Constructing Face Image Logs that are Both Complete and Concise

Adam Fourney, and Dr. Robert Laganire
School of Information Technology and Engineering
University of Ottawa
Ottawa, Ontario, K1N 6N5, Canada
{afour005, laganier}@uottawa.ca

Abstract

This paper describes a construct that we call a face image log. Face image logs are collections of time stamped images representing faces detected in surveillance videos. The techniques demonstrated in this paper strive to construct face image logs that are complete and concise in the sense that the logs contain only the best images available for each individual observed. We begin by describing how to assess and compare the quality of face images. We then illustrate a robust method for selecting high quality images. This selection process takes into consideration the limitations inherent in existing face detection and person tracking techniques. Experimental results demonstrate that face logs constructed in this manner generally contain fewer than 5% of all detected faces, yet these faces are of high quality, and they represent all individuals detected in the video sequence.

1 Introduction

It is often useful to summarize surveillance video by collecting the images of faces that are visible in the original video sequence. This paper will refer to these collections as face image logs, and we assume that the face images are time stamped. Whether reviewed by security personnel, or an automated system; processed in real time, or upon request; these logs allow investigators to determine who was in the vicinity of the surveillance camera at any particular moment in time.

In order to be useful, face image logs should be complete in the sense that they contain, at the very least, one high quality image for each individual whose face appeared unobstructed in the original video. High quality images are important because they maximize the probability that individuals will be correctly identified. We would also like the face image logs to be concise. By concise, we mean that the log need not contain every instance of every face; usually only one high quality image is required to identify an individual. There are at least two approaches to constructing complete logs, and they are described in the sections that follow.

1.1 The brute force approach to face log construction

The most direct approach to constructing complete face image logs involves using existing face detection technologies to extract face images directly from video sequences. Simply stated, a traditional face detector is applied to the individual video frames, and all detections are immediately appended to the log. In this scenario, one face may be detected per person per frame. Surveillance footage, captured at 15 frames per second, could potentially capture 900 face images per person per minute. The high rate of detections could easily overwhelm any human operator or automated biometric face recognition system that might be trying to process the face image log in real time. Real time or not, much of this processing is wasteful since the resulting logs are not concise; each individual may appear in the log hundreds of times.

1.2 Person oriented face log construction

In order to reduce the number of faces added to the face log, person oriented face log construction seeks to store one (or perhaps a few) high quality image(s) for each individual. To accomplish this, faces need to be tracked in addition to simply being detected. By doing so, a face image history can be compiled for each individual. When an individual leaves the scene, one or more high quality images are selected from his or her face history, and these images are added to the face log. The number of faces selected depends on the confidence in the face tracking results; more faces are selected when the tracker results are poor. By constructing person oriented face logs, it should be possible to avoid overwhelming face recognition systems, when attempting to process these logs in real time.

There are many uncertainties when developing a person oriented face logging system. For instance, it is not obvious how the quality of face images should be assessed and compared so that high quality images can be selected. It is also unclear how the face selection process should proceed in situations where there is low confidence in the tracker results. The approach presented in this paper attempts to resolve these difficulties.

1.3 Previous research

The proposed method of constructing face image logs can be divided into four major tasks: face detection, face tracking, face quality analysis, and the selection of high quality face images to be appended to the face image log. Regarding the detection of faces for the purpose of tracking, many researchers [4] [9] [5] have suggested techniques that involve the use of skin color segmentation to locate candidate face regions. Others [1] [3], have detected frontal faces using cascades of simple classifiers. Some strategies for tracking the resulting face regions include techniques based on the overlapping of bounding boxes [4] [9], techniques using partial Kalman filtering for motion prediction [4], and techniques which use the mean shift method for tracking [5]. While the topics discussed by this paper do not assume the use of any one particular face detection or tracking strategy, our results were obtained using cascading Haar classifiers (as implemented in the Intel ®OpenCV library) to detect frontal faces. Faces were then tracked using a strategy very similar to that described in [9].

Unfortunately, few publications in the literature describe methods for assessing the quality of face images. The most relevant work in this area seems to be a face image validation system described by Subasic, Loncaric, Petkovic *et al.* in [8]. This system analyses images of faces and determines if they are suitable for use in identification documents such as passports. While their technique did provide a numeric quality score for input images, their focus was mainly on detecting images that do not meet the criteria established by the International Civil Aviation Organization; their decisions were inherently binary. In contrast, a more continuous appraisal system is required when selecting high quality face images from video sequences in which there may be vast variations in face resolution, pose, illumination and sharpness.

Finally, to the best of our knowledge, no existing literature adequately discusses the problem of combining face tracking and quality appraisal for the purpose of selecting high quality faces from video sequences. Consequently, we believe that we are presenting a novel approach to the problem of face log construction.

2 Assessing the quality of face images

In order to select high quality face images, it is necessary to develop a procedure for their appraisal. Of course, the quality of a face image is rather subjective. In this paper, face image quality corresponds roughly to an image's potential to lead to a correct identification when using existing face recognition software. It is assumed that any image useful for biometric face recognition would also be useful when manually attempting to identify an individual.

Many criteria, weighted to varying degrees of importance, are considered when determining the quality of a face image. These criteria include: resolution, pose, illumination, skin content and sharpness. The following sections describe how each of these criteria can be measured, and how they contribute to the overall quality score of an image. With the exception of skin detection (which requires color images), each of the following scoring procedures expects a grayscale image of an upright face as input. It is important to note that, depending on the method by which faces are detected and tracked, many of the intermediate results used for assessing quality may be readily available. For example, the location of the eyes, the contour of the face, and the location of skin colored pixels may have been computed when the face tracker was locating candidate face regions. Additionally, it is also important to recognize that the scoring procedures do not directly distinguish between face and non-face images (except for in certain extreme situations) since this is precisely what the face tracker is expected to provide. In general, however, high quality face images should out perform non-face images.

2.1 Pose estimation in face images

One of the most challenging problems encountered by face recognition systems involves properly identifying individuals despite variations in pose [10]. Generally, faces can experience out-of-plane rotations of up to 60^o before important facial features are no longer visible [10]. [1] It is expected that a robust face detector should return face images whose pose falls within the slightly narrower range of $\pm 45^o$. Consequently, it is important that the quality score be able to distinguish between various rotations, and award higher scores to the least rotated images.

In order to estimate pose, we begin by locating three columns on the face image: the first two columns, defined by $x = l$ and $x = r$, estimate the locations of the left and right visible edges of the face, respectively. The third column, $x = c$, approximates the face's axis of natural horizontal symmetry. Importantly, the face image is symmetric about the line $x = c$, only when the face is not rotated.

[1] As usual, an out-of-plane rotation of 0^o occurs when an individual looks directly at the camera, providing a perfectly frontal view of the face.

The values of l, r and c are determined using a technique developed by Kun Peng, Liming Chen, Su Ruan *et al.* [6]. This technique involves an analysis of the gradient image in order to locate the left and right sides of the face, as well as the vertical location of the eyes. From these values the approximate location of the eyes can be estimated, and the brightest point between the eyes is expected to lie on the face's axis of symmetry. Unfortunately, this method is not effective when subjects are wearing glasses, or when faces are not upright.

If the face has not experienced an out-of-plane rotation, and the values l, r, and c are accurate, then it is expected that c is equidistant from l and r. As a face experiences rotation, c deviates from its expected position c^*. The following equation is used to estimate the angle of rotation:

$$\theta = \begin{cases} 90^o & \text{if } \left|\frac{2(c-c^*)}{r-l}\right| > l \\ \frac{180^o}{\pi} \sin^{-1}\left(\frac{2(c-c^*)}{r-l}\right) & \text{otherwise} \end{cases} \tag{1}$$

This is not a particularly accurate estimator because it models the human head as a cylinder. However, high accuracy is not needed since θ will not be used in any computations other than the following equation:

$$S_1 = \begin{cases} 0 & \text{if } \theta > 45^o \\ 1 - \left|\frac{\theta}{45^o}\right| & \text{otherwise} \end{cases} \tag{2}$$

where S_1 represents the quality score awarded to the image for pose. This score has the desirable property that it decreases linearly as the estimated angle of rotation increases. Notice that values of θ greater than 45^o are considered to be inaccurate, and result in a score of zero.

2.2 Measuring the quality of illumination

Variations caused by changes in illumination constitute yet another significant challenge encountered by automated face recognition systems [10]. In fact, research by Adini, Moses, and Ullman, has shown that certain methods of face identification are more sensitive to differences in lighting than they are to the differences between distinct individuals [10]. In order to compensate for different lighting conditions, face identification systems may use histogram equalization, or similar histogram dependent techniques, in order to normalize an image before processing. For this reason, it is very important to begin with images which make the best (maximum) use of the available dynamic range. This utilization U is estimated by determining the smallest range of gray intensities to which at least 95% of an image's pixels can be attributed. The score S_2 is simply the percentage of the total dynamic range represented by U. For example, if the input is an 8-bit grayscale image, then $S_2 = U/256$.

Utilization of available dynamic range is not the only desirable property of properly illuminated face images; we would also like faces to be evenly lit. In other words, one side of the face should not appear brighter than the other. Using the l, r, and c measurements obtained when estimating pose, the evenness of the illumination can be determined by comparing the histograms of the opposing halves of the face. Let L and R be the grey intensity histograms attributed to the left and right halves of the face, normalized so that the integrals over their respective bins are both equal to one. The score S_3, which represents the evenness of the illumination, is then equal to the integral of the histrogram resulting from the intersection $L \cap R$.

2.3 Determining the sharpness of an image

It is important that a face image's score reflect its sharpness. Images marred by motion blur, for example, should score quite poorly. The main assumption when measuring the sharpness of face images is that these images should have comparable power spectra. Consequently, a simple global measure of sharpness can be used. Such a measure was described by Doron Shaked and Ingeborg Tastl in [7], and it proceeds as follows:

For a given image represented by the function $a(x, y)$, let $A(u, v)$ be its corresponding frequency domain representation. Given the frequencies f_1 and f_2, where $f_1 < f_2$, define:

$$\begin{aligned} H &= \{(u,v) \mid \|(u,v)\|_2 > f_2,\ (u,v) \in A\} \\ L &= \{(u,v) \mid f_1 < \|(u,v)\|_2 \leq f_2,\ (u,v) \in A\} \end{aligned}$$

$$Sh = \frac{\int_{(u,v)\in H} |A(u,v)|^2}{\int_{(u,v)\in L} |A(u,v)|^2} \tag{3}$$

where Sh measures the global sharpness of the image. While this measure provided excellent results (when using $f_1 = 2$ and $f_2 = 8$), we found it useful to further refine the procedure by omitting the frequency terms that occur at orientations within 10^o of the u or v-axis. This effectively masks out the unnatural step function that occurs at the boundaries of the image as a result of the cyclic nature of the Fourier transform. In practice, the frequencies f_1 and f_2 are chosen so that it would be highly improbable for a natural face to achieve a sharpness score greater than 1. Thus, an image's official sharpness score is defined as:

$$S_4 = min\,\{1, Sh\} \tag{4}$$

2.4 Detecting the presence of human skin

Images of faces are expected to be composed mostly of flesh-toned pixels. Lack of such pixels could indicate that an image's color temperature is not properly balanced, or that the color has been washed out due to overly harsh lighting conditions. In either case, such images should score poorly in overall quality. Research conducted by Margaret Fleck, David Forsyth and Chris Bregler reveals that pixels representing flesh tones are tightly clustered in a small region of hue-saturation color space [2]. Results for this paper were obtained using the region containing all hues between -30^o and 30^o, saturated between 5% and 95%. An image's skin score S_5 is computed as the percentage of its pixels that occur within this region. This measure is most useful when skin color segmentation is not already being used for locating candidate face regions.

2.5 Image resolution

An image's resolution score is perhaps the easiest of the aforementioned criteria to measure. The resolution of a face image is defined to be the area of its bounding rectangle. In general, high resolution images are preferred over low resolution images. This trend is accurate to a limit, beyond which it becomes no longer useful to achieve higher resolutions. In our case, this limit was 60×60 pixels. Consequently, the resolution score of an image having dimensions $w \times h$, is computed as follows:

$$S_6 = \min\left\{1, \frac{\sqrt{wh}}{60}\right\} \tag{5}$$

2.6 Combining the criteria into a general score

Each of the six criteria discussed in the previous sections can score in the range $[0, 1]$, but these criteria need not contribute equally to an image's overall quality score. For this reason, they are combined according to the following weighted sum

$$S = \frac{\sum_{i=1}^{6} W_i S_i}{\sum_{i=1}^{6} W_i} \tag{6}$$

where the coefficients W_i determine the impact each of the quality criteria have on the final score. Currently, these coefficients have been chosen using a process of trial and error and seem to yield scores consistent with our expectations of image quality. Their values are summarized in table 1.

Unfortunately, when using a weighted sum alone, it becomes difficult to develop a single set of weights that works both when an image scores moderately well in all criteria, as well as when an image scores in the extremes of one or more criteria. In an attempt to resolve this problem, a threshold T_i is associated to each of the individual quality criteria. An image is determined to be useful for identification purposes, with respect to the i^{th} image criteria, exactly when $S_i > T_i$. For each criteria where this occurs, a value of 1.0 is added to the final score S. In this way, an image's score primarily reflects the number of criteria satisfied, while the original weighted sum is used to break ties between images satisfying an equal number of criteria.

3 Selecting high quality face images

With a reliable face tracking algorithm, at most one face should be detected per person per frame. In these cases, selecting the best face image for each individual is rather trivial; as each frame is processed, simply retain the highest quality face detected for the individual thus far. Unfortunately, there are many real world situations in which the face tracker's results are ambiguous. In such situations, care must be taken to ensure that all individuals are represented in the face image log.

Before addressing the aforementioned problem, it will be useful to introduce the concept of a person group. A person group (or more simply, a "group") is defined as a set of people whose detected faces cannot be distinguished by the face tracker [2]. Person groups occur naturally in face tracking schemes that measure the overlapping of bounding boxes in order to track individuals (such as in [9]) ; in any given frame, overly large bounding regions may contain or partially overlap numerous faces. In these cases, it is understood that the tracker is following a group of people rather than a single individual. More generally, face groups can be constructed in situations where the face tracker has indicated low confidence in its results. Grouping faces is a welcome alternative compared to the introduction of track-

i	Weight (W_i)	Threshold (T_i)
1	2	0.8
2	1	0.2
3	1	0.4
4	$70/3$	0.13
5	$10/7$	0.4
6	2	0.5

Table 1. Weights and thresholds used to combine the six quality criteria into a single overall score.

[2] Single individuals can be considered person groups of size one, and do not need special consideration.

ing errors.

For the purpose of face log construction, it is assumed that a group's membership remains constant throughout its existence. This assumption can be justified if a group is assigned a new identity whenever its membership is suspected to have changed. Sudden and significant changes in the dimensions of a group's bounding box might indicate that such an event has recently occurred. So too can persistent changes in the number of faces associated to the group on a frame by frame basis. Alternatively, some trackers may be able to detect merge and split events between moving objects, and these cues could be useful for detecting changes in group membership.

Additionally, the software is expected to maintain a face history for each person group observed by the face tracker. A face history is essentially a smaller face image log, containing all faces that were ever detected and associated to the group. As with the global face log, images in the face history should be labeled by the frame in which they originated. A face history is considered to be complete only after the person group has exited the scene.

In order to estimate a group's size, it is assumed that some frame exists in which the faces of all group members are clearly visible. Consequently, the group size can be estimated as the maximum number of faces ever detected and associated to the group in any single frame. This can be determined by searching a group's face history at the moment that the face history becomes complete. Call this maximum M.

Once a group's size can be estimated, it becomes possible to identify the historical frames in which all M group members were detected. From this subset of frames, one frame will best represent the group. This will be the frame with the best minimum quality face. Let Q be this quality score. All faces in the group's face history having a quality less than Q should be discarded, and any faces that remain should be added to the global face log.

Unfortunately, falsely detected faces can artificially inflate the estimated group size M. Since false positives are likely to score quite poorly in quality, the situation could result in a low quality threshold Q. With a low Q, many faces will be considered of sufficient quality to be added to the global face log. One might suspect that the group size has been overestimated when fewer than 10% of all frames in the group's face history contain M (or more) faces. In these situations, the estimated group size should be reduced to a level that satisfies this condition. Extreme caution must be employed in order to avoid underestimating the group size.

Another main concern is that a group's face history continues to grow until the group leaves the scene. If the group remains onscreen for an extended period of time, storing the face history could become prohibitively expensive. Thus, low quality face images may be purged whenever the face history's memory requirements approach a predetermined limit. Alternatively, low quality images may be purged on regular periodic intervals. In either case, the purge operation proceeds by using the same algorithm described above, exactly as if the group has recently exited the scene; however, the remaining faces are not added to the global log at this time. Care must be taken to ensure that the faces which are purged are still represented in any statistics used to estimate group size. Additionally, images from frames containing more than M faces are never discarded before the group has exited. This ensures that faces from these frames will be available in case the estimate of group size is increased in future invocations of the algorithm. By definition, fewer than 10% of all frames will contain faces that are kept for this precautionary measure.

4 Experimental results

The proposed system has been used to construct face image logs of numerous short video sequences. These sequences were captured at a rate of 30 frames per second, at a resolution of 640×480 pixels and a color depth of 24 bits per pixel. In one representative video, four participants were observed passing a stationary camera (sample frames from this video are presented in figure 1). In this video, a total of 221 faces were detected, of which four were eventually added to the face log (presented in figure 2). These four images were of excellent quality, and they represented all four of the video's participants. The results of the remaining six tests are summarized in table 2. In all cases, face logs contained few images when compared to the total number of images detected – yet, none of the logs were incomplete, by our definition.

The results summarized above do not tell the whole story, since they can not convey the quality of the images that found their way into the logs. In order to visualize the accuracy of face quality analysis, figure 3 presents a subset of the faces detected in the representative video described above. These faces were scored using our six quality criteria, and are listed from lowest quality to highest quality. While these results appear promising, it will be useful to conduct a more organized perceptual experiment to compare how automatic scoring compares to a manual approach. Such an experiment has not yet been performed.

In terms of efficiency, the processing of video sequences was conducted offline, at a leisurely rate of 7 frames per second on an Apple iMac ® (late 2006 model). This machine was equipped with a 2.16 GHz Intel Core 2 Duo ® processor and 1GB of random access memory. Despite the availability of two processing cores and 64-bit instructions, the software implementation used only a single thread along with 32-bit instructions. The majority of the processing time (99%) was spent on detecting and tracking faces.

	Detections			Log Contents		
Video	People	Faces	False Positives	People	Faces	False Positives
1	***4***	***221***	***0***	***4***	***4***	***0***
2	2	114	0	2	2	0
3	2	168	1	2	11	0
4	2	96	0	2	4	0
5	2	190	0	2	3	0
6	1	51	2	1	1	0
7	1	4	0	1	1	0

Table 2. Results from processing seven test video sequences. Emphasis has been placed on the results from the representative video which was discussed above.

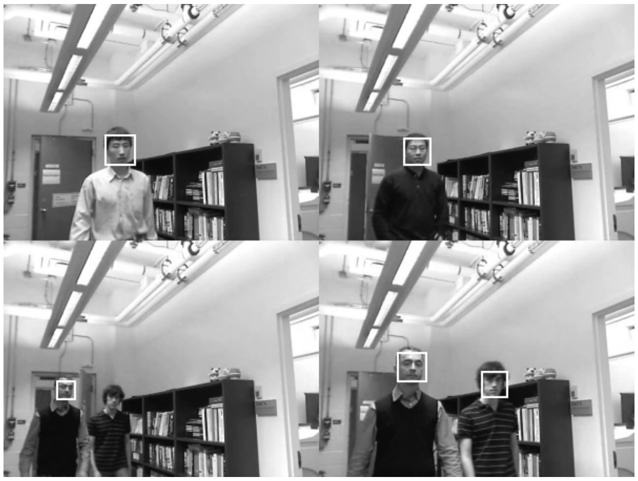

Figure 1. Sample frames from the representative test video sequence.

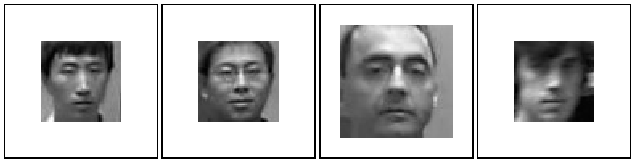

Figure 2. The contents of the face image log after the processing of the representative video.

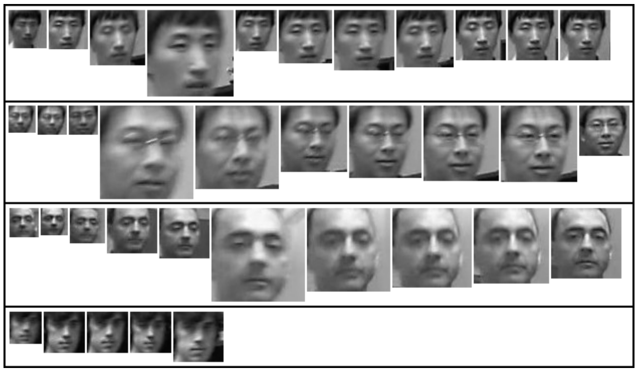

Figure 3. A sample of the face images detected in the representative test video sequence. Images are arranged from lowest quality to highest quality, when read from left to right.

As mentioned above, the face quality analysis and selection procedure required less than 1% of the total processing time. In order to test the throughput of these tasks alone, one test involved the processing of a video sequence in which the motion and locations of all faces were known in advance. The processing of this video achieved a rate of 30 frames per second. Consequently, we expect that an efficient face tracker may allow face logs to be constructed in near real time.

5 Conclusion

We have successfully demonstrated a method for constructing face image logs that are both complete and concise. This has been achieved by developing a measure for assessing the quality of face images detected from video, as well as a procedure for intelligently selecting faces to include in the final face log. Importantly, this selection process automatically adjusts so that more faces are logged in situations where a tracker's results are ambiguous. Finally, we have shown that quality appraisal and face image selection can be implemented efficiently, and these tasks do not add much overhead to the underlying face detection and tracking subsystems.

References

[1] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... buffy – automatic naming of characters in tv video. In *Proceedings of the British Machine Vision Conference*, 2006.

[2] M. Fleck, D. Forsyth, and C. Bregler. Finding naked people. In *ECCV (2)*, pages 593–602, 1996.

[3] L. Gagnon, F. Laliberté, S. Foucher, A. Branzan Albu, and D. Laurendeau. A system for tracking and recognizing pedestrian faces using a network of loosely coupled cameras. In *Visual Information Processing XV. Edited by Rahman, Zia-ur; Reichenbach, Stephen E.; Neifeld, Mark A.. Proceedings of the SPIE, Volume 6246, pp. (2006).*, jun 2006.

[4] V. Girondel, A. Caplier, and L. Bonnaud. Real time tracking of multiple persons by kalman filtering and face pursuit for multimedia applications. In *6th IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 201–205, 2004.

[5] C. Lerdsudwichai and M. Abdel-Mottaleb. Algorithm for multiple faces tracking. In *Proceedings of the International Conference on Multimedia and Expo*, volume 2, pages II 777–80, 2003.

[6] K. Peng, L. Chen, S. Ruan, and G. Kukharev. A robust algorithm for eye detection on gray intensity face without spectacles. *Journal of Computer Science and Technology*, 5(3):127–132, 2005.

[7] D. Shaked and I. Tastl. Sharpness measure: Towards automatic image enhancement. Technical Report HPL-2004-84R2, Hewlett-Packard, June 2005.

[8] M. Subasic, S. Loncaric, T. Petkovic, H. Bogunovic, and V. Krivec. Face image validation system. In *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, pages 30–33, 2005.

[9] H. X. Zhao and Y. S. Huang. Real-time multiple-person tracking system. In *Proceedings of the 4th International Symposium on Pattern Recognition*, volume 2, pages 897–900, 2002.

[10] W. Zhao and R. Chellappa, editors. *Face Processing: Advanced Modeling and Methods*. Academic Press, International, 2006.