Robust Speaking Face Identification for Video Analysis

Yi Wu[1, 2], Wei Hu[2], Tao Wang[2], Yimin Zhang[2], Jian Cheng[1], Hanqing Lu[1]

[1]National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Science
{ywu, jcheng, luhq}@nlpr.ia.ac.cn
[2]Intel China Research Center, Beijing, P.R. China
{wei.hu, tao.wang, yimin.zhang}@intel.com

Abstract. We investigate the problem of automatically identifying speaking faces for video analysis using only the visual information. Intuitively, mouth should be first accurately located in each face, but this is extremely challenging due to the complicated condition in video, such as irregular lighting, changing face poses and low resolution etc. Even though we get the accurate mouth location, it's still very hard to align corresponding mouths. However, we demonstrate that high precision can be achieved by aligning mouths through face matching, which needs no accurate mouth location. The principal novelties that we introduce are: (i) proposing a framework for speaking face identification for video analysis; (ii) detecting the change of the aligned mouth through face matching; (iii) introducing a novel descriptor to describe the change of the mouth. Experimental results on videos demonstrated that the proposed approach is efficient and robust for speaking face identification.

Keywords: SIFT, watershed, speaking face identification, change detection, mouth alignment, video analysis

1 Introduction

Speaker identification is a crucial step in many video analysis problems such as automatic annotation of characters [1], [10], [11], audio-visual speech recognition [2], user interfaces based on vision [3], [9], etc. In this paper, we address the speaking face identification problem in teleplay or movie video, only using the visual information. It is a challenging problem due to the following reasons: 1) face pose and expression change; 2) lip deformation; 3) changing illumination; 4) background clutters and 5) other factors, such as motion of the camera.

In recent years, many techniques have been proposed for speaker identification. Saenko et al. [2] use SVM to train a discriminative classifier to locate the lip and then train another strong classifier to detect the subclass of lip appearance corresponding to the presence of speech. However, they only consider frontal and upright faces under the controlled environment which is not practical in teleplay or movie video. In [3] Murphy et al. use Bayesian network model as an attractive statistical framework for cue fusion to detect speaker. The model combines four simple vision sensors: face detection, skin color, skin texture, and mouth motion. Their aim is to build a human-

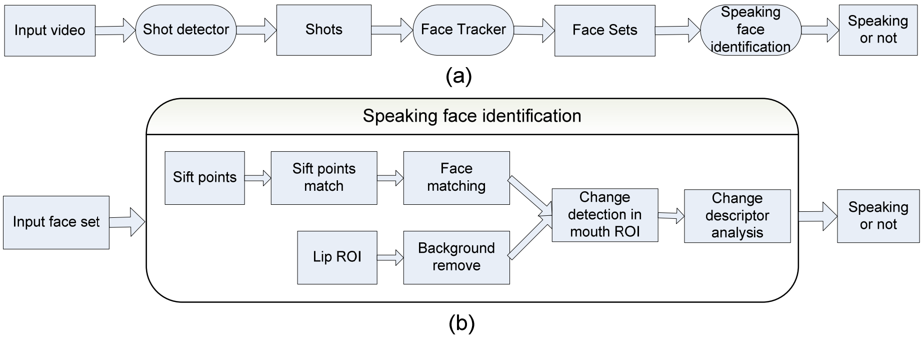


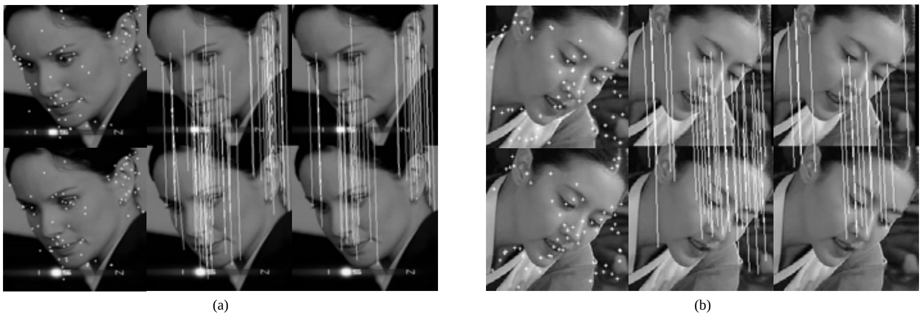
Fig. 1. (a) The flowchart of the proposed speaking face identification system; (b) The flowchart of the speaking face identification module.

centered user interface and they assume that the face detector can only detect frontal faces. The assumption may be useful to construct a robust user interface, but again it is not true when dealing with teleplay or movie video.

Everingham et al. [1] use speaking face identification as one module of automatic naming of characters in TV video. They achieved this by finding face detections with significant lip motion. A rectangular mouth region within each face is identified using the located mouth corners, and a mean squared difference of the pixel values in consecutive mouth regions is computed to determine if the shape of the mouth is changing or not. To achieve translation invariance the difference is computed over a search region around the mouth in the current frame and the minimum is taken. Two thresholds on the difference are set to classify face into ‘speaking’, ‘non-speaking’ or ‘refuse to predict’. There are many constraints in their approach. First, they detect and track frontal upright faces which only occupy about 40% of the total faces in telefilm videos. The statistical number is got in our experiments by using multi-view face detector. Second, the detected mouth corners are used to locate the mouth and align the mouths, but to locate mouth corners precisely and stably is still a difficult problem, especially in profile face and moving mouth. Finally, they only consider translate transformation between two consecutive mouths. In this paper, we will try to resolve these problems and construct a robust speaking face identification system.

Despite many works have been proposed for speaker detection, most of the existing methods limit their use in indoor situations with controllable lighting condition, and their experiments are based on full frontal upright faces in good quality images. Few of them mentioned possible solutions to robust multi-view speaking face identification in real media such as teleplay or movie.

To address the problems mentioned above, we propose a framework for speaking face identification in this paper. The proposed framework is illuminated in Fig. 1. Video is first segmented into shots [4]. Then face sets in each shot are got by multi-view face detector and tracker [5]. Finally each face is labeled to be speaking or not roughly as following steps. First, the mouth Region-of-Interest (ROI) is located using the information got from the face detection and tracking module. The watershed segmentation [8] is then applied to remove non-face background pixels. Second, SIFT feature points are extracted in current face image and previous one, then these two



(a) (b)

Fig. 2. Left column of each entry: the result of SIFT feature points detection; Middle: feature matching result before RANSAC; Right: matching result after RANSAC.

sets of SIFT points are matched. Third, we use the matched SIFT points to calculate the transformation model to wrap the current face to the previous face image plane. The change in the aligned mouth ROI can be used to judge if the face is speaking. Here, we use a novel descriptor, which we call *Normalized Sum of Absolute Difference* (NSAD), to describe the change in the mouth ROI. Thus we get a vector of NSAD for each face set and use it to label if the face is speaking. Finally, we post-process the label vector and get the final smooth identification result. Experimental results on videos demonstrate that the proposed approach is efficient and robust for speaking face identification.

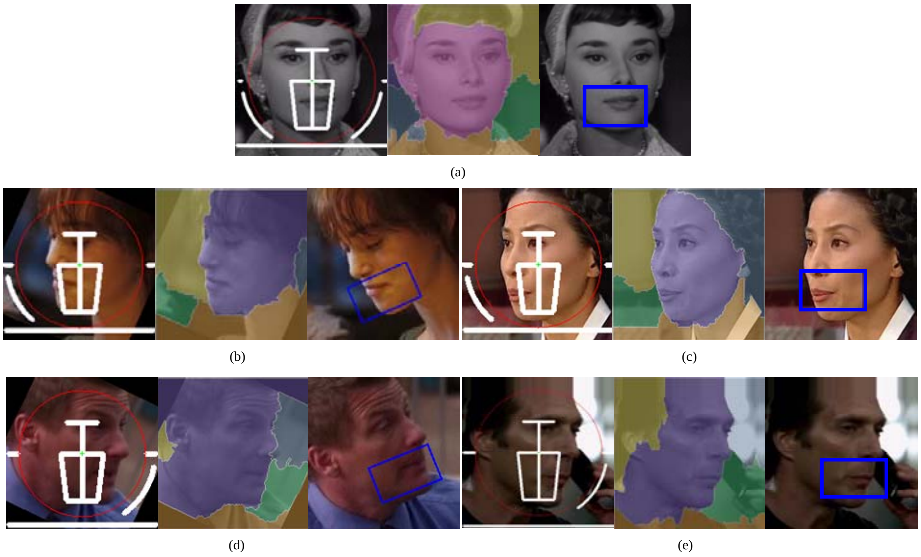
The paper is organized as follows. Section 2 describes our approach for speaking face identification for video analysis. The experimental results are demonstrated in Section 3, followed by the conclusions in Section 4.

2 Speaking face identification

The speaking face identification is performed in two steps. The first step aligns the consecutive mouths through face matching. At the same time, the mouth ROI is extracted and background pixels in the ROI are removed. In the second stage the speaking face identification is achieved by capturing the change of the mouth using a novel descriptor.

2.1 Mouth alignment

We observed the fact: it is much easier to match two consecutive whole faces robustly than to match two mouths directly, because face is more informative than mouth, and with less deformation. It is true especially when the resolution is low or the face size is small. Through matching two faces, the corresponding mouth alignment can be achieved naturally: when the two consecutive faces are well matched, the mouth on each face is also well aligned. Moreover, the translation limitation will be eliminated by using a four parameter affine transformation model when matching the faces.



(a)

(b)

(c)

(d)

(e)

Fig. 3. Mouth ROI extraction and background remove. Left image of each example: the green point is the center of the face and the red circle reflects the face scale. The white lines are connected components for watershed segmentation. Middle image of each example: the segmentation result of the watershed algorithm. Right image of each example: the blue rectangle is the mouth ROI.

2.1.1 SIFT feature points detection

For each face set, the current face image and its previous one are fed to the feature detector. The feature detector should be able to work reliably in demanding natural environments. It should be robust against illumination variations, imaging noise, image rotation and scaling. We tested different approaches presented in the literature [12], [13], and found that the SIFT feature [7] performs best. See [7] for details of SIFT points detection.

Left columns of Fig. 2(a) and (b) illustrate detected SIFT points using the SIFT detector. After the feature points have been detected, they are forwarded to the feature matching stage.

2.1.2 SIFT feature points matching

For the two sets of feature points got in the consecutive faces, respectively, we seek for two closest by using the Euclidean distance as the similarity measure. If the two distances are too close to each other, the matching cannot be done reliably and the feature is discarded. Otherwise, the closest match is included to the match set. This procedure effectively removes the duplicate matches. In our experiments, we ignore the feature if the ratio of the two closest distances is bigger than 0.6. The feature matching stage outputs a set of feature matches between the current face image and the previous one. See Fig. 2 for examples of SIFT points matching.

2.1.3 Estimation of Geometric Transformation

After the feature matching stage, we have a set of feature correspondences between the current face image and the previous one. Most of the duplicate features are removed during the matching process, but there is still a possibility that some outliers, such as mismatched feature points, are included in the set. In order to achieve a reliable estimate for the transformation model, these outliers need to be removed. We adopt a well known and robust algorithm, the RANdom SAmple Consensus (RANSAC) [6]. The matching result after RANSAC is shown in Fig. 2.

In this work, a four parameter affine model is used. It is considered as sufficient for approximating transformation between consecutive faces as it can represent 2-D transformation consisting of translation, rotation, and scaling:

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} s\cos\theta & -s\sin\theta & x_0 \\ s\sin\theta & s\cos\theta & y_0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$$

where (x, y) is a coordinate in the current face image and (x', y') is its correspondence in the previous image. x_0, y_0 are related to common translational motion and s, θ are 2-D scale and rotation respectively.

The transformation model can now be used to transform the current face image to the previous face image plane. After that, the mouth would be aligned.

2.2 Mouth ROI extraction

To detect the change of the mouth on consecutive images, we need to locate the mouth on each face first. However, accurate mouth location directly is a challenging task, especially in teleplay or movie video, because the environment is so complicated that there is no uniform color space to describe the lip/mouth, and the difference between lip and face may be very small in color or intensity. Fortunately, we can make use the face information acquired from the face detector and tracker to help to locate the mouth. Although the mouth location is somewhat coarse, it works very well for our purpose of speaking face identification.

The multi-view face detector and tracker [5] we used can provide the following information for each face: the center location of the face, the scale, in-plane rotation angle and out-of-plane rotation angle. From observation, some simple heuristic rules can be outlined and used to locate the mouth ROI. Example of location result is shown in Fig. 3 (the blue box).

Let (x_{lf}, y_{lf}) denote the left top point of the mouth ROI and w, h denote the width and the height of the ROI respectively. The center and the radius (scale) of the face are denoted by (x_c, y_c) and r , respectively. We employ a simple empirical formulas to

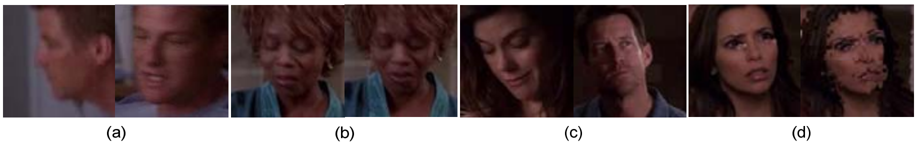


Fig. 4. Example images that can not be aligned well. (a) (b) motion blur; (c) shot detection error; (d) the poor quality of the video.

locate the mouth ROI as follow:

$$w = r, h = 0.6 \times r, y_{lf} = y_c + 0.2 \times h$$

$$x_{lf} = \begin{cases} x_c - 0.67 \times w & left-profile \\ x_c - 0.5 \times w & frontal \\ x_c - 0.33 \times w & right-profile \end{cases}$$

x_{lf} is changed according to the different face poses (out-of-plane rotation). If the face has in-plane rotation the mouth ROI should rotated according to the degree, as illustrated in the left of Fig. 3(b), (d).

2.3 Background remove

The mouth ROI acquired in above step may contain some non-face region such as background clutters, especially for profile face. This will greatly influence the result of the change detection in the mouth ROI. Thus, these non-face regions should be removed before the mouth change detection. However, face segmentation is a challenging problem, especially in teleplay or movie video, due to the complicated illumination and background clutters. In this paper, face is segmented by watershed algorithm [8] which can easily make use the prior knowledge, e.g. the face information got from the face detector and tracker.

Connected component region selection is the most critical stage of the watershed method. Based on the information from the face detector and tracker, an empirical connected component mask is designed to separate the face pixels from the clutter backgrounds, for each of the three face poses respectively, as illustrated in Fig. 3. Although this kind of segmentation is somewhat coarse, it works well. After the segmentation, most non-face background pixels near the mouth are removed.

2.4 Mouth change description

After consecutive mouths are aligned and non-face background has been removed, we now describe the change in the mouth ROI. This change is a strong cue for speaking face identification. In [1], mean squared difference of the pixel values in the mouth region is computed between the current and previous frame to describe the change. To achieve translation invariance the difference is computed over a search region around the mouth region in the current frame and the minimum is taken. There are two main limitations of this approach: 1) the descriptor is not normalized according to the

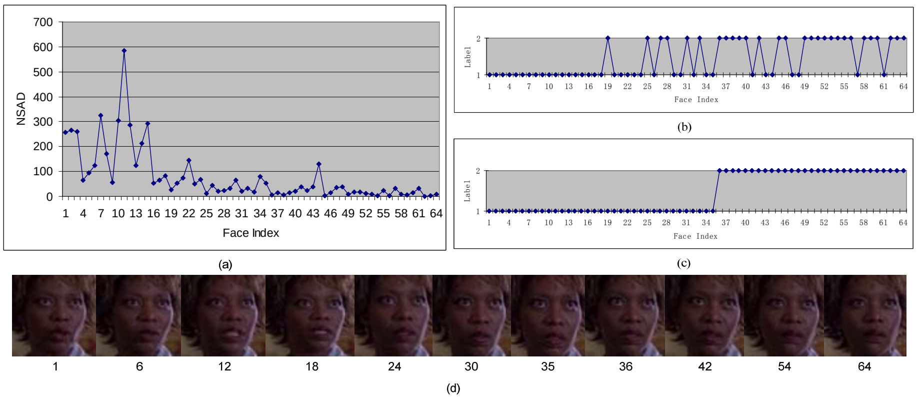


Fig. 5. (a) Illustration of the vector of change descriptor; (b) Illustration of speaking face labeling; (c) Label smoothing; (d) sample images from the face set. The character is speaking between frames 1-35 and remains silent for the rest of the track.

illumination and the scale of the face or the mouth; 2) it only considers translation transformation of the face. However, the motion of the face is not that simple, especially in teleplay or movie video. We have solved the second problem by the geometric transformation which can represent 2-D transformation consisting of translation, rotation, and scaling. Here, we will describe our proposed change descriptor which is illumination and scale normalized.

Denote the intensity in the previous face image and the current face image by I_p and I_c respectively. We can get the ordinary *Sum of Absolute Difference* (SAD) in the mouth ROI as follows:

$$d_{SAD} = \sum_{ROI} \left| I_p - I_c \right|$$

Then the average and standard deviation for the previous face region $\left(\mu_p, \sigma_p\right)$ and $\left(\mu_c, \sigma_c\right)$ for the current face region are calculated. The *Illumination Normalized SAD* (INSAD) and *Scale Normalized SAD* (SNSAD) are calculated as follows:

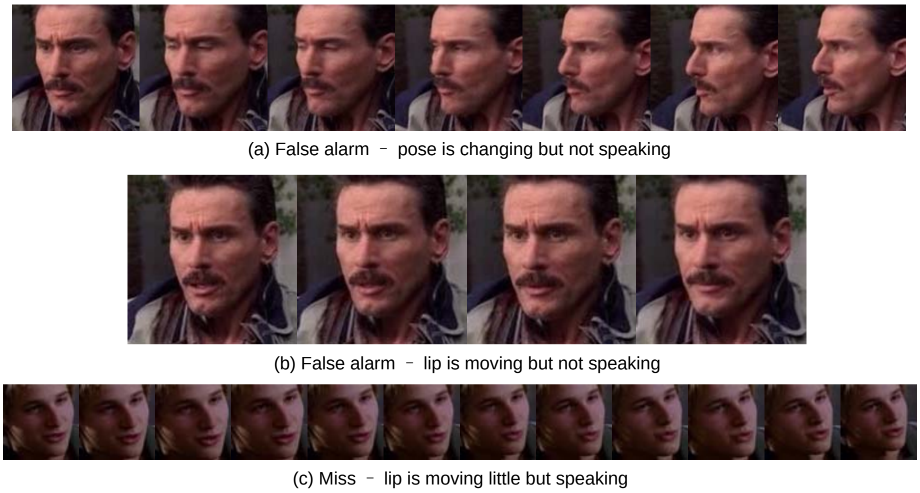
$$d_{INSAD} = \sum_{ROI} \left| \frac{I_p - \mu_p}{\sigma_p} - \frac{I_c - \mu_c}{\sigma_c} \right|$$

$$d_{NSAD} = d_{SNSAD} = d_{INSAD} \times s_0 / s$$

$$s = w \times h$$

We normalize the INSAD from scale s to scale s_0 and let the *Normalized SAD* (NSAD) $d_{NSAD} = d_{SNSAD}$. w, h stand for the width and the height of the previous face image respectively. In our experiments, we let $s_0 = 160 \times 160$.

In real-world environment, there are cases that the number of matched SIFT points in the consecutive faces is small, thus there is no reliable alignment between these two images. This always occurs when there are motion blur or shot detection



(a) False alarm – pose is changing but not speaking

(b) False alarm – lip is moving but not speaking

(c) Miss – lip is moving little but speaking

Fig. 6. The example face images of false alarm and misses.

error, as shown in Fig. 4. In these cases, we let the NSAD equal to 2000 and refuse to judge if the face is speaking.

After normalization, we can label each face if it is speaking according to the NSAD value, regardless of the illumination or the scale changes.

2.5 Speaking face labeling

In a face set, we calculate a NSAD value for each pair of consecutive face images, and then we get a vector of change descriptor, as illustrated in Fig. 5(a). When the NSAD is small, which means there is a reliable alignment between two face images, and at the same time the mouth is not moving, we can label it as 'non-speaking'. If the NSAD is relatively large, we label it as 'speaking'. If the NSAD is too large, we label it as 'refuse to predict'. In most cases, the too large NSAD values come from the wrong alignment of faces or out-of-plane rotation of the face. Following is the labeling criterion.

$$L = \begin{cases} 1 & t_1 \le d_{NSAD} \le t_2 \\ 2 & d_{NSAD} < t_1 \\ 3 & d_{NSAD} > t_2 \end{cases}$$

1: speaking; 2: non-speaking; 3: refuse to predict.

In all of our experiments, $t_1 = 30, t_2 = 700$. Fig. 5(b) shows the labeling result.

In the label vector of each face set, some label value may be different from its neighbors. This does not make sense in practice. To solve this problem, we can smooth the NSAD vector before labeling. Here we smooth the label vector instead, since it is semantic and meaningful. We use a median filter with the window size setting to five. The result is illustrated in Fig. 5(c).

3 Experimental results

Experiments are carried out on five videos, including *episode 22 of Prison Break season 2, episode 21 of Desperate Housewives season 2, episode 31 of Dae Jang Geum, 25-minute clip of Roman Holiday* and *one hour clip of Pride and Prejudice.* (In the following we use *PB, DH, Dae, Roman* and *Pride* for short respectively) We do not care those face sets that are not long enough (less than 40 frames), or contain very small faces, e.g., the scale (radius) of the face is less than 35 pixels. About ninety percent of the ignored face sets are non-speaking. Details of the data are shown in Table 1.

In order to quantitatively evaluate the proposed method we randomly select twenty face sets of each video to label each face if it is speaking, totally 13,527 faces, and 6,348 faces among them are labeled as speaking. Table 2 shows the precision/recall result. The term 'precision' and 'recall' are defined as follows:

$$precision = \frac{\#correctly\ identified\ speaking\ \ face}{\#identified\ speaking\ \ face}$$

$$recall = \frac{\#correctly\ identified\ speaking\ \ face}{\#total\ ground\ truth\ speaking\ \ face}$$

Note that the criterion is stricter than that used in [1]. We evaluated the result in 'frame' level while their evaluation is in 'track' level.

In our experiments, false alarm usually happens when the character's head is out-of-plane moving or the lip is moving but the character is not speaking. Miss detection of the speaking face always happens when the character is speaking but the lip doesn't move or move slightly. These incorrect cases are hardly eliminated through only the visual information, even if we can get the accurate contour of the lip. Some false identification examples are shown in Fig. 6.

When we fuse visual information with audio cues, the false alarm and miss would mostly be eliminated. This is our on-going work and will be reported elsewhere.

Table 1. The information of each video.

	Frames	Resolution	Face Sets	Faces	Filtered sets	Filtered faces
PB	62127	608*336	953	34415	243	17113
DH	60856	608*336	908	53167	285	25669
Dae	67023	352*288	822	67444	346	38980
Roman	37372	528*384	507	58960	107	17440
Pride	104458	640*272	959	85069	320	37697

Table 2. The precision/recall result of each video.

	PB	DH	Dae	Roman	Pride
precision	81.2%	89.1%	91.5%	85.2%	86.8%
recall	88.3%	81.9%	86.7%	83.4%	85.1%

4 Conclusion

Automatically identifying speaking faces for video analysis based solely on visual input is a challenging problem. In this paper, the speaking face identification is formulated as a change detection problem. We align the mouths through face matching and propose a novel change descriptor which is illumination and scale normalized. It can describe the change of the mouth effectively and we can get accurate speaking face identification through the analysis of the NSAD. The proposed method is tested on five videos and the experimental results demonstrate that the approach is reliable and robust.

5 Acknowledgements

This work is partially supported by the National Hi-Tech Research and Development Program of China (863 Program, Grant No. 2006AA01Z315), and the National Natural Science Foundation of China (Grant No. 60121302).

References

1. M. Everingham, J. Sivic and A. Zisserman: “Hello! My name is... Buffy” – Automatic Naming of Characters in TV Video. In *Proc. of the BMVC*, pp. 889-908, 2006.
2. K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell: Visual Speech Recognition with Loosely Synchronized Feature Streams. In *Proc. of the ICCV*, 2005.
3. J. M. Rehg, K. P. Murphy, P. W. Fieguth: Vision-Based Speaker Detection Using Bayesian Networks. In *Proc. of the CVPR*, 1999.
4. J. Yuan, W. Zhang, and et al.: Shot boundary detection and high-level feature extraction. In *TRECVID Workshops*, 2004.
5. Y. Li, H. Z. Ai, C. Huang, S. H. Lao: Robust Head Tracking with Particles Based on Multiple Cues Fusion. In *HCI/ECCV*, pp. 29-39, 2006.
6. M. A. Fischler and R. C. Bolles: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM, 24(6):381-395, 1981.
7. D. G. Lowe: Distinctive Image Features from Scale-Invariant Keypoints. In *International Journal of Computer Vision*, 2004.
8. Meyer, F.: Color image segmentation. In *Proc. of the ICIP*, pp. 303-306, 1992.
9. K. Waters, J. M. Rehg, M. Loughlin, S. B. Kang, and D. Terzopoulos: Visual sensing of humans for active public interfaces. In *Computer Vision for Human-Machine Interaction*, pp. 83-96, 1998.
10. O. Arandjelovic and A. Zisserman: Automatic face recognition for film character retrieval in feature-length films. In *Proc. of the CVPR*, pp. 860-867, 2005.
11. M. Everingham and A. Zisserman: Identifying individuals in video by combining generative and discriminative head models. In *Proc. of the ICCV*, pp. 1103-1110, 2005.
12. J. Shi and C. Tomasi: Good features to track. In *Proc. of the CVPR*, pp. 593-600, 1994.
13. C. Harris and M. Stephens: A combined corner and edge detector. In *4th Alvey Vision Conference*, pp. 147-151, 1988.