A Framework for Combined Recognition of Actions and Objects

Ilktan Ar[1,2] and Yusuf Sinan Akgul[2]

[1] Kadir Has University, Cibali, Istanbul 34083, Turkey,
ilktana@khas.edu.tr
[2] Gebze Institute of Technology, Gebze, Kocaeli 41400, Turkey,
GIT Vision Lab: http://vision.gyte.edu.tr
akgul@bilmuh.gyte.edu.tr

Abstract. This paper proposes a novel approach to recognize actions and objects within the context of each other. Assuming that the different actions involve different objects in image sequences and there is one-to-one relation between object and action type, we present a Bayesian network based framework which combines motion patterns and object usage information to recognize actions/objects. More specifically, our approach recognizes high-level actions and the related objects without any body-part segmentation, hand tracking, and temporal segmentation methods. Additionally, we present a novel motion representation, based on 3D Haar-like features, which can be formed by depth, color, or both images. Our approach is also appropriate for object and action recognition where the involved object is partially or fully occluded. Finally, experiments show that our approach improves the accuracy of both action and object recognition significantly.

Keywords: Action and object recognition, Bayesian Network, motion pattern

1 Introduction

Object recognition is still a challenging problem in Computer Vision. There has been considerable research on object detection and recognition [1, 2]. Although most of the object recognition methods are appearance-based (shape, texture, etc.) some of them use contextual information such as text, scene, and human interactions. Carbonetto et al. [3] developed a caption/text guided object detection system which attaches meaningful labels to specific regions of an image and learns spatial relationships between objects. Torralba [4] introduced a framework for object recognition by using the correlation between scenes and the objects.

The idea of using the relationship between actions and objects has been exploited before. Filipovych and Ribeiro [5] recognized primitive actor-object interactions with the concept of actor-object states. But modeling high-level actions such as taking a picture with a digital camera is more complex than a primitive action such as grasping a spoon. Wu et al. [6] developed a dynamic

Bayesian network model which combines RFID and video data to jointly infer the most likely activity and object labels. Kjellstrm et al. [7] investigated object categorization according to its function. They simultaneously segment and classify hand actions with the detection and classification of the involved objects. But their method is dependent on tracking and reconstructing the hand pose.

Human action recognition is one of the most challenging topics in computer vision. The aim of human action recognition frameworks is to recognize human actions from offline/live videos. This task is very difficult because it deals with illumination and view point variations, perspective effects, scene variations, occlusions, individual variations of people due to appearance, clothing, motion, etc [8, 9].

In this paper, we examine the role of action understanding in object classification and vice-versa. With the assumption of one-to-one relation between object and action type exists, we propose a novel approach that recognizes actions and objects within the context of each other. Towards this approach, a Bayesian network based framework which combines motion patterns and object usage information is developed. Moreover, the motion patterns which describe motion information in an image sequence are formed by a novel motion representation. This representation is based on 3D Haar-like features and can be created using depth, color, or both image sequences. Finally, using the proposed approach, the accuracy of both action and object recognition are improved.

The rest of this paper is organized as follows. Section 2 presents the dataset. Section 3 describes the proposed framework with related feature representations. Section 4 demonstrates experimental results. Finally, Section 5 concludes the paper.

Table 1. Actions with the related objects.

Action	**Object**
Pour liquid in a cup	Pitcher
Drink	Cup
Use a brush	Brush
Use a remote control	Remote control
Use a roller	Roller
Use a calculator	Calculator
Make a phone call	Phone
Wear headphones	Headphones
Play with a videogame	Gameboy
Take a picture	Camera
Use a pen	Pen

2 Dataset

The public dataset recorded by Gall et al. [10] is modified and used for evaluation. This dataset consists of video sequences acted by 6 different actors. These actors perform 13 different actions. Actions are: 'Pour liquid in a cup', 'Drink with the left hand', 'Drink with the right hand', 'Use a brush', 'Use a remote control', 'Use a roller', 'Use a puncher', 'Use a calculator', 'Make a phone call', 'Wear headphones', 'Play with a videogame', 'Take a picture', and 'Use a pen'. Image sequences are stored as 640x480 RGB images and 144x88 256 gray-level depth images. The sample RGB images are displayed in Fig. 1b.

In this work, we focus on the actions which involve interaction with one object in each image of the image sequences. Therefore, we modify the original dataset in [10]. For example 'Use a puncher' action is discarded because this action involves a paper and a puncher interaction at the same time. 'Drink with the left hand' and 'Drink with the right hand' actions are merged as 'Drink'. The relationship between actions and objects are summarized in Table 1.

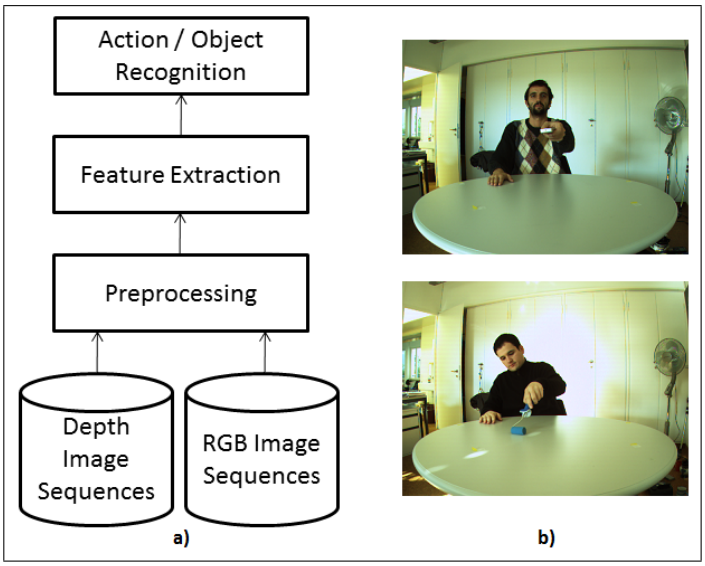

Fig. 1. a) The design of our framework, b) The sample RGB images taken from 'Use a remote control' and 'Use a roller' image sequences.

3 Our Framework

Our framework includes 3 main stages: Preprocessing, Feature extraction, and the Action/Object recognition as shown in Fig. 1a. In the preprocessing stage

RGB images are converted to 256 gray-level images (named as G-RGB in this paper). Then a 3x3 median filter is applied to remove noise in G-RGB images. In the feature extraction stage, object and motion information is obtained by forming the related representations. In the action/object recognition stage, a Bayesian network is built to recognize actions/objects by using object and motion representations which are obtained at feature extraction stage.

3.1 Feature Extraction

Feature extraction in image sequences defines representations of motion and object information. Motion information is built by motion patterns in image sequences. Depth, G-RGB, or both images can be used to extract motion information. Object information represents object availability in still images.

Representation of Motion Information Motion information in image sequences is the main element of action recognition. To extract motion patterns we propose a novel approach by adopting 3D Haar-like features which are used to detect pedestrians in [11].

In our approach, the motion information contained in the whole image sequence is represented by 16 different cubic filters as shown in Fig. 2. 3D Haar-like features are extracted by applying these cubic filters to depth and/or G-RGB image sequences with a convolution process. 3D Haar-like features are normalized to 0-255 interval for efficiency. Local motion information, LMI, between

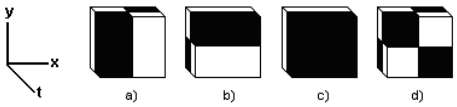

Fig. 2. Cubic filters which are used in the extraction of 3D Haar-Like features (8x8 and 4x4 pixel size in spatial domain, 4 and 8 frames in temporal domain).

consecutive images is calculated by the histogram of 3D Haar-like features as

$$LMI_w(t, f) = Histogram[HF_w(x, y, t, f)], \tag{1}$$

where HF describes 3D Haar-like features, x and y are 2D coordinates, t is the order of image, f is the filter id of our 16 different cubic filters, and w is the image sequence id (it can be either a G-RGB or a depth sequence). To obtain a global motion representation GMR for whole image sequence, variance (var) and mean (μ) are employed as

$$GMR_1(w, f) = \mu(LMI_w(t, f)), \tag{2}$$

$$GMR_2(w, f) = var(LMI_w(t, f)). \tag{3}$$

Finally, the motion information MI in a given image sequence w is represented with the concatenation ($||$) process as

$$MI(w) = GMR_1(w, f_1)||...||GMR_2(w, f_n), \tag{4}$$

where n is the maximum filter id. Representation of motion information can be built for either a G-RGB or depth image sequence separately or combination of them by enlarging the concatenation process defined in (4).

Representation of Object Information Object information in the image sequences reveals important cues about the type of the action. Although the depth image sequences can be used to detect and classify objects, we prefer using G-RGB image sequences to use state of the art methodologies effectively.

Images are selected at predefined uniform time-intervals (one image out of 20 images) in order to represent object information for a given image sequence. Then the object detection algorithm in [12], which uses bag of words models to detect and classify objects, is adopted to check the availability of the corresponding object in the selected images. The count of images which include the corresponding object are calculated and divided by the total number of selected images. The corresponding objects with related actions are given in Table 1. Note that these ratios, $OI(w, o)$ (where w is the sequence id, o is the object id), represent the object information in the image sequence.

3.2 Action/Object Recognition

In Action/Object recognition stage we aim to classify the action/object in the given image sequence by using the representations obtained at the feature extraction stage. For this classification problem, we need a classifier that assigns an action/object label $a \in A$ to the image sequence. There are two general approaches available to the classification problem as generative or discriminative. Examples of discriminative classifiers are Neural Networks, Additive Models, and Logical Regression. Examples of generative classifiers includes Hidden Markov Models, Fisher Discriminant Analysis, and Bayesian Networks [13]. We prefer to use a Bayesian network structure for this classification problem because of the robustness of Bayesian networks for representing of joint distributions and encoding conditional independence assumptions.

The Bayesian network uses the graphical model in Fig. 3. to represent conditional independence relationships between random variables: action/object type (A), object availability (O), motion information (M), and representation (R). Label assignment process $L(r)$ is defined as

$$L(r) = \underset{a \in A}{argmax} \sum_{M,O} P(A, M, O, R), \tag{5}$$

where r is the representation ($r \in R$) of the given image sequence and $P(A, M, O, R)$ is the joint probability distribution table. $P(A, M, O, R)$ is defined by using the conditional dependencies in the graphical model (Fig. 3.)

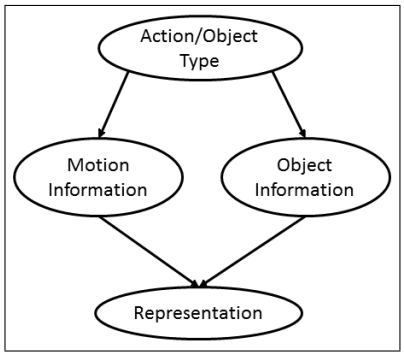

Fig. 3. The graphical model of our framework

as

$$P(A, M, O, R) \propto P(A)P(M|A)P(O|A)P(R|M, O), \tag{6}$$

where $P(A)$ is constant, and $P(O|A)$ term can be calculated easily by Table 1. For example, $P(O = cup|A = drink)$ is 1 because the drink is done only with cup according to Table 1. $P(R|M, O)$ term needs to be converted by using axioms as

$$P(R|M, O) = \frac{P(M, O|R)P(R)}{P(M, O)}. \tag{7}$$

$P(R)$ and $P(M, O)$ values in (7) are neglected because these are the same for a given image sequence. $P(M, O|R)$ is efficiently represented as

$$P(M, O|R) = P(M|R)P(O|R). \tag{8}$$

Finally, the values of $P(M|R)$ and $P(O|R)$ are needed to obtain label $L(r)$. $P(O|R)$ is equal to the $OI(w_r, o)$ obtained at the end of representation of object information process. $P(M|R)$ is related to MI. For this relation, linear kernel Support Vector Machines (SVMs) are trained by using MIs as inputs. These SVMs assign scores to each MI. The Gibbs distribution is used to translate an SVM score into prediction as

$$P(M|R) = \frac{1}{Z} exp(-Q_m(M, R)), \tag{9}$$

where the potential function $Q_m(M, R)$ carries information about the motion pattern m ($m \in M$) and Z is the normalizing constant (taken as 0.5). Note that the SVMs are used in one-to-all manner with binary fashion.

4 Experimental Results

Evaluation of the framework is conducted with various experiments on the modified [10] dataset. 60 different image sequences (6 for each action/object type)

are used. The proposed framework is tested in leave-one-actor-out procedure in each experiment. The obtained results are same for object recognition and action recognition.

Table 2. Action/Object recognition accuracy with different approaches.

Table 2	[12]	[12] + Action(G-RGB)	[12] + Action(G-RGB+Depth)
Recog. Acc.	61.7%	71.7%	83.3%

Table 2 shows the overall results with different approaches. Using only [12] as an object recognition method, we labeled 37 out of 60 image sequences correctly in terms of action and object recognition with a 61.7% recognition accuracy. It is important to mention that the object recognition method suffers from occlusions. Then combination of object availability information with the motion information in the proposed framework improved the recognition accuracy to 71.7%. Note that only G-RGB image sequences are used as a source of motion information in this process. Finally, our framework recognized 50 out of 60 image sequences with an accuracy rate of 83.3% by using depth image sequences along with G-RGB image sequences for motion representation.

Table 3 shows the detailed action/object recognition results with and without using depth image sequences for motion representation. Without depth image sequences, our framework recognizes actions/objects within 3D environment: x-y space and time. The addition of depth image sequences increases this environment to 4D: x-y-z space and time. Recognition accuracy of actions which include hand movements in the z-axis such as using a brush on the table (moving brush towards and away from the camera) and using a remote control (holding and pointing remote control to the camera) improved significantly by addition of depth information. The general misclassified sequence in terms of action and object recognition is 'playing a videogame with gameboy'. In this sequence, gameboy is similar to remote control and cell phone in shape, and the action is similar to using a calculator.

5 Conclusions

In this work, we focused on the role of action understanding in object classification and vice-versa. Then we proposed a framework that recognizes actions and objects within the context of each other. The proposed framework combines information about object availability and motion patterns in image sequences with a Bayesian network. The current framework does not require any temporal segmentation, body-part segmentation, and hand tracking methods. The experimental results demonstrated that the use of action and object context together improved recognition accuracy. Additionaly, we observed that our framework benefited from the depth image sequences efficiently.

Table 3. Detailed results for each action-object type with/without using depth image sequences. The values on the left side of the / indicates the count of image sequences which are recognized with using depth images and the values on the right side of the / indicates the count of image sequences which are recognized without using depth images.

Table 3	Pour	Drink	Brush	R.Cont.	Calc.	Phone	Headp.	V.game	Pict.	Pen
Pour	5/4	1/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/1
Drink	0/0	6/6	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
Brush	0/0	0/0	6/4	0/0	0/1	0/0	0/0	0/1	0/0	0/0
R.Cont.	0/0	0/0	0/0	5/3	0/0	0/1	0/0	1/1	0/1	0/0
Calc.	0/0	0/0	0/0	0/0	5/4	0/0	0/0	0/1	0/0	1/1
Phone	0/0	0/0	0/0	0/0	0/0	5/5	0/0	1/1	0/0	0/0
Headp.	0/0	0/0	0/0	0/0	0/0	0/0	6/6	0/0	0/0	0/0
V.game	0/0	0/0	0/0	1/2	0/0	1/1	0/0	4/3	0/0	0/0
Pict.	0/0	0/0	0/0	1/1	0/0	1/1	0/0	0/0	4/4	0/0
Pen	0/0	0/0	0/0	0/0	2/2	0/0	0/0	0/0	0/0	4/4

References

1. Ullman, S.: High-level vision: object recognition and visual cognition. MIT Press. (1996)
2. Fei-Fei, L., Fergus, R., Torralba, A.: Recognizing and learning object categories: short course. ICCV. (2009)
3. Carbonette, P., Freitas, N., Barnard, K.: A statistical model for general contextual object recognition. ECCV. (2004) 350-362
4. Torralba, A.: Contextual priming for object detection. Int. J. Comput. Vision **53** (2003) 169-191
5. Filipovych, R., Ribeiro, E.: Recognizing primitive interactions by exploring actor-object states. CVPR. (2008)
6. Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., Rehg, J.M.: A scalable approach to activity recognition based on object use. ICCV. (2007) 1–7
7. Kjellström, H., Romero, J., Kragić, D.: Visual object-action recognition: interfering object affordances from human demonstration. Comput. Vis. Image Underst. **115** (2011) 81-90
8. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. **104** (2006) 90–126
9. Poppe, R.: A survey on vision-based human action recognition. Image Vision Comput. **28** (2010) 976–990
10. Gall, J., Fossati, A., Gool, L.J.V.: Functional categorization of objects using real-time markerless motion capture. CVPR. (2011) 1969–1976
11. Ciu, X., Liu, Y., Shan, S., Chen, X., Gao, W.: 3D Haar-like features for pedestrian detection. ICME07. (2007) 1263–1266
12. Fei-Fei, L.: Bag of words models: recognizing and learning object categories. CVPR07. (2007)
13. Rubinstein, Y., Hastie, T.: Discriminative vs. informative learning. Proc.of the 3th Int. Conf. on Knowledge Discovery and Data Mining (1997) 49–53