Real-Time Skin Labelling in Active Camera Images

Paul Chippendale
ITC-irst, SSI Division. Via Sommarive, 18. 38050 POVO (TN)
Chippendale@itc.it

Abstract. A real-time method for deriving accurate body-part labels and basic
human posture from monocular, active camera image sequences is presented.
Bootstrapping parameters, such as human clothing models or a background
model of the environment are not required. The ultimate goal of the research is
to track human gestures using active PTZ cameras and hence the human seg-
mentation must operate in real-time. Since such cameras are free to pan, tilt and
zoom, a background model of the scene is not available. By fusing the outputs
from various temporal filters and a statistical skin colour model, and then ap-
plying novel spatial and temporal skin linkage stages, the system is capable of
reliably associating and labelling hands and heads, as well as providing clues as
to the basic forms of multiple people within its field of view after just a few
seconds of observations.

1 Introduction

The tracking of human body parts from video sequences has received a great deal of
attention in recent years [1][2][3][4], with the majority of work being conducted us-
ing static cameras in controlled environments. However, for applications such as the
labelling of body features and moreover to understand certain gestures created from
their temporal displacements, there are inherent minimum observable size constraints;
tiny movements, potentially providing valuable information, could otherwise go un-
detected. To meet this visual coverage requirement effectively, a room has either to
be filled with high resolution static cameras, or active cameras must be employed and
controlled in an automatic way so as to track individuals. However, the problems
associated with an optimum selection of pan, tilt and zoom parameters are not trivial;
if there are multiple moving objects in the viewing space, or an extreme close-up of a
single face is required, camera deployment is complex. This paper explores a means
of analysing human presence within image streams produced from active cameras that
are often re-orientated.

The system introduced in this paper fuses the results from a series of temporal fil-
ters and a statistical skin filter; a spatial and temporal skin-to-skin linkage association
stage then follows and thus human activity within the scene can be deduced. As the
system is intended for the surveillance of humans, it is essential that all of the algo-
rithms operate in real-time.
2 Background

Body tracking from a monocular sequence is an extremely challenging task due to the high number of degrees of freedom present in the articulated body structure, some of which are unobservable (due to the ambiguity inherent in the 3D to 2D projection). Furthermore, the complexity of the indoor environment makes the accurate labelling of body parts even more challenging.

At present, there are two fundamental approaches that can be adopted to locate and track people: contour-based and articulated model-based [1][2]. Although articulated model-based approaches, which often use motion models, can achieve more robust and accurate results and are able to recover high-level information such as joint angles, their high level of complexity restricts their deployment in real-time situations. Conversely, contour-based approaches can be very fast and are able to operate in real-time, since only regional information is considered. The contour-based methods used in Ghost [3] utilise background subtraction to create a silhouette that is then used to determine the location of body parts in generic postures. It combines hierarchical body pose estimation, a convex hull analysis, and a partial mapping from the body parts onto the silhouette using a distance transform method. Although this method works well for a limited posture dataset, its deployment is very restrictive and it requires the creation of a good silhouette to function effectively. An improvement on the Ghost approach is presented in [4]. Here, a method for integrating appearance-based body part labelling is presented which temporally combines the posture estimates of Ghost by determining a best path through a posture-time lattice. Although their system recognises human postures more robustly and with a greater flexibility, background subtraction is still required.

Efforts have been made to create panoramic background models synthesized from mosaics of images taken from a large variety of pan and tilt configurations [5]. This technique considers only static, non-moving pixels when creating the background model, it is not necessary for the scene to be empty during the modelling phase. Although this approach can create a complete background, the issue of camera zooming is not addressed.

In [6], another contour-based approach is investigated employing motion, colour and head-template matching in combination to provide a means of detecting humans present in an indoor meeting room type environment. In order to enable the rapid detection of people in a large mosaic image (created from several cameras mounted back-to-back in a star configuration), frame differencing is employed to extract small sub-regions likely to contain moving entities. Frontal faces are searched for and tracked in the full image using mean shift based colour tracking. The head-shape template used for matching is a simple Ω shape and, when combined with a human skin colour model, confidence values of possible matches are returned.

Temporal differencing and template matching are employed in [7] to locate and track vehicles as well as pedestrians. The targets appear small in the image, and their relative pixel speed compared to the camera’s frame rate is high, thus the target’s displacement frame-to-frame is large, minimising self-occlusion. However, this is not the case for indoor applications involving human tracking, where the goal is to ideally
make the target as large as possible in the field of view. Moreover, the target’s location is often near-stationary resulting in temporal self-occlusion.

Skin colour segmentation can also be employed for human tracking [8] [9]. Skin models can provide clues as to where a human face or hands may be, but without a frontal camera aspect, assigning reliable labels to the hands and face is prone to errors. If an open hand, for example, is raised above the head, its pixel area can be larger than the face; as it is also the highest skin region in the image, it is generally incorrectly assumed to be the head. Hand and head identification is made even more difficult when there are multiple people in the room with the further ambiguity of which hand should be associated with which face.

One of the standard approaches employed to associate separate skin regions to the correct individuals uses background subtraction, creating a discrete blob or silhouette for each person. In this way, skin blobs that lie within the same silhouette belong to the same person by default. Again, this approach can be limiting as it is often the case that from a camera’s point of view, one person may obscure another and consequently create an ambiguous, homogeneous blob. As a model for the background is not available with active cameras, we had to seek a different approach.

3 Human Segmentation and Labelling

3.1 Activity Filters

The basic assumption in this research is that motion potentially indicates the presence of a human. Such regions are then substantiated by locating skin-coloured pixels in their near vicinity. Therefore, moving skin-coloured pixels which are of sufficient density to form cohesive blobs are assumed to be moving heads or hands within the image. Skin blobs are verified as belonging to the same entity through spatial linkages rather than through template matching to determine whether a skin blob actually resembles a face or a hand.

The filters, used for detecting the presence of human activity within the camera’s field of view, run in parallel and are applied every frame. Temporal inconsistency, applying a two-frame differencing algorithm to the luminance component, is one of the simplest cues that can be utilised to detect human presence (see Fig. 1).

Given that in practice even stationary people move a little, only a few centimetres of movement are required to reveal a telltale movement edge around the

Fig. 1 Two-frame difference image; intensity indicates the level of difference
target. As the output from the temporal filter is noisy, it is thresholded and then filtered to remove isolated pixels.

Inter-frame motion is insufficient to indicate the presence of humans all of the time. When a person remains stationary, the detectable motion is greatly reduced making segmentation very difficult; hence additional complementary filters are required. The light grey trace in Fig. 2 shows the percentage of temporal difference pixels on a frame-by-frame basis over a test video sequence*. It can clearly be seen that there are several periods, some over 50 frames in length, where there is negligible temporal motion.

![Fig. 2 Pixel activity of various filters during the test sequence](image)

To compensate for these detection troughs, a form of Motion History Image (MHI) filter [10] is employed, in which moving objects are slowly absorbed into a background model after only a few seconds. After initially taking the first frame of a sequence as a reference, depending on a pixel’s value in subsequent frames, a given step is either added or subtracted to the corresponding pixel in the background model. To provide an estimate of the ‘foreground’ for each frame, the current frame and model are then compared. Fig. 3a illustrates one of the typical background model images, while Fig. 3b shows the resulting extracted foreground.

By referring back to Fig. 2, the dashed trace (labelled ‘Foreground’) shows that the percentage of active pixels in the test sequence is a more consistent indicator than the temporal filter; it does not decay as quickly. This hysteresis however has a price; there is now a wake following any activity due to the absorption of the person into the background (see the area behind the head in Fig. 3b). The background model recovery time once the area is devoid of human presence is also delayed. This residue decays quickly and its impact is minor as wake pixels have low confidence.

---

* The test video sequence is 1800 frames in length, recorded at 15fps with a resolution of 320x240 and is JPEG compressed. A single person was asked to move around the room stopping from time-to-time to explore and point at various objects.
The final filter used to create the activity map (see Fig. 6) is a simple statistical skin model. This model, using the Cb and Cr colour channels (as in [11] and [12]), was created through observation and manual segmentation of hundreds of skin regions from our camera images. The YCbCr colour space was chosen for two reasons: firstly because it is inherent to the JPEG video camera codecs (hence colour space transforms are not required); and secondly, because the Cb and Cr channels provide luminance independence. To create the statistical model, the Cb and Cr values of all of the manually segmented skin-pixels were plotted on the same graph forming a CbCr cluster. Then an ellipse was fitted over the cluster as in [12]. The cluster is thus represented by a line segment representing its kernel and a radius representing its size. The skin likelihood of a pixel is then calculated by taking the Euclidean distance from a pixel’s Cb and Cr value to the segment and normalising this by a factor relating to the radius and the desired output bit-depth (here eight-bits). E.g. a skin likelihood of 255 is given if a pixel’s Cb and Cr values lie on the segment, 127 if the CbCr values lie on the ellipse, and 0 if they lie further than twice the radius away from the segment. The application of this filter to one of the test sequence frames can be seen in Fig. 4.
As can be seen, Fig. 4 is very noisy. To obtain the image in Fig. 5, pixels are only kept if they meet one of the following conditions:

1. Their likelihood is above 127 (i.e. inside the ellipse model) and they lie closer than 5 pixels to either temporal or foreground pixels
2. Their likelihood is above 127 and there has been qualified skin activity nearby in the past half a second.

Skin pixel activity for the test sequence is shown in Fig. 2 by the dotted line. As can be seen, motion qualified skin is consistently detected over the entire sequence, it is thus a good indicator of human presence.

The final stage of activity detection is to fuse the above three filters into a composite image which offers as much information as possible as to where a person could be. This fusion of the filters is achieved by normalisation and addition. Fig. 6 shows the result of this fusion, where the brightness of each pixel indicates its likelihood of belonging to a person.

The remaining information required to pass to the skin blob linkage stage are co-ordinates of the centroids of the skin regions seen in Fig. 5. These are obtained quite easily through the application of a morphological operator to the image to form homogeneous blob regions, which then have their centre of masses computed.

### 3.2 Spatial Skin Blob Association

For each frame containing human movement an activity map and a list of skin blob centroids are created. If this list contains two or more centroids, the activity map is explored to discover the best linkage between blob pairs. The maximum number of linkage searches, $S$, per image is expressed in Equation (1), where $n$ is the number of blobs.

$$ S = \frac{n!}{(n-2)!} $$

However as this search is exhaustive, for example if there are 3 people each showing 3 skin blobs then 36 linkage tests would have to be conducted, large numbers of tests could cause problems for real-time execution. Consequently, to make this exploration tractable, a blob-pair is only explored if it is likely to belong to the same person. To determine whether this is the case, a crude silhouette is created by temporarily superimposing activity maps using a morphological operator and a MHI filter. For each frame the activity map is dilated and then added to the constantly decaying MHI of activity.
The rate of decay from an active state to an inactive state for a MHI pixel is in the order of a second, bringing rise to a considerable temporal wake. Fig. 7 shows a snapshot of the activity map MHI, after thresholding, with the skin blob centroids superimposed. Where two or more centroids lie inside the same activity blob, spatial linkages are explored. Of course, if two people are close to each other in the image then a homogeneous blob is formed. However this only results in more blob pair linkage tests and does not imply that the skin blobs actually belong to the same person.

The human model used for the purposes of understanding skin blob linkages spatially is a very simple one. Hands are connected to faces via an elliptical arc travelling through the arms and neck. Its major diameter is defined by the line connecting the two centroids (or nodes) and the minor radius is defined by the perpendicular distance from the major’s bisect to the elbow. Fig. 8 shows the range of minor radii that are explored for a given blob pair in order to locate the most probable/strongest link. In order to minimise the execution time of the spatial linkage exploration, the number of minor radii possibilities explored is reduced to a reasonable level. A step of 3 pixels in between arc radii was empirically selected as a good compromise.

The strength, \( S \), of an arc linkage is calculated by looking at the pixel values from the activity map at the coordinates of each point along the arc. Thus the intensity of each pixel and the number of pixels on the arc are used in the calculation. Equation (2) shows how the strength of each test linkage arc is calculated, where \( l_j \) represents the luminance value of a list of corresponding pixels from the activity map that lie underneath the prescribed arc of length \( n \), and \( T=2^{b-1} \), where \( b \) is the luminance bit depth.

\[
S = \frac{\sum_{j=0}^{n} l_j/T}{n}
\]
The results obtained from applying the spatial linkage tests to the three centroid combinations are shown in Fig. 9. As a quantised radius size is employed, the strength curves often exhibit spikes that can distort linkage interpretation. Therefore, a sliding mean filter is passed over the data series before further analysis. From Equation (1), we know that three linkage combinations are required for a three skin blob configuration. In our case these linkages are: left arm to head, right arm to head and left arm to right arm. The x-axis in Fig. 9 represents radius length and the y-axis indicates the strength of the linkage from Equation (2).

![Fig. 9 Spatial linkage curves from frame 457 of the test sequence](image)

The strongest radius is selected from each pairing through statistical analysis of the resulting curve. Firstly, the weakest linkage is selected from either the most negative or positive radius sizes. This weakest linkage provides an indication as to which side of the zero-radius line the human lies (i.e. the higher the strength the more likely it is that the body of the person is present). From the lowest strength, the first substantial peak, whilst moving towards the centre of the graph, is located. This peak signifies the strongest possible linkage, but is only validated if all of the following conditions are met:

1. The gradient over the previous 10 pixels of minor radius is steeper 10%,
2. The best linkage radius strength is more that twice that of the weakest,
3. The best linkage strength is greater than $T$.

In our example, looking at the left arm to head curve in Fig. 9 shows us that the weakest is at +42 pixels with a strength of 21 and the first peak is reached at +3 with a strength of 217, therefore the strongest linkage between the hand and the left arm can be expressed (+3, 217). The right arm to head reports (-12, 230). However, as all of the above criteria are not met for the left to right arm curve (as the gradient is too shallow), no strongest match is forthcoming. The final step of the spatial linkage stage deduces the strongest linkage from each centroid. This entails examining all of the successful linkage tests from the exploratory pairings and keeping the strongest for each node. For the test case in Fig. 9 the best linkages are:

- With left arm inclusion: (+3, 217) going to the head,
- With right arm inclusion: (-12, 230) going to the head,
With head inclusion: (-12, 230) going to the right arm.

As is apparent, the head is the common node for both left and right hands. This node commonality is the indicator employed for recognising which blobs are heads and which are hands. However, a single snapshot in time is insufficient to guarantee correct labelling, hence a stage of temporal linkage examination follows.

3.3 Temporal Skin Blob Association

The output from the spatial linkage stage consists of a list of skin centroids and each centroid’s strongest linkage to another. To understand what is happening in the current scene, a tri-plane MHI (containing hand, arm and face planes) is created from the stream of spatial linkages. The assimilation of this stream is referred to as the temporal linkage stage, and attempts to provide not only an idea of human presence and orientation, but also which blobs are likely to be heads and which hands are linked to them.

A hypothesis as to whether a skin blob is a hand, face or something inconclusive is made upon the observation of ‘Λ’ shape patterns in localised areas of the arm-plane MHI. In each frame, the contents of the arm-plane MHI are updated by incrementing the values of pixels that lay under spatial linkages and decrementing those that do not. Each spatial linkage is represented as an ellipse connecting the two skin blob centroids, with the minor radius determined as a ratio of the major.

Over time, pixels in the arm-plane MHI which often lay underneath spatial linkages will strengthen, revealing the distinctive ‘Λ’ shape patterns, as illustrated in Fig. 10 by green ellipses. A skin centroid from the current frame is added to the head plane as a small vertical ellipse, if it is connected via a spatial linkage to two others that connect to it exclusively. In this case, the two other nodes in the group are designated as hands and are added to the hands plane of the MHI. Fig. 10 shows the results after a few frames of such a validation; the large green ellipses show the arm position history, the yellow circles (actually red circles over green arms) indicate the hand history, and the...
blue oval (mainly cyan due to the overlaying of the green arms) shows the head history.

In order to maintain a good estimate of head and hand labelling, even under less than ideal conditions, each of the skin centroids is tracked by searching recent detections and either continuing an existing track, or starting a new one where appropriate (Fig. 11 shows an example of three skin blob tracks). In this way, once a skin blob has been labelled as a face, it will be tracked as a face even if the ‘Λ’ shape is not observed for a few seconds. However, to ensure that incorrect labels are not propagated, the assumption that a current blob track is a face deteriorates over time if no new positive observations are made.

4 Results

To evaluate the performance of the system, the aforementioned test sequence was manually labelled. For each frame, the skin centroid positions were noted and labelled ‘hand’ or ‘face’. The automatic labelling system processed the image sequence and reported the position and designation for each skin blob recognised. Fig. 12 illustrates system performance compared to ground truth data. For clarity, the tracks show only the x coordinate of the face. Similar results were also obtained for the y coordinate of the face, and for the hands.

![Fig. 12 Face labelling performance for the test sequence](image)

As the graph above shows, the correlation between the automatic labelling and the ground truth labels is excellent. As expected, there is a delay of a few frames at the start after the head appears; this is due to the inertia of the temporal linkage MHI. There are also a few false head detections in the sequence. These are the results of disjoint skin blobs being formed during periods of almost zero motion, resulting in momentary false ‘Λ’ shapes being created. Once the tracking algorithm is included, the detection of the head is even better. The brief loss of tracking from frame 280 to
300 is down to a few seconds of near-zero movement by the protagonist, resulting in the absence of motion-verified skin blobs. We intend to solve this by adding further premises into the tracking algorithm (e.g. if there was a face at a given location, then this face remains likely despite lack of motion. This would result in tracking-verified skin blobs as well as motion-verified).

Fig. 13 and Fig. 14 show results from a couple of frames taken from two different test sequences. Red ovals on the two images represent labelled faces, the red squares illustrate labelled hands and the connecting green arcs show an indication of arm location.

5 Conclusions

The integration of simple and efficient filters combined with novel spatial and temporal skin linkage stages enables system execution at speeds in excess of 15fps on a 3GHz Pentium4 PC. At such frame rates, accurate labelling is possible after only a second of stable human observation. Thus, once integrated into a 3D tracking system, any active camera orientation changes only result in the loss of a few frames of actual observation results.

One of the main shortcomings in the current scheme is the skin model. It was not the focus of this research and at present is tailored to our cameras and the specific lighting condition in the laboratory. However, it would be a simple task to utilise a more generic skin model in our automatic labelling system; all that the system requires is a skin likelihood value for each pixel.

The results in this paper have shown that heads and hands can be accurately identified without prior knowledge of the environment. Instances of head-label loss or false positives will be addressed in future work, with the integration of further premises into the tracking stage.
References