STATISCAL-BASED SKIN CLASSIFIER FOR OMNI-DIRECTIONAL IMAGES

Asaf Shupo, Bill Kapralos and Miguel Vargas Martin
University of Ontario Institute of Technology
2000 Simcoe Street North, Oshawa, Ontario, Canada. L1H 7K4

Keywords: Skin detection, surveillance, omni-directional video sensing, maximum likelihood estimator.

Abstract: This paper describes the development of a simple, video-based system capable of efficiently detecting human skin in images captured with an omni-directional video sensor. The video sensor is used to provide a view of the entire visual hemisphere thereby providing multiple dynamic views of a scene. Color models of both skin and non-skin were constructed with images obtained with the omni-directional video sensor. Using a stochastic weak estimator coupled with a linear classifier, the system is capable of distinguishing omni-directional images that contain human skin from those that do not. Results indicate that the system is able to accomplish this task in a simple and computationally efficient manner. The ability to obtain an image of the entire scene from a single viewpoint using the omni-directional video sensor and determine whether the image contains human skin (e.g., one or more humans) in a simple and efficient manner is practical as a precursor for a number of applications including teleconferencing, remote learning, and video surveillance, the application of interest in this work.

1 INTRODUCTION

The area of human detection in the visual domain is rather large, well investigated, and has many practical applications including surveillance (Boult et al., 1998), video teleconferencing (Kapralos et al., 2003), and face detection (Herpers et al., 1999). A good, economical human detection/tracking system must be able to locate humans quickly and reliably in the presence of noise and other objects in the environment. It must run fast, efficiently (e.g., run in real-time), and operate using inexpensive camera and computer equipment (Bradski, 1998). Skin color is often proposed as an economical and efficient cue to detecting humans in the visual domain. Color is the simplest attribute in a set of pixels comprising the image (Jones and Rehg, 1998) and does not require extensive computational processing to compute. This allows a system utilizing color cues to operate in real-time. In addition, the color of an object may be used as an identifying feature that is local to the object and largely independent of the view and resolution. As a result, the use of color information may be used to detect objects from differing viewpoints (Swain and Ballard, 1991). In general, color cues are invariant to partial occlusion, rotation in depth, scale and resolution changes (Raja et al., 1998). Furthermore, there are various fast and simple color-based human detection and tracking systems available (Chai and Ngan, 1999; Chopra et al., 2006; Herpers et al., 1999; Jones and Rehg, 1998; Kapralos et al., 2003; Shupo et al., 2006; Stiefelhagen et al., 1999; Hans et al., 1999; Yang and Waibel, 1996).

Although skin detection is itself a simple and efficient process, many of the existing systems employ traditional cameras with a limited field of view. Using such cameras, in order to capture a view of the entire visual hemisphere, multiple stationary cameras may be used or a single camera may be panned to different directions. Furthermore, in various applications (such as video teleconferencing) the subject (person) may physically move into the view of a stationary camera. Both approaches can greatly increase the computational and time requirements potentially making these approaches impractical for real-time operations. Rather than having a user move into the camera's field of view, having multiple cameras or focusing the camera in different directions, an omni-

Shupo A., Kapralos B. and Vargas Martin M. (2007).
STATISCAL-BASED SKIN CLASSIFIER FOR OMNI-DIRECTIONAL IMAGES.
In *Proceedings of the Second International Conference on Computer Vision Theory and Applications*, pages 11-17
DOI: 10.5220/0002060800110017

directional video sensor (Cyclovision's ParaCamera system (Nayar, 1997; Baker and Nayar, 1999)) can be utilized instead. The ParaCamera captures a 360° (*hemispherical*) view from a single viewpoint.

In this paper we describe an approach that is used to classify ParaCamera images as either i) containing skin or ii) not containing skin. Color models of both skin and non-skin were constructed with images obtained with the ParaCamera. Using a stochastic estimator coupled with a linear classifier, results suggest the system is capable of distinguishing images that contain human skin from images that do not. This work is part of an ongoing research project investigating the fundamental issues related to the development of a video-based surveillance and monitoring system capable of locating humans within a scene. Humans within a scene may represent potential intruders and allows them to be automatically detected in an efficient manner can allow further, more complex actions to be taken. Such actions may include focusing a high resolution pan-tilt camera on to the potential intruder, and alerting a human operator.

The remainder of the paper is organized as follows. Section 2 provides greater details regarding the proposed system. In particular, further details regarding the ParaCamera and a detailed description of the statistical skin detection method is provided. Results of several experiments conducted to provide an indication of the effectiveness of method are provided in Section 3. Finally, concluding remarks and plans for future research are presented in Section 4.

2 THE APPROACH

2.1 ParaCamera Omni-Directional Camera System

Cyclovision's ParaCamera omni-directional camera consists of a high precision paraboloidal mirror and a combination of special purpose lenses (see Figure 1). By aiming a suitably equipped camera at the face of the paraboloidal mirror, the optics assembly permits the ParaCamera to capture a 360^o *hemispherical* view from a single viewpoint. Once the hemispherical view has been obtained, it may be easily un-warped (Peri and Nayar, 1997) producing a *panoramic* view. From this panoramic view, a *perspective* view of any size corresponding to portions of the scene can be easily extracted. An example of both hemispherical and panoramic images are illustrated in Figure 2. Examination of the images shown in Figure 2 illustrates the distortion and varying resolution inherent with images obtained using a ParaCamera. In the hemispherical view (see Figure 2(a)), the resolution decreases as we move away from the center of the image. Further examination also reveals that given the low resolution inherent in these images, fine details (e.g., facial features) cannot generally be detected. This can be a limitation with many of the existing computer vision/image processing algorithms currently available (e.g., face detection by locating both eyes). However, the low resolution inherent with ParaCamera images does not pose a problem for this application. As will

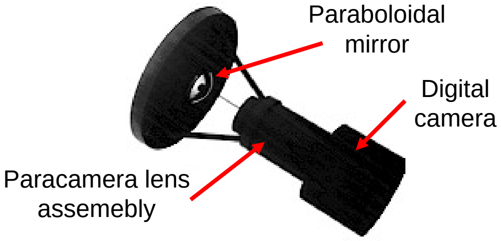
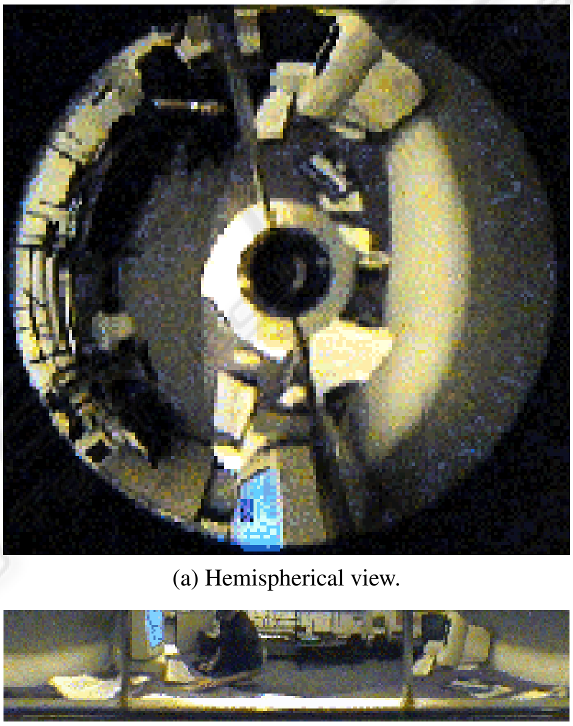


Figure 1: Cyclovision's ParaCamera omni-directional video sensor. The camera is aimed at the face of a paraboloidal mirror allowing it to capture a view of the entire visual hemisphere from a single viewpoint.


(a) Hemispherical view.

(b) Panoramic view

Figure 2: Sample ParaCamera images. (a) Hemispherical view. (b) A simple procedure allows the hemispherical view to be transformed into a panoramic view.

be described below, in this work the goal is to employ a ParaCamera to obtain a fast overview of the scene while flagging potential areas of human activity. Other higher resolution imaging sensors can then be focused on these potential areas of interest in order to provide further information.

The ability of an omni-directional video sensor such as the ParaCamera to capture an image of the entire visual hemisphere makes it very attractive for a variety of applications. Such applications include the capture the simultaneous video of each participant in a small group meeting (Stiefelhagen et al., 1999; Yong et al., 2001), surveillance (Boult et al., 1998; Gutchess et al., 2000), autonomous robot navigation (Zheng and Tsuji, 1992), virtual reality (Yasushi, 1999), telepresence (Yasushi, 1999), video-teleconferencing (Kapralos et al., 2003) and pipe inspection (Basu and Southwell, 1995).

2.2 Skin Classification

Skin detection is performed using an estimator for feature extraction coupled with a linear classifier. The estimators considered in this work are the *stochastic learning weak estimator* (SLWE) (Oommen and Rueda, 2006) and the *maximum likelihood estimator* (MLE) (Duda et al., 2000). The SLWE is considered to be more accurate in dealing with non-stationary data (e.g., a sequence of images capturing the motion of a subject), which is a relevant consideration when considering video surveillance. The classification process consists of two phases: i) the *training phase* and ii) the *testing phase*. In the training phase, two feature vectors are used to classify the *skin* and *non-skin* images (e.g., images that contain human skin regions and images that do not respectively). By extracting statistical properties from the labeled images (i.e., those that have been sorted into the skin and non-skin vectors), we are able to conduct the experiments using this as a basis of comparison. The features that were extracted from the training phase with any needed adjustments were input into the classifier that we used in the validation phase of the SLWE. The experiments were then repeated for the MLE (the estimators are discussed in the following sections). The training phase, as pointed out above, aims first to extract the statistical features of the images corresponding to all images in the training dataset, producing one feature vector for each class. The procedure shown in Algorithm 1 produces these two feature vectors when it is run for each dataset (e.g., skin and non-skin).

Algorithm 1 Procedure for determining the feature vectors.

```
Initialize an array B of counters to zero
For each image $I$ of the training dataset of class $j$:
    For each 8-bit byte $b_j$ of $I$:
        Increment $B[b_j]$ by 1
Initialize an array $V_j$ of probabilities to zero.
For $k = 0$ to 255
    Set $V_j[k] = B[b_j]$ / total number of 8-bit bytes of the set of images
```

The algorithm is used separately for the skin and non-skin training datasets. The output of the algorithm is a feature vector, an array $V_\circ$ or V_n, one for the skin and non-skin dataset, respectively. The appropriate statistical characteristics are initially extracted and the feature vectors $V_\circ$ and V_n are formed. The next step is to use an estimator to extract the features of the image to be classified, namely a vector V'. The classification rule consists of assigning an unlabeled package to the class, skin or non-skin, that minimizes the distance between V' and the trained arrays $V_\circ$ or V_n. Two metrics have been used for this purpose with both the ML and SLWE algorithms (described later in greater detail): i) the Euclidean distance, and ii) the weighted Euclidean distance. Both metrics are used to calculate the distance between the two labeled vectors, and once the distance has been determined, a classification is made as to whether an image contains skin or non-skin (e.g., the image is classified as being either "skin" or "non-skin"). Before using a classification metric, statistical characteristics of the datasets must be extracted. In the following sections, the MLE and SLWE used to extract such features are described in greater detail. For this purpose, we obtain the frequency of occurrence for each of the symbols (from 0 to 255) for a given image that has to be classified.

2.2.1 The Maximum Likelihood Estimator

The maximum likelihood estimator (MLE) is a traditional technique that aims to maximize the likelihood that a given sample generates using a specific probabilistic model, either parametric or non-parametric. We assume that we are dealing with a multinomial random variable with 256 possible realizations (one symbol for each 8-bit ASCII value). It has been shown that the likelihood is maximized when the estimate for each symbol is given by the frequency counters divided by the total number of bytes in the image (Duda et al., 2000) (see Algorithm 2).

The algorithm produces an array V' that contains the estimates for each (8-bit) byte in the testing image H. That vector V' is then input to the classification rule, which decides on the class based on a distance function and the trained feature vectors.

Algorithm 2 The maximum likelihood estimator.

```
For each image $H$ captured by the ParaCamera:
    Initialize an array $C$ of counters to zero
    For each 8-bit byte $b_j$ of $H$:
        Increment $C[b_j]$ by 1
Initialize an array $V'$ of probabilities to zero
For $k$ = 0 to 255
    Set $V'[k]$ = $C[b_j]$ / total number of 8-bit bytes of
    this image.
```

Algorithm 3 The stochastic learning weak estimator.

```
For each image $H$ captured by a ParaCamera:
    Initialize each entry of the feature
    vector $V'$ to 1/256
    For each 8-bit byte $b_j$ of H:
        For $k$ = 0 to 255
            If $i \neq bi$ then
                $V'[k] = \lambda \times V'[k]$
            Else
                $V'[b_i] = V'[b_i] + (1-\lambda) \sum_{k \neq i} V'[k]$
```

2.2.2 The Stochastic Learning Weak Estimator

Estimators like the one described by the MLE algorithm outlined in the previous section suffer from a lack of ability to capture quick changes in the distribution of the source data (e.g., dealing with non-stationary data, that is, data from different types of scenarios). Oommen and Rueda (2006) proposed a stochastic learning weak estimator (SLWE). The SLWE combined with a linear classifier has been successfully used to deal with problems that involve non-stationary data and has been effectively used to classify television news into business and sports news (Oommen and Rueda, 2006). In this work, each image to be classified is read from the testing dataset, and is used to feed the classification rule by means of extracting statistical features into a feature vector. The source alphabet contains n symbols ($n = 256$), which represent the possible realizations of a multinomial random variable, and whose estimates are to be updated by using the SLWE rules. While this rule requires a ”learning” parameter (λ), it has been found that a good value for multinomial scenarios should be close to 1 (e.g., $\lambda = 0.999$ (Oommen and Rueda, 2005)) (see Algorithm 3).

The classification rule is validated using labeled images and adjustments are made if necessary. Note that in the actual classification process the label of each image is not known. To classify the complete image, several distance metrics are employed. Greater details regarding these metrics are provided in the following section.

2.2.3 Distance Metrics

The choice of a distance function (also referred to as “metric”) is not a trivial task. Often, different components of the feature vectors may have different weights in classification of an arbitrary image. Some entries of the feature vector may be more important than other entries, or some entries may have more noise than other entries. Therefore, the choice of a metric plays an important role in the performance of the algorithm.

Euclidean Distance In this metric, it is assumed that all entries in the feature vector have equal weight. The Euclidean distance $d(V,V')$ between two feature vectors V and V' is defined as

$$d(V,V') = \sqrt{\sum_{i=0}^{255} (V[i] - V'[i])^2} \qquad (1)$$

Weighted Euclidean Distance This metric is also known as the Mahalanobis distance when the covariance matrix is considered as a diagonal matrix. It is assumed that different entries in the feature vector have different importance in the classification of images. It is also assumed that an entry in the feature vector is of less importance than another entry if its variance is greater than the variance of another entry. The weighting factor w is defined as $w = 1/\sigma^2$, and the weighted Euclidean distance is given as

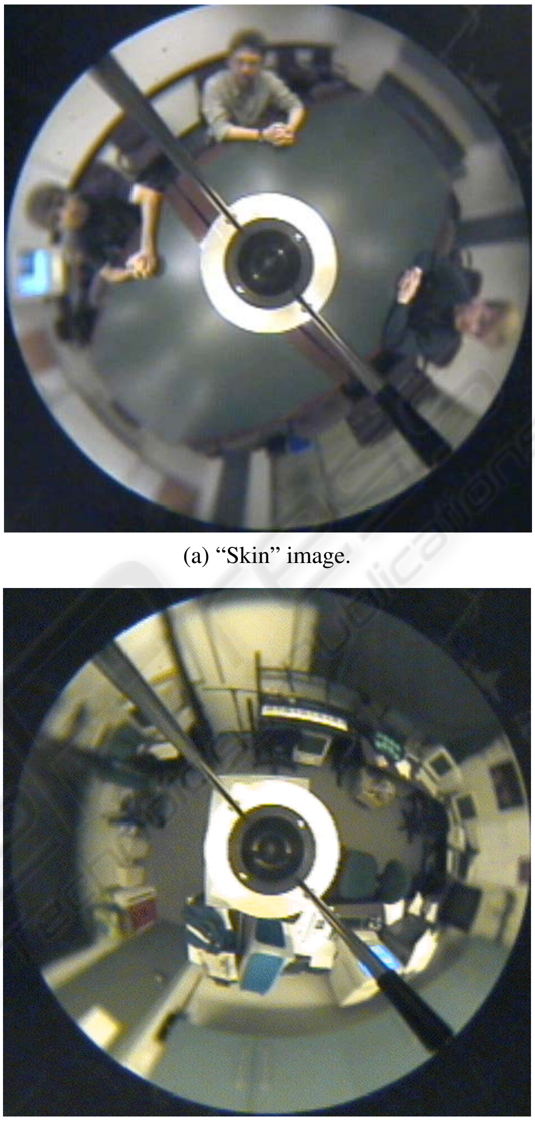
$$d(V,V') = \sqrt{\sum_{i=0}^{255} \frac{(V[i] - V'[i])^2}{\sigma^2}} \qquad (2)$$

A discussion regarding some of the issues related to the weighting factor w are presented in (Shupo et al., 2006).

3 EXPERIMENTAL RESULTS

In this section, the results of two experiments are presented. For both experiments two datasets of images obtained with the ParaCamera were used. The first dataset contained 43 images that contained one or more humans in the scene and therefore skin regions (e.g., “skin images”). From these images, 24 were used in the learning phase and 19 were used in the testing phase. The second dataset contained 47 images without any humans present in the scene and therefore no skin regions within the images (e.g., “non-skin images”). From these images, 24 of them were used in the learning phase and 23 were used in the testing phase. The images in both datasets were of type JPEG. JPEG images are compressed images where each "byte segment" contains the information needed to reconstruct the original image. For example, a byte or group of bytes may represent the encoding of some coefficient that results from the discrete cosine transform, or other transformation. In some cases, however, the encodings result in variable length codes (i.e., of length not necessarily multiple of 8) which are spread out in a number of bytes. Sample images from both the skin and non-skin datasets are provided in Figure 3. Both the ML and SLWE employ the use of a distance measure and appropriate *distance threshold*in order to classify an image as either "skin" or "non-skin". In general, different distance metrics and different threshold settings may lead to different results leading to variations in the number of false positives and false negatives. In order to draw a meaningful comparison between the results of the ML and SLWE algorithms using the Euclidean and weighted Euclidean distance metrics, results are presented as a comparison between the probability of false positives (e.g., the probability of incorrectly classifying an image as "skin" when in fact it does not contain any skin) and the probability of false negatives (e.g., the probability of incorrectly classifying an image as "non-skin" when in fact it does contain skin). In other words, setting the algorithmic parameters such that the algorithm results in a particular number of false negatives implies the algorithm will also result in the corresponding number of false positives. The resulting graphs illustrate the probability of false negatives vs. the probability of false positives. A summary of the results for the ML and SLWE algorithms using the weighted Euclidean distance metric are illustrated in Figure 4(a),(b). The resulting graphs illustrate the probability of false negatives vs. the probability of false positives.

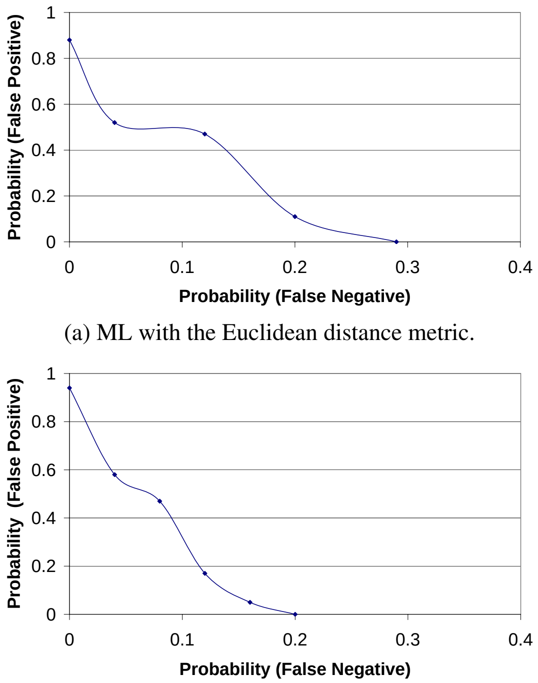
Since there is a trade-off between the probability of false positives and false negatives with each algorithm, one algorithm may outperform the other (e.g., provide "better" results, where "better" is defined as minimizing either the number of false positives, the number of false negatives or having an equal number of false positives and false negatives). When considering the weighted Euclidean distance, the ML algorithm with a 0% false positive rate and a 12% false negative rate, outperforms the SLWE when a minimal false positive rate is desired. With respect to the Euclidean distance, the SLWE with a false positive rate of 17% and false negative of 12% provides superior results over the ML algorithm when an equal number of false positive and false negatives is desired. A summary of the results for the ML and SLWE algorithms using a weighted Euclidean distance metric are illustrated in Figure 5(a),(b). The resulting graphs illustrate the probability of false negatives vs. the probability of false positives. Since there is a trade-off between the probability of false positives and false negatives with each algorithm, one algorithm may outperform the other (provide "better" results where "better" is with respect to minimizing the number of both false positives and false negatives). With respect to the Euclidean distance metric, the SLWE with a false positive rate of 17% and false negative of 12% pro-



(a) "Skin" image.

(b) "Non-skin" image.

Figure 3: Sample images form the "skin" and "non-skin" image datasets. (a)"Skin image" (e.g., the image contains regions of human skin) and (b) "non-skin image" (the image does not contain any regions of human skin).



(a) ML with the Euclidean distance metric.

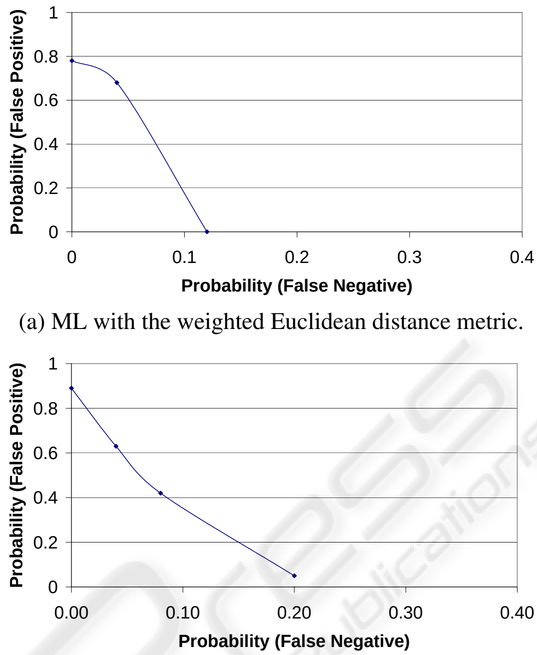
(b) SLWE with the Euclidean distance metric.

Figure 4: Decision strategies. Probability of false positives (e.g., the probability of incorrectly classifying an image as “skin” when in fact it does not contain any skin) vs. the probability of false negatives (e.g., probability of incorrectly classifying an image as “non-skin” when in fact it does contain skin) for the ML and SLWE algorithms employing the Euclidean distance metric. (a) ML algorithm and (b) SLWE algorithm.

vides superior results over the ML algorithm. When considering the weighted Euclidean distance metric, the ML algorithm with a weighted Euclidean distance with a 0% false positive rate and a 12% false negative rate provides superior results over the SLWE algorithm.

4 CONCLUSION

In this paper a description of a color-based human skin detector was presented. The system is capable of accurately and efficiently classifying images obtained with an omni-directional video sensor as either containing skin or not containing skin. The system is part of a larger ongoing project whose goal is to develop an autonomous surveillance system that can monitor an area for human activity. The work described here presented one aspect of this system. In particular, the development of a video sensor to obtain a fast overview of the scene and identify potential areas of human activity. Various other techniques can then be focused to these potential areas of interest in order to obtain further details. Two algorithms were presented: the maximum likelihood estimator and the stochastic learning weak estimator and for each algorithm, two distance metrics (Euclidean and weighted Euclidean) were experimented with.



(a) ML with the weighted Euclidean distance metric.

(b) SLWE with the weighted Euclidean distance metric.

Figure 5: Decision strategies. Probability of false positives (e.g., the probability of incorrectly classifying an image as “skin” when in fact it does not contain any skin) vs. the probability of false negatives (e.g., probability of incorrectly classifying an image as “non-skin” when in fact it does contain skin) for the ML and SLWE algorithms employing the weighted Euclidean distance metric. (a) ML algorithm and (b) SLWE algorithm.

During a training phase, ParaCamera images are classified as either skin (contain regions of skin) or non-skin (do not contain any regions of skin). The features that are extracted from the training phase with any needed adjustments are input into the classifier that was later used in the validation phase to classify incoming images as skin or non-skin. Results were presented that compared both algorithms and both distance metrics based on the number of false positives and false negatives. The ML algorithm provided superior results using the Euclidean distance metric while the SLWE provided superior results using a weighted Euclidean distance metric. Results also indicate that the system is capable of accurately classifying ParaCamera images as either skin or non-skin in a very efficient manner even when considering the poor resolution inherent with ParaCamera images.

Future work will include greater, more extensive experimentation using much larger training datasets and a larger number of test images obtained that have been obtained under a variety of scenarios and lighting conditions. Future work will also examine the processing of skin classified images in order to obtain greater details regarding the photographed scene. Processing can include locating and grouping skin regions within the image and inferring their position in the real world. This information is of importance to any surveillance application.

ACKNOWLEDGEMENTS

Asaf Shupo and Miguel Vargas Martin gratefully acknowledge the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

Baker, S. and Nayar, S. (1999). A theory of single viewpoint catadioptric image formation. *International Journal of Computer Vision*, 35(2):1–22.

Basu, A. and Southwell, D. (1995). Omni-directional sensors for pipe inspection. In *IEEE Transactions on Systems Man and Cybernetics*, volume 25, pages 3107–3112.

Boult, T., Michaels, R., Gao, P., Lewis, C., Yin, W., and Erkan, A. (1998). Frame rate omni-directional surveillance and tracking of camouflaged and occluded targets. http://www.eecs.lehigh.edu/ tboult/TRACK/LOTS.html.

Bradski, G. (1998). Computer vision face tracking for use in a perceptual user interface. Technical report, Intel Corp., Santa Clara, CA USA.

Chai, D. and Ngan, K. (1999). Face segmentation using skin-color map in videophone applications. *IEEE Transactions on Circuits Systems and Video Technology*, 9(4):551–564.

Chopra, M., Vargas Martin, M., Rueda, L., and Hung, P. (2006). A source address reputation system to combating child pornography at the network level. In *Proceedings of the IADIS International Conference on Applied Computing*, San Sebastian, Spain.

Duda, R. O., Hart, P. E., and Stork, D. E. (2000). *Pattern Classification*. Wiley Interscience, Hoboken, NJ. USA, second edition.

Gutchess, D., Jain, A., and Cheng, S. (2000). Automatic surveillance using omni-directional and active cameras. In *Proceedings of the 2000 Asian Conference on Computer Vision*.

Hans, S., Anderson, H., and Granum, E. (1999). Skin color detection under changing lighting conditions. In *Proceedings of the 7th Symposium on Intelligent Robotics Systems,*, pages 187–195, Columbia, Portugal.

Herpers, R., Derpanis, K., Topalovic, D., and Tsotsos, J. (1999). Detection and tracking of faces in real environments. In *Proceedings of the International Workshop on Recognition, Analysis and Tracking of Faces in Real-Time Systems*, pages 96–104, Korfu, Greece.

Jones, M. and Rehg, J. (1998). Statistical color models with applications to skin detection. Technical Report CRL 98/11, Compaq Computer Corp., Cambridge, MA USA.

Kapralos, B., Jenkin, M., and Milios, E. (2003). Audiovisual localization of multiple speakers in a video teleconferencing setting. *Journal of Imaging Science and Technology*, 13(1):95–105.

Nayar, S. (1997). Omnidirectional video camera. In *Proceedings of the DARPA Image Understanding Workshop*, pages 235–241, New Orleans, LA.

Oommen, B. J. and Rueda, L. (2005). On utilizing stochastic learning weak estimators for training and classification of patterns with non-stationary distributions. In *Proc. 28th German Conf. on AI*, pages 107–120, Koblenz, Germany.

Oommen, B. J. and Rueda, L. (2006). On utilizing stochastic learning weak estimators for training and classification of patterns with non-stationary distributions. *Journal of Pattern Recognition*, 39:328–341.

Peri, V. and Nayar, S. (1997). Generation of perspective and panoramic video from omnidirectional video. In *Proceedings of the DARPA Image Understanding Workshop*, pages 243–245, New Orleans, LA USA.

Raja, Y., McKenna, J., and Gong, S. (1998). Segmentation and tracking using color mixture models. In *Proceedings of the Third Asian Conference of Computer Vision*.

Shupo, A., Vargas Martin, M., Rueda, L., Bulkan, A., Chen, Y., and Hung, P. (2006). Toward efficient detection of child pornography in the network infrastructure. *IADIS International Journal on Computer Science and Information Systems*, 1(2):15–31.

Stiefelhagen, R., Yang, J., and Waibel, A. (1999). Modeling focus of attention for meeting indexing. In *Proceedings of the ACM Multemedia ’99*, pages 3–10, Orlando, FL USA.

Swain, M. and Ballard, D. (1991). Color indexing. *International Journal of Computer Vision*, 7:11–32.

Yang, J. and Waibel, A. (1996). A real time face tracker. In *Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision*, Sarasota, FL, USA.

Yasushi, Y. (1999). Omni-directional sensing and its applications. *IEEE Transactions on Information and Systems*, E82-3.

Yong, R., Gupta, A., and Cadiz, J. (2001). Viewing meetings captured by an omni-directional camera. In *ACM Transactions on Computer-Human Interaction*.

Zheng, J. and Tsuji, S. (1992). Panoramic representation for route recognition by a mobile robot. *International Journal of Computer Vision*, 9(1):55–76.