doi:10.3772/j.issn.1006-6748.2016.02.014

Intelligent outdoor video advertisement recommendation system based on analysis of audiences' characteristics[①]

Liu Peng (刘　鹏)[* **], Li Songbin[②][* **], Deng Haojiang[**], Wang Jinlin[**]
([*] Haikou Laboratory, Institute of Acoustics, Chinese Academy of Sciences, Haikou 570105, P. R. China)
([**] National Network New Media Engineering Research Center, Institute of Acoustics,
Chinese Academy of Sciences, Beijing 100190, P. R. China)

Abstract

An integrated implementation framework of an intelligent recommendation system for outdoor video advertising is proposed, which is based on the analysis of audiences' characteristics. Firstly, the images of the scene and the people who view the video advertisements are captured by the network camera deployed on the video advertising terminal side. Then audiences' characteristics can be obtained by applying computer vision technologies: face detection, face tracking, gender recognition and age estimation. Finally, an intelligent recommendation algorithm is designed to decide the most fitting video ads for each terminal according to multi-dimensional statistical information of its audiences' characteristics. The experimental results show that the proposed system can effectively improve the audience arrival rate of the video advertisements by an average growth of 27.04%. Moreover, a novel face detection method and a new face tracking method have been proposed to meet the practical requirements of the system, of which the average F1-score is 0.988 and 0.951 respectively.

Key words: face detection, face tracking, intelligent recommendation system, outdoor video advertisement

0 Introduction

Since the 1990s, the outdoor advertising industry has been developing rapidly, resulting in the emergence of various kinds of outdoor video advertising terminals. As more widely the applications of the outdoor advertising terminals grow, one problem has stood out: the information blocking between merchants and advertising audiences. On one hand, once video ads are released on terminals, it is hard for merchants to evaluate the advertising effectiveness due to the lack of information of whether there are audiences or not, how many audiences they have and what kinds of audiences they are; on the other hand, and the location of advertising terminals could lead to differences between gender and age among audiences. Audiences of different terminals may have various concerns and interests, so they seldom obtain interested information after watching ads.

To overcome this problem, this work presents an intelligent recommendation system for outdoor video advertising based on analysis of audiences' characteristics. To date, the scientific literature contains no paper that presents a similar integrated implementation framework of an intelligent recommendation system for outdoor video advertising. But there exist many relevant researches on some of the key technologies used in the system.

At present, face detection is dominated by scanning window classifiers[1-3], most ubiquitous of which is the Viola Jones detector[4]. Viola and Jones[4] propose a simple and effective face detector by using haar-like features to train weak classifiers with AdaBoost algorithm. It can achieve high detection rates that meet real-time requirements. However, when detecting faces against complex background, the detection accuracy declines obviously. Recent studies on common object detection have shown that incorporating parts during detection helps to capture the object class spatial layout better[5-7], which will significantly improve the detec-

[①] Supported by the National High Technology Research and Development Program of China (No. 2011AA01A102), the Important Science & Technology Project of Hainan Province (No. JDJS2013006, ZDXM2015103) and the Young Talent Frontier Project of Institute of Acoustics, Chinese Academy of Sciences.

[②] To whom correspondence should be addressed. E-mail: lisb@dsp.ac.cn

Received on June 23, 2015

tion accuracy. Motivated by it, this paper presents a landmark localization based face detection method.

Object tracking is a challenging problem due to appearance changes caused by illumination, occlusion, pose and motion. An effective appearance model is of crucial importance for the success of a tracking method that has received widespread attention in recent years[8-10]. Zhang, et al.[10] present a real-time compressive tracking method using a very sparse measurement matrix to extract the features for the appearance model. It performs well in terms of efficiency, accuracy and robustness. However, since all the features are extracted by the same sparse measurement matrix, it can't adapt to the object scale changes. To overcome this problem, this study adds a scale standardization step of features on the basis of Zhang's work[10] and presents a face tracking method directing at the audiences' behaviors on viewing the ads.

Compared with the major gender recognition methods, Makinen, et al.[11] point that the performance difference of current methods is not obvious. This paper adopts Lian and Lu's method[12]. The research on age estimation develops slowly due to its complexity and difficulty. Recent researches have achieved some progress[13,14]. This study adopts Luu, et al.'s method[13].

Intelligent recommendation algorithms are widely used in online video recommendation[15,16] and other systems. Almost all of them are human-target algorithms which are inconsistent with the actual needs of the proposed outdoor video advertisement recommendation system. For this, a terminal-target intelligent recommendation algorithm is designed to decide the most fitting video ads for each terminal according to multi-dimensional statistical information of its audiences' characteristics.

1 Overall system architecture

This work aims to build an integrated implementation framework of an intelligent recommendation system for outdoor video advertising based on analysis of audiences' characteristics. The overall architecture of the proposed system is shown in Fig. 1. Firstly, the images of the scene and the people who view the video ads are captured by the network camera deployed on the video advertising terminal side. By conducting face detection and face tracking on the massive video data which contain audiences, the system can obtain some basic information such as the face images of audiences, the length of audiences' watching time and the overall number of audience. Then it will perform gender recognition and age estimation on these face-images to get audiences' genders and ages. Combining all the information above, the system will obtain multi-dimensional statistical characteristics of audiences, through which the merchants will be able to assess the effectiveness of their ads.

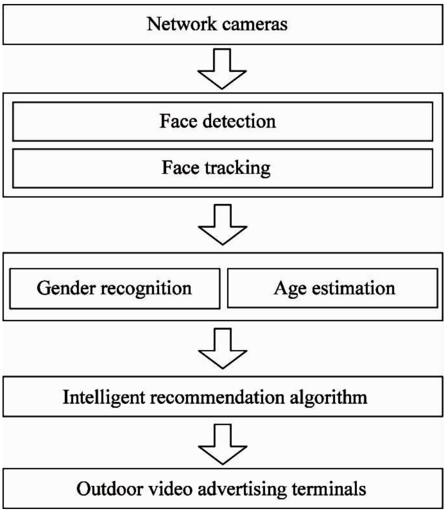


Fig. 1 Overall architecture of the proposed system

Evaluation of the effectiveness of ads is not the final goal that this system aims to achieve. It will be eventually able to push targeted video ads on different terminals to attract more audiences. An intelligent recommendation algorithm is designed to decide the most fitting video ads for each terminal according to multi-dimensional statistical information of its audiences´ characteristics.

2 Face detection

In order to obtain the information of audiences which are viewing the ads, face detection on the original captured videos should be first conducted. The face detection step is the basis of the following analysis procedures such as face tracking, gender recognition and age estimation, which needs to meet two requirements: (1) Since face detection step is aimed to the analysis of audiences' behavior of viewing the ads, a frontal face detection method is needed; (2) Unlike the face detection under limited conditions, this method should be able to cope with the complex environment of the live scene of the outdoor advertising, and has high precision ratio.

Before face detection, the facial landmark detector should be trained. The deformable part model (DPM) is used for the modeling of facial landmark localization problem. Let $I \in R^{M \times N}$ denote a candidate image with

$M \times N$ pixels where R denotes a set of pixel values. Face model can be denoted by a graph $G = (V, E)$. $V = \{v_1, \cdots, v_n\}$ denotes the n parts of a face and the edge $(v_i, v_j) \in E$ denotes the connection between two parts. A landmark configuration can be expressed as $l = \{l_1, l_2, \cdots, l_n\} \in L$, where l_i represents the location of part v_i and L is the set of all the configurations. For image I, $p_i(I, l_i)$ is used to measure the model matching degree of part v_i in l_i position, which can be called local appearance model; $q_{i,j}(l_i, l_j)$ is used to measure the model deformation degree of part v_i and v_j in position l_i and l_j, which can be called deformation cost model. The quality of the landmark configuration can be measured by a scoring function:

$$f(I, \boldsymbol{l}) = \sum_{i=1}^{n} p_i(I, l_i) + \sum_{(v_i, v_j) \in E} q_{i,j}(l_i, l_j) \tag{1}$$

When $f(I, l)$ takes the maximum value $S_{\max}$, the landmark configuration has a best quality. So the best landmark configuration can be obtained by

$$\boldsymbol{l}^* = \underset{L}{\operatorname{argmax}}\left(\sum_{i=1}^{n} p_i(I, l_i) + \sum_{(v_i, v_j) \in E} q_{i,j}(l_i, l_j)\right) \tag{2}$$

Seven landmarks are trained with the labeled faces in the wild (LFW) database: the tip of the nose, the corners of the mouth, and the canthi of the left and the right eye. The landmark detector can be achieved by the structured output support vector machine (SVM) method[17].

The proposed face detection method is shown in Fig. 2. Firstly, it uses the AdaBoost based frontal face detector to yield coarse face detection results. Then it uses the facial landmark detector to get the localization result $\boldsymbol{l}^*$ and the quality score $S_{\max}$. If $S_{\max}$ is larger than the preset threshold S_δ, the candidate image is considered to be a face. However, the quality score may be affected by expression and other factors. When $S_{\max} \leqslant S_\delta$, in order to prevent mistakes, nose area is used for validation which appears least affected by interference factors. The coordinates localized by the facial landmark detector can be denoted as: $(x^{\text{eye}}_{\text{L-in}}, y^{\text{eye}}_{\text{L-in}})$, $(x^{\text{eye}}_{\text{L-out}}, y^{\text{eye}}_{\text{L-out}})$, $(x^{\text{eye}}_{\text{R-in}}, y^{\text{eye}}_{\text{R-in}})$, $(x^{\text{eye}}_{\text{R-out}}, y^{\text{eye}}_{\text{R-out}})$, $(x^{\text{mouth}}_{\text{L}}, y^{\text{mouth}}_{\text{L}})$, $(x^{\text{mouth}}_{\text{R}}, y^{\text{mouth}}_{\text{R}})$ and $(x^{\text{nose}}, y^{\text{nose}})$. The upper-left coordinate $(x_{\text{rect}}, y_{\text{rect}})$, the width $width_{\text{rect}}$ and the height $height_{\text{rect}}$ of the nose area can be obtained by Eq. (3).

When the image of the nose area is obtained, AdaBoost based nose detector is used to detect the nose. If a nose is detected, the candidate image is considered to be a face.

$$\begin{cases} x_{\text{rect}} = x^{\text{nose}} - width_{\text{rect}}/2 \\ y_{\text{rect}} = \min(y^{\text{eye}}_{\text{L-out}}, y^{\text{eye}}_{\text{R-out}}) - (\max(y^{\text{mouth}}_{L}, y^{\text{mouth}}_{R}) \\ \qquad - y_{\text{nose}})/2 \\ width_{\text{rect}} = (x^{\text{eye}}_{\text{R-in}} + x^{\text{eye}}_{\text{R-out}})/2 - (x^{\text{eye}}_{\text{L-in}} + x^{\text{eye}}_{\text{L-out}})/2 \\ height_{\text{rect}} = (\max(y^{\text{mouth}}_{L}, y^{\text{mouth}}_{R}) - y^{\text{nose}})/2 - y_{\text{rect}} \end{cases} \tag{3}$$

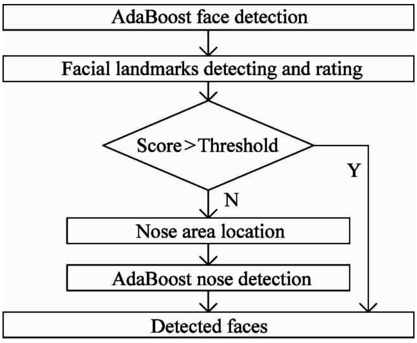


Fig. 2 The flowchart of the proposed face detection method

3 Face tracking

To capture the audiences' behaviors of viewing each ad on each advertising terminal, the audiences' viewing processes need to be tracked. So this paper proposes a face tracking method based on compressive sensing and image recognition. This method is based on the results of face detection, and is specific to the need of information collecting on the terminal side.

The implementation of the proposed face tracking method is shown in Fig. 3. After face detection, the trained online Naive Bayesian Classifier in frame t is used to classify the faces in frame $t+1$. And the multiple individual tracking is realized by selecting the < Classifier, Image > pair that has the highest score in sequence. Suppose n faces are detected in frame t, and m faces in frame $t+1$. When $n > m$, it is considered there exists face exiting. Since the result of face detection has major effect on the result of face tracking, for example, when a target face cannot be detected due to illumination changes or occlusion, it will lead to losing track of target. This work brings in the thought of face recognition to assist face tracking. It will guarantee no loses of track taking place due to the illumination changes or occlusion during one viewing process of audiences. At this point, the target's nearest γ pictures are chosen as the positive samples, while the presupposed γ pictures as the negative samples. Both types of samples' histogram equalized pixels are extracted as features to train a SVM classifier, and the classifier is only effective in σ seconds. If the already tracked pictures of the target are less than γ but more than threshold $\gamma_{\min}$, they would all be used for training; if less than $\gamma_{\min}$, none would be used. When $n < m$, it is considered there exists face entering. Then the trained SVM classifier can be used to judge whether the entering face already exists or not. If not, sampling around the face region to obtain positive and negative samples, and then extract features out of the samples based on compressive sensing to train Naive Bayesian Classifiers. If exists, the face is used to update the current online Naive Bayesian Classifier. When $n = m$, it is considered no face entered or exited, and then these m faces in frame $t + 1$ will be used to update the current online Naive Bayesian Classifier.

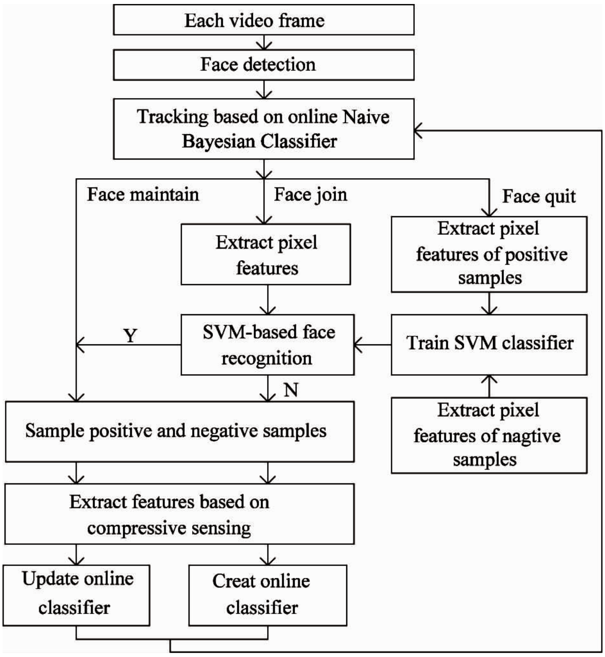

Each video frame
Face detection
Tracking based on online Naive Bayesian Classifier
Face maintain
Face join
Face quit
Extract pixel features
Extract pixel features of positive samples
Y
SVM-based face recognition
Train SVM classifier
N
Sample positive and negative samples
Extract pixel features of nagtive samples
Extract features based on compressive sensing
Update online classifier
Creat online classifier

Fig. 3 The block diagram of the face tracking method

Recent research has shown that using the histogram equalized pixels as input to SVM classifier can achieve better performance when the image size is 36 × 36[11]. It's because the pixels themselves contain the entire information while any feature extraction on them will lead to information loss in some degree. Inspired by this, this study uses the histogram equalized pixels from the resized 36 × 36 image as input when training a SVM classifier.

Zhang, et al.[10] present a compressive sensing based tracking method. The positive and negative samples are obtained around the target and projected with the same sparse measurement matrix to train an online simple naive Bayesian classifier. However, since all the features are extracted by the same sparse measurement matrix, it can't adapt to the object's scale changes. To overcome this problem, this paper adds a scale standardization step of features on the basis of Zhang's work. The feature extraction process is as follows:

An image with $M \times N$ pixels is represented by convolving with a set of rectangle filters $\{h_{1,1}, \cdots, h_{M \times N}\}$ defined as

$$h_{i,j}(x, y) = \begin{cases} 1, & 1 \leqslant x \leqslant i,\ 1 \leqslant y \leqslant j \\ 0, & \text{otherwise} \end{cases} \qquad (4)$$

where i and j are the width and height of a rectangle filter, respectively. Then, the image can be represented as a $(M \times N)^2$-dimension column vector $\boldsymbol{x}$. Since the vector is in high dimensions, random projection is used to reduce dimensions.

Firstly, generate an $n \times (MN)^2$ sparse measurement matrix $\boldsymbol{R}$[10]. Then an n-dimension column vector $\boldsymbol{f}$ can be obtained by $\boldsymbol{f} = \boldsymbol{Rx}$. In this paper, n is set to 70. In order to ensure that this method can be applied to different image scales, a scale standardization step of $\boldsymbol{f}$ should be taken. Suppose that the sparse matrix $\boldsymbol{R}$ is generated based on an image with $M \times N$ pixels, the width and height of h_{kj} are $width_{kj}$ and $height_{kj}$, respectively. Suppose that the current image is of $M \times N$ pixels. Then scale standardization step of f_k can be conducted as

$$f_k = \sum_j r_{kj} \frac{p(\delta_x \cdot x_{kj}, \delta_y \cdot y_{kj}) \otimes h'_{kj}}{width_{kj} \cdot height_{kj}} \qquad (5)$$

where $\delta_x = M'/M$ and $\delta_y = N'/N$ denote the scale coefficients of x-axis and y-axis. The width and height of h'_{kj} are $\delta_x \cdot width_{kj}$ and $\delta_y \cdot height_{kj}$, respectively. When solving, the actual location of rectangular window is first to be determined by the image's size and the relative location of rectangular window. Then calculate the cumulative sum of the gray values of the pixels in the rectangular window and divide it by the number of the pixels to get the average gray value of pixel of one rectangular window. At last, add up all the average gray values of pixel according to their weights to obtain f_k.

4 Intelligent recommendation algorithm

The block diagram of the proposed intelligent recommendation algorithm is shown in Fig. 4. Suppose the audience is divided into K categories by gender and age, and there are L different ads played on $\boldsymbol{T}$ advertising terminals. Then $\boldsymbol{T}$ User-Ad Matrixes of $K \times L$ dimensions can be created. The "User" in User-Ad Matrix denotes each category of audience. The advertisement acceptance rate is expressed as $\delta = t_{\text{watch}}/t_{\text{ptime}}$, where t_{watch} denotes the duration of users watching the ad and t_{ptime} denotes the duration of the ad. When one person's advertisement acceptance rate of some ad δ is larger than threshold δ_g, it can be considered he or she is attracted to the ad, and the value in the corresponding position of the User-Ad Matrix that represents his or her category plus 1. When all the elements in $\boldsymbol{T}$ User-Ad matrixes are calculated, the system can conduct the global relevance analysis and the local popularity analysis of the ads.

In the global relevance analysis of the ads, all the $\boldsymbol{T}$ matrixes are added together to generate a global User-Ad matrix. In this global matrix, ad a_i corresponds to a K-dimensional vector $\boldsymbol{V}_i = (u_1, \cdots, u_m, \cdots, u_K)$, where u_m denotes there are u_m people in m-th audience category attracted to this ad. For any two ads a_p and a_q, the cosine correlation coefficient between them can be calculated based on vector $\boldsymbol{V}_p$ and $\boldsymbol{V}_q$, which is : $\boldsymbol{V}_p \cdot \boldsymbol{V}_q / \| \boldsymbol{V}_p \| \| \boldsymbol{V}_q \|$. For each ad, its $(L-1)$ cosine correlation coefficients corresponding to the other ads can be calculated and arranged in the descending order. And the first Nads corresponding to the first N coefficients are considered to be most relevant. Now an Ad-Ad matrix of $L \times N$ dimensions can be generated, in which each row denotes TOP N recommendation list of each ad.

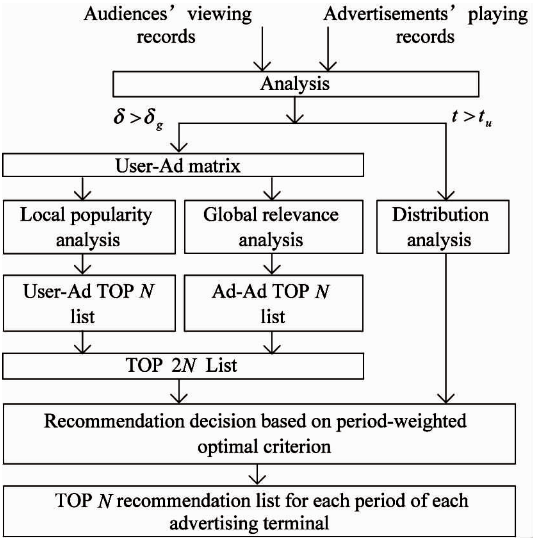
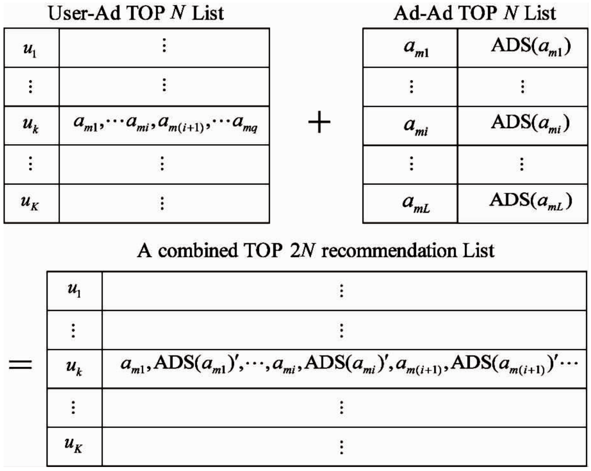

Audiences' viewing records
Advertisements' playing records
Analysis
$\delta > \delta_g$
$t > t_u$
User-Ad matrix
Local popularity analysis
Global relevance analysis
Distribution analysis
User-Ad TOP N list
Ad-Ad TOP N list
TOP 2N List
Recommendation decision based on period-weighted optimal criterion
TOP N recommendation list for each period of each advertising terminal

Fig. 4 The block diagram of the proposed recommendation algorithm

The local popularity analysis of the ads means analyzing the popularity of ads released on each terminal. Each ad is arranged according to the viewer numbers along with the users' category based on the User-Ad Matrix of each terminal. Then the favorite TOP N ads list of each category of audience can be got in front of each advertising terminal, which is called the User-Ad TOP N recommendation list.

After steps described above, a combined TOP 2N recommendation list can be created based on the Ad-Ad TOP N recommendation list and the User-Ad TOP N recommendation list as shown in Fig. 5. Suppose the favorite N ads of some category of audience u_k of some terminals T_j are $a_{m1}, a_{m2}, \cdots, a_{mi}, \cdots, a_{mq}$, as shown in the User-Ad TOP N list in Fig. 5. Then the nearest ads of a_{mi} can be found in the Ad-Ad TOP N recommendation list in Fig. 5 and can be denoted as ADS (a_{mi}). If $a_{m(i+1)}$ appears in ADS (a_{mi}), ADS ($a_{m(i+1)}$) (which can also be found in the Ad-Ad TOP N List) are then used to replace the following ads of $a_{m(i+1)}$ in ADS (a_{mi}) and the ads before $a_{m(i+1)}$ in ADS (a_{mi}) are remained , which can be denoted as ADS (a_{mi})$'$. As long as the number of ads for a_{m1} does not reach to 2N, ADS(a_{m2}), ADS(a_{m3}), $\cdots$, ADS (a_{mL}) are selected for a_{m1} one after another in the similar way. When the selecting ads add up to 2N (including a_{m1} itself), those 2N ads are decided to be the most suitable ads to recommend for the audience u_k of some terminals T_j to continue to select ads for more audiences of more terminals to complete the TOP 2N list. A list created in this way includes individual information of each terminal as well as the global relevant information of all ads.



User-Ad TOP N List

u_1	⋮
⋮	⋮
u_k	$a_{m1}, \cdots a_{mi}, a_{m(i+1)}, \cdots a_{mq}$
⋮	⋮
u_K	⋮

+

Ad-Ad TOP N List

a_{m1}	ADS(a_{m1})
⋮	⋮
a_{mi}	ADS(a_{mi})
⋮	⋮
a_{mL}	ADS(a_{mL})

=

A combined TOP 2N recommendation List

u_1	⋮
⋮	⋮
u_k	a_{m1}, ADS(a_{m1})$'$, $\cdots$, a_{mi}, ADS(a_{mi})$'$, $a_{m(i+1)}$, ADS($a_{m(i+1)}$)$'$ $\cdots$
⋮	⋮
u_K	⋮

Fig. 5 The formation of the combined TOP 2N recommendation list

At last, different weights are to be given to these 2N ads respectively. Suppose w_{ai}^k be the k-th weight of a_i:

$$w_{ai}^k = 2N - i + 1, \quad 1 \leqslant i \leqslant 2N \qquad (6)$$

Meanwhile, distribution table of audience categories in each period of each advertising terminal can be generated. Suppose M periods are calculated from start time of a terminal to the end time of it when half an hour is as one period of time. Suppose t is the length of watching time of some person captured by some terminal. Define another threshold t_u, which is very small.

A person is seen as an audience when $t \geqslant t_u$. As soon as there is audience in some period, its category can be determined by gender and age, and then the overall number of individuals of this category can be calculated; a histogram of all categories of audiences can be created through counting the overall number of individuals of each category. Now suppose in the i-th period of time P_i, the overall number of individuals of the k-th category of audience u_k is N_k, the corresponding weight to be w_k^{pi}:

$$w_k^{pi} = \frac{N_k}{\sum_{i=1}^{K} N_i}, \quad 1 \leqslant k \leqslant K \tag{7}$$

At last, the final TOP N recommendation list specific to each period of each advertising terminal can be generated by combining these two lists. For any terminal T_o, the weight w_k^{pi} of any category of audience in period P_i is known; meanwhile the favorite $2N$ ads of this category of audience along with their weights can be found in the TOP $2N$ recommendation list. If the left $(L - 2N)$ ads have 0 weight values, then the scores of all L ads in this period on this terminal can be calculated. For the n -th ad, its score S_n is calculated as

$$S_n = \sum_{j=1}^{K} w_j^{pi} \times w_{an}^{j}, \quad 1 \leqslant n \leqslant L \tag{8}$$

When these L ads are arranged in the descending order by their scores, the first N ads are exactly the TOP N ads in this period on this terminal. A complete TOP N ads recommendation list specific to each period of this terminal can be generated when the same procedure is carried on in all M periods of time. Further more, the final recommendation list of the whole system can be generated by applying the procedure above on all terminals.

5 Experiments

5.1 Performance evaluation of face detection

Since the results of face detection are decisive for the procedure followed, the performance evaluation of it is first considered. Three datasets are used for performance evaluation: frontal subset of CAS-PEAL-R1, self-collected high resolution images (SCH) dataset and self-collected frontal-face images (SCF) dataset. CAS-PEAL-R1 dataset is built by the Institute of Computing Technology, Chinese Academy of Sciences. All images are collected in limited conditions. It contains the head-shoulder images of 1040 Chinese individuals, of which the frontal subset contains exactly 360 × 480 resolution pictures of the 1040 individual frontal faces. In consideration of the actual work environment of the system being non-limited, two self-collected datasets are created through gathering and sorting out web pictures to assess the performance more precisely. The self-collected high resolution images dataset contains 207 pictures, of which the resolutions are between 300 million and 2400 million pixels. And it contains 247 human faces in total. The self-collected frontal-face images dataset contains 227 pictures and 223 human faces.

The proposed face detection method consists of 3 main steps: 1) Using the classical Viola-Jones cascade detector in OpenCV[4] for primary detection; 2) Selecting based on the matching score of the facial landmarks; 3) Using the nose detector as supplement. Two comparison methods are implemented to analyze the influence on the detection result of each step. Comparing Method No. 1 (CM-I) adopts only the classical Viola-Jones cascade detector in step 1. Comparing Method No. 2 (CM-II) includes step 1 and step 2. Besides, the effective part-model based detector of Zhu and Ramanan[6] is also adopted as comparison method. The results are shown in Table 1. The recall ratio σ_r and precision ratio σ_p are used as the measure standard. To note that the parameters of each detector are adjusted to the optimal state. And CM-I, CM-II adopt the same parameters as the proposed method.

Conclusion can be drawn from Table 1 that four methods can all acquire ideal detection results on the frontal subset of CAS-PEAL-R1, of which the images are collected in limited conditions. The recall ratios and precision ratios of these four methods have slight differences on this dataset. However, four methods performed distinctly differently on the two other non-limited datasets. On the SCH dataset, the precision ratios of CM-I and Zhu&Ramanan detector both have remarkable declines when compared to the proposed method in this paper. This is because the background of the pictures in this dataset appear to be more complex, and the images vary a lot in quality, as well as have a lot of noise. Besides, when the size of the minimum scan window is fixed, the increase in resolution will significantly raise the amount of the windows. So it will be more likely to expose the weakness of low precision ratio of one method. On the SCF dataset, a major drop of precision ratio of CM-I turns up. It is because the dataset contains a lot of ties of complex texture, which will lead to many misjudgments. Some samples are shown in Fig. 6. Through the comparison between CM-I and CM-II, it can be seen that step 1 can boost the robustness of the complicated background and texture, which leads to improvement in detection accuracy. However, as the precision ratio increases, some real faces are misclassified, which will cause the recall ratio drop. Through the comparison between CM-II and the proposed method, step 3 can rectify some misjudgments brought by step 2, thus it can raise the recall ratio under the premise of the precision ratio stays almost the same.

Table 1 Face detection result

Methods	CAS-PEAL-R1		SCH		SCF	
	σ_r	σ_p	σ_r	σ_p	σ_r	σ_p
CM-I	100%	96.4%	98.4%	33.1%	100%	70.3%
CM-II	99.8%	100%	92%	96.6%	94.7%	100%
Proposed method	100%	99.9%	97.6%	96.4%	99.1%	100%
Zhu&Ramanan	99.8%	99.2%	96.4%	53.8%	94.2%	96.8%

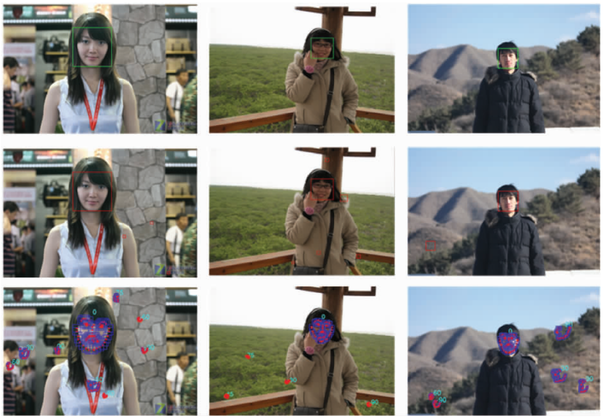


Fig. 6 Some examples of the output of the face detectors on images from the two self-collected dataset. The results of the proposed method, Viola&Jones method and Zhu&Ramanan method are in the first, second and third line respectively

In order to measure the experimental result, the average F1-scores of the three methods are calculated as the evaluation standard. The mathematical formula of F1-score is

$$F1 = \frac{2 \cdot \sigma_p \cdot \sigma_r}{\sigma_p + \sigma_r} \tag{9}$$

The average F1-score of the four methods is 0.768, 0.9714, 0.988 and 0.88, respectively. It can be seen that the proposed method in this study has obvious advantage in F1-score.

5.2 Performance evaluation of face tracking

The performance evaluation of face tracking is carried out as below:

1) Annotate the frame number of the beginning frame and ending frame of each audience's viewing process in video manually.

2) From the appearance of one's frontal face to the disappearance of it can be seen as an audience's viewing process.

3) Suppose there're n audience's viewing processes through manually annotating, and m audience's viewing processes are tracked by face tracking. If q processes out of the m audience's viewing processes match to those of the manually annotated, then the recall ratio is q/n, and the precision ratio is q/m.

Three sections of video were collected for test and their information is shown in Table 2.

Table 2 The information of the test videos

Video name	Time	Resolution	Frame
Video1	14:00	352 × 288	17037
Video 2	12:15	352 × 288	13507
Video 3	00:38	352 × 288	687
Video4	00:27	352 × 288	519
Video5	00:45	352 × 288	641

Since the face tracking method applied in this system's condition is quite different from traditional face tracking, these are not compared with. In experiment three methods are compared instead: (Ⅰ) Face tracking with SVM assisted; (Ⅱ) Face tracking with naive bayesian classifier assisted; (Ⅲ) Face tracking with nothing assisted.

It can be seen from Table 3 that face tracking with SVM assisted has both obviously higher recall ratio and precision ratio. The average F1-score of the three methods is 0.951, 0.785 and 0.206, respectively. It can be seen that the proposed method is able to meet the actual requirement.

5.3 Performance evaluation of recommendation

When someone views the ad in some period of time more than threshold t_u, which is set to 1s in this study, he or she is deemed to be audience. When some audience acceptance rate δ specific to some ad is higher than the preset threshold δ_g, which is set to 0.4, this specific ad can be deemed to effectively arrive at the audience.

Five advertising terminals, A, B, C, D and E, which have different locations, are involved in the experiment. When analyzing the arrival rate of audience, the period from 7:50 am to 8:20 am of the terminal A、

Table 3 Face tracking result

Video name	I		Ⅱ		Ⅲ	
	σ_r	σ_p	σ_r	σ_p	σ_r	σ_p
Video1	95.7%	93.8%	74.5%	58.3%	36.2%	14.4%
Video2	91.7%	88%	62.5%	37.8%	29.2%	8%
Video3	100%	100%	100%	100%	60%	37.5%
Video4	100%	100%	100%	100%	0%	0%
Video5	83.5%	100%	66.7%	100%	33.3%	18.2%

B、C and the period from 6:00 pm to 6:30 pm of the terminal D、E are selected. There are 150 ads released on these terminals. To insure the number of viewers of each advertisement on each terminal, the total number of display advertisements on every terminal should not be too large. Therefore, 150 ads are divided into 5 groups. Terminal A will release the ads in group 1, group 2 and group 3; terminal B will release the ads in group 2, group 3 and group 4; terminal C will release the ads in group 3, group 4 and group 5; terminal D will release the ads in group 1, group 4 and group 5; and terminal E will release the ads in group 1, group 2 and group 5. 7 ads are played on each terminal during the selected period respectively, which means N equals to 7. And audiences are divided into 10 categories by gender and age.

The proposed recommending algorithm first adopts the synergy recommending algorithm to generate the TOP $2N$ recommendation list of each category of audience. The synergy recommending algorithm combines the local popularity and the global relevance. Then the TOP N recommendation list can be obtained based on the period-weighted optimal criterion. The process of generating the TOP $2N$ recommendation list of each category of audience is the key of the proposed recommending method in this paper. Since no other recommending algorithm which specifically aims at the application scenarios of the outdoor video advertising has been proposed, basic synergy recommending algorithms are introduced for comparison in this paper. Two reference methods are implemented. Reference Method No. 1 (RM-I) adds all the $\boldsymbol{T}$ User-Ad matrixes according to the global popularity analysis, and then it can obtain the most popular $2N$ ads in each category of audience to generate the TOP $2N$ recommendation list. Reference Method No. 2 (RM-II) generates the TOP $2N$ recommendation list according to both the local popularity analysis and the User-Ad matrix of each terminal.

The phase of collecting data from advertising terminals and then generating new push strategy on advertising is defined as the data collecting phase. And the phase of assessing the effectiveness of advertising after the new push strategy is defined as effectiveness analysis phase. The duration of the data collecting phase is 20 days, while the duration of the effectiveness analysis phase is 30 days. In the effectiveness analysis phase, every 3 days are set as one test unit. In each test unit, the terminals release ads according to recommendation result of the proposed recommending method in this paper in the first day, the result of the RM-I in the second day, and the result of the RM-II in the third day.

The human traffics of the 5 terminals in the data collecting phase are 7440, 4752, 5528, 3359 and 3662. The display distributions of the ads on the five terminals are shown in Fig. 7. And the comparison of advertisement arrival rate in the two phases is shown in Table 4. The experimental results show that the proposed method can effectively improve the audience arrival rate of the video ads by an average growth of 27.04%. The recommending method based on the local popularity outperforms the recommending method based on the global popularity. This is due to the difference in the audiences' interests with the difference in the location of terminals. Besides, although the display distribution of the ads on each terminal shows uniformity, there exists hardly uniformity on the global basis. It will affect the recommendation result of the recommending method based on the global popularity greatly. The proposed method outperforms the recommending method based on the local popularity. This is because the latter cannot recommend ads to the audience of one terminal

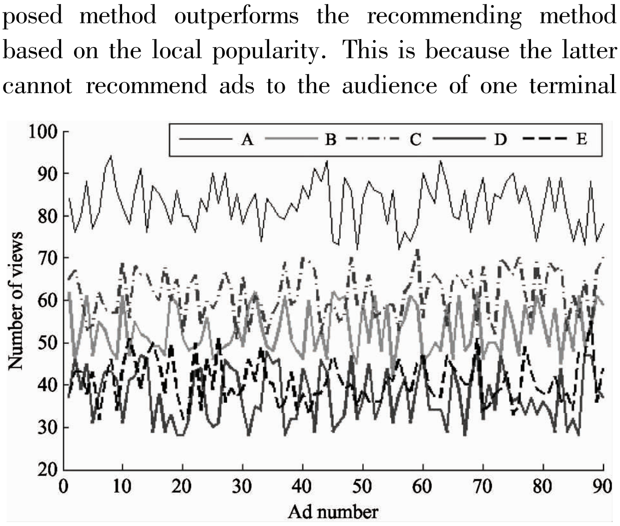


Fig. 7 View numbers of each Ad

Table 4 The result on the advertisement arrival rate

Terminal name	Data collecting phase	The proposed method	RM-Ⅰ	RM-Ⅱ
A	31.34%	58.19%	39.15%	48.29%
B	33.56%	61.43%	44.38%	52.1%
C	32.45%	59.15%	41.52%	47.24%
D	35.68%	64.58%	43.66%	53.76%
E	34.77%	59.67%	42.12%	48.53%

that are not displayed on it. However, even though some ads are not released on one terminal, they may have strong relevance to the popular ads on this terminal. The proposed method can exactly deal with this situation by combining the result of the global relevance analysis. Furthermore, the cosine correlation coefficient is insensitive to the vector length, so the negative influence on the recommending result due to the nonuniformity of the global display distribution of the ads is largely reduced.

6 Conclusion

As more widely the applications of the outdoor advertising terminals grow, one problem has stood out that is the information blocking between merchants and advertising audiences. To overcome this problem, this paper proposes an integrated implementation framework of an intelligent recommendation system for outdoor video advertising based on the analysis of audiences' characteristics. Computer vision technologies: face detection, face tracking, gender recognition and age estimation are applied to obtain the multi-dimensional statistical information of audiences' characteristics.

A landmark localization based face detection method is presented. The average F1-score of this method is 0.988, which is the highest among the four face detection methods.

A face tracking method is proposed directing at the audiences' behaviors of viewing the advertisements. The average F1-score of it is 0.951, which means it is able to meet the actual requirement.

An intelligent recommendation algorithm is designed to push targeted video ads on different terminals according to the multi-dimensional statistical information of audiences. The experimental results show that the proposed system can effectively improve the audience arrival rate of the video ads by an average growth of 27.04%.

References

[1] Li S Z, Zhang Z. FloatBoost learning and statistical face detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2004, 26(9):1112-23

[2] Anvar S M H, Yau W Y, Teoh E K. Multiview face detection and registration requiring minimal manual intervention. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2013, 35(10):2484-97

[3] Ma K, Ben-Arie J. Vector array based multi-view face detection with compound exemplars. In: Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012. 3186-3193

[4] Viola P, Jones M J. Robust real-time face detection. *International Journal of Computer Vision*, 2004, 57(2):137-54

[5] Cevikalp H, Triggs B, Franc V. Face and landmark detection by using cascade of classifiers. In: Proceedings of the 10th IEEE International Conference on Automatic Face and Gesture Recognition, Shanghai, China, 2013. 1-7

[6] Zhu X X, Ramanan D. Face detection, pose estimation, and landmark localization in the wild. In: Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012. 2879-2886

[7] Zhu L, Chen Y, Yuille A, et al. Latent hierarchical structural learning for object detection. In: Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA, 2010. 1062-1069

[8] Learned-Miller E G, Lara L S. Distribution fields for tracking. In: Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012. 1910-1917

[9] Oron S, Bar-Hillel A, Levi D, et al. Locally orderless tracking. In: Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012. 1940-1947

[10] Zhang K, Zhang L, Yang M H. Real-time compressive tracking. In: Proceedings of the 12th European Conference on Computer Vision, Firenze, Italy. 864-77

[11] Makinen E, Raisamo R. Evaluation of gender classification methods with automatically detected and aligned faces. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2008, 30(3):541-7

[12] Lian H C, Lu B L. Multi-view gender classification using local binary patterns and support vector machines. In: Proceedings of the 3rd International Symposium on Neural Networks, Chengdu, China, 2006. 202-209

[13] Luu K, Ricanek K, Bui T D, et al. Age estimation using active appearance models and support vector machine regression. In: Proceedings of the IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, Washington, USA, 2009. 1-5

[14] Wu T, Turaga P, Chellappa R. Age estimation and face verification across aging using landmarks. *IEEE Transactions on Information Forensics and Security*, 2012, 7(6): 1780-1788

[15] Davidson J, Liebald B, Liu J, et al. The YouTube video recommendation system. In: Proceedings of the 4th ACM Conference on Recommender Systems, Barcelona, Spain, 2010. 293-296

[16] Park J, Lee S J, Lee S J, et al. Online video recommendation through tag-cloud aggregation. *IEEE Transactions on MultiMedia*, 2011, 18(1):78-86

[17] U řičář M, Franc V, Hlaváč V. Detector of facial landmarks learned by the structured output SVM. In: Proceedings of the 7th International Conference on Computer Vision Theory and Applications, Rome, Italy, 2012. 547-56

Liu Peng, born in 1989. He is a Ph. D. student in the Haikou Laboratory, Institute of Acoustics, Chinese Academy of Sciences. He received his BS degree from Hainan University in 2011, majoring in communication engineering. His research interests mainly focus on multimedia signal processing and machine learning.