Bottom-up Attention Improves Action Recognition Using Histograms of Oriented Gradients

Go Tanaka[1] Yukie Nagai[1] Minoru Asada[1,2]

[1]Graduate School of Engineering, Osaka University
[2]JST ERATO Asada Synergistic Intelligence Project
2-1 Yamadaoka, Suita, Osaka, 565-0871 Japan
{go.tanaka, yukie, asada}@ams.eng.osaka-u.ac.jp

Abstract

When recognizing others' action, we pay attention to their body parts and/or objects they are manipulating rather than observing their whole body movement. Bottom-up saliency is a promising cue to determine where to attend and hence to identify what the persons are doing because their body parts acting on objects become more conspicuous when contributing to the action. This paper proposes an architecture for action recognition that integrates bottom-up saliency with Histograms of Oriented Gradients (HOG). The HOG extracts the local features of others' action while the saliency gives an attentional weight to the HOG descriptor. Our experiments demonstrate that the saliency-based attention improves the performance of action recognition by emphasizing the HOG features relevant to the action.

Figure 1. A sample scene in which a person is reading a book. Our system recognizes his action using a camera attached on a wall.

1 Introduction

Building an action recognition system is still an open challenge. Although many architectures have been developed (refer to [1] for the review), none of them, to our knowledge, can be used in an unrestricted environment. For example, motion capturing systems strictly limit the environment because of special cameras and markers attached to a target person. Applying a 3D body-model to a normal camera image achieves as high accuracy in posture estimation as motion capturing systems do [2]; However, a small disturbance could cause a significant error in the fitting of the body model. Extracting skin color is another solution to estimate a person's posture [3]. The position of a person's head and hands provides a rough estimation of his posture. This method as well as the above, however, does not seem robust against environmental changes (e.g., different lighting conditions) or occlusion since they require the whole body (or the whole upper body) of a target person to be observed.

Other research on action recognition points out the importance of objects manipulated by a target person. Some models incorporate the information about the objects separately from the person's movement (e.g., [4]). A Hidden Markov Model (HMM) has also been proposed to represent the temporal change both in a person's posture and in objects [5]. The model is supposed to know what features are important and should be represented by the HMM. However, systems are often exposed to enormous data in an unrestricted environment. An open challenge is to make systems properly select important features in order to reduce the complexity and ambiguity in the information. An attentional mechanism to highlight the action-relevant information would improve action recognition.

To address the issues mentioned above, we propose an action recognition model that integrates bottom-up attention with Histograms of Oriented Gradients (HOG). The HOG [6] represents primitive features of a person's posture. The histograms of image gradients are calculated for local areas, which increases robustness against occlusion or disturbances unlike 3D body-models. Bottom-up attention based on visual saliency [7] adds an attentional weighting to the HOG feature. Since a person's body parts acting on an object often become more salient due to their motion, the saliency-based attention is supposed to emphasize the action-relevant information. Our system is verified in a daily life situation, where a target person reads a book, calls by phone, etc., as shown in Figure 1.

Section 2 describes our proposed model, which integrates HOG with saliency-based attention. Experiments presented in Section 3 demonstrate the advantage of employing bottom-up attention over a model without such attention. The paper is concluded with future directions in Section 4.

2 Action recognition model integrating HOG with saliency-based attention

The proposed model recognizes actions generated with the upper body of a target person (e.g., reading a book and drinking from a cup). Figure 2 shows the architecture of the model. It first calculates a saliency map (see Figure 2 (b)) and a HOG descriptor (c) from an input image (a), and then integrates (c) with a

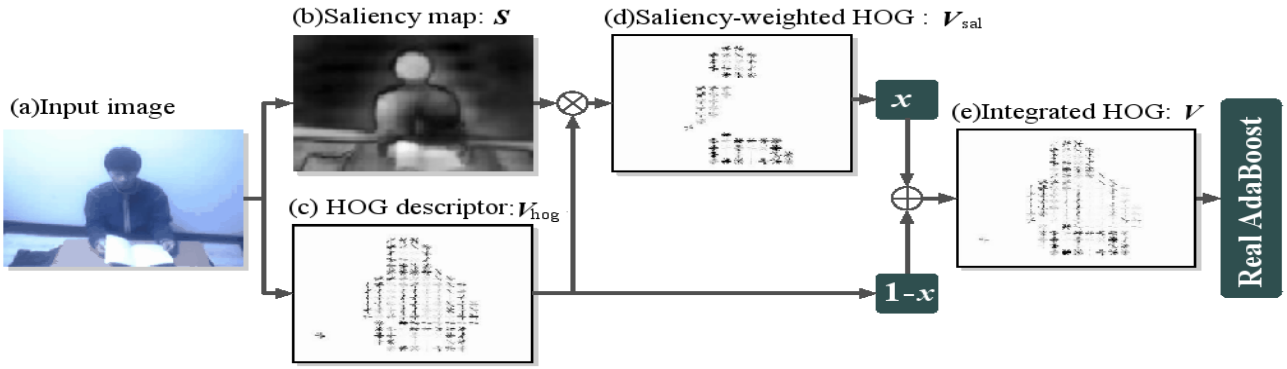


Figure 2. The proposed model for action recognition integrating HOG with saliency-based attention. A saliency map (b) and a HOG descriptor (c) are calculated from an input image (a). (c) is then integrated with a saliency-weighted HOG (d) to generate (e), which highlights action-relevant information. The categories of actions are finally learned using a Real Adaboost algorithm.

saliency-weighted HOG (d) to emphasize the action-relevant features as shown in (e). The details are described in the following sections.

2.1 Extracting a HOG descriptor

A HOG descriptor (see Figure 2 (c)) is calculated from the input image (a). The image is divided into 5 $\times$ 5 pixels of cells, where a histogram of intensity gradients is computed with respect to nine orientations (i.e., 0°, 20°, ..., 160°). The proposed model uses the HOG feature extracted only from the foreground image because the object involved in the action as well as the person performing the action is supposed to be included in the foreground. As shown in Figure 2 (c), the contour of the person as well as the book is well extracted in the HOG image. Refer to [6] for the detailed mechanism.

HOG has often been used to detect people or vehicles from a static image [6]. Although the descriptor does not represent an accurate shape of targets, it can robustly extract them against noises or changes in environmental conditions. *No* need for special equipments or for high computational power is another advantage in real-time action recognition.

2.2 Calculating a saliency map

The proposed model then calculates bottom-up saliency to pay stronger attention to the action-relevant information (see Figure 2 (b)). Our model employs an architecture proposed by Itti et al. [7], which defines the saliency for image areas as the difference from the surroundings. The following four features are used to compute the saliency: color (red/green and blue/yellow), intensity (black/white), orientation (0, 45, 90, 135 [deg]), and temporal change in the intensity (on/off).

In recognizing actions, the use of motion information is crucial as well as that of objects' information. It is assumed that the stronger the movement of a body part, the more it contributes to the action.

2.3 Integrating HOG with saliency

The model next integrates the HOG with the salience information. Let $\boldsymbol{V}_{hog}$ and $\boldsymbol{S}$ the HOG descriptor and the saliency map derived from the input image, respectively. The integration procedure is as followings:

1. Calculate a saliency-weighted HOG $\boldsymbol{V}_{sal}$ by multiplying each histogram of $\boldsymbol{V}_{hog}$ by $\boldsymbol{S}$:

$$\boldsymbol{V}_{sal}(i,\ j) = \boldsymbol{S}(i,\ j) \cdot \boldsymbol{V}_{hog}(i,\ j), \tag{1}$$

   where $(i,\ j)$ is the position of a histogram in the image.

2. Integrate $\boldsymbol{V}_{sal}$ and $\boldsymbol{V}_{hog}$ with a weight $x : (1 - x)$

$$\boldsymbol{V} = x\boldsymbol{V}_{sal} + (1 - x)\boldsymbol{V}_{hog}, \tag{2}$$

   where $0.0 \leq x \leq 1.0$, so that $\boldsymbol{V}$ emphasizes action-relevant locations. The higher x is, the more highlighted the salient locations are.

Figures 2 (d) and (e) show the HOG descriptors weighted by saliency and then integrated with the original, respectively. Comparing (e) with (c) indicates the higher attention to the action-relevant locations (i.e., the person's head, right arm, and the book).

2.4 Learning to recognize actions

The proposed model learns to recognize actions using a Real Adaboost algorithm [8]. The algorithm creates many weak classifiers, which achieve higher accuracy for the recognition than do algorithms using a single classifier.

3 Experiments of action recognition

3.1 Five actions to recognize

We targeted five actions: reading a book, drinking from a cup, calling by phone, writing on a book, and doing nothing. The same person demonstrated the five actions in front of a camera. To vary the samples, he sometimes changed the arm to perform the actions as

(a) Drinking from a cup

(b) Calling by phone

(c) Writing on a book

Figure 3. Sample images of three actions: (a) drinking, (b) calling, and (c) writing. From left to right, the input image, the saliency map, and the HOG descriptor integrated with the saliency-weighted HOG.

well as his clothes. 250 images for each action were randomly selected from the video to train with the proposed model.

Figure 3 shows sample images of the three actions: (a) drinking, (b) calling, and (c) writing. From left to right, the input image, the saliency map, and the HOG descriptor integrated with the saliency-weighted HOG are presented. We can see that the HOG descriptors extracted from the action-relevant locations are more strongly emphasized by the salience information. For example, the cup in Figure 3 (a) shows higher saliency because of its outstanding color and motion, which is important in recognizing the "drinking" action. The book in Figure 3 (c) also shows higher saliency due to the moving hands on the book as well as the conspicuous color and intensity of it, which contributes to recognizing the "writing" action. In the case of "calling," the telephone in Figure 3 (b) is not as salient as the other objects. However, the right arm holding the phone is salient enough to be distinguished from the other actions. These examples qualitatively demonstrate the validity of the saliency-based attention to emphasize the action-relevant locations.

3.2 Effect of integration ratio of saliency

The first experiment assessed the effect of the integration ratio x used in Eq. (2). We compared the performance of action recognition when x changed from 0.0, 0.1, ..., to 1.0. When $x = 0.0$, the HOG descriptor

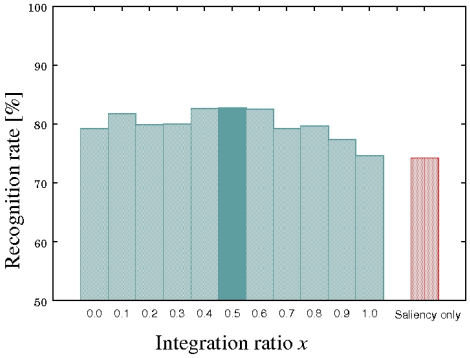


Figure 4. Performance of action recognition with different integration ratio x (green bars) and only using saliency map (red bar)

without attentional-weighting was used. When $x = 1.0$, in contrast, the HOG descriptor *only* in the salient area was used.

Figure 4 shows the result of the recognition rate. The green bars on the left side show the results with different integration ratio, whereas the red bar on the right most shows the result of using *only* the saliency map (i.e., *no* HOG descriptor). The models were tested with 2500 images (500 images for each action) which were not used in the training. This result suggests that weighting the HOG features with bottom-up saliency improves the performance of action recognition. The saliency model properly highlighted the person's body parts and the object involved in the action, despite *no* context knowledge given to the model. The ratio $x = 0.5$ achieved the best performance in this experiment. Regarding the performance for each action, however, there was variation in the best ratio: $x = 0.1$ for "reading," $x = 0.4$ for "drinking," $x = 0.6$ for "calling," and $x = 0.6$ for "writing." A reason for the low ratio for "reading" could be the strong contour of the book. The model might not need to farther emphasize it with the salience. We will discuss in the future how the system determines an appropriate ratio of the salience information.

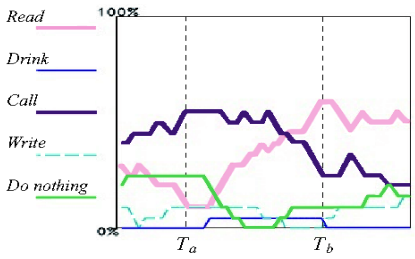
3.3 Recognition of multiple simultaneous actions

In the first experiment, the subject was supposed to perform only one action with the relevant object. However, people sometimes execute multiple actions or maintain multiple objects while doing a single action. The second experiment verified whether the saliency-based attention can appropriately shift the attentional target to identify the actions.

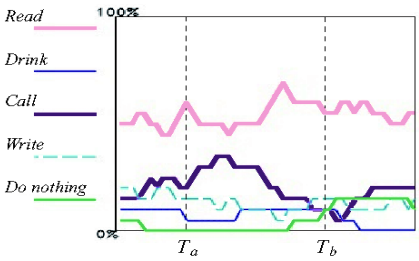
Figure 5 shows sample images and the recognition rate when the subject gradually shifted his action from "calling" to "reading." He held the two objects (i.e., a phone and a book) through the demonstration. Figures 5 (a) and (b) are the images captured at time T_a and T_b, respectively. Figure 5 (c) shows the transition of the recognition rate when the model applied the saliency-based attention with $x = 0.5$ (the best performance in Figure 4), whereas (d) shows the result for the model without attention (i.e., $x = 0.0$). For each condition, the recognition rate was calculated from the last 20 time steps to prevent an unnecessary change

(a) Calling by phone at T_a

(b) Reading a book at T_b



(c) Action recognition *with* saliency-based attention



(d) Action recognition *without* any attention

Figure 5. Recognition for multiple actions. (a) and (b) are the images captured when the subject was calling by phone at T_a and reading a book at T_b, respectively. The transition of the recognition rate *with* and *without* saliency are shown in (c) and (d).

caused by a momentary movement of the subject.

The results demonstrate that the model with the saliency properly estimated the subject's action. It recognized "calling" at T_a (the purple line in Figure 5 (c)) and "reading" at T_b (the pink line). The proposed model could emphasize the action-relevant location with the bottom-up saliency because the subject produced more prominent movement to the locations. When calling by phone, for example, the subject's left arm was more visible because of the posture and the movement while the book was in no motion. When he started reading the book, his right hand as well as the book became more salient due to the movement of turning the pages. In contrast, the model without attention could not detect the change in the action. (see Figure 5(d)) It always recognized the subject's action as "reading" (the pink line) over the experiment. The clear contour of the book might cause the false recognition of the action. This experimental result verifies the importance of integrating the attention.

4 Conclusion

This paper proposed an architecture for action recognition that integrates a HOG descriptor with bottom-up saliency. As people pay attention to action-relevant locations, the model weighted the HOG descriptor using a saliency map. Our underlying hypothesis was that when a person performs an action, his/her body parts and objects involved in the action become so conspicuous as to be detected by a saliency model. The first experiment verified that the saliency-based attention properly highlighted the action-relevant information. The following experiment for multiple actions also showed that the proposed model outperformed the model without an attentional mechanism. It properly recognized two actions as well as a single action using bottom-up attention.

A future issue is to have the model autonomously determine the integration ratio x in Eq. (2). Appropriate ratios would take more advantages of saliency-based attention. The attention enables us to take surrounding objects into account in action recognition. It is expected to accommodate new environments using unsupervised learning. We will investigate to what extent the attentional mechanism can deal with changes in the position and/or the orientation of subjects.

Acknowledgments

This work is partially supported by cooperative research with Daiwa House Industry Co., Ltd. We thank Prof. Yoshio Iwai for his kind advices.

References

[1] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28:976–990, 2010.

[2] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3d exemplars. In *Proc. of the 11th IEEE International Conference on Computer Vision*, pages 1–7, 2007.

[3] P. Buehler, A. Zisserman, and Everingham. Learning sign language by watching tv (using weakly aligned subtitles). In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2961–2968, 2009.

[4] D. Lee, H. Kunori, and Y. Nakamura. Association of whole body motion from tool knowledge for humanoid robots. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2867–2874, 2008.

[5] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–385, 2002.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[7] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[8] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37:297–336, 1999.