HEAD DETECTION FOR VIDEO SURVEILLANCE BASED ON CATEGORICAL HAIR AND SKIN COLOUR MODELS

Zui Zhang, Hatice Gunes and Massimo Piccardi

Faculty of Engineering and Information Technology, University of Technology, Sydney (UTS), Australia

ABSTRACT

We propose a new robust head detection algorithm that is capable of handling significantly different conditions in terms of viewpoint, tilt angle, scale and resolution. To this aim, we built a new model for the head based on appearance distributions and shape constraints. We thus construct a categorical model for hair and skin, separately, and train the models for four categories of hair (brown, red, blond and black) and three categories of skin representing the different illumination conditions (bright, standard and dark). The shape constraint fits an elliptical model to the candidate region and compares its parameters with priors based on the human anatomy. The experimental results validate the usability of the proposed algorithm in various video surveillance and multimedia applications.

Index Terms— Head detection, hair color, skin color, shape constraints

1. INTRODUCTION

The purpose of head detection is to determine the presence and position of heads within an image or a video frame Some remarkable applications include human tracking in video surveillance systems [7, 11], intelligent environments [4, 8], and others. Depending on the type of input images (e.g., passport photo-like vs. scene at distance) and applications, the approaches used for head detection vary significantly. For applications with high resolution and frontal face input images, generally face detection is adopted instead of head detection. Face detection refers to any technique relying on the visibility of the main facial features such as the eyes, eyebrows, nose and mouth, and sometimes hair. However, in this paper we target the more general case of head detection where facial features' visibility is limited or not available at all due to either limited image resolution or non-facial views of the head.

The most popular approaches for head or face detection are the appearance-based approaches where skin colours are modelled to provide a rough estimate of the facial region [6]. However, a major disadvantage of these approaches is that they are constrained to detect quasi-frontal view faces only. A solution to this issue is the introduction of hair colour detection parallel to the skin colour detection when detecting the human head (e.g., [5]).

Accordingly, this paper presents a new robust head detection algorithm that is effective in a range of operating conditions including viewpoint (i.e. frontal, profile, back view, from -180 degrees to +180 degrees), tilt angle (i.e. from horizontal to aerial), scale and resolution. Our method is capable of handling head detection in images with challenging illumination conditions and low resolutions. We construct a categorical model for hair and skin, separately, and train the models for four categories of hair (brown, red, blond and black) and three categories of skin representing the different illumination conditions (bright, standard and dark).

Sections 2 and 3 describe the hair colour and skin colour models used for head detection respectively. Section 4 proposes the principle for head shape modelling and fitting based on the detected head region. Section 5 presents the experiment results from two different environments while Section 6 provides the concluding remarks.

2. HAIR COLOR DISTRIBUTION MODEL

Hair colours of humans seem to cluster around a few colour categories, namely black, blond, brown, and red. Our hair colour training dataset consists of manually extracted image clips from a range of head images with different illumination conditions collected from both the WWW and the colour FERET database [2]. Figure 1 shows some samples from our training set.

We construct a categorical model for hair and train it for four categories of hair. In our approach, we have decided to adopt a combination of two distinct colour spaces (XYZ and HSV) in an attempt at increasing the overall representation accuracy as researchers in image processing have not reached a consensus as to which colour space is more efficient for representation.

The pdf of each category is estimated using a Gaussian mixture whose parameters are estimated through the EM algorithm. For each mixture, the number of components is set manually by examining the histogram of the training samples. Experimental observations show that the estimated Gaussian mixture models fit head images from various datasets with good accuracy.

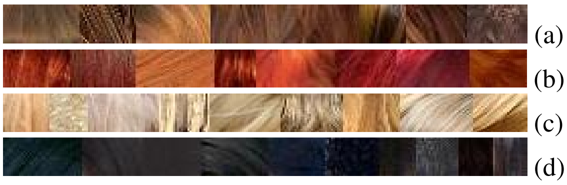

Figure 1. Example of hair colour training samples manually extracted from a range of head images: (a) brown, (b) red, (c) blond and (d) black.

For a pixel to be labelled as a hair pixel, the likelihood of its colour in the hair distribution models must be high. Let us assume that a pixel, x, is represented as $\{x_X, x_Y, x_Z\}$ and $\{x_H, x_S, x_V\}$ in the XYZ and HSV colour spaces, respectively, and that G is a variable representing the hair colour category, $G \in \{black, blond, brown, red\}$. The likelihood of x given G in the XYZ and HSV spaces is then estimated as follows:

$$p^{XYZ}(x \mid G) = p(x_X \mid G) * p(x_Y \mid G) * p(x_Z \mid G) \tag{1}$$

$$p^{HSV}(x \mid G) = p(x_H \mid G) * p(x_S \mid G) * p(x_V \mid G) \tag{2}$$

The right members of equations (1) and (2) are a simplification of the joint probability assuming statistical independence between colour channels. The combined probability of x given G for the two colour spaces is then estimated as:

$$P_{ha}(x \mid G) = w_1 p^{XYZ}(x \mid G) + w_2 p^{HSV}(x \mid G) \tag{3}$$

In an ideal case, $p^{XYZ}(x \mid G)$ and $p^{HSV}(x \mid G)$ should have very similar values; therefore, an assumption of independence in equation (3) would be inappropriate and the combined probability is instead approximated by a weighted sum criterion. The final probability of pixel x being a hair pixel $P_{ha}(x)$ is estimated as:

$$P_{ha}(x) = \underset{G}{arg\,max}(P_{ha}(x \mid G)) \tag{4}$$

In equation (4), we deliberately avoid the use of category priors in order to not trade off individual accuracy for higher accuracy over the entire population. Eventually, a binary decision is made for x based on inequality $P_{ha}(x) > th_{ha}$, where th_{ha} is a threshold determined experimentally. Weights in equation (3) are also chosen experimentally so as to maximize correct hair pixel detection over a training set.

3. SKIN COLOR DISTRIBUTION MODEL

The colour distribution of human skin is spread more continuously in the colour space than hair colour distribution. Thus, it is not possible to model skin colour based on colour categories. Instead, for the algorithm to be flexible for different environments and/or lightning conditions, we model the skin colour in three categories based on different illumination conditions, namely,.bright, standard, and dark.

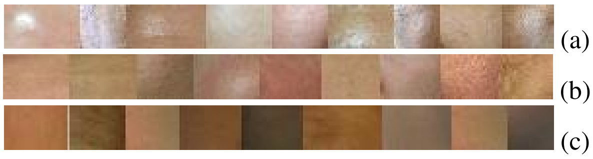

Figure 2. Example of skin colour training samples manually extracted from a range of head images under different illumination conditions: (a) bright, (b) standard, (c) dark

Our skin colour training dataset also consists of manually extracted image clips from a range of head images with different skin colours and different illumination conditions collected from both the WWW and the colour FERET database [2]. Figure 2 shows samples of skin image clips. The categorical modelling of the skin colour distribution by a Gaussian mixture was performed in a similar manner as for each category of the hair colour distribution model. Since the general principal of the categorical skin colour model is the same as hair colour model, the equations for deriving the final probability of a pixel x being a skin pixel is very similar to those explained in Section 2. Eventually, pixel x is labelled as a head pixel simply if it is labelled as a hair pixel or a skin pixel

4. HEAD SHAPE MODELING AND FITTING

As the proposed method is purely based on colour, the detection accuracy is highly dependent upon the differences of lighting conditions between the training and the test sets. In order to increase the flexibility of the proposed method, a lighting compensation technique (similar to [6]) is performed before the actual hair/skin detection takes place. In addition to the appearance criterion, we also impose a shape constraint on the candidate head region.

Thus, on the set of points returned by the hair and skin detection algorithms, we first apply a morphological closure, a connected-component labelling algorithm and a noise filtering. Then, on each of the N connected components found, $O_j, j = 1 \ldots N$, a set of ellipses is fit in order to find the more appropriate one that models the detected head according to certain criteria explained below. For that purpose, certain parameters are computed:

- *T*: defined as the top-most position on the head, and it can be computed easily by scanning through the region vertical downwards from the top
- *a*: the major axis of the ellipse is assumed to be always vertical with no tilt as this makes it faster to find the best-fit-ellipse of the connected component. It was proven that this assumption did not imply worse performance.
- The best fitting ratio $r(a)$ is defined as the occupancy of the ellipse over the connected component, times the proportion of the ellipse's region compared to the connected component's region. By trying different size

ellipses, the ellipse that maximizes $r_{(a)}$ is selected as the best-fit-ellipse of the connected component. This ellipse is then tested against a set of rules constraining the position and occupancy ratio of the candidate head region to confirm it as head or discard it. The best fitting ratio R is defined as follows:.

$$R = \max(r_{(a)} \mid (0.65 I_0 \leq a{=}0.65 I_0 + i \leq 1.35 I_0 , i{=}1,2,\ldots))=$$

$$R = \max(\frac{H}{E^1_{(a)}} \bullet \frac{E^2_{(a)}}{B}) \qquad (9)$$

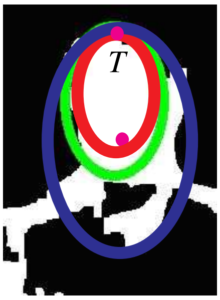

Figure 3. Illustration of the head shape modelling and fitting procedure. Red ellipse is the initial ellipse, blue ellipse is the ending ellipse, and green ellipse is the best-fit ellipse.

I_0 (a constant) is the approximation of length (in pixels) of the major axis of the ellipse that best fits the head in the image. H is the number of skin/hair pixels within the ellipse, $E^1_{(a)}$ is the area of the ellipse, $E^2_{(a)}$ is the size of the best-fit-rectangle of the ellipse, and B is the size of the best-fit-rectangle of the connected component. The operation of the algorithm for finding the best fitting ellipse is illustrated in Figure 3.

5. EXPERIMENTS

As the main goal of the proposed head detection algorithm is to assist human tracking in realistic environments and scenarios, we carried out experiments to evaluate the proposed algorithm for different settings:

- Experiment 1: a laboratory environment vs surveillance environment. We used the AVSS 2007 multiple faces dataset for laboratory environment and abandoned baggage scenario dataset for surveillance environment. First dataset contains four targets repeatedly occluding each other while appearing and disappearing from the field of view of the camera [1] (see Figure 4) and the second was created for event detection in CCTV footage [1 (see Figure 4).
- Experiment 2: comparison between our algorithm with the Haar-like features face detection algorithm [9] provide by OpenCV. A variety of video sequences from AVSS 2007 and ETISEO[3] dataset.

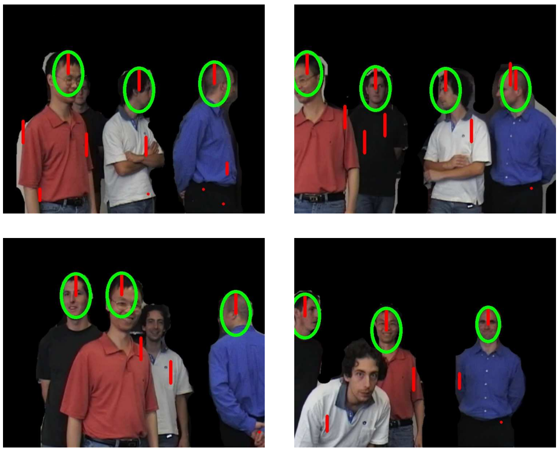

Figure 4. Lab environment: results on manually segmented AVSS 2007 sequences.

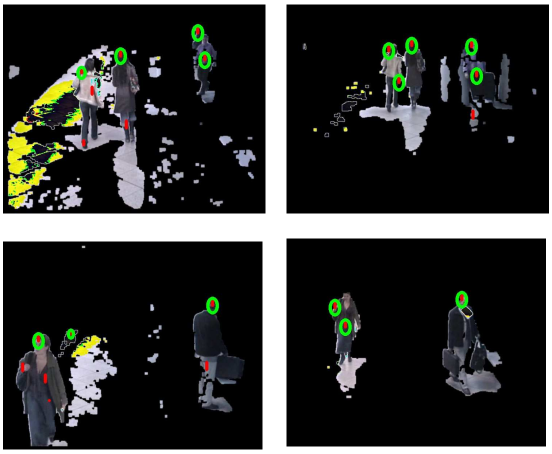

Figure 5. Surveillance environment: results on surveillance video sequences taken from AVSS 2007 dataset.

The experimental results are summarized in Table 1-3. Accuracy or hit rate (HR), is calculated as the number of correctly detected heads divided by the number of actually present heads, while here we define the false alarm rate (FAR) as the number of incorrectly detected heads divided by the total number of detected heads. In addition, we analyse the reasons why the proposed head-detection method may fail: changes of illumination, occlusions, heads that are very close to each other (adjacent heads) and colour similarity (many non-objects can have similar colours to the hair colour categories).

For Exp 1, given the list above, in Table 2 we present an analysis of false alarm rate (FAR) and miss detection under different reasons. We observe that 100% of the false detection cases are due to some foreground regions having similar colours to either skin or hair colour (the black hair category). Miss detection, instead, is mainly caused by adjacent heads and/or head occlusions. At times miss detection is encountered due to combination of reasons such as colour similarity and adjacent heads. Table 2 reveals that in a surveillance video, it is unlikely to find occurrences of head occlusions as well as chances of adjacent heads but illumination plays an important role

Table 1. Hit rate (HR) and false alarm rate (FAR) for both experiments.

	Exp.1	Exp.2
No. of frames	252	379
Ground truth (total no. of heads)	778	803
Total no. of detected heads	684	354
Total no. of falsely detected heads	15	613
Average HR	79%	44%
Average FAR	2%	78%

Table 2. Analysis of false alarm rate (FAR) and miss detection under different reasons.

Reasons	FAR		Miss detection	
	Exp.1	Exp.2	Exp.1	Exp.2
Illumination	0%	57%	10%	100%
Occlusion	0%	0%	47%	0%
Adjacent heads	0%	0%	38%	0%
Colour similarity	100%	43%	5%	0%

Table 3. Comparison between our algorithm and standard face detection algorithm

Dataset	Frame	HR		FAR	
		Our	Haar	Our	Haar
AVSS multiple face	252	79%	83%	2%	12%
AVSS surveillance	379	45%	0%	70%	0%
ETISEO corridor	1240	70%	0%	50%	0%
ETISEO footpath	1122	62%	0%	25%	0%

Experiment 3 is used to demonstrate the differences between the head detector and face detector. Face detector is mainly used to detect frontal faces in high quality images, while head detector is used to detect heads in low quality videos for video surveillance purposes only. It is clearly shown in Table 3 that using a face detector for surveillance videos is inappropriate as it cannot detect any head due low resolution and abnormal illumination. Therefore, we find the need to apply a head detector which is only targeted for those environments. Despite the detection rate being not very high, the proposed algorithm is still perfectly suitable for video surveillance applications. It can be efficiently used for as anchor for a model-based human tracking that can tolerate frequent occlusions (e.g., [10]). The false alarm rate can also be significantly eliminated using spatial and temporal constraints such as a head location constraint, or various motion features However, enforcing those constraints is only possible if the previous history and the prediction of the head of all tracking targets in the current frame are available.

6. CONCLUSIONS

We presented a novel algorithm for human head detection in color images suitable for surveillance videos. We constructed and trained a novelty model for four categories of hair (brown, red, blond and black) and three categories of skin representing the different illumination conditions (bright, standard and dark). The shape constraint fits an elliptical model to the candidate region and compares its parameters with priors based on the human anatomy. The proposed head detection algorithm is used in a people tracking application for aligning targets' models with new observations. The hit rate is in the order of 45-79% across the experiments, allowing a tracker to detect a head of a tracked person every second or third frame. This is sufficient to maintain the target's model up to date and disambiguate even challenging data association. The false alarm rate is remarkably high, with a maximum of 70% in one scenario. However, such extra matches do not compromise the application since they can be pruned by the following stage of model-observation comparison.

7. REFERENCES

[1] AVSS 2007 dataset
http://www.elec.qmul.ac.uk/staffinfo/andrea/avss2007_d.html

[2] Colour FERET database
http://www.itl.nist.gov/iad/humanid/colorferet/home.html

[3] ETISEO dataset
http://www-sop.inria.fr/orion/ETISEO/

[4] M. Chen and S. Kee, "Head Tracking with Shape Modeling and Detection," *in* Proc. of the CRV, pp. 483 – 488, 2005

[5] W. Haiyuan, C. Qian, and M. Yachida, "Face detection from colour images using a fuzzy pattern matching method," IEEE Trans.on PAMI, vol. 21, no. 6, pp. 557-563, 1999.

[6] R. L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in colour images," *IEEE Trans. on PAMI*, vol. 24, no. 5, pp. 696-706, 2002.

[7] Y. Ishii, H. Hongo, K. Yamamoto and Y. Niwa, "Face and Head Detection for a Real-Time Surveillance System*," in Proc. of ICPR, vol. 3, pp. 298-301, 2004.*

[8] S.J. Krotosky, S.Y. Cheng, and M.M. Trivedi, "Real-time stereo-based head detection using size, shape and disparity constraints," in Proc. of IEEE IVS, pp.550-556, 2005.

[9] R. Lienhart and J. Maydt. "An Extended Set of Haar-like Features for Rapid Object Detection", in Proc. of IEEE ICIP, *Vol.* 1, pp. 900-903, 2002.

[10] Z. Zhang, H. Gunes, and M. Piccardi, "Tracking People in Crowds by a Part Matching Approach", Proc. of the IEEE AVSS, pp. 88-95, 2008.

[11] Z. Zhang and M. Piccardi, "A Review of Tracking Methods under Occlusions", Proc. of the IAPR Conf. on MVA, pp. 146-149, 2007.

© [2009] IEEE. Reprinted, with permission, from [Zui Zhang; Gunes, H.; Piccardi, M., Head detection for video surveillance based on categorical hair and skin colour models, 16th IEEE International Conference on Image Processing (ICIP), 2009]. This material is posted here with permission of the IEEE. Such permission of the IEEE does not in any way imply IEEE endorsement of any of the University of Technology, Sydney's products or services. Internal or personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by writing to pubs-permissions@ieee.org. By choosing to view this document, you agree to all provisions of the copyright laws protecting it