35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

A Filtering Approach for Human Detection in Bionic Eye

Fang Wang, Yi Li, and Nick Barnes

Abstract— **Current bionic eye hardware implants have limited number of separately perceivable electrodes. This challenges the processing algorithms for the visualization unit in bionic eye. Therefore, it is important to ensure that the critical information must remain perceivable in downsampled images. Since the main goal for any bionic eye device is for interaction in the user's everyday life, the location of human subjects is one of the important sources, and detecting humans in natural scenes becomes fundamental in the process of extracting critical visual information.**

This paper presents an approach to localizing human subjects using a filtering approach. Filtering is one of the most efficient approaches for processing images. In addition, it is very safe in biomedical devices because the behaviour of a filter is well understood. Thus it is easier to pass the regulatory authorities. Our goal is to automatically generate a set of filters, each of which describes a representative pose of human bodies. This is achieved by grouping visual features from images directly. In this process, we use Support Vector Machine followed by an efficient post-processing procedure. Then, these filters naturally define human bodies in a fine granularity for effective image processing. Experiments on a large number of images suggest our approach is practical for bionic eye.

I. Introduction

In recent years, a number of algorithms have been developed for displaying visual information for bionic eye world wide. These algorithms in the visual processing unit aim at providing assistive modules based on computer vision techniques. Ground plane segmentation [1], saliency [2], and just noticeable difference [3] have demonstrated effective assistance for individuals with vision impairment.

The primary goal in all visual processing modules is to visualize critical information in real world using a low resolution display device. Current hardwares for prosthetic vision have both limited dynamic range and levels of separately perceivable brightness [4]. This means information may be lost during downsampling. Therefore, visual processing unit must be properly designed to ensure that critical information must remain perceivable.

Thus, the concept of "importance maps" serves as an important concept in bionic eye. For instance, objects of interest must be given higher priorities to be perceivable, in order to improve the user experience. This essentially means a visual unit must be able to detect these objects such that any augmented reality device can display these regions of interest properly.

Fang Wang is with Nanjing University of Science and Technology and National Information and Communication Technology Australia (NICTA). Yi Li and Nick Barnes is with National Information and Communication Technology Australia (NICTA) and College of Engineering and Computer Science at the Australian National University, Canberra, ACT Australia 2601. This work is done when Fang Wang is visiting NICTA. Email: {fang.wang, yi.li, nick.barnes}@nicta.com.au

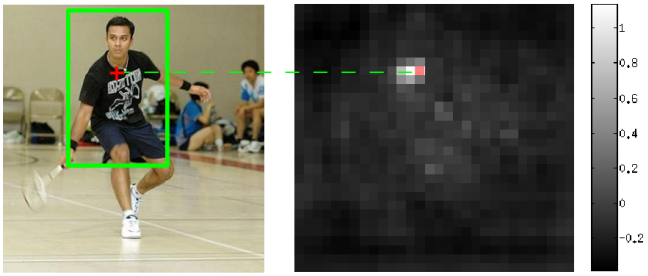

Fig. 1. A humans in an indoor environment and the ideal detection (center of the upper body) in visual processing units for bionic eye.

Since the main goal for any bionic eye device is for interaction in the user's everyday life, the location of the surrounding human subjects is one of the important sources, and detecting humans in natural scenes becomes fundamental in the process of extracting critical visual information.

Fig. 1 presents this idea. The input of a visual processing unit is usually obtained from standard video stream. The unit must be able to detect the human locations (e.g., "upper body" in this example), and visualizing them in a lower resolution device. In this example, we use 30×30 because it is a reasonable configuration in the near future. Once detected, other visualization algorithms may either highlight the region automatically, or the user can have an option for manually zooming in.

In this paper, we present an approach to localizing human subjects using a filtering approach. Filtering refers to a linear operation that performs convolution on a pixelized discrete digital image to retrieve important information. This is one of the most efficient approach for processing images. In addition, any processing based on simple filtering is safe for biomedical devices because the behavior of filters can be analyzed quantitatively and qualitatively. Thus it is easier to pass the regulatory authorities, such as the Food and Drug Administration (FDA).

Our goal is to automatically generate a set of filters for detecting humans of various poses in everyday life. Each filter may be regarded as a description of a representative pose of human bodies. This is achieved by grouping visual features from images directly. In this so called "clustering" process, we use an effective pattern recognition and machine learning method called Support Vector Machine to initialize the grouping, and then we adopt an efficient post-processing to select the best filters.

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

Due to the articulated nature of human body, we choose stable parts in this filtering process. Ideal choices include the combination of head and shoulders, or that of head and torso. These filters naturally define various human bodies in a fine granularity for effective image processing. Experiments on a large dataset suggest our approach is able to practically detect humans in images.

II. Related work

Bionic eye aims at restoring the sense of vision to people living with blindness and low vision. The surrounding world contains a tremendous amount of visual information. Therefore, human tends to focus on only a few parts while ignore others in a scene [5]. As a result, modelling regions of interest becomes an important topic in bionic eye.

Humans are one of the most important "objects" in real world interaction. Describing and detecting humans have been a classical topic in psychology and computational vision. Decades ago, psychologist Marr proposed a hierarchical organization of body parts for understanding humans [6]. Each model in this representation must be a complete unit that is appropriate for recognition. His ideas motivated a number of approaches in computational psychology.

In many detectors for recognition, a number of appearance models were adopted extensively to estimate human body. Powerful features, such as Histogram of Gradient (HOG) [7] is frequently used to improve the performance of detection. Based on these low level features, researchers also seek to higher level visual representations. Bourdev et al [8] proposed the idea of poselets for human recognition, which refers to combined parts that are distinctive in images.

In many localization algorithms, sliding window technique is frequently adopted. If linear classifiers are used, this technique becomes a linear convolutional operation in two dimensional space. Fast filtering techniques, such as those based on multi-threading, can be used to significantly reduce the computational costs.

III. Detecting humans using filtering

We present three important components in our visual processing algorithm in this section: 1) image features, 2) human detectors, and 3) post-processing procedure for automatically selecting optimal filters.

Fig. 2 illustrates our idea. Given a set of instances of a part, our approach clusters these instances to subsets that are meaningful in visual appearance, and summarize it to a set of linear weights.

A. Image features

Histogram of Oriented Gradients (HOG) is a powerful feature descriptor used in detecting objects in images. Given an intensity images, the essential idea is that local object appearance and shape can be described by the distribution of intensity gradients (edge orientation and strength).

This descriptor can be efficiently implemented. First, an image is divided into small regions on a regular grid called cells. Then, a histogram of gradient for the pixels within the cell is computed. The histograms are combined and used to represent the local properties of an image.

Fig. 2. Generating filters for human detection. Given a set of instances of a compositional part, our approach clusters these instances and generates a set of linear filters for these clusters.

HOG features has a number of advantages over the original intensity image. It is invariant to geometric and photometric transformations. Coarse spatial sampling allows the grouping a class of objects, which is particularly suitable for human detection in images.

Fig. 3 shows some examples of HOG features. Please refer to [7] for details.

Fig. 3. Examples of the HOG feature.

B. Filtering humans

Humans have large variations in appearance and poses. Therefore, it is very challenging to localize humans in images by a single detector. Our goal is to automatically generate a set of filters. Each filter is responsible for a representative pose.

Assuming we have manually labelled human skeletons in images, we attempt to group the visual features, and generate filters for different groups of features. This is achieved by automatically fitting parameters of a multi-class classifier.

Preprocessing dataset

The dataset, such as the LSP [11], contains the human skeletons labelled manually. For the joints we are interested

in (e.g., head and torso), we can pre-process the labelled data by simple clustering algorithms. We choose k-means, because it is efficient and reasonably accurate.

Fig. 2a shows some results after k-means. Image patches with similar body poses are grouped together. This greatly reduces the complexity and variation in human poses, and allow us to learn one filter for each category independently.

Computing filters

Denote p as the body part of interest in human subjects. For instance, in Fig. 2 it is head and torsos. Let $\phi(p)$ denotes the visual feature of p in the image. We aim at learning linear classifiers to group visual features automatically.

We followed [9] and built a Support Vector Machine model for learning visual subcategories. Given N positive instances of a compositional part and negative instances, we learn a filter of each positive set. Our objective function for each group is as follows

$$\begin{aligned} \arg\min_{w} \frac{1}{2} \parallel w_k \parallel_2^2 + C \sum_{i=1}^{N} \epsilon_i, \\ y_i(w_k \phi(p_i) - b) \geq 1 - \epsilon_i, \\ \forall i \in \text{class } k, \epsilon_i \geq 0, \end{aligned} \tag{1}$$

where $y_i = \{1, -1\}$ denotes whether y_i is from positive or negative sample sets, and w is the weights of the feature map for each part.

Due to the page limit of the paper, please refer to [9] for solving Eq. 1. Fig. 2b shows some examples of the filters for head and torso. Please note that each block in the filter images corresponds to a cell in HOG templates (best viewed in color).

C. Postprocessing

To achieve effective training when the number of labelled examples is small, we use a post-processing [10] to select the best filters and fine tune the performance.

The main idea of post-processing is to split the training set to two subsets, S_{tr} and S_{ev}. First, a training step is performed on S_{tr}. Then, each filter is evaluated on the S_{ev}, and weak filters with few detected samples will be removed from the filter set.

The whole training process is conducted by iteratively switching S_{tr} and S_{ev} in each iteration. After training the remaining classifiers are more effective and meaningful.

D. Filtering for detection

Eq. 1 outputs sets of weight vectors. The linear combination between a weight vector and the HOG features results in the detection results. When this applies to the whole image, this is a convolution process.

Fig. 4 gives an example of the filters and their response maps. The red dashed bounding box denotes the ground truth. In this example, we have eight filters, and the filter response maps are showed on the right hand side of the input figure. The filter that has the highest response is denoted by green color and the corresponding bounding box is labeled in the input image. Then, one can perform further processing on the detected region.

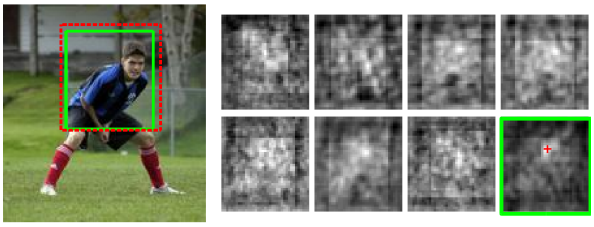

Fig. 4. Filtering as a detection process.

IV. Experiments

We first show results of our algorithm. Then, we demonstrate the performance using part detection accuracy. Our input is an image from various indoor and outdoor scenes, and the output is a 30×30 grid by uniformly downsampling the response maps.

In annotated datasets (e.g., the LSP [11]), the joints of human bodies are manually labeled. Since our goal is to detect the main part of the body, we use the head and torso to reasonably approximate human body. We then train our filters using the LSP dataset and use the trained filters in this section.

A. Demonstration

We demonstrate the effectiveness of filtering procedure using a video sequence [12] in this section. Fig. 7 shows a daily scene in street. We show the localization results (bounding boxes of different colors) in the input images. Then, we visualize the local maxima in each detection result map in the reduced resolution.

This example shows that our method is able to reliably detect humans in a natural scene. Please note that the subjects in this image has various poses, and our method is capable of capturing this dynamics by filtering.

B. Statistical results

We show the statistical results on the LSP dataset. In this section, we do not show the downsampled images for visualization in bionic eye.

First, we show some images and our detection results in Fig. 5. One can see the main bodies of the humans are consistently detected. We further show the precision-recall curve in Fig. 6. In evaluation, the part is detected if the detected bounding box overlaps with the ground truth bounding box over 50%.

C. Runtime

We used eight filters in our experiments. The filtering process is very efficient. Since the code is written in C++, the run time for each frame is approximately $0.05s$ on a computer of 2.2 GHz CPU and 4G memory.

Ultimately, this needs to be implemented in a portable device. Therefore, we should explore options such as FPGA, where time consumption of convolution operations can be further reduced significantly.

Fig. 5. Examples and their detection results in the LSP dataset dataset.

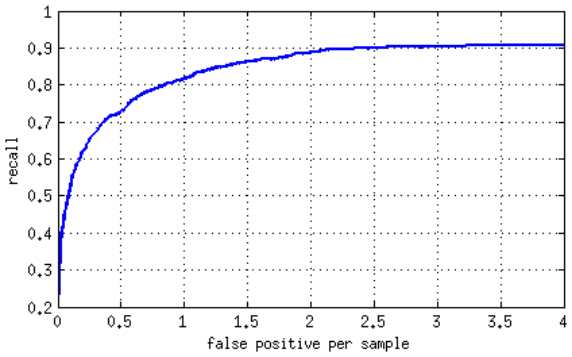

Fig. 6. Statistical results.

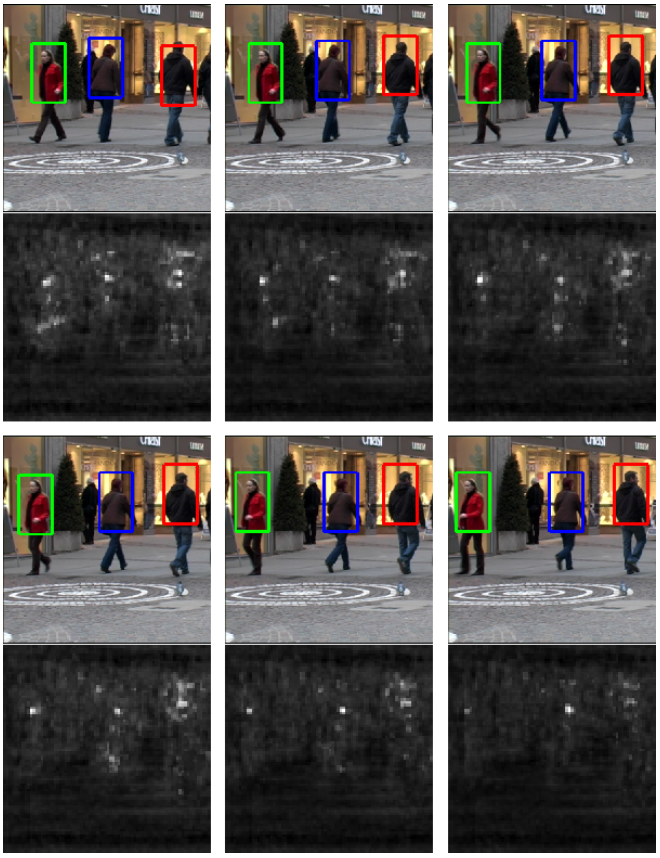

Fig. 7. Examples of the filtering process.

V. Conclusion

This paper presents a novel approach to localizing humans in images and video streams for bionic eye. The main contribution is the visual filters that result in fast and accurate detection. We demonstrate that our method's efficacy on a large dataset, and suggest the method is applicable to visual processing units for bionic eye.

VI. Acknowledgement

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications, and the Digital Economy, and the Australian Research Council (ARC) through the ICT Centre of Excellence Program. This research was also supported in part by ARC through its Special Research Initiative (SRI) in Bionic Vision Science and Technology grant to Bionic Vision Australia (BVA).

References

[1] C. McCarthy, N. Barnes, and P. Lieby, "Ground surface segmentation for navigation with a low resolution visual prosthesis," in *Proceedings of IEEE EMBC 2011*. IEEE, 2011, pp. 4457–4460.

[2] A. Stacey, Y. Li, and N. Barnes, "A salient information processing system for bionic eye," in *Proceedings of IEEE EMBC 2011*. IEEE, 2011.

[3] Y Li, C. McCarthy, and N. Barnes, "On just noticeable difference for bionic eye," in *EMBC 2012*.

[4] P. Lieby, N. Barnes, C. McCarthy, N. Liu, H. Dennet, J. Walker, V. Botea, and A Scott, "Substituting depth for intensity and real-time phosphene rendering: Visual navigation under low vision conditions," in *EMBC*, 2011.

[5] J. Tsotsos, "Analyzing vision at the complexity level," *Behavioral and Brain Sciences*, vol. 13, pp. 423–445, 1990.

[6] David Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, Henry Holt and Co., Inc., New York, NY, USA, 1982.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 886–893.

[8] L. Bourdev, S. Maji, T. Brox, and J. Malik, "Detecting people using mutually consistent poselet activations," *Computer Vision–ECCV 2010*, pp. 168–181, 2010.

[9] S. Divvala, A. Efros, and M. Hebert, "How important are deformable parts in the deformable parts model?," *CoRR*, vol. abs/1206.3714, 2012.

[10] Saurabh Singh, Abhinav Gupta, and Alexei A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Proceedings of the 12th European conference on Computer Vision - Volume Part II*.

[11] Sam Johnson and Mark Everingham, "Learning effective human pose estimation from inaccurate annotation," in *CVPR*, 2011, pp. 1465–1472.

[12] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1014–1021.