Symmetry **2015**, *7*, 427-449; doi:10.3390/sym7020427

OPEN ACCESS

symmetry

ISSN 2073-8994

www.mdpi.com/journal/symmetry

Article

Unsupervised Object Modeling and Segmentation with Symmetry Detection for Human Activity Recognition

Jui-Yuan Su [1,2], Shyi-Chyi Cheng [1,*] and De-Kai Huang [1]

[1] Department of Computer Science and Engineering, National Taiwan Ocean University, 2 Pei-Ning Road, Keelung 202, Taiwan; E-Mail: dreamdk215@gmail.com

[2] Department of New Media and Communications Administration, Ming Chuan University, 250 Sec. 5 Zhong Shan North Road, Taipei 111, Taiwan; E-Mail: rysu@mail.mcu.edu.tw

* Author to whom correspondence should be addressed; E-Mail: csc@mail.ntou.edu.tw; Tel.: +886-2-2462-2192 (Ext. 6653), Fax: +886-2-2462-3249.

Academic Editor: Christopher Tyler

Received: 29 November 2014 / Accepted: 16 April 2015 / Published: 23 April 2015

Abstract: In this paper we present a novel unsupervised approach to detecting and segmenting objects as well as their constituent symmetric parts in an image. Traditional unsupervised image segmentation is limited by two obvious deficiencies: the object detection accuracy degrades with the misaligned boundaries between the segmented regions and the target, and pre-learned models are required to group regions into meaningful objects. To tackle these difficulties, the proposed approach aims at incorporating the pair-wise detection of symmetric patches to achieve the goal of segmenting images into symmetric parts. The skeletons of these symmetric parts then provide estimates of the bounding boxes to locate the target objects. Finally, for each detected object, the graphcut-based segmentation algorithm is applied to find its contour. The proposed approach has significant advantages: no *a priori* object models are used, and multiple objects are detected. To verify the effectiveness of the approach based on the cues that a face part contains an oval shape and skin colors, human objects are extracted from among the detected objects. The detected human objects and their parts are finally tracked across video frames to capture the object part movements for learning the human activity models from video clips. Experimental results show that the proposed method gives good performance on publicly available datasets.

Keywords: object detection and segmentation; Hough voting; human activity recognition; symmetry detection

---

1. Introduction

Part-based object detection and segmentation is an important problem in computer vision. Classical object detection methods often use learned models to detect and recognize the targets [1,2]. Quality-conscious object segmentation spans a new way to build the most discriminative models, compared with classical object modeling schemes, which often delimit the training objects with inaccurate bounding boxes. Recently, segmentation-based tracking, incorporating temporal information of object movements to improve the detection accuracy, has attracted great attention in the field of video object segmentation [3–5] due to its potential for many vision-based applications, such as video surveillance, man-machine interfaces, sports analysis, and authoring of video games [6]. To incorporate the spatial and temporal information for improving the accuracy of object segmentation is particularly important and remains a challenge.

Object segmentation is generally far more difficult than low-level image segmentation, which groups pixels of similar features, *i.e.*, colors, textures, and optical flows, into regions, without inferring the complete image understanding models. During the past three decades, intensive research works have been carried out in the automatic segmentation domain [7–12]. These techniques achieve efficient segmentation by subdividing an image into a number of moving objects and the background according to a homogenous low-level feature criterion and object tracking. This homogenous grouping almost extracts semantically incomplete objects, each of which perhaps consists of multiple parts with different homogeneous features. Moreover, using a tracking or body pose estimation in real world videos is generally not reliable due to object occlusion, distortion and changes in lighting. Semi-automatic semantic object segmentation algorithms [13–15] are thus proposed to tackle these difficulties. In the common first step of these methods, users initially identify a semantic object by using tracing interface and the computer automatically tracks the segmented object for the successive frames.

Recent approaches suggest using pre-learned object models to detect, segment, track, and recognize the target objects in images [1,16–18]. For instance, in [1], parts arranged in a deformable configuration are modeled to capture the local property of objects. The use of visual patterns of local patches in object modeling is related to several ideas, including the approach of local appearance codebooks [19] and the generalized Hough transform (GHT) [20] for object detection. At training time, these methods learn a model of the spatial occurrence distributions of local patches with respect to object centers. At testing time, based on the trained object models, the visual patterns of patches, with points of interest as their centers, are matched to visual codebooks to locate the targets using the Hough voting framework. However, the effectiveness of visual pattern grouping by Hough voting is heavily dependent on the quality of the learned visual model, the ability to precisely locate the target objects, and the features extracted from training samples.

Many object detection approaches are limited by the ill-defined object models, which are trained from a set of limited views and deficient in characterizing the texture in local parts and their spatial constraints [1,2]. The performance of these methods degrades dramatically when the input image has enormous deformation compared with the training images. Symmetry, however, is an essential characteristic of man-made or natural objects. Accordingly, the motivation of this paper is to integrate symmetry detection into classical object detection and segmentation to construct a model-free approach. Instead of learning a complex object model using a large amount of training samples, our approach

defines the part-based object detection and segmentation to be the task of decomposing an image into constitute salient symmetric parts, each of which is characterized by a common set of local features, *i.e.*, symmetric skeletons, dominate colors, and shape descriptors. Thus, our approach first detects salient symmetries in the test image with the Hough voting framework. The patches that constitute each of the detected symmetries are then determined by the inverse Hough transformation. The clusters of symmetries are generated to locate potential objects, each of which is specified with a bounding box. Finally, performing classical image segmentation on each bounding box, the target object is segmented.

Object classifiers can be further used to annotate, check and interpret the detected objects. Traditional object classifiers are trained from a set of weakly annotated sample objects, each of which is specified by a bounding box with undesirable background information. Instead, the proposed object detection and segmentation would introduce less noise from the targets and help avoid performance degradation in both the learning and recognition of object classifiers. To verify the effectiveness of the object detection and segmentation, we perform the face detection algorithm [21] on all detected parts to locate human objects. The detected human objects and their parts are then tracked across video frames to capture the object part movements for learning the poselet-like models, which had been verified to be effective in human activity recognition [22]. Experimental results show that the proposed method gives good performance on publicly available datasets in terms of detection accuracy and recognition rate.

The remainder of this paper is organized as follows. Section 2 presents the related work for the semantic object segmentation and symmetry detection. Section 3 describes the approach to deal with the object segmentation based on the detected results of the salient symmetric parts. Section 4 presents the application on human activity recognition. Section 5 describes the experimental tests to illustrate the effectiveness of the proposed method. Finally, conclusions are drawn in Section 6.

2. Related Work

Segmentation-based object recognition has been extensively studied with many algorithms available [12,23–25] in computer vision. Among them, the most interesting approach related to object recognition is semantic segmentation, which assigns each pixel in an image to one of several pre-defined semantic categories [23]. Compared to classical low-level unsupervised segmentation, which groups pixels of similar features, such as color, texture, or optical flows into homogeneous regions, semantic segmentation uses a supervised learning algorithm to build up semantic object models.

State-of-the-art semantic segmentation algorithms often use the local appearance model of an object to estimate the score of a pixel, a patch, or a region belonging to the target category [12,23,26–28]. To address the labeling consistency between neighboring local appearances, the local consistency model is then used to further group pixels, patches or regions into parts, though these parts still need merged to capture an object as a whole [1,2,29,30]. Therefore, a global consistency model is finally used to enforce global consistencies, *i.e.*, at a region or image level [30,31]. Girshick *et al.* have shown that rich feature hierarchies are very useful for accurate object detection and semantic segmentation [32].

Recently, object segmentation in videos spans a way to estimate the object boundaries by tracking pixels, patches, or regions to obtain their trajectories. Local elements with similar trajectories are then grouped into parts and objects [3–5,7,9–15]. However, the accuracy of any boundary estimate is limited by a number of systemic factors such as image resolution, noise, motion skew and the object occlusion.

For example, formulating object segmentation as motion segmentation using optical flow rests on the assumption of brightness constancy, which is violated at moving boundaries, resulting in poor estimates of object contours [33]. Object segmentation also tries to detect and segment the observed motions into semantic meaningful instances of particular activities from videos [17]. To reach this goal, recent approaches consider the detection and recognition of the video object as an extension of 2D object detection with higher dimensionality.

Many human-made objects, human bodies, natural scenes, or animals have symmetric parts. Several feature-based approaches have been proposed in the literature to detect symmetries in images for object detection and segmentation [34–36]. The common process in these approaches is that they dedicate the design of the reliable features for patch correspondences. For instance, Hsieh *et al.* designed a symmetric transformation to provide a framework for finding pairs of symmetric patches in vehicle images [36]. A recent survey of the symmetry in 3D geometry can be found in [37]. Although the symmetries provide a natural way to group low-level patches into middle-level parts, the combination of symmetric parts into high-level objects remains a challenging problem. Some methods depend on a prior global consistency model about the target object to perform top-down detection [29]. On the contrary, unsupervised object detection and segmentation, which does not rely on either human input, or top-down information, is important due to its potential in a variety of applications.

3. Unsupervised Object Detection and Segmentation

In this section, we present a probabilistic symmetry-based framework for combined object detection and segmentation. First we outline the notations to define the problem, and then emphasize the symmetry detection and clustering to estimate object locations. This is followed by image segmentation to obtain precise object boundaries. Finally, we describe a generative model that sets the foundation of our proposed object detection and segmentation.

3.1. Notations and System Overview

Let I and O, respectively, denote the image frame and the object frame (a bounding box in I, shown in Figure 1b). Let $\mathrm{x} = \{\mathrm{x}_n\}_{n=1}^{N}$ denote the set of centers of the sampled patches $\mathrm{P} = \{P_n\}_{n=1}^{N}$ in O, and $\mathrm{F} = \{f_n\}_{n=1}^{N}$ be the set of feature vectors to describe P. The object being segmented is represented by its shape C, the bounding box B, and the set S of symmetries determined by the set M of symmetric patch pairs. The bounding box B can be used to intersect the segmentation result obtained by performing image segmentation on I [24] to obtain the final object segmentation. The feature of an 8 × 8 patch used in this study is the well-known histogram of gradients (HOG) [38] though other complex features such as scale-invariant feature transform (SIFT) [39] or speeded up robust features (SURF) [36] descriptors can also be used as the replacement. A patch pair is in M if their HOG distance is less than a predefined threshold. The optical flow of a patch can also be used as the supplementary feature to improve the detection accuracy of symmetric parts when it is available.

The unsupervised approach consists of six pipelining steps, shown in Figure 2, to automatically locate multiple objects in an image, I. To perform the well-known Canny edge detection on I, we divide I into multiple 8 × 8 patches, each of which is described by the center (an edge point) and the HOG feature

vector. Next, based on a distance function in terms of HOG, patches in I are grouped into multiple clusters, each of which determines a set of symmetric patch pairs with the symmetry detection by Hough voting to follow. These detected symmetries are then used to model the object structures with a graph representation, which is optimally partitioned with the domain sets algorithm [39]. Each symmetry sub-graph estimates the bounding box B of an object. Finally, to use the graph cut algorithm [24] on B, the approach locates an object, which contains as less background as possible. A significant contribution of our approach is, at the moment of object detection, no tedious object models need learned in advance.

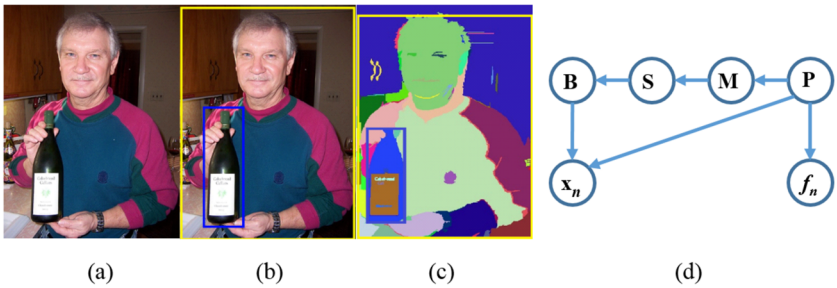

Figure 1. An example to illustrate the generative model: (**a**) the original image; (**b**) two bounding boxes to locate the target objects, *i.e.*, a person and a bottle; (**c**) perform the Graph cut segmentation algorithm [24] to obtain the segmentation results; (**d**) graphical representation of the generative model used in our method.

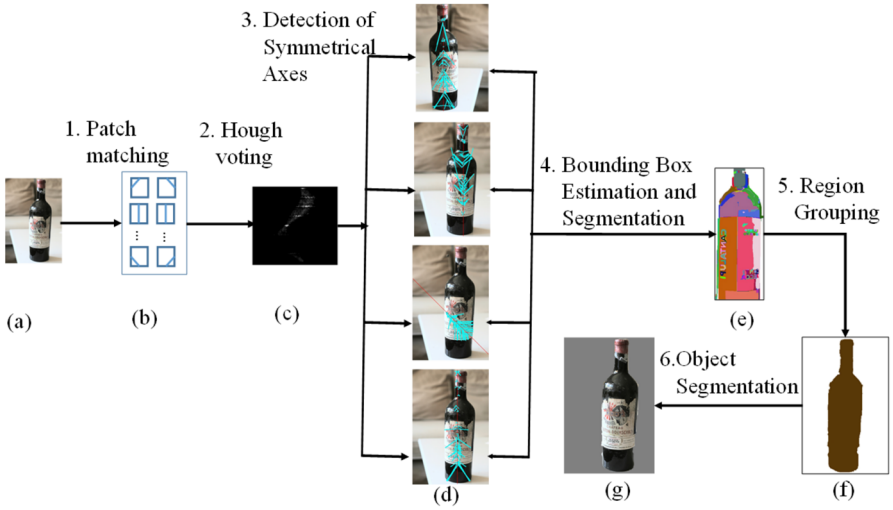

Figure 2. The overall procedure for the object detection and segmentation: (**a**) the input image is first partitioned into multiple patches; (**b**) the set of candidate symmetric patch pairs generated by matching patches in (**a**) with each other; patch pairs in (**b**) are used to generate the Hough voting image (**c**) whose peaks locate the salient symmetric axes and parts, shown in (**d**); (**e**) these detected symmetries are then used to estimate the bounding box of the target object; the sub-image constrained by the bounding box is segmented to obtain the segmentation mask; and the result, shown in (**f**) and (**g**), respectively.

3.2. Discovery of Symmetric Patch Pairs

An image, I, is first partitioned into multiple overlapping 8 × 8 patches P_i, $i = 1, \ldots, N$ with edge points as the centers x_i. For each patch, an 8-bin HOG descriptor with the quantization angles $j \times 45°$, $j = 0, \ldots, 7$ is used to represent its local appearance [1,38]. However, HOG lacks the capability of defining symmetric patch pairs, and thus we should firstly define the symmetric relations between patches in terms of HOG descriptors. Figure 3 shows that a small patch sampled from the contour of an object could contain a line edge, and the peak bin angle of the corresponding HOG approximates the gradient direction of the line in the patch.

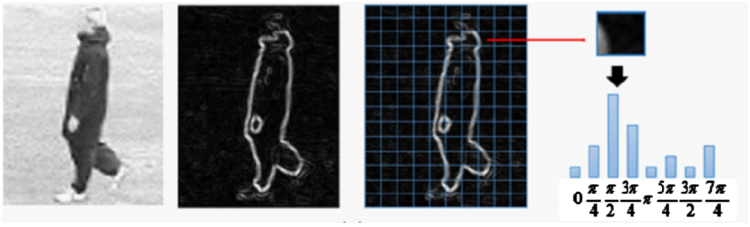

Figure 3. Using line edges to approximate the contour of an object.

Let f_i be the HOG of patch P_i. The first step to discover all symmetric patch pairs in I is to cyclically right shift f_i to obtain the normalized $\overline{f}_i$ with the peak being on the bin 0. We search the symmetric patches of P_i in the $L \times L$ window surrounding P_i, where L is the maximal distance between two patches belonging to the same object. The similarity measurement, based on the normalized HOGs $\overline{f}_i$ and $\overline{f}_j$, measures the similarity between patches P_i and P_j as follows:

$$Sim(P_i, P_j) = \delta(\| x_i - x_j \| < L)(\overline{f}_i \cdot \overline{f}_j) \quad (1)$$

where $\delta(\| x_i - x_j \| < L)$ is the delta function that returns 1 when the geometric distance between the centers of P_i and P_j is less than L, otherwise it returns 0; $(\overline{f}_i \cdot \overline{f}_j)$ is the inner product to measure the similarity between $\overline{f}_i$ and $\overline{f}_j$. Using (1) and the k-means clustering [40], patches in I are grouped into k clusters $\{PC_i\}_{i=1}^{k}$.

As mentioned above, two patches belonging to the same cluster form a pair of symmetric patches. Thus, the set of symmetric patch pairs can be defined as:

$$\mathbf{M} = \left\{ (P_i, P_j) \mid P_i \in PC_n \wedge P_j \in PC_n, i = 1, \ldots, N, j \neq i, n = 1, \ldots, k \right\} \quad (2)$$

Note that the value of k could not be large to preserve most of the potential symmetric patch pairs, and this brings fast convergence to the k-means clustering. Thus, the computational complexity to execute the patch clustering on-the-fly is not high.

3.3. Discovery of Symmetric Parts

Let $\{P_i, P_j\}$ be a patch pair in **M**. The pairwise patches of **M** can determine the skeleton K of the corresponding symmetric part shown in Figure 4a. Also let (l_i, m_i) and (l_j, m_j) be the normal vectors of gradient direction of P_i and P_j, respectively. These two normal vectors determine two lines $L_i = x_i + t_i (l_i, m_i)$ and $L_j = x_j + t_j (l_j, m_j)$. The intersection point (X, Y) of L_i and L_j can be obtained by

$$(X,Y)=(x_i,y_j)+\lambda(l_i,m_j),\lambda=[m_j(x_j-x_i)-l_j(y_j-y_i)]/(l_im_j-l_jm_i) \tag{3}$$

We can also compute the included angle ψ between L_i and L_j by

$$\psi=\tan^{-1}\frac{\phi_j-\phi_i}{1+\phi_j\phi_i} \tag{4}$$

where $(\phi_i,\phi_j)=(\tan^{-1}m_i/l_i,\tan^{-1}m_j/l_j)$. Next, as shown in Figure 4b, we compute the skeleton K characterized by two parameters (r, θ):

$$(r,\theta)=(\sqrt{X^2+Y^2}\cos(\theta-\tan^{-1}Y/X),-\frac{\pi}{2}+\frac{\psi}{2}+\phi_j) \tag{5}$$

The local similarity measurement for $\{P_i,P_j\}$ then casts a vote on the 2D (r, θ) space V:

$$\mathrm{V}(r,\theta)=\mathrm{V}(r,\theta)+Sim(P_i,P_j). \tag{6}$$

We collect the votes from all symmetric patch pairs in **M** to generate the Hough voting image V. In what follows is the peak detection on V to define the skeletons of salient symmetries with the criterion:

$$K:x\cos\theta+y\sin\theta=r \text{ if } \mathrm{V}(r,\theta)>\gamma \text{ for all } (r,\theta) \tag{7}$$

where γ is a pre-defined threshold. The member patch pair $P_{ij}=(P_i,P_j)$ to constitute a symmetry S with skeleton K characterized by (r, θ) can thus be defined as:

$$S=\{P_{ij}\mid P_{ij}\in \mathrm{M}\wedge P_s\in INV(r,\theta)\} \tag{8}$$

where INV(r, θ) is the inverse Hough transform on V(r, θ) that returns the set of patch pairs casting votes on (r, θ). Multiple peaks can be detected from V to locate multiple salient symmetric parts for the input image I. Note that the patch pairs not in **M** are supposed to be less similar and are excluded from casting a vote on the Hough voting image V. This avoids generating spurious peaks. Figure 5 shows an example to illustrate the Hough voting framework for symmetry detection.

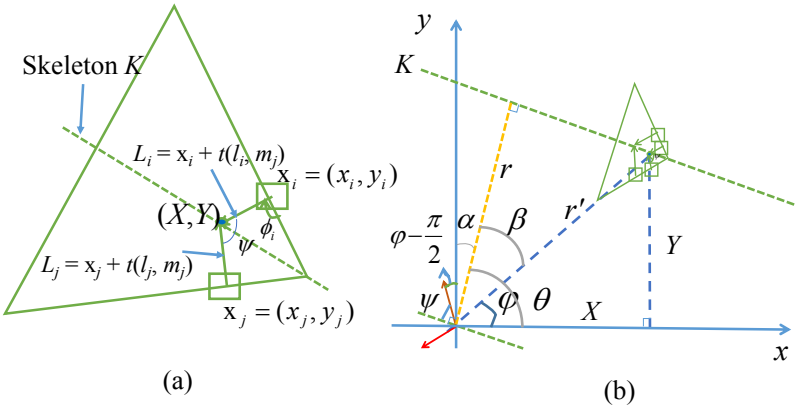

Figure 4. Determining the skeleton of a symmetric part using pairwise symmetric patches $\{P_i,P_j\}$. (**a**) The skeleton K of the corresponding symmetric part determined by the pairwise patches; (**b**) The skeleton K characterized by two parameters (r, θ).

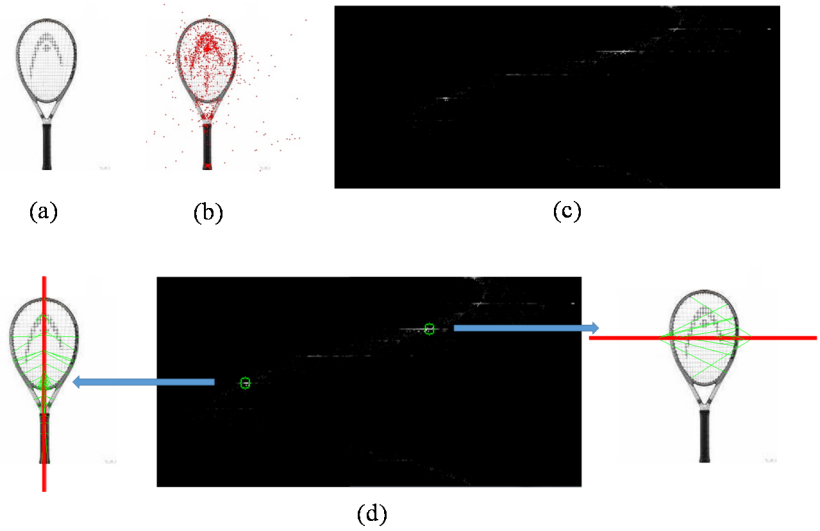

Figure 5. An example to illustrate the Hough voting framework for symmetry detection: (**a**) the original image; (**b**) the intersection points of symmetric patch pairs; (**c**) the Hough voting image; and (**d**) the peak detection and inverse Hough voting to compute the skeletons and symmetries.

3.4. Object Detection with Symmetry Graph Partitioning

The set of detected symmetries $\mathbf{S} = \{S_k\}_{k=1}^{s}$ can be used to locate multiple objects in I by merging highly correlated symmetries. Let $\mathbf{K} = \{K_k\}_{k=1}^{s}$ the set of skeletons to describe the symmetric axes of **S**. Every skeleton K_k is a line and characterized by two parameters (r_k, θ_k). As mentioned above, using (3), the (i,j)-th patch pair P_{ij} in S_k defines an intersection point $\bar{x}_{ij} = (X_{ij}, Y_{ij})$. These intersection points defined by patch pairs in S_k can be used to estimate the bounding rectangle that locates the corresponding symmetric part. To achieve this goal, we first compute the part center $\bar{x}^{(k)} = (X^{(k)}, Y^{(k)})$ as the mean of $\bar{x}_k$:

$$\bar{x}^{(k)} = \frac{1}{|S_k|}\sum\nolimits_{P_{ij} \in S_k} \bar{x}_{ij} \tag{9}$$

$$d_{ij} = \sqrt{(X_{ij} - X^{(k)})^2 + (Y_{ij} - Y^{(k)})^2} \tag{10}$$

where $|S_k|$ is the cardinality of S_k. We also compute the distances d_{ij} to measure the part elongation along the skeleton K_i, which is characterized by the line parameters (r_k, θ_k). Using (10), the potential outliers in S_k are defined as:

$$S_k^{(e)} = \left\{ P_{ij} \mid d_{ij} > 2\sigma_k, P_{ij} \in S_k \right\} \tag{11}$$

where $\sigma_k^2 = \frac{1}{|S_k|}\sum_{P_{ij} \in S_k} (d_{ij} - \frac{1}{|S_k|}\sum_{P_{ij} \in S_k} d_{ij})^2$ is the distance variance of S_k. To have a better estimation of the symmetry using S_k, we eliminate the outliers from the original S_k, *i.e.*, $S_k^{(new)} = S_k^{(old)} - S_k^{(e)}$.

We also define the line $K_k^{\perp}$ passing $\bar{x}^{(k)}$ and being orthogonal to K_k as:

$$K_k^{\perp} : -x\sin\theta_k + y\cos\theta_k = r_k^{\perp} \tag{12}$$

where $r_k^{\perp} = \sqrt{(X^{(k)} - r_k \cos\theta_k)^2 + (Y^{(k)} - r_k \sin\theta_k)^2}$. The line $K_k^{\perp}$ divides S_k into two parts according to the following rule:

$$P_{ij} \in \begin{cases} S_k^b & if \quad -X_{ij}\sin\theta_k + Y_{ij}\cos\theta_k - r_k^{\perp} < 0 \\ S_k^u & otherwise \end{cases} \tag{13}$$

The patch pairs to define the top and bottom boundaries of the bounding box B_k of S_k can thus be defined as:

$$(P_b, P_u) = (\arg\max_{P_{ij} \in S_k^b} d_{ij}, \arg\max_{P_{ij} \in S_k^u} d_{ij}) \tag{14}$$

where the distance function d is defined in (10). The lines L_b and L_u that are passing through the centers of P_b and P_u and parallel to $K_k^{\perp}$ then define the top and bottom boundaries of B_k, respectively.

Similarly, the skeleton K_k of S_k divides the patches in S_k into two parts:

$$P_i \in \begin{cases} S_k^l & if \quad x_i\cos\theta_k + y_i\sin\theta_k - r_k < 0 \\ S_k^r & otherwise \end{cases} \tag{15}$$

where (x_i, y_i) is the center of the patch P_i in S_k. The patches to define the left and right boundaries B_k of S_k can thus be defined as:

$$(P_l, P_r) = (\arg\max_{P_i \in S_k^l} d_i, \arg\max_{P_i \in S_k^r} d_i) \tag{16}$$

where $d_i = | x_i\cos\theta_k + y_i\sin\theta_k - r_k |$ is the distance between P_i and K_k. The lines L_l and L_r that are passing through the centers of P_l and P_r and parallel to K_k then define the left and right boundaries of B_k, respectively. Figure 6c shows the bounding boxes of detected symmetries in Figure 6b.

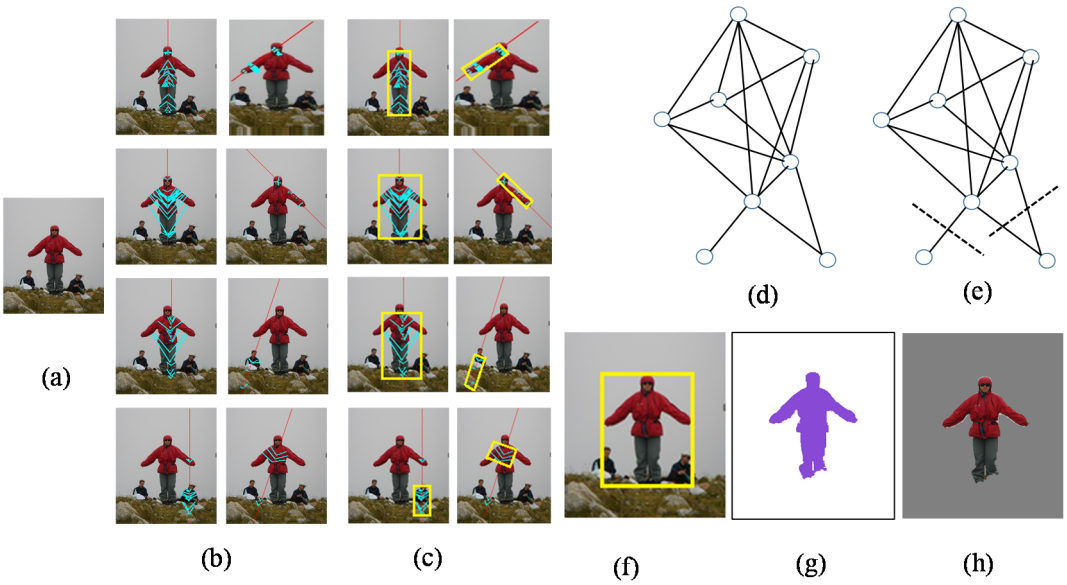

Figure 6. Object detection with the symmetry graph representation: (**a**) the original image; (**b**) the detected skeletons and symmetries; (**c**) the estimated bounding boxes of the symmetries in (b); (**d**) the symmetry graph; (**e**) graph partitioning by the dominant sets algorithm [28]; (**f**) the merged bounding box for the person object; (**g**) the contour of the object using the graphcut segmentation algorithm [24] on (f); and (**h**) the segmentation result.

The bounding boxes belonging to the same object might heavily overlap with each other. To locate objects in the input image based on the symmetry graph representation, this paper uses the well-known dominant sets algorithm [39] to merging symmetries into objects. We first construct a weighted symmetry graph G = (S, E) where S is the set of detected symmetries (the set of nodes) and E is the set of edges. The weight on an edge between nodes *i* and *j* is defined as:

$$w_{ij} = \frac{B_i \cap B_j}{B_i \cup B_j} \tag{17}$$

where B_i and B_j are the bounding boxes of symmetries *i* and *j*, respectively. Modeling an object as a dominant sets, the graph partitioning algorithm optimally divides the symmetry graph G into multiple sub-graphs, each of which merges its member symmetries into an object [39]. Obviously, the bounding box of an object might contain background information, which would degrade the performance of the resulting object classification. To tackle this difficulty, the graphcut algorithm [24] can be further used to eliminate the irrelevant background in the bounding boxes of the detected objects.

3.5. The Generative Model

We are now ready to describe the probabilistic generative model, which derives the foundation of object detection and segmentation. The underlying concept behind the graphical model is that, given the set of symmetric patch pairs **M**, we can sample an object patch $\{x_n, f_n\}$ where the patch center is x_n and the patch feature is f_n. The graphical model shown in Figure 1d tells us the joint distribution for a patch is

$$\mathrm{P}(x_n, f_n, \mathbf{B}, \mathbf{S}, \mathbf{M}, \mathbf{P}) = \mathrm{P}(x_n \mid \mathbf{B}, \mathbf{P})\,\mathrm{P}(f_n \mid \mathbf{P})\,\mathrm{P}(\mathbf{B}|\mathbf{S})\,\mathrm{P}(\mathbf{S}|\mathbf{M})\,\mathrm{P}(\mathbf{M} \mid \mathbf{P})\,\mathrm{P}(\mathbf{P}) \tag{18}$$

We first condition on x_n and f_n and assume both P(x_n) and P(f_n) to be constant. Then we condition on **P**, so the prior term P(**P**) is removed. Dividing both sides of (18) by P(x_n), P(f_n) and P(**P**), we get the following expression:

$$\mathrm{P}(\mathbf{B}, \mathbf{S}, \mathbf{M} \mid x_n, f_n, \mathbf{P}) = \mathrm{P}(x_n \mid \mathbf{B}, \mathbf{P})\,\mathrm{P}(f_n \mid \mathbf{P})\,\mathrm{P}(\mathbf{B}|\mathbf{S})\,\mathrm{P}(\mathbf{S}|\mathbf{M})\,\mathrm{P}(\mathbf{M}|\mathbf{P}) \tag{19}$$

To take product over the patch-wise posterior, the posterior probability to be maximized is

$$\mathrm{P}(\mathbf{B}, \mathbf{S}, \mathbf{M} \mid x_n, f_n, \mathbf{P}) = \prod_{n=1}^{N} \left\{ \mathrm{P}(x_n \mid \mathbf{B}, \mathbf{P})\,\mathrm{P}(f_n \mid \mathbf{P}) \right\} \mathrm{P}(\mathbf{B}|\mathbf{S})\,\mathrm{P}(\mathbf{S}|\mathbf{M})\,\mathrm{P}(\mathbf{M}|\mathbf{P}) \tag{20}$$

Now we explain each of the distribution terms in (20) in details. $\mathrm{P}(x_n \mid \mathbf{B}, \mathbf{P})$ is the probability of the pixel location x_n given the bounding box **B** and the set of sampled patches **P**. The function of this term is to select patches belonging to **P** and constrained by **B**. Similarly, $\mathrm{P}(f_n \mid \mathbf{P})$ is the probability of the patch feature f_n belonging to **P**. P(**B**|**S**) represents the probability of the bounding box **B** given the set of detected symmetries **S**, which is determined by **M** with the probability P(**S**|**M**). Finally, P(**M**|**P**) is the probability of patch pairs that are symmetrical with each other. The goal of our method is to seek the parameters of **B**, **M**, and **S** that maximize the posterior probability $\mathrm{P}(\mathbf{B}, \mathbf{S}, \mathbf{M} \mid x_n, f_n, \mathbf{P})$. To achieve the goal, a pre-learned object model should be built up using a generic training approach. However, the learning approach to build up a high-precision object model is obviously not a trivial work. Instead of the usage of the object model, the approach uses a greedy method to optimize $\mathrm{P}(\mathbf{B}, \mathbf{S}, \mathbf{M} \mid x_n, f_n, \mathbf{P})$.

The value of P(**M**|**P**) in (20) can thus be estimated by

$$\hat{\mathrm{P}}(\mathbf{M} \mid \mathbf{P}) = \frac{2 \mid \mathbf{M} \mid}{N^2} \tag{21}$$

where |**M**| is the size of **M**. Obviously, this value depends on the number of patch clusters. The probability P(**S**|**M**) can be further decomposed into

$$\mathrm{P}(\mathbf{S} \mid \mathbf{M}) = \sum\nolimits_{i=1,\ldots,k} \mathrm{P}(S_i \mid \mathbf{M})\,\mathrm{P}(S_i) \tag{22}$$

since we can detect k peaks in the Hough voting image V to locate the corresponding salient symmetries $\mathbf{S} = \{S_n\}_{n=1}^{k}$. To apply the inverse Hough voting, we can estimate the value of P(S_i) with the ratio of the number of patch pairs to construct S_i to the size of **M** and the value of P(S_i|**M**) by the voting value of the i-th peak in V. That is, the estimated value of P(S_i|**M**) can be computed by

$$\hat{\mathrm{P}}(\mathbf{S} \mid \mathbf{M}) = \frac{1}{\mid \mathbf{M} \mid \sum_{(r,\theta)} \mathbf{V}(r,\theta)} \sum_{i=1}^{k} \mid S_i \mid \mathbf{V}(r_i, \theta_i) \tag{23}$$

Finally, the dominant sets algorithm and the graph cut segmentation are used to optimize the terms P(**B**|**S**) and $\prod_{n=1}^{N} \{\mathrm{P}(x_n \mid \mathbf{B}, \mathbf{P})\,\mathrm{P}(f_n \mid \mathbf{P})\}$, respectively.

4. The Application to Human Activity Recognition

One obvious deficiency of unsupervised object detection and segmentation is that the semantic lack of detected objects. To tackle the difficulty, in constructing a real-world application, the object semantics could be augmented by a model. We use poselet models [22], shown in Figure 7, to explore the degree of the quality-conscious object detection and segmentation in improving the performance of human activity recognition.

To train a human activity classifier using SVMs, a dataset $\mathbf{D} = \{(V_i, y_i)\}_{i=1}^{T}$ is collected, where V_i is a video sample and y_i is the label of V_i. To build the multi-class activity model based on the symmetries-based object detection, we firstly perform a generic key frame detection [41] on the input video to obtain a compact video representation. Next, the proposed object detector divides every frame into multiple objects, in which the human objects are identified by a fast facial detection algorithm [42]. The detected human objects in key frames are then divided into J poselets, which localize discriminative parts of the body and are proven to be effective for human activity recognition [22]. Inspired from the work of [22] and based on a few weak annotations on a sparse set of frames, shown in Figure 8, two types of poselet features, including the HOG descriptors and the BoW features, are used for training the poselet detector. The BoW features, quantized dense descriptors (SIFT [43], histogram of optical flow (HOF) [44], and motion boundaries (HoMB) [45]), are used to augment the HOG descriptors for capturing the motion information of poselets. In this paper, the background information is removed from the poselets by the segmentation scheme, which, in turn, improves both the quality of the poselet models in the learning phase and the recognition accuracy in the testing phase.

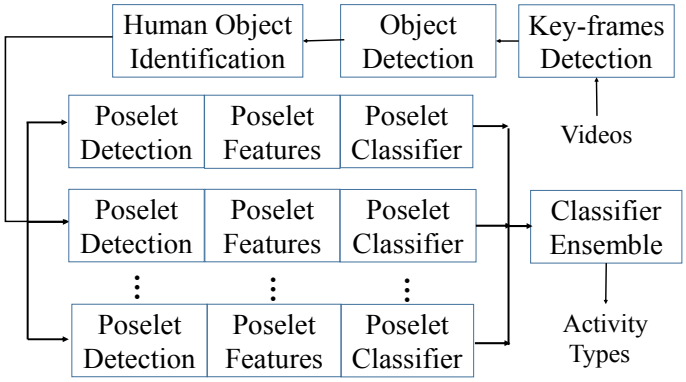

Figure 7. System flowchart of the application to human activity recognition.

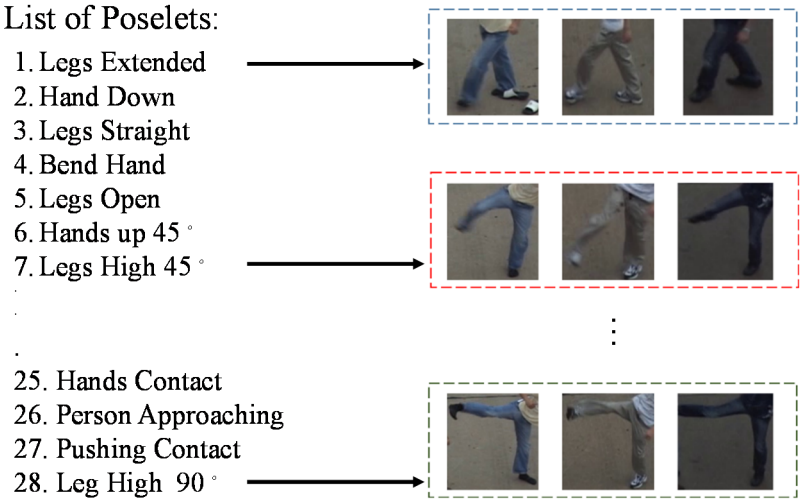

Figure 8. Examples of annotated poselets.

In the training phase, the annotated training samples are trained to learn poselet-specific HOG and BoW templates. In the testing phase, these poselet templates are used to locate the poselets in the human objects of a frame. For each video frame, we collect the highest scores from both HOG-based and BoW-based poselet templates by performing the branch-and-bound techniques [46] on the detected human objects to represent the frame as a poselet activation vector [47]. Our feature representation represents a video as a HOG-based feature sequence and three BoW-based feature sequences. Finally, for each poselet model, a SVM classifier with a multi-channel string kernel [48] is trained to form a part-based weak classifier. The multi-channel string kernel is defined as:

$$g(F,F') = \exp(-\sum_i \frac{1}{A_i} D_P(F_i, F_i')) \tag{24}$$

where F and F' are two multi-channel histogram-based feature sequences; (F_i, F_i') are the i-th channel feature sequences for (F,F') ; $D_P(F_i,F_i')$ is the distance between F_i and F_i' using dynamic programming; A_i is the average of D_p distances using the i-th channel features of training samples. These poselet SVMs are then bootstrapped to constitute an ensemble classifier for human activity recognition. The rule to classify the input video clip V, which is represented by k key frames, is thus

$$c(V) = \arg\max_{c \in C}(Score(c)), Score(c) = \sum_{j=1,\ldots,J} \alpha_j \delta(s_j(V) == c) \tag{25}$$

where A is the set of activity classes and s_j is the j-th poselet SVM classifier with the weighting factor α_j which is proportional to the accuracy of activity recognition using s_j. α_j was determined in the training phase.

5. Experimental Results

A series of experiments was conducted on an Intel CORE i7 3.0GHz PC and three datasets, The INRIA dataset [49], the PASCAL VOC 2012 dataset [50], and the UT-Interaction dataset [51], are constructed to evaluate the performance of the human object detection and activity recognition system. The INRIA dataset has been used in many static person detection studies. It annotates a training dataset including 614 positive samples and 1218 negative samples. Multiple poses are included in both the training and testing datasets. Also many different natural scenes are used to construct the set of negative examples. The size of the image in the INRIA dataset is 64 × 128. The PASCAL VOC 2012 dataset contains 20 object classes with all images taken from natural scenes. The train and validation dataset has 11,530 images containing 27,450 region of interest (ROI) annotated objects and 6929 segmentations. Among them, the person class has 632 images. The UT-Interaction dataset contains 20 videos of continuous executions of six classes of human-human interactions: hands shaking, pointing, hugging, pushing, kicking and punching. Ground truth labels for these interactions are provided, including time intervals and bounding boxes. Every video sequence taken with the resolution of 720 × 480, 30 fps, and the height of a person in the video is about 200 pixels. The lengths of video sequences are around one minute. Each video contains at least one execution per interaction, providing us eight executions of human activities per video on average. Several participants with more than 15 different clothing conditions appear in the videos. Furthermore, the dataset is divided into two sets. Set 1 is composed of 10 video sequences taken on a parking lot. The videos of set 1 are taken with slightly different zoom rate, and their backgrounds are mostly static with little camera jitter. Set 2 (*i.e.*, the other 10 sequences) are taken on a lawn in a windy day. Background is moving slightly (e.g., tree swaying), so they contain more camera jitters. Each set has a different background, scale, and illumination. Figure 9 shows several images of these three datasets.

First of all, to clarify the differences between the proposed unsupervised object segmentation method and the standard image segmentation method, the graph cut algorithm [24] is implemented, which is used to segment images into regions. Notice that the segmentation results of both the proposed and graph cut algorithms contain multiple objects in an image. However, the latter does not group regions into objects. On the contrary, our method spans a new way to group detected symmetries into objects using a symmetry graph partition algorithm. The contour of the target object can also be obtained by intersecting the detected object with the segmented regions. Thus, the proposed method solves the problem of image segmentation in object segmentation. Incorporating the segmentation results of the graph cut algorithm into the object detection approach, Figures 10–12 show examples of the object detection and segmentation using the three datasets. To compare the performance between the proposed method and regions with CNN features (R-CNN) method [32], the detection quality judged subjectively for both methods is compatible. Note that R-CNN trains high-capacity convolutional neural networks (CNNs) in advance to the bottom-up region proposals in order to localize and segment objects. Accordingly, the symmetry detection unequivocally facilitates effective object detection and segmentation without the object models.

Figure 9. Example images of the datasets: (**a**) the positive and negative training samples of the INRIA dataset; (**b**) examples of the "person" class of the PASCAL VOC 2012 dataset; (**c**) sample interactions in the UT-Interaction dataset.

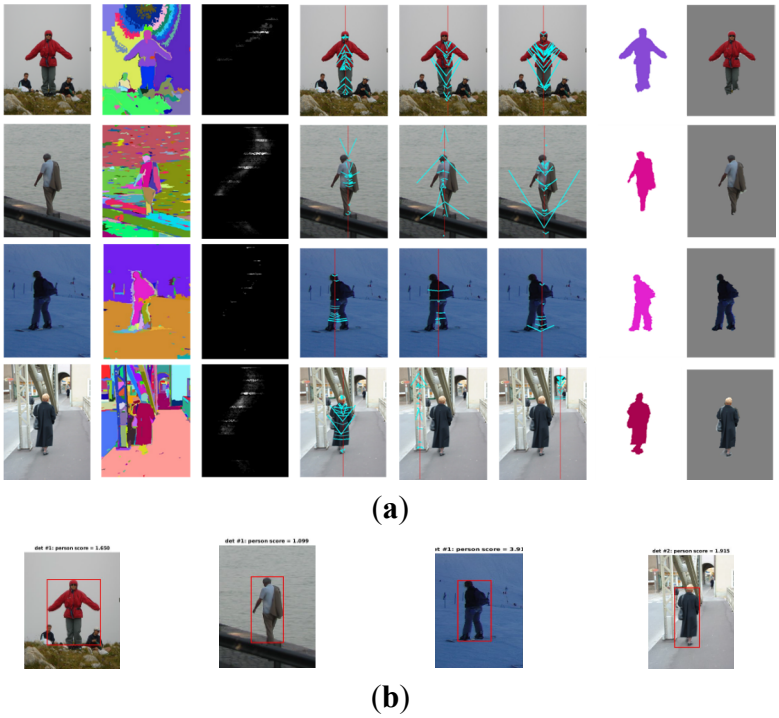

Figure 10. The object detection and segmentation results of the compared approaches using the dataset INRIA: (**a**) each row from left to right: original image; segmentation result of the graph cut algorithm; Hough voting image; major detected symmetric parts; object mask; and segmentation result of the proposed method; (**b**) the detection results [52] of R-CNN [32] on images from Flickr. The training data are from PASCAL VOC.

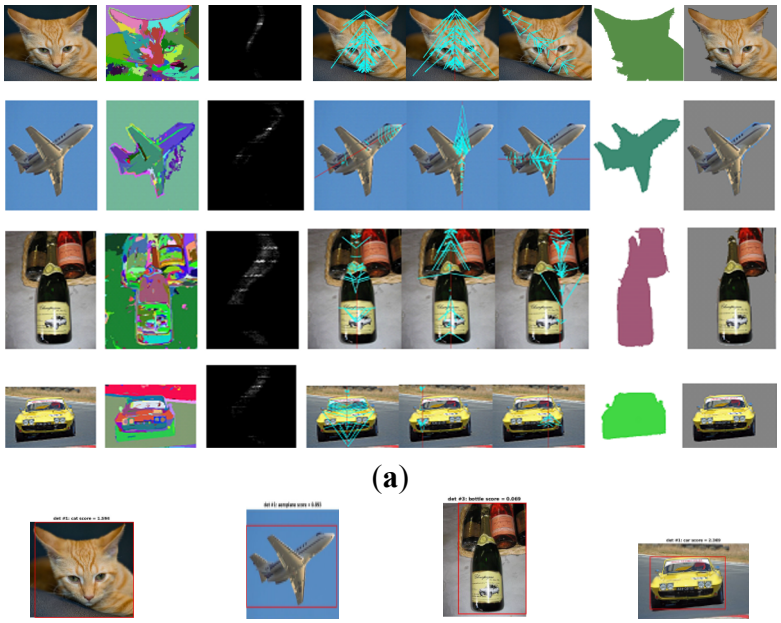

Figure 11. The object detection and segmentation results of the compared approaches using the dataset PASCAL 2012: (**a**) each row from left to right, original image; segmentation result of the graph cut algorithm; Hough voting image; major detected symmetric parts; the object mask; and segmentation result of the proposed method; (**b**) the detection results [52] of R-CNN [32] on images from Flickr. The training data are from PASCAL VOC.

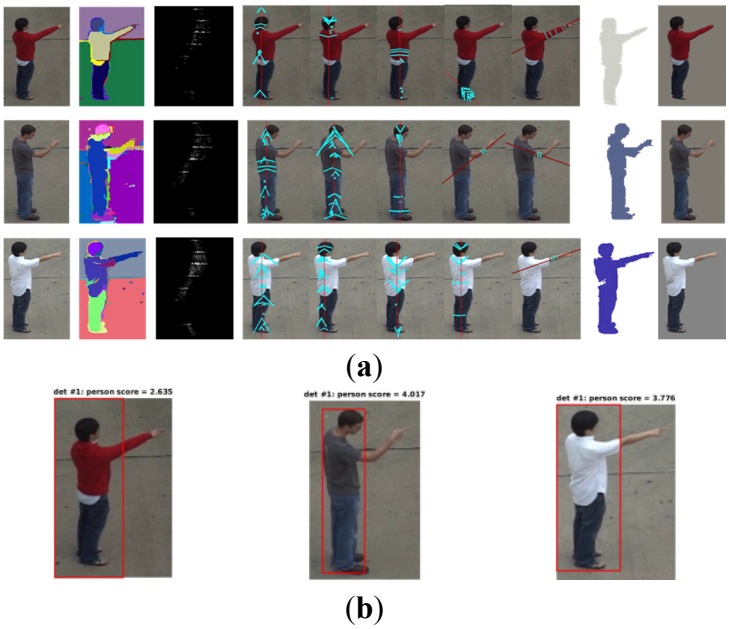

Figure 12. The object detection and segmentation results of the compared approaches using the dataset UT-Interaction: (**a**) each row from left to right: original image; segmentation result of the graph cut algorithm; Hough voting image; major detected symmetric parts; object mask; and segmentation result of the proposed method; (**b**) the detection results [52] of R-CNN [32] on images from Flickr. The training data are from PASCAL VOC.

The class labels, as ground truth for images in the test datasets, are used to determine the accuracy of human object detection. For the proposed approach, the problem of human object detection is tackled by automatically locating objects with facial parts in the detected object set of an image. That is, we do not need constructing a person classifier, which is necessary for many existing person detectors. To test the effectiveness of the person detector, classification results are shown in Table 1 for the proposed and compared state-of-the-art recognition systems [1,53–59]. The proposed approach outperforms the compared methods since symmetric properties are salient features in person objects.

We follow the same localization evaluation rule in [22]: a detection is considered correct if, (1) the poselets in a human object are correctly classified, and (2) the intersection-union ratio of the detection and ground truth bounding box is not less than a threshold θ. For the UT-Interaction dataset, selected frames were hand-annotated with bounding boxes, and the bounding boxes for the frames in between were generated by linear interpolation. Table 2 shows the performance comparison in poselet localization accuracy using the dataset UT-Interaction. The proposed method has a better result compared to [22] because our features in constructing the poselet detectors are from more accurate results of human object detection and segmentation. Thus, our final features contain less irrelevant background in representing the corresponding poselets. Moreover, we detect poselets from human objects detected in the previous step of the approach. Consequently, this increases the robustness in the poselet detection. Figure 13 also shows examples of poselet detection using the UT-Interaction dataset.

Table 1. Performance comparison in person detection using the datasets INRIA and PASCAL 2012.

Dataset	Methods	Detection Accuracy	Dataset	Methods	Detection Accuracy
INRIA	Proposed	78.1%	PASCAL 2012	Proposed	50.1%
	HOG-LBP [53]	61.5%		CVC_CLS [54]	42.3%
	LARSVM-V2 [1]	77.3%		NEC [55]	32.8%
	MULTIFER+CSS [56]	75.0%		SYSU_DYNAMIC[57]	37.5%
	FEATSNTH [58]	69.0%		OXFORD [59]	46.1%

Table 2. Performance comparison in poselet detection using the dataset UT-Interaction. A detection is considered correct if, (1) the poselets in a human object are correctly classified, and (2) the intersection-union ratio of the detection and ground truth bounding box is not less than a threshold θ.

Threshold	Methods	Accuracy
θ = 0.25	Proposed	**100%**
	Raptis *et al.* [22]	86.7%
θ = 0.50	Proposed	**100%**
	Raptis *et al.* [22]	86.7%
θ = 0.75	Proposed	**85.4%**
	Raptis *et al.* [22]	83.3%
θ = 1	Proposed	**81.3%**
	Raptis *et al.* [22]	80.0%

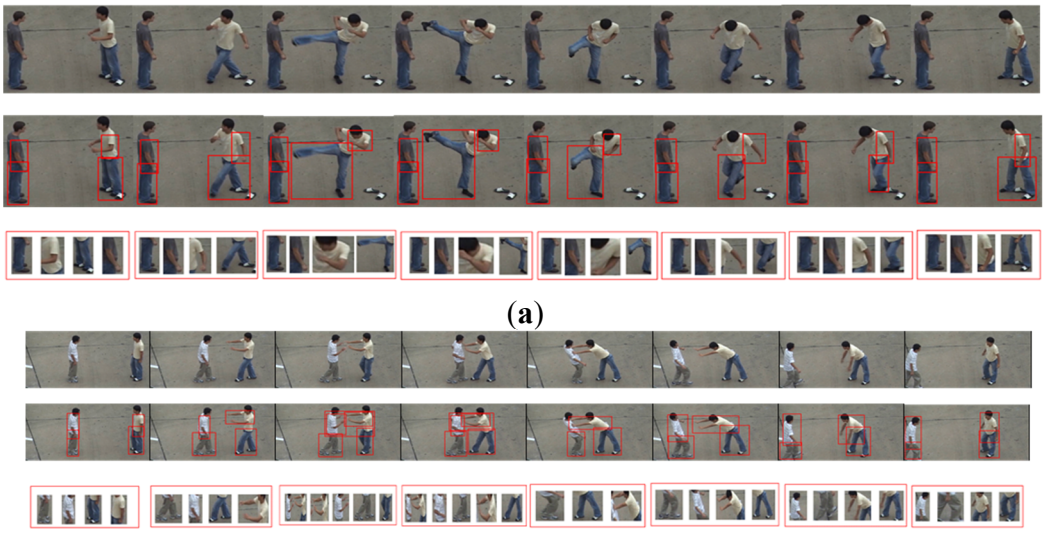

Figure 13. Examples of poselet detection using the UT-Interaction dataset: (**a**) a "Kick" activity; (**b**) a "Push" activity. The bounding boxes locate the detected poselets in individual frames.

Evaluations of our approach in human activity recognition are carried out with a leave-one-out cross-validation method. Classification results are shown in Table 3 and compared with state-of-the-art recognition systems [22,60–68]. Accordingly, the proposed method has a great improvement in classification accuracy. Note that both the poselet models and the feature setting to describe poselets in the approach of Raptis *et al.* [22] are adopted in our human activity recognition. However, the proposed approach has better performance in terms of classification accuracy. This is because the detected poselets are more accurate compared to those of [22]. Figure 14 shows the confusion matrices of the UT-Interaction dataset for the proposed and the method by Raptis *et al.* Both matrices show similar confusion patterns. This shows that poselet models are effective in human activity recognition. The detection of symmetries is not always accurate in the class "Kick" because the symmetries to constitute the poselets in this class are often occluded with each other. This degrades the accuracy to recognize "Kick" activities.

Table 3. Comparison of UT-Interaction classification with other methods. "–" indicates the data is not provided in the original papers.

Method	Set 1	Set 2	Total
Proposed	**96.6%**	**91.6%**	**94.1%**
Patron-Perez *et al.* [60]	84%	86%	85%
Waltisberg *et al.* [61]	88%	77%	82.5%
Vahdat *et al.* [62]	93%	90%	91.5%
Yu *et al.* [63]	–	–	91.7%
Burghouts *et al.* [64]	–	–	88.3%
Raptis *et al.* [22]	–	–	93.3%
Mukherjee *et al.* [65]	85%	73.3%	79.17%
Ryoo [66]	–	–	85%
Kong *et al.* [67]	–	–	88.3%
Zhang *et al.* [68]	95%	90%	92.5%

Class	Hand Shake	Hug	Kick	Point	Punch	Push
Hand Shake	100	0	0	0	0	0
Hug	0	100	0	0	0	0
Kick	0	0	85	15	0	0
Point	0	0	0	100	0	0
Punch	0	0	10	0	90	0
Push	0	0	0	0	10	90

(a)

Class	Hand Shake	Hug	Kick	Point	Punch	Push
Hand Shake	100	0	0	0	0	0
Hug	0	100	0	0	0	0
Kick	0	0	90	10	0	0
Point	0	0	0	100	0	0
Punch	0	0	20	0	80	0
Push	0	0	0	0	10	90

(b)

Figure 14. Confusion matrices for UT-Interaction data: (**a**) proposed method; (**b**) Raptis *et al*. [22].

6. Conclusions

In this paper, we have presented an interesting approach for unsupervised object detection and segmentation, based on the fusion of symmetries detection, dominate sets clustering, and image segmentation. To use the object detection and segmentation as a processing, we also have presented a systematic way to construct a bank of poselet SVM classifiers for human activity recognition. The proposed activity recognition modeling encodes every video as a sequence of multi-channel histogram-based feature sequences. Multi-channel string kernels are thus introduced to improve the recognition accuracy of week classifier with individual poselet models. For each class, a set of training videos is also used to train an ensemble classifier, which verifies the correctness of the candidate detected human activities at testing time.

Compared with related human object detection and activity recognition methods, the proposed method makes a significant contribution: this paper formulates the problem of object detection through symmetries detection. Not only can the dynamic programming process model the activities of training videos as multi-channel poselet feature sequences, the procedure can also be used to detect and recognize human objects from the input video clip automatically. Our system presents an approach to detect multiple human objects from a video clip. Experimental results show that the proposed method performs well on several publicly available datasets in terms of detection accuracy and recognition rate.

The proposed method, however, suffers from the following limitations. The computational complexity of our approach using class-specific model matching through dynamic programming and Hough voting is essentially high. Future work will focus on implementing the system on parallel architecture, e.g., a GPU servers and cloud computing platforms.

Acknowledgments

This work was supported in part by Ministry of Science and Technology, Taiwan under Grant Number MOST 103-2221-E-019-018-MY2.

Author Contributions

Jui-Yuan Su and Shyi-Chyi Cheng designed and performed experiments, analyzed data and wrote the paper; De-Kai Huang designed and performed experiments. All authors contributed equally to the paper.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645.
2. Leibe, B.; Leonardis, A.; Schiele, B. Robust object detection with interleaved categorization and segmentation. *Int. J. Comput. Vis.* **2008**, *77*, 259–289.
3. Brendel, W.; Todorovic, S. Video object segmentation by tracking regions. In Proceedings of 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 833–840.
4. Brox, T.; Malik, J. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 500–513.
5. Brox, T.; Malik, J. Object segmentation by long term analysis of point trajectories. In *Computer Vision—ECCV 2010*; Daniilidis, K., Maragos, P., Paragios, N., Eds.; Springer Berlin Heidelberg: Berlin, Germany, 2010; Volume 6315, pp. 282–295.
6. Turaga, P.; Chellappa, R.; Subrahmanian, V.S.; Udrea, O. Machine recognition of human activities: A survey. *IEEE Trans. Circ. Syst. Video* **2008**, *18*, 1473–1488.
7. Tsaig, Y.; Averbuch, A. Automatic segmentation of moving objects in video sequences: A region labeling approach. *IEEE Trans. Circ. Syst. Video* **2002**, *12*, 597–612.
8. Carreira, J.; Sminchisescu, C. Constrained parametric min-cuts for automatic object segmentation. In Proceedings of 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 3241–3248.
9. Chien, S.Y.; Huang, Y.W.; Chen, L.G. Predictive watershed: A fast watershed algorithm for video segmentation. *IEEE Trans. Circ. Syst. Video* **2003**, *13*, 453–461.
10. Cheng, S.C. Visual pattern matching in motion estimation for object-based very low bit-rate coding using moment-preserving edge detection. *IEEE Trans. Multimed.* **2005**, *7*, 189–200.
11. Cheng, S.C.; Wu, T.L. Scene-adaptive video partitioning by semantic object tracking. *J. Vis. Commun. Image Represent.* **2006**, *17*, 72–97.
12. Angelova, A.; Shenghuo, Z. Efficient object detection and segmentation for fine-grained recognition. In Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 811–818.
13. Arbelaez, P.; Maire, M.; Fowlkes, C.; Malik. J. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 898–916.

14. Bibby, C.; Reid, I. Real-time tracking of multiple occluding objects using level sets. In Proceedings of 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 1307–1314.
15. Yilmaz, A.; Javed, O.; Shah, M. Object tracking: A survey. *ACM Comput. Surv.* **2006**, *38*, 1–45.
16. Zhang, H.; Fritts, J.E.; Goldman, S.A. Image segmentation evaluation: A survey of unsupervised methods. *Comput. Vis. Image Underst.* **2008**, *110*, 260–280.
17. Chuang, C.H.; Cheng, S.C.; Chang, C.C.; Chen, Y.P.P. Model-based approach to spatial-temporal sampling of video clips for video object detection by classification. *J. Vis. Commun. Image Respresent.* **2014**, *25*, 1018–1030.
18. Xie, C.J.; Tan, J.Q.; Chen, P.; Zhang, J.; He, L. Collaborative object tracking model with local sparse representation. *J. Vis. Commun. Image Respresent.* **2014**, *25*, 423–434.
19. Li, X.; Hu, W.M.; Shen, C.H.; Zhang, Z.F.; Dick, A.; van den Hengel, A. A survey of appearance models in visual object tracking. *ACM Trans. Intell. Syst. Technol.* **2013**, *4*, 1–48.
20. Ballard, D.H. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognit.* **1981**, *13*, 111–122.
21. Viola, P.; Jones, M.J. Robust real-time face detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154.
22. Raptis, M.; Sigal, L. Poselet key-framing: A model for human activity recognition. In Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; IEEE Computer Society: Washington, DC, USA, 2013; pp. 2650–2657.
23. Csurka, G.; Perronnin, F. An efficient approach to semantic segmentation. *Int. J. Comput. Vis.* **2011**, *95*, 198–212.
24. Boykov, Y.; Funka-Lea, G. Graph cuts and efficient N-D image segmentation. *Int. J. Comput. Vis.* **2006**, *70*, 109–131.
25. Rubinstein, M.; Joulin, A.; Kopf, J.; Liu, C. Unsupervised joint object discovery and segmentation in internet images. In Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; IEEE Computer Society: Washington, DC, USA, 2013; pp. 1939–1946.
26. Verbeek, J.; Triggs, B. Region classification with Markov field aspect models. In Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
27. Gu, C.; Lim, J.J.; Arbeláez, P.; Malik, J. Recognition using regions. In Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 1030–1037.
28. Arbelaez, P.; Hariharan, B.; Gu, C.; Gupta, S.; Bourdev, L.; Malik, J. Semantic segmentation using regions and parts. In Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 3378–3385.
29. Borenstein, E.; Ullman, S. Combined top-down/bottom-up segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 2109–2125.
30. Boix, X.; Gonfaus, J.M.; van de Weijer, J.; Bagdanov, A.D.; Serrat, J.; Gonzàlez, J. Harmony potentials fusing global and local scale for semantic image segmentation. *Int. J. Comput. Vis.* **2012**, *96*, 83–102.

31. Lucchi, A.; Yunpeng, L.; Boix, X.; Smith, K.; Fua, P. Are spatial and global constraints really necessary for segmentation? In Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 9–16.
32. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
33. Nicolescu, M.; Medioni, G. A voting-based computational framework for visual motion analysis and interpretation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 739–752.
34. Xiang, Y.; Li, S. Symmetric object detection based on symmetry and centripetal-sift edge descriptor. In Proceedings of the 2012 21st International Conference on Pattern Recognition (ICPR), Tsukuba, Japan, 11–15 November 2012; pp. 1403–1406.
35. Loy, G.; Eklundh, J.-O. Detecting symmetry and symmetric constellations of features. In Proceedings of the 9th European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; Volume 2, pp. 508–521.
36. Hsieh, J.W.; Chen, L.C.; Chen, D.Y. Symmetrical surf and its applications to vehicle detection and vehicle make and model recognition. *IEEE Trans. Intell. Transp.* **2014**, *15*, 6–20.
37. Mitra, N.J.; Pauly, M.; Wand, M.; Ceylan, D. Symmetry in 3D geometry: Extraction and applictions. *Comput. Graph. Forum* 2013, 32, 1–23.
38. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893.
39. Hsiao, P.C.; Chang, L.W. Image denoising with dominant sets by a coalitional game approach. *IEEE Trans. Image Process* **2013**, *22*, 724–738.
40. Kanungo, T.; Mount, D.M.; Netanyahu, N.S.; Piatko, C.D.; Silverman, R.; Wu, A.Y. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 881–892.
41. Liu, T.C.; Kender, J.R. Computational approaches to temporal sampling of video sequences. *ACM Trans. Multimed. Comput.* **2007**, *3*, 1–23.
42. Dornaika, F.; Ahlberg, J. Fast and reliable active appearance model search for 3-d face tracking. *IEEE Trans. Syst Man Cybern. Part B* **2004**, *34*, 1838–1853.
43. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110.
44. Chaudhry, R.; Ravichandran, A.; Hager, G.; Vidal, R. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 1932–1939.
45. Raptis, M.; Kokkinos, I.; Soatto, S. Discovering discriminative action parts from mid-level video representations. In Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 1242–1249.
46. Lampert, C.H.; Blaschko, M.B.; Hofmann, T. Efficient subwindow search: A branch and bound framework for object localization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 2129–2142.

47. Maji, S.; Bourdev, L.; Malik, J. Action recognition from a distributed representation of pose and appearance. In Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 3177–3184.
48. Ullah, M.M.; Parizi, S.N.; Laptev, I. Improving bag-of-features action recognition with non-local cues. In Proceedings of the British Machine Vision Conference, Aberystwyth, UK, 31 August–3 September, 2010; Labrosse, F., Zwiggelaar, R., Liu, Y., Tiddeman, B., Eds.; British Machine Vision Association: Durham, UK, 2010; pp. 1–11.
49. Inria Person Dataset. Available online: http://pascal.inrialpes.fr/data/human/ (accessed on 27 November 2014).
50. The PASCAL Visual Object Classes Homepage. Available online: http://pascallin.ecs.soton.ac.uk/challenges/VOC/ (accessed on 27 November 2014).
51. Ryoo, M.S.; Aggarwal, J.K. UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA). Available online: http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html (accessed on 27 November 2014).
52. Playing around with RCNN, State of the Art Object Detector. Available online: http://cs.stanford.edu/people/karpathy/rcnn/ (accessed on 16 March 2015).
53. Wang, X.; Han, T.X.; Yan, S. An HOG-LBP human detector with partial occlusion handling. In Proceedings of 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 32–39.
54. Khan, F.S.; van de Weijer, J.; Vanrell, M. Modulating shape features by color attention for object recognition. *Int. J. Comput. Vis.* **2012**, *98*, 49–64.
55. Russakovsky, O.; Lin, Y.; Yu, K.; Li, F.-F. Object-centric spatial pooling for image classification. In *Computer Vision–ECCV 2012*; Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., Eds.; Springer Berlin Heidelberg: Berlin, Germany, 2012; pp. 1–15.
56. Walk, S.; Majer, N.; Schindler, K.; Schiele, B. New features and insights for pedestrian detection. In Proceedings of 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; IEEE: San Francisco, CA, USA, 2010; pp. 1030–1037.
57. Wang, X.; Lin, L.; Huang, L.; Yan, S. Incorporating structural alternatives and sharing into hierarchy for multiclass object recognition and detection. In Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 3334–3341.
58. Bar-Hillel, A.; Levi, D.; Krupka, E.; Goldberg, C. Part-based feature synthesis for human detection. In Proceedings of the 11th European Conference on Computer Vision: Part IV; Daniilidis, K., Maragos, P., Paragios, N., Eds.; Springer Berlin Heidelberg: Berlin, Germany, 2010; Volume 6314, pp. 127–142.
59. Hoai, M.; Ladicky, L.; Zisserman, A. Action recognition from weak alignment of body parts. In Proceedings of the British Machine Vision Conference, Nottingham, UK, 1–5 September 2014; Valstar, M., French, A., Pridmore, T., Eds.; BMVA Press: Durham, England, UK, 2014; pp. 1–12.
60. Patron-Perez, A.; Marszalek, M.; Reid, I.; Zisserman, A. Structured learning of human interactions in TV shows. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2441–2453.

61. Waltisberg, D.; Yao, A.; Gall, J.; Gool, L.V. Variations of a hough-voting action recognition system. In Proceedings of the 20th International Conference on Recognizing Patterns in Signals, Speech, Images, and Videos, Istanbul, Turkey, 23–26 August 2010; Springer-Verlag: Istanbul, Turkey, 2010; pp. 306–312.
62. Vahdat, A.; Gao, B.; Ranjbar, M.; Mori, G. A discriminative key pose sequence model for recognizing human interactions. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 1729–1736.
63. Yu, G.; Yuan, J.; Liu, Z. Predicting human activities using spatio-temporal structure of interest points. In Proceedings of the 20th ACM international conference on Multimedia, Nara, Japan, 29 October–2 November 2012; ACM: New York, NY, USA, 2012; pp. 1049–1052.
64. Burghouts, G.J.; Schutte, K.; Bouma, H.; den Hollander, R.J.M. Selection of negative samples and two-stage combination of multiple features for action detection in thousands of videos. *Mach. Vis. Appl.* **2014**, *25*, 85–98.
65. Mukherjee, S.; Biswas, S.K.; Mukherjee, D.P. Recognizing interaction between human performers using "key pose doublet". In Proceedings of the 19th ACM international conference on Multimedia, Scottsdale, AZ, USA, 28 November–1 December, 2011; ACM: New York, NY, USA, 2011; pp. 1329–1332.
66. Ryoo, M.S. Human activity prediction: Early recognition of ongoing activities from streaming videos. In Proceedings of 2011 IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 1036–1043.
67. Kong, Y.; Jia, Y.; Fu, Y. Learning human interaction by interactive phrases. In *Computer Vision—ECCV 2012*; Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., Eds.; Springer Berlin Heidelberg: Berlin, Germany, 2012; Volume 7572, pp. 300–313.
68. Zhang, Y.; Liu, X.; Chang, M.-C.; Ge, W.; Chen, T. Spatio-temporal phrases for activity recognition. In *Computer Vision—ECCV 2012*; Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., Eds.; Springer Berlin Heidelberg: Berlin, Germany, 2012; Volume 7574, pp. 707–721.