2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing

A NOVEL VIEWER COUNTER FOR DIGITAL BILLBOARDS

Duan-Yu Chen and Kuan-Yi Lin

Department of Electrical Engineering, Yuan Ze University, Taiwan

ABSTRACT

This paper presents a novel viewer counter for an environment in which a stationary camera can count the number of people watching an electronic billboard without counting the repetitions in real time video streams. The potential buyers actually watching an advertisement or merchandise are captured via frontal face detection techniques. To count the number of viewer precisely, the problem of occlusions between viewers is tackled. Besides, a complementary set of features is extracted from the torso of a viewer due to the fact that the part of the body contains relatively rich discriminative information than other body parts. In addition, for conducting robust viewer recognition, an online classifier trained by AdaBoost is developed. Our experiment results demonstrate the robustness of the proposed system for the viewer counting task.

Index Terms— Viewer Counting

1. INTRODUCTION

Counting the number of viewers over time is very important in many real-world applications. For instance, counting the number of viewers in a shopping mall may provide valuable information for optimizing trading hours, as well as evaluating the attractiveness of some shopping areas or shopping items. In this paper, we focus on counting the number of people watching an electronic billboard or merchandise. With the advent of intelligent cameras and the increasing capabilities of video surveillance [1], automation of people-counting is now technically possible.

In recent years, a great deal of research has been directed at providing more accurate people counting methods. Generally, the developed methods can be categorized into two types: people detection-based and feature-based approaches. In people detection-based approaches, once people have been detected, they can be counted easily. For instance, in the W4 system proposed by Haritaoglu et al. [7], shape information is used to identify individuals; while Viola et al. [8] employ boosted classifiers to detect pedestrians by using appearance and motion clues. The main problem with these approaches is that their applicability is limited. In some cases, such as when people walk next to each other and/or occlude each other, the detection/tracking process may fail. In contrast, feature-based approaches do not include a people detection step, but try to transform the people-counting problem into some feature space using computer vision techniques. Typically, these methods extract features based on edge density [9], the number of moving pixels [9], blob size [2] or multiple clues [10] to estimate the number of people in a scene. Then, a classifier, such as a trained neural network, is applied to perform classification based on the extracted features.

In existing approaches, the number of people counted is an approximation of the actual number of people in the field of view of the camera. For instance, Chan et al. [2] first localize motion areas and then separate the areas into individual blobs. Then, the sizes of blobs are estimated to determine the number of people. However, for an electronic advertisement billboard, the advertising agent may ask: "How many people actually watched the billboard in some specific advertisement?" Current people-counting systems cannot answer this question. The exclusive feature of the proposed viewer-counting system is that it does not repeatedly count an individual if a viewer watches an advertisement for a long period. Clearly, to solve the problem, a face recognition system must be built. This requirement is very different from existing people-counting systems.

In this work, the system performs a face detection step to identify people who are actually watching the advertisement, rather than simply standing in the area. Since some non-facial regions may be detected accidentally, we design a face filtering process to verify that a detected region is in fact a face part. Then, we extract features directly from that region. However, since the information in the face portion is insufficient due to the low resolution of surveillance videos, we extract features from the torso region to compensate for the deficiency. To ensure robust people recognition we have developed an online training classifier based on AbaBoost. By using the features extracted from the part of the torso region, the system can effectively execute the viewer counting task.

The remainder of this paper is organized as follows. In Section 2 we explain how face detection and face filtering are performed. Section 3 describes the feature set used for viewer recognition. In Section 4, we discuss the online training classifier, which is trained by AdaBoost. We then detail our experiment results in Section 5, and summarize our findings in Section 6.

2. FACE DETECTION

In this phase, we first use the support vector machine based (SVM-based) face detector developed by Kienzle et al. [3]

978-0-7695-3762-7/09 $26.00 © 2009 IEEE
DOI 10.1109/IIH-MSP.2009.211

to perform the face detection task. Then, we apply the proposed filtering process to remove false positives detected by the above face detector.

2.1 SVM-based Face Detector

To count the number of people watching an advertisement on a TV billboard, it is necessary to detect frontal part of faces because the people are "really watching" the advertisement. Therefore, face detection is the first important step to be accomplished. To satisfy the real-time requirement, we adopt the SVM-based face detector developed by Kienzle et al., which can provide fast approximations of support vector decision functions. The detail of the face detector can be found in [3].

Most of the faces can be detected correctly based on this approach, but there are some false positives. Removing false positives from the detected frontal face set is necessary since the detected results (no matter whether they are correct or incorrect) will be further analyzed to compute the number of viewers. Since the subsequent face recognition module will consume a great deal of computational power, a pre-processing step to filter out false positive faces is very important. In the next section, we describe an effective approach for removing false positives from the set of detected face candidates.

2.2 Filtering False Positive Faces

Face detection has attracted a great deal of attention in the past decade. Well-developed face detection techniques, such as OpenCV and Kienzle et al. [3], can achieve success rates of 80%-90%. However, their detection rate of false positives is in the 10%-20% range. In our system, we adopt a temporal filter to remove those detected candidates that appear and disappear suddenly. Detected non-faces usually behave in this manner and would not continually appear in a short duration. Therefore, a probability that measures the spatiotemporal characteristic of a candidate is as follows:

$$P(o_i) = e^{-MSSD(o_i, o_j)\Delta t(i,j)}, \quad (1)$$

where $MSSD(o_i, o_j)$ denotes the minimum sum of square difference between candidates o_i and o_j in different time instant in a short duration. $\Delta t(i, j)$ means the time interval in terms of the number of frames between o_i and o_j. The measure not only captures the appearance similarity between detected candidates in different time instant but also evaluates the temporal consistency between them.

2.3 Viewers Blocking Caused by Occlusions

A viewer that stays behind another viewer with a certain distance in both horizontal and vertical direction has to be blocked for counting. For example, as shown in Fig.1, viewers marked by yellow boxes are the ones that are blocked temporally by our system since the feature set used is extracted from the part of torso. To prevent extracting features from the part of other viewers, these viewers are blocked for counting until their frontal space is clear. This assumption is reasonable for real scenarios in which the camera usually mounted above the viewers. The lower the face position in the two dimensional space usually denotes the closer to the camera. Therefore, the higher the position of a face means the corresponding person stays further to the camera.

Fig. 1. Viewers are blocked for counting due to the occlusion by their frontal viewers.

3. FEATURE EXTRACTION

When counting the number of viewers actually watching an advertisement on a billboard, the chosen feature set should have strong discriminative power so that correct assessments can be made. In this work, the features extracted are the edge-based shape context [4] and our proposed kernel weighted region saliency that is a kind of region-based feature. The purpose of using these different kinds of features is to compensate for each other when edge features are significant for one viewer and region-based color features are dominant for the other. We describe these features in Section 3.1 and Section 3.2, respectively.

3.1 Shape Context

The shape context descriptor for a point on a shape is the histogram of relative polar coordinates of all other points on the shape. Basically, this descriptor provides global discrimination. The corresponding points on two similar shapes usually have similar shape contexts. This characteristic enables us to solve the shape correspondence problem as an optimal assignment problem. Point correspondences between two shapes are thus established by minimizing the point matching costs, i.e., the x^2 test statistic for histograms. Global optimal correspondences can be found by minimizing the sum of the individual matching errors. The above-mentioned correspondence matching problem can be solved by a bipartite graph matching algorithm that enforces a one-to-one point matching process. Therefore, the shape distance, D, [4] is estimated as the weighted sum of the image appearance distance D_{ac}, the shape context distance D_{sc}, and the bending energy D_{be} as follows:

$$D = w_1 D_{ac} + w_2 D_{sc} + w_3 D_{be}, \quad (2)$$

where w_i denotes the weighting of its corresponding distance. D_{ac} is the appearance cost, defined as the sum of squared brightness differences in Gaussian windows around corresponding image points:

$$D_{ac}(P,Q)=\frac{1}{n}\sum_{i=1}^{n}\sum_{\Delta\in Z^2}G(\Delta)\left[I_P(p_i+\Delta)-I_Q\left(T\left(q_{\pi(i)}\right)+\Delta\right)\right]^2, \quad (3)$$

where I_P and I_Q are the gray-level images corresponding to P and Q, respectively; Δ denotes some differential vector offset; G is a windowing function, which is usually a Gaussian and $\{P_i\}, i\in[1,n]$ is a point set of P. The distance is computed after the thin plate spline (TPS) transformation T has been applied to warp the images into alignment as much as possible; and $\pi(i)$ is the permutation of points $q(i)$ of Q resulting from minimizing the costs of all pairs of points of P and Q.

D_{sc} is used to measure the shape context distance between shapes P and Q as the symmetric sum of the shape context matching costs over the best matching points, i.e.,

$$D_{sc}(P,Q)=\frac{1}{n}\sum_{p\in P}\arg\min_{q\in Q}C(p,T(q))+\frac{1}{r}\sum_{q\in Q}\arg\min_{p\in P}C(p,T(q)), \quad (4)$$

where

$$C(p_i,q_j)=\frac{1}{2}\sum_{k=1}^{K}\frac{[h_i(k)-h_j(k)]^2}{h_i(k)+h_j(k)}. \quad (5)$$

$h_i(k)$ and $h_j(k)$ denote the K-bin normalized histogram at p_i and q_j, respectively. The distance of the bending energy D_{be} corresponds to the minimal amount of transformation needed to align the shapes P and Q; thus, it is equivalent to minimizing the bending energy I_f

$$I_f=\iint_{\Re^2}\left[\left(\frac{\partial^2 f}{\partial x^2}\right)^2+2\left(\frac{\partial^2 f}{\partial x\partial y}\right)^2+\left(\frac{\partial^2 f}{\partial y^2}\right)^2\right]d_x d_y, \quad (6)$$

where

$$f(x,y)=a_1+a_x x+a_y y+\sum_{i=1}^{n}\eta_i U\left(\left\|(x_i,y_i)-(x,y)\right\|\right),$$

and the kernel function $U(s)=s^2\log s^2$ and $U(0)=0$. η_i is a weighting for the point (x_i,y_i) and $\|\cdot\|$ denotes the 2-norm operation. The detailed derivation of the above method can be found in [5].

3.2 Block-Based Region Saliency

In addition to the edge-based shape context feature, a region-based color feature is extracted by computing both the global region saliency and the local color information. We propose a technique that performs template matching by matching the salient regions between distinct templates. We use the region saliency to generate a compact template signature. First, the original image I is thresholded to a binary image by using a non-parametric and unsupervised method of automatic threshold selection [6]. An optimal threshold is selected by maximizing the discriminant measure of separability of the resultant gray level histogram. We then apply the Euclidean distance transform on the binary image as follows:

$$\tilde{I}=\min_{p(x,y)\in P}\left\{(x-i)^2+(y-j)^2\right\}, \quad (7)$$

where P is the region with intensity value 0 and pixel (i,j) belongs to the region with intensity value 1. The salient regions are the areas with binary intensity value 1; thus, we compute a new image whose pixel value is inversely proportional to that of $\tilde{I}$ by

$$\hat{I}(x,y)=e^{-(\tilde{I}(x,y))}\cdot I(x,y), \quad (8)$$

where $I(x,y)$ is the original pixel intensity of image I. It is used as a weighting function to characterize the color information. The resultant new image, $\hat{I}$, i.e., the kernel weighted region saliency.

To match shapes by using salient regions, a distance measure d_{rs} is defined as

$$d_{rs}=\sum_{i=1}^{b}\left\{\left|\hat{I}_p(x,y)-\hat{I}_q(x,y)\right| \ (x,y)\in B_i\right\}, \quad (9)$$

where $\hat{I}_p$ and $\hat{I}_q$ are the templates obtained after applying Eq.(8). B_i is the block obtained by first normalizing a template to a pre-defined size and then partitioning it into b blocks of equal size. By using the distance measure, the difference in the spatial relationships of region saliency between a pair of templates, i.e., $\hat{I}_p$ and $\hat{I}_q$, can be computed. In addition, by controlling the parameter b, we can adjust the degree of tolerance in translation and scaling.

4. ONLINE LEARNING-BASED VIEWER COUNTING

An intrinsic characteristic of a real-time viewer counting system is that the pose of a people inevitably changes over time. Most existing methods need to address the pose change problem before performing template matching, i.e., they must pre-define a set of thresholds to ensure good functionality of the feature set. However, it is extremely difficult to find an appropriate feature set that can fit all changes. To resolve this problem, we propose a template matching algorithm that has online appearance learning ability. In our proposed method, a reliable set of templates for each viewer must be collected for online training. A template can be regarded as reliable enough if its spatiotemporal features are similar to one previous template. The similarity between two templates is measured by

$$\eta(O_{i,t},O_j)=e^{-(\alpha_1\cdot D_s+\alpha_2\cdot D_t)}, \quad (10)$$

where α_i is the weight for its i^{th} corresponding distance. D_s is the spatial distance defined by $D_s=\beta_1\cdot D+\beta_2\cdot d_{rs}$, which is a linear combination of the distance of the shape context D and the kernel weighted region saliency d_{rs} with weights β_1 and β_2, respectively. The weighting sets α and β are determined empirically based on extensive experiments; and D_t is the time interval between $O_{i,t}$ and O_j. If $\eta(\cdot,\cdot)$ is larger than a pre-defined threshold, a new template, i.e., a new viewer, is found.

Once the number of collected templates is larger than a threshold, the online training is executed.

The algorithm [11] of online feature selection of our feature set is adopted. Using this approach, the edge and color region features can be adaptively selected according to the appearance characteristics of viewers. In this work, one-against-all classifiers are trained for each viewer.

5. EXPERIMENTAL RESULTS

We used an empirical method to determine the number of separable filters that should be used in face detection. In our experiment, we used between 1 and 5 filters. Fig. 2 shows the performance of the face detector in terms of precision and recall when different numbers of separable filters were used. Considering the tradeoff between the precision and recall rates, it is clear that the best performance was achieved when both precision and recall were 87% and the number of separable filters was 3.

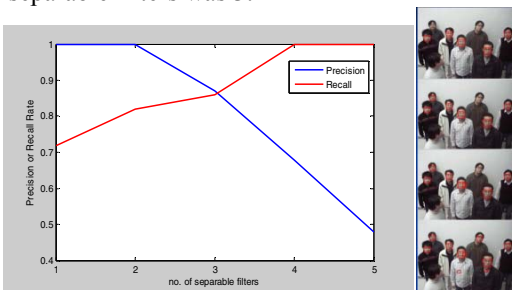

Fig. 2. Determining the number of separable filters in the face detector by evaluating the precision and recall rates

To evaluate the overall performance of the proposed people counting system, we used a long test video that contained many events. In the test video, a large number of human subjects moved frequently in the field of view, and mutual occlusions between the subjects occurred frequently. We implemented the proposed system using Matlab 7.0 with a 1.83GHz Intel CPU. The frame rate of the video was approximately 3-5 fps.

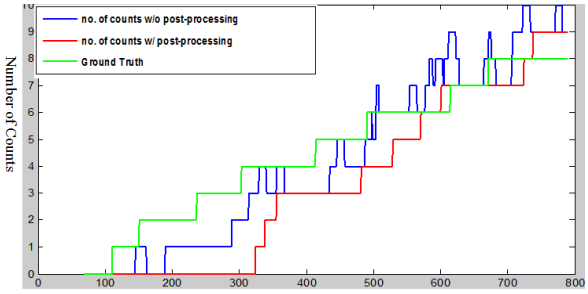

Fig. 3. A snapshot of the performance of our proposed viewer counting system.

A snapshot of people counting process is shown in Fig. 3. In Fig.3, the horizontal axis is the frame number of the test dataset and the vertical axis represents the accumulated number of people counts. Three different curves indicating the counts without post-processing, the counts with post-processing, and the ground truth are shown, respectively. Without using post-processing to filter out noises, it is clear that the accumulated number of people counts would vary abruptly by a sudden increase and decrease of counts. For instance, in the interval between frame 300 and 400, two peaks appeared almost consecutively. While applying the post-processing, the accumulated number of people counts increased smoothly without abrupt variations since the candidates who only appeared for a short period of time were considered as noises and were filtered out. Therefore, in our system, the people counting task can be executed stably and accurately.

6. CONSLUSION

In this paper, a novel viewer counter has been proposed for an environment in which a stationary camera can count the number of people watching an electronic billboard without counting the repetitions in real time video streams. To count the number of viewer precisely, the problem of occlusions between viewers has been tackled. Besides, a complementary set of features is extracted from the torso of a viewer. In addition, for conducting robust viewer recognition, online classifiers trained by AdaBoost are developed. Our experiment results have demonstrated the robustness of the proposed system for the viewer counting task.

7. REFERENCES

[1] R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade, "Algorithms for Cooperative Multisensor Surveillance," *Proc. of IEEE*, Vol. 89, No.10, pp. 1456-1477, Oct. 2001.

[2] A. B. Chan, Z. S. Liang, and N. Vasconcelos, "Privacy Preserving Crowd Monitoring: Counting People without People Models or Tracking," Proc. *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2008.

[3] W. Kienzle, G. Bakir, M. Franz and B. Scholkopf, "Face Detection - Efficient and Rank Deficient," *Advances in Neural Information Processing Systems*, Vol. 17, pp. 673-680, 2005.

[4] S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 24, pp. 509-522, April 2002.

[5] F. L. Bookstein, "Principal Warps: Thin-Plate Splines and Decomposition of Deformations," *IEEE Transactions on Pattern Analysis and Machine Learning*, Vol. 11, No. 6, pp. 567-585, June 1989.

[6] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 9, No. 1, pp. 62-66, Jan. 1979.

[7] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: Real-Time Surveillance of People and Their Activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No.8, August 2000.

[8] P.Viola, M.J. Jones, and D. Snow, "Detecting Pedestrians Using Patterns of Motion and Appearance," *International Conference on Computer Vision*, 2003.

[9] S.-Y. Cho, T. W. S. Chow, and C.-T. Leung, "A Neural-Based Crowd Estimation by Hybrid Global Learning Algorithm," *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, Vol. 29, No. 4, August 1999.

[10] C. S. Regazzoni, A. Tesei, "Distributed data fusion for real-time crowding estimation," *Signal Processing*, Vol. 53, pp 47-63, 1996.

[11] Y. J. Yeh, and C. T. Hsu, "Online Selection of Tracking Features Using AdaBoost," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 19, No. 3, pp. 442-446, March 2009.