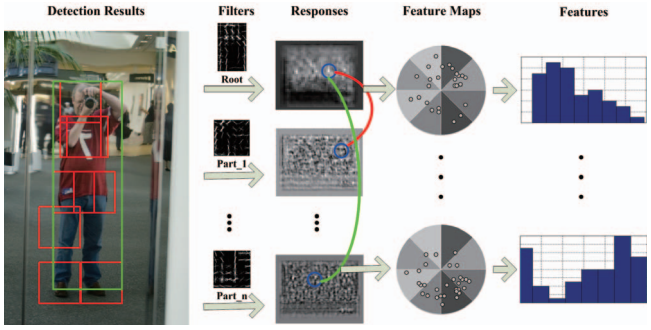


Fig. 2. Work flow of feature formation from detection results.

part positions and body appearances for the purpose of recognition. To model body part positions, we calculate the relative position vectors between the detected root location and part locations, by which the orientations of relative position vectors can be computed. For the body appearance, we extract the HOG features as we do in detection procedure. Then the filter responses are calculated by convoluting the HOG features with root and part filters.

We sample the filter responses around detected locations of root and parts within a boundary. Our action features can be generated by the orientations of the relative position vectors and sampled responses together. The work flow for feature formation is shown in Fig. 2. The details are introduced as follows.

Through the detection procedure, we can get the root detection set D and the whole detection set Z with root and part locations defined in Eq. (2)

$$z = \{(p_0, ..., p_n) | p_0 \in D\}, \tag{6}$$

where $p_i = (x_i, y_i, l_i)$ specifies the position at (x_i, y_i) in level l_i of feature pyramid H. p_0 is the root location and the others are the part locations.

The responses are computed by convolution of a filter F_i and the feature H_i at level l_i specified in p_i in the feature pyramid H,

$$R_i = H_i \otimes F_i, \qquad i = (0, ..., n). \tag{7}$$

We sample R_i around the position specified by (x_i, y_i) in p_i with a sampling radius s,

$$S_i = \{(x, y, r) | (x - x_i)^2 + (y - y_i)^2 \leq s^2\}, \tag{8}$$

where (x, y) is position in R_i and r is the response value. We get one root set S_0 and n part set $(S_1, ..., S_n)$. Because the root filter and part filters are in different level of feature pyramid, we map the position of root set to the same level as the part sets and get the root set S'_0.

Our features are extracted from the root set and part sets. Let (x^0, y^0, r^0) be an element in the root set S'_0, and (x^i, y^i, r^i) be an element in a part set S_i, $i \in \{1, ..., n\}$. Thus we can obtain an element pair, $\{(x^0, y^0, r^0), (x^i, y^i, r^i)\}$, between the root and a part. The relative position vector between the element pair is,

$$(x, y) = (x^i - x^0, y^i - y^0). \tag{9}$$

Let $\theta(x, y)$ be the vector orientation. The orientation for each element pair is discretized into one of p bins,

$$B = \text{round}\left(\frac{p \cdot \theta(x, y)}{2\pi}\right) \mod p. \tag{10}$$

We define a pair-level feature map. Let $b \in \{0, ..., p-1\}$ range over orientation bins. The feature vector is,

$$F_b^i = \begin{cases} r^0 + r^i, & \text{if } b = B, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

For all the possible pairs between S'_0 and S_i, we can aggregate features to F^i. For n parts, we obtain the features,

$$F = (F^1, ..., F^n). \tag{12}$$

As can be seen from the whole procedures of feature formation, the action feature specifies both body part configurations and body appearances therefore is quite suitable for recognition. In the next section, we conduct experiments on three datasets for action recognition. SVM is used as classifier.

3. EXPERIMENTS

3.1. Datasets

We evaluate the proposed method on three publicly available datasets: (i) **YouTube** video dataset [10], (ii) Dataset of human attributes (**HAT**) [11], and (iii) **Willow action** dataset [9]. The sample images from the three datasets are shown in Fig. 3. The original YouTube videos do not include action annotations and Ikizler-Cinbis *et al.* [12] annotated 11 videos from the dataset. So in total 775 frames for five actions are included. HAT dataset is built for learning semantic human attributes. It has annotation for 27 classes based on age (*e.g.* kid, baby), appearance (*e.g.* swimsuit, longskirt) and pose (*e.g.* sitting, crouching). Willow actions is a challenging dataset for action classification, in which 7 classes of common human actions: "Interacting with computers", "Photographing", "Playing a musical instrument", "Riding bike", "Riding horse", "Running" and "Walking" are included. It's also designed to investigate the effect of background noise.

3.2. Experimental Setting

Parameters selection for detection. In our experiments, the threshold t in Eq. 3 is set to -0.5 considering both *precision* and *recall*. The label of a successful detection, which satisfies Eq. (5), is given by the annotations provided by the datasets.

(a) (b) (c)

Fig. 3. Example images on three datasets: (a) YouTube, (b) HAT and (c) Willow action dataset.

The obtained data set and data labels have smaller size than that of the annotated data in the datasets due to the detection *recall*. The false positive human detections (false detections or not labeled positive detection) are discarded (not included in the data set) for classification task.

Classifier setup. We use libSVM [13] for multi-class classification. The data set is randomly divided into training set and testing set. We use RBF kernel, in which two parameters, C and γ need to be set beforehand. In order to identify good (C, γ) for high accuracy, 5-folder cross-validation is carried out on the training set. Sequentially one subset is tested using the classifier trained on the other 4 subsets.

Performance metrics. We use several metrics [14] to evaluate the performance. Four common performance metrics, *accuracy*, *precision*, *recall* and *F-measure* (*F-score*), can be calculated by the equations shown in Fig. 4. Performance measure *per-class accuracy* we used in this paper is calculated by the equation across all classes.

We also use *overall accuracy* (multi-SVM accuracy). overall accuracy $= c/n$, where c is the number of correctly classified samples, and n is the number of all samples.

Another typical performance measure for multiple classification tasks is *classification accuracy* which is the average of the confusion table diagonal [9]. The diagonal of confusion matrix is also the per-class recall as defined in Fig. 4. The *classification accuracy* that Delaitre *et al.* [9] used is also the mean of per-class recall in our experiments.

3.3. Experimental Results

Feature bins and sample size. As described in Section 2.2, there are two parameters, number of bins b and sampling size s, which have significant impacts on the performance. b determines the feature dimensions and s reflects how much information is used around the root and part positions. Since there are only two parameters, the grid-search is straightforward. In order to make the results clear, we analyze one parameter while keeping the other fixed.

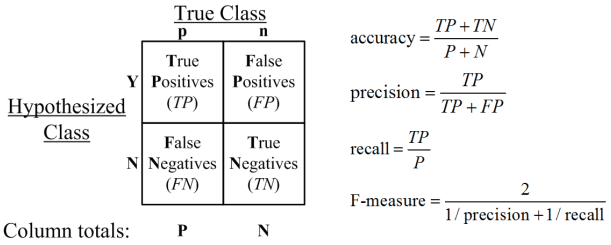
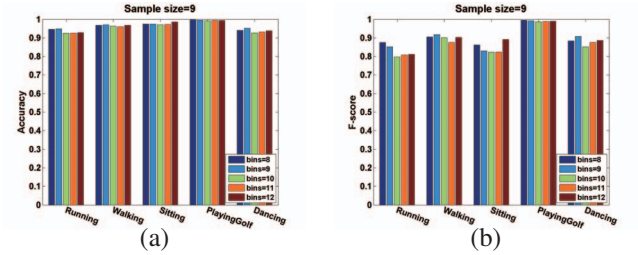


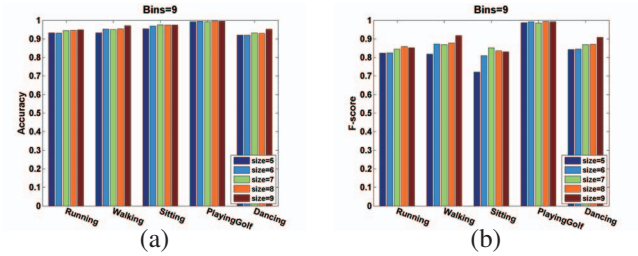
Fig. 4. Performance metrics defined in [14].



(a) (b)

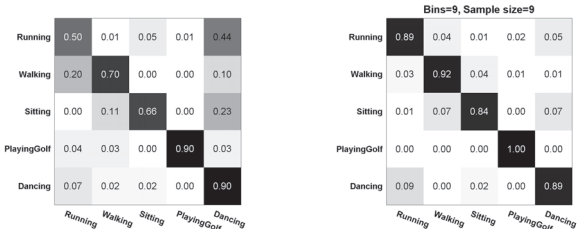
Fig. 5. Per-class *Accuracy* and *F-score* under different bins when **sample size** is 9 on YouTube dataset. (a) shows *accuracy* and (b) shows *F-score*.

More feature bins can provide more details, but it also increase the feature dimensions. With a certain amount of data, high dimension data is more likely to suffer from overfitting problem. As shown in Fig. 5, the sample size is 9, and the average *accuracy* and average *F-score* are highest when the number of bins is 9. We test our approach on all three dataset with different sampling sizes, and we get high average *accuracy* and *F-score* when $b \in [9, 10]$. The filter responses are sampled around the detection positions of root and parts. Small samples reflect 'local' feature while large samples contain more 'background' information. The best sample size s varies on different datasets. Fig. 6 shows that when the number of bins is 9, the average *accuracy* and average *F-score* are highest if sample size is 9 on YouTube dataset. In our experiences on three datasets, we get high *accuracy* and high



(a) (b)

Fig. 6. Per-class *Accuracy* and *F-score* under different sample sizes when the **number of bins** is 9 on YouTube dataset. (a) shows *accuracy* and (b) shows *F-score*.



(a) (b)

Fig. 7. Confusion Matrix on YouTube dataset. (a) is results of Ikizler-Cinbis *et al* [12] and (b) is our results.

Table 1. Overall performance comparison. Our result is reported when the number of bins is 9 and sample size is 9.

	Ikizler-Cinbis [12]	Our method
Overall accuracy	0.7587	0.9200

F-score when $s \in [7, 9]$.

Performance on YouTube dataset. We compare our method with the method proposed in [12]. The confusion matrixes are shown in Fig. 7. The confusion between "Running" and "Dancing" is reduced by our method. As well, "Sitting" and "Dancing" are less likely to be mixed with each other by our method. Table 1 is the comparison of overall accuracies of method of Ikizler-Cinbis *et al.* and ours. Our overall accuracy in all classes got improvement up to 16%.

To demonstrate the detailed results on each class, per-class *accuracy*, *precision*, *recall* and *F-score* on YouTube dataset are shown in Table 2. The F-score is over 0.9 on actions "Walking", "Playing Golf" and "Dancing". "Playing Golf" is the easiest to be distinguished from the others.

Performance on HAT dataset. As we focus on body pose configuration for action recognition, four classes based on pose are evaluated on HAT dataset. Table 3 shows the results of four classes, "Standing", "Runwalk", "Sitting", and "Crouching". The high accuracy on "Sitting" and "Crouching" does not mean that our method fits for those two classes, because their F-scores are low. This is due to the imbalanced size of positive and negative datasets. For "Crouching", there are more negatives than the positives, and negatives contributes more to the accuracy than positives. Our results of person detection on HAT dataset include 4992 instances for "Standing" class, 1419 for "Runwalk", 873 for "Sitting" and 185 for "Crouching". Some of the data have been given multiple labels. This is a challenging classification task, and we only get an *overall accuracy* of 64.95%.

Table 2. Per-class *Accuracy*, *precision*, *recall* and *F-score* when the number of bins is 9 and sample size is 9 on YouTube dataset.

Actions	Accuracy	Precision	Recall	F-score
Running	0.9489	0.8236	0.8855	0.8516
Walking	0.9700	0.9206	0.9155	0.9175
Sitting	0.9744	0.8269	0.8440	0.8299
Playing Golf	0.9956	0.9898	0.9960	0.9928
Dancing	0.9511	0.9335	0.8858	0.9080

Table 3. Per-class *Accuracy* and *F-score* when the number of bins is 9 and sample size is 7 on four classes of HAT dataset.

Actions	All dataset (imbalanced)		Data subset (balanced)	
	Accuracy	F-score	Accuracy	F-score
Standing	0.6431	0.7717	0.7612	0.5484
Runwalk	0.8029	0.2305	0.6660	0.2757
Sitting	0.8676	0.0968	0.6874	0.4416
Crouching	0.9690	0.0498	0.6660	0.2501

Table 4. The mean of average precision (mAP) on HAT dataset. Our results are based on detection and recognition.

	SPM ([15])	EPM ([16])	Our Method (Det. & Rec.)
mAP	0.555	0.587	0.321

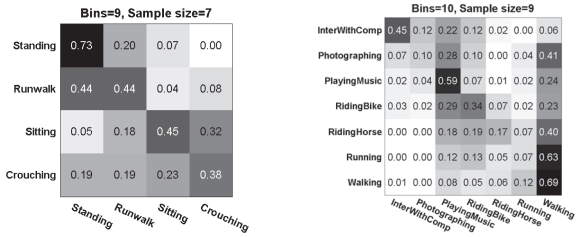
To get rid of the problem of imbalanced data and illustrate how our action features work on classification, a subset of balanced data on all the four classes is selected, with 129 instances for each class. The improvement on F-score of "Sitting" and "Crouching" is shown in Table 3. The confusion matrix for the subset is shown in Fig. 8 (a). The best performance is on "Standing". "Runwalk" is often confused with "Standing", and "Sitting" is often confused with "Crouching". We compare our results with those of Lazebnik *et al.* [15] and Sharma *et al.* [16] in Table 4. The mAP of our results is over 20% lower than theirs. One significant reason is that our results are based on recognition-by-detection but their results are obtained with ground truth bounding box.

Performance on Willow action dataset. We test our method on Willow action dataset and the *overall accuracy* is only 39.70%. The confusion matrix is shown in Fig. 8 (b). "Photographing", "Riding bike", "Riding horse" and "Running" are often confused with "Walking". The results of Delaitre *et al.* [9] and our results are compared in Table 5. The *classification accuracy* (average of the diagonal of the confusion matrix) is 34.43% which is over 20% lower than 57.05% of 'LSVM' method of Delaitre *et al.* [9] and much lower than 68.76% of their 'LSVM+C2' method.

Results Analysis. Our method combines human detection and action classification. The final performance depends on both accuracy of detection and effectiveness of action features. Our method performs better on YouTube dataset than HAT dataset and Willow action dataset. We take YouTube

Table 5. The classification accuracy on Willow action dataset. Our results are based on detection and recognition.

	LSVM ([9])	LSVM+C2 ([9])	Our Method (Det. & Rec.)
Classification accuracy	0.5705	0.6876	0.3443



(a) (b)

Fig. 8. Confusion Matrix of: (a) the balanced subset of data on HAT dataset and (b) Willow action dataset. "InterWithComp" is the short for "InteractionWithComputers".

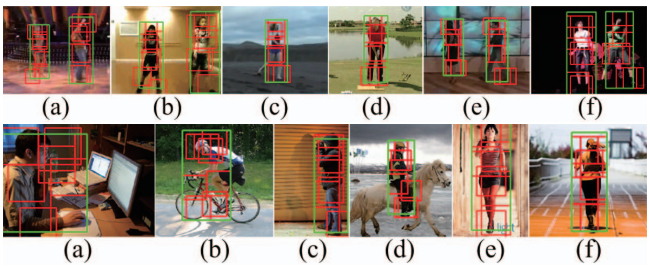


Fig. 9. Human detection results on YouTube dataset (upper row) and Willow action Dataset (lower row).

and Willow action dataset for the analysis.

From the precision and recall curves shown in Fig. 1, we can find the complexity of Willow action dataset where some instances may not be detected, while the *precision* is taken into consideration. However, the final results severely depend on the detection accuracy. That's why the results on YouTube dataset are better than Willow action dataset. Fig. 9 shows some example images of human detection on YouTube and Willow action datasets. As can be seen, most bounding boxes in (a) and (b) of the Willow action dataset are away from human body, which result in false features. Another reason is closely related with the action classes. YouTube dataset consists of "Running", "Walking", "Sitting", "Playing Golf" and "Dancing" actions, which can be represented by the pose configurations. In Willow action dataset, the interactive objects (e.g. computer, camera, bike) play important roles. But our method mainly focus on classifying actions by pose configurations. This also explains why several actions are confused with "Walking" (Fig. 9 (c), (e) and (f) in the second row).

4. CONCLUSION

In this work, we propose a recognition-by-detection framework for human motion recognition and achieve remarkable performance in the related dataset. Because our method is designed for pose specific actions, there is much space for improvement in future by taking the context information (e.g. horse, bike) into account.

5. REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, pp. 16, 2011.

[2] X. Zhao, Y. Liu, and Y. Fu, "Exploring discriminative pose sub-patterns for effective action classification," in *ACM MM*, 2013.

[3] C. Desai and D. Ramanan, "Detecting actions, poses, and objects with relational phraselets," in *ECCV*. 2012.

[4] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *CVPR*, 2011.

[5] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results," .

[7] A. Ess, B. Leibe, K. Schindler, and L. van Gool, "A mobile vision system for robust multi-person tracking," in *CVPR*, 2008.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

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

[10] J. Niebles, B. Han, A. Ferencz, and Li Fei-Fei, "Extracting moving people from internet videos," in *ECCV*. 2008.

[11] G. Sharma and F. Jurie, "Learning discriminative spatial representation for image classification," in *BMVC*, 2011.

[12] N. Ikizler-Cinbis, R.G. Cinbis, and S. Sclaroff, "Learning actions from the web," in *ICCV*, 2009.

[13] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM TIST*, vol. 2, pp. 27:1–27:27, 2011.

[14] T. Fawcett, "Roc graphs: Notes and practical considerations for researchers," *Machine Learning*, vol. 31, pp. 1–38, 2004.

[15] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.

[16] G. Sharma, F. Jurie, and C. Schmid, "Expanded parts model for human attribute and action recognition in still images," in *CVPR*, 2013.