

Machine learning based chroma phase offset detection and correction in motion video

Advait Mogre, Shekhar Madnani

Digital Media Group, Video Quality Assessment, Interra Systems Incorporated, Cupertino, USA

Article Info

Article history:

Received Jan 5, 2022

Revised May 19, 2022

Accepted Jun 17, 2022

Keywords:

Chroma phase offset
Correction
Facial shot recognition
Hue
Machine Learning
Region of Interest
Skin tone reference

ABSTRACT

Generally, chroma phase or hue offset issues within a scene are hard to detect, without a reference or context (i.e. some apriori knowledge about how certain objects within the scene should actually appear in terms of their hue). Moreover, when it comes to skin/flesh tones, hue deviation can be noticeable and can markedly degrade the viewer quality of experience (QoE), whenever it does occur. However a lot of research has gone into flesh tone detection, specifically, the color gamut within which flesh tone is present. This topic has been well documented in the literature with respect to various color spaces: red, green, blue (RGB) and YIQ. Therefore, overall issues with chroma offset or hue within the video content could potentially be approached by extracting and analyzing a reliable reference, such as skin or flesh tone (if present), within some allowable deviation. This involves machine learning (ML) based facial recognition and tracking followed by skin tone region recognition within the detected facial sequence (i.e. Region of Interest). The skin region serves as a 'self-reference' in order to discern any inherent phase offset within the content. Finally, the angular chroma deviation discerned can then be used for subsequent correction as well.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Advait Mogre

Digital Media Group, Research and Development, Interra Systems Incorporated

1601 S. De Anza Blvd., Cupertino, CA 95014, USA

Email: advait@interrasystems.com

1. INTRODUCTION

Deviation in chroma phase, or hue, in received video content can be visually noticeable and degrade the viewer's quality of experience (QoE). Moreover, this is not uncommon and can be found even in contemporary media content, apart from file based legacy material. The key to handling this problem is being able to discern the offset in the chroma phase, or hue, in the absence of having the original unaffected video reference at the receiver or viewer's end. Even for an intelligent human viewer, it is not possible to perceive an anomaly in the hue or tint, till a portion of the video contains objects or patterns that are recognizable for what they are, and more importantly, for how they should 'normally' appear to be and the context within which they should be viewed. In this regard, hue/tint changes within skin tones can be easily discerned by the human observer, (since that knowledge is acquired and ingrained by the viewer over years of experience). However, the goal here is to be able to have the receiver (i.e. the 'intelligent machine') be able to detect any hue/tint offset within the received content and be able to correct for the same, so as to improve the viewer's QoE. In order to do so, one makes use of the fact that skin tones have a well-established color gamut, (within a given color space), that is startlingly similar in scope across the entire human race [1]–[5] and it is this property that is critical (when broadly interpreted), in order to extract a self-reference to discern any inherent hue offset within the video content. Here, the term self-reference implies that one can use certain regions,

such as skin tones, within the received video sequence to discern any anomalies present, without having the need to refer to, or have apriori knowledge of, the original/undistorted video content. The presumption of course being that human subjects and hence skin-tones are available at some stage, (at least for a limited period of time), within the video sequence that is to be analyzed.

Skin tone detection has been studied for several years, for applications ranging from adult content detection [6], [7], to detecting news anchors in television broadcast content for the sake of video automatic annotation [8], [9], archival, and retrieval and also for rudimentary flesh tone only correction [10], [11]. In the application in [10], for example, pixels converted to YIQ that were detected in and around the flesh tone region were moved closer to the “ideal” flesh tone by halving the Q component. (I-Q space has been discussed in section 2. In such applications, it was typically assumed that the face and the hands of the anchor person comprised the most significant flesh-tone region in a given video frame. Furthermore, news programs in studios are typically shot in environments with backgrounds that are designed to make them easily separable from the face/skin regions of the anchor or the panel [3], thus mitigating the issue of falsely detecting skin-tone like pixels, i.e. false positives.

However, the issue of extracting a flesh-tone self-reference (were it to be present) within a received video sequence becomes challenging when the background is not controlled, or is in a ‘natural’ setting. Here, a crucial step is to first extract regions of interest (RoIs) within which the likelihood of detecting skin tones is high, such as faces within a frame, and over time, spatially collocated facial sequences across consecutive frames. Note that several machine learning (ML) based intra-frame face detection techniques often (though not always) rely upon relevant features extracted from the grayscale image [12]–[15]. Therefore in our case, the derived grayscale facial shot using an available convolutional neural network (CNN) [16], [17] based facial tracking platform basically provides a ROI in space and time which is less susceptible to false positives, owing to the inherent spatio-temporal correlation of the facial shot sequence. Thus the ROI in each frame has the likelihood of facial regions detected within the given frame, and these regions are then analyzed or segmented for their flesh tone, allowing for some variation due to any inherent chroma offset. Ultimately, an offset so detected is used to generate a compensatory hue angle that is applied to each pixel of every frame, by first converting it to hue, saturation, value (HSV) space [10], [18], in order to correct for the offset within the hue (H) component. The corrected image sequence is then converted back to red, green, blue (RGB) for subsequent transcoding to a chroma phase corrected stream, or to drive a display. This overall methodology in handling the chroma phase or hue correction issue using such a ML based approach has not, to the authors’ knowledge, been dealt with or discussed in the available literature, and so has been pursued and presented in this work.

The paper is organized as follows: section 2 describes some of the different color spaces used for skin detection described in the literature. Section 3 covers the chroma phase offset detection method implemented, using ML based facial shot tracking to obtain RoIs. In section 4 we discuss the technique used to correct for any detected chroma phase offsets. Finally, in section 5, we draw conclusions from our current work.

2. FLESH TONE GAMUT IN VARIOUS COLOR SPACES

As studies show, the human skin has a well-defined (and narrow) gamut of hues and is not highly saturated. This is attributed to the fact that the skin pigmentation is determined by melanin, in particular, the balance between its two key components, (Eumelanin-brown and black, Pheomelanin-red and yellow) [19]. This narrow gamut property, as we shall see in the subsequent plots, makes skin tone detection viable across a range of color spaces. One notes that if the available content is in a given color space, such as RGB, it can typically be converted to another 3-component color space by a linear (3x3 matrix) transformation. So the choice of a color space for chroma phase analysis will be governed by the availability of skin-tone gamut rules applicable to that color space, as well as the convenience with which chroma phase values can be extracted.

So let us first review the literature on skin-tone color gamut with respect to various color spaces. To begin with, [3] has done a survey of color spaces widely used for skin tone representation, namely RGB, YCrCb, Yuv and Lab, described in a 4x3 matrix of plots shown in Figure 1(a) to Figure 1(l). Here, each of the three columns represent the skin tone gamut of an ethnicity across all the four color spaces, where each row represents the skin tone color gamut for a given color space, across the various ethnicities. The three ethnicities/columns are {Asian, African, Caucasian} and the four color spaces/rows are {RGB, YCrCb, Yuv, Lab}. Note that while these are 3-D spaces, only the two principal chroma components are used for the 2-D representation and subsequent analysis. Thus, the first row of Figure 1 shows the flesh tone gamut within the R-G space of Asian skin in Figure 1 (a), African skin in Figure 1 (b) and Caucasian skin in Figure 1 (c), The second row of Figure 1 shows the flesh tone gamut within the Cb-Cr space of Asian skin in Figure 1 (d), African skin in Figure 1 (e) and Caucasian skin in Figure 1 (f). The third row of Figure 1 shows the flesh tone gamut within the u-v space of Asian skin in Figure 1 (g), African skin in Figure 1 (h) and Caucasian skin in

Figure 1 (i). Finally, the last row of Figure 1 shows the flesh tone gamut within the a-b space of Asian skin in Figure 1 (j), African skin in Figure 1 (k) and Caucasian skin in Figure 1 (l).

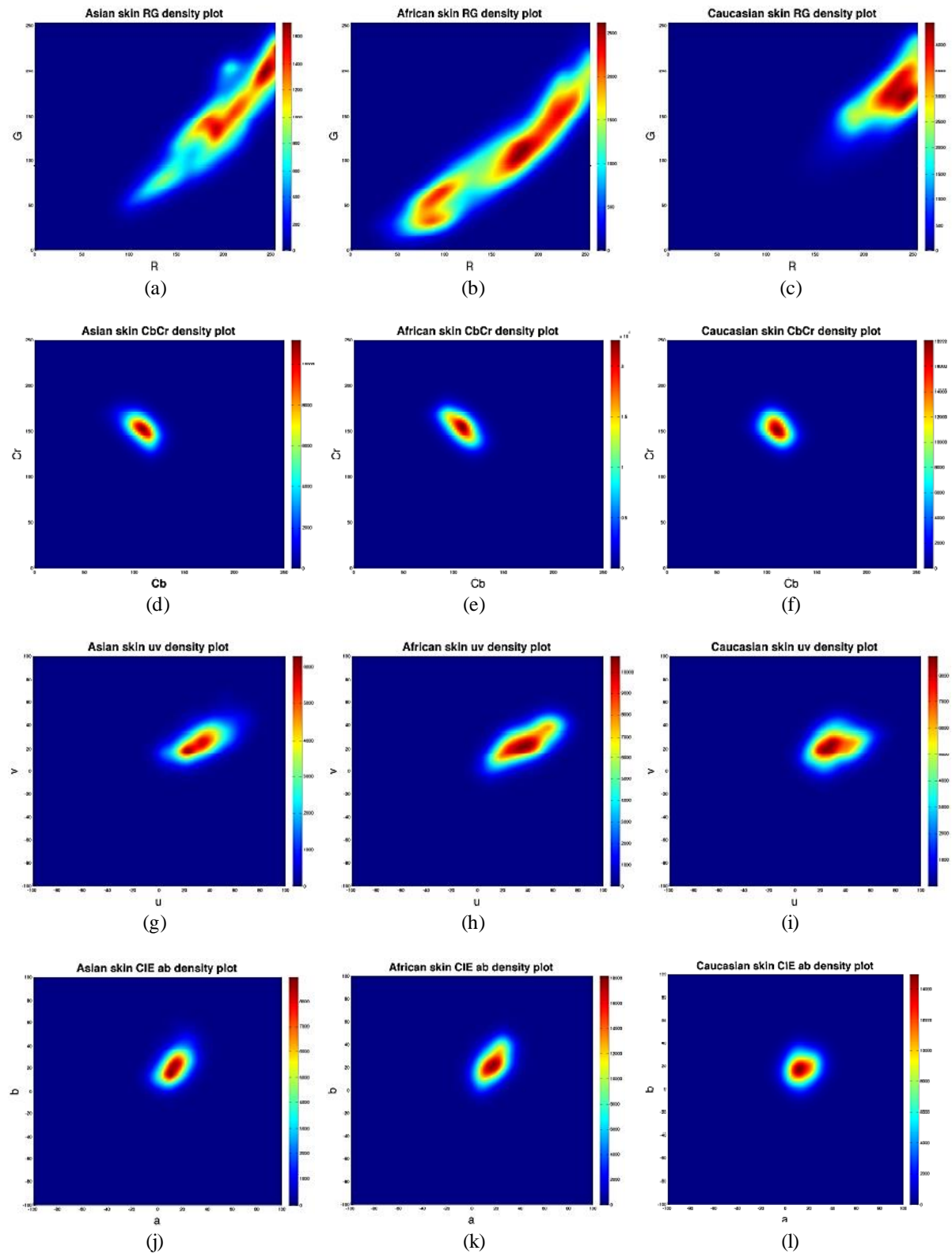


Figure 1. Skin tone color gamut plots in various color spaces [3] in (a) Asian R-G, in (b) African R-G, in (c) Caucasian R-G, in (d) Asian Cr-Cb, in (e) African Cr-Cb, in (f) Caucasian Cr-Cb, in (g) Asian u-v, in (h) African u-v, in (i) Caucasian u-v, in (j) Asian a-b, in (k) African a-b and in (l) Caucasian a-b

Furthermore, one more color space considered here is $\{Y, I, Q\}$, which was first used in the early years of NTSC color television. For skin-tone detection, color spaces such as $\{Y, I, Q\}$ separate the luma and chroma information, allowing us to analyze the color information in two dimensional chroma space, in this case the I-Q space, represented in an equivalent polar coordinate form. This is convenient, as can be seen when we consider the skin color gamut described in the I-Q color plane in Figure 2, and the corresponding model in section 2.2.

2.1. RGB skin-tone gamut model

The RGB color space is amongst the most common and has been reliably used to come up with a set of heuristic rules that define the skin tone gamut. Therefore, it shall be used in our work as a segmentation step, after having extracted the facial shot RoIs, (if available). A given pixel in (R, G, B) 8-bits per component format is classified as 'skin' under the following compound conditions [1]:

$$\begin{aligned} & \text{If } (R > 95 \text{ and } G > 40 \text{ and } B > 20 \text{ and} \\ & \max\{R, G, B\} - \min\{R, G, B\} > 15 \text{ and} \\ & |R - G| > 15 \text{ and } R > G \text{ and } R > B) \text{ is 'True'} \\ & \text{Then } (R, G, B) \text{ is a skin-pixel} \end{aligned} \quad (1)$$

2.2. I-Q skin tone gamut model

Note that the RGB model in section 2.1 requires all the three components for skin-tone classification and while it is well established and reliable, it does not directly lend itself to the concept 'chroma phase'. So we then perform a $\{R, G, B\} \rightarrow \{Y, I, Q\}$ color space conversion of the segmented skin-tone pixels classified in section 2.1, in order to obtain their representation in the 2D I-Q plane as described in Figure 2. In the latter scenario, the gamut of the skin pixels need to be within $\pm 30^\circ$ ($\pi/6$) of the I axis, as represented by the shaded region in Figure 2 [10].

$$-\pi/6 \leq \arctan\left(\frac{Q}{I}\right) \leq \pi/6 \quad (2)$$

Note that by itself, (2) is necessary, but not sufficient. We will first be using the skin-tone segmentation rules (such as described by (1)) within the extracted facial shot ROI's, to discern if the video content has undergone any chroma phase offset, and then if so, to what extent will be based on the non-conformance of (2). This is described in detail in the section 3.

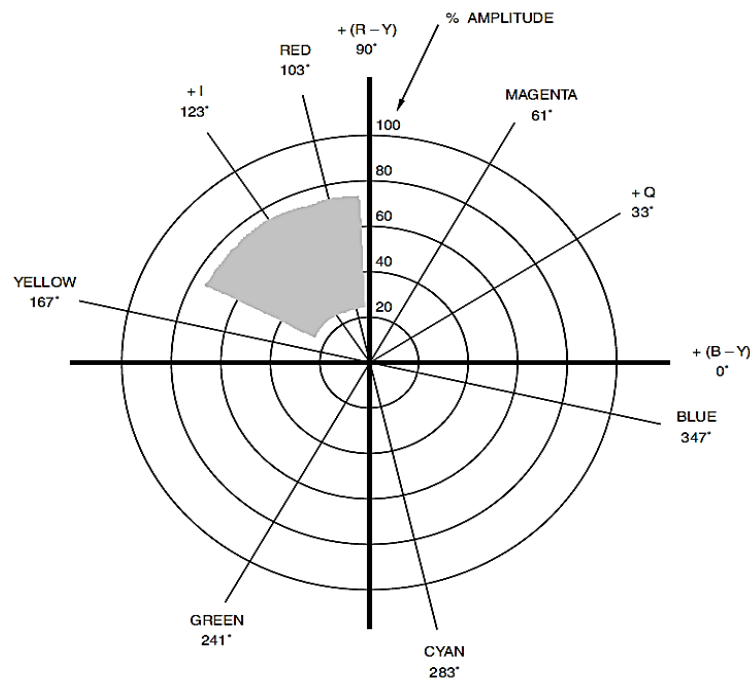


Figure 2. Shaded region indicates typical flesh tone color range in I-Q space [10]

3. METHOD

This technique first detects and extracts a self-reference within the given video content, (i.e. a facial sub-shot or ROI that can then be used to focus on the flesh tone), which is based upon ML based facial recognition routines implemented by a CNN. Next, we see how much deviation there is within the flesh tone pixels from the 'normal' gamut on a per frame basis and generate the corresponding confidence. This confidence value is accumulated at the end of each frame in order to generate an average color phase metric at the end of the given shot, which in turn is filtered (i.e. smoothened) to generate a hue compensation angle in the HSV space to correct for the phase anomaly, if present. In the following sections, we first describe the system in terms of a flow chart that captures the overall end-to-end approach, and then describe them in greater detail in the subsequent sections.

3.1. Algorithm outline

A system block diagram/flow chart is shown in Figure 3. It summarizes the algorithmic steps of the chroma phase detection and correction approach. The following sections 3.1.1 to 3.1.5, describe each of the main steps in greater detail.

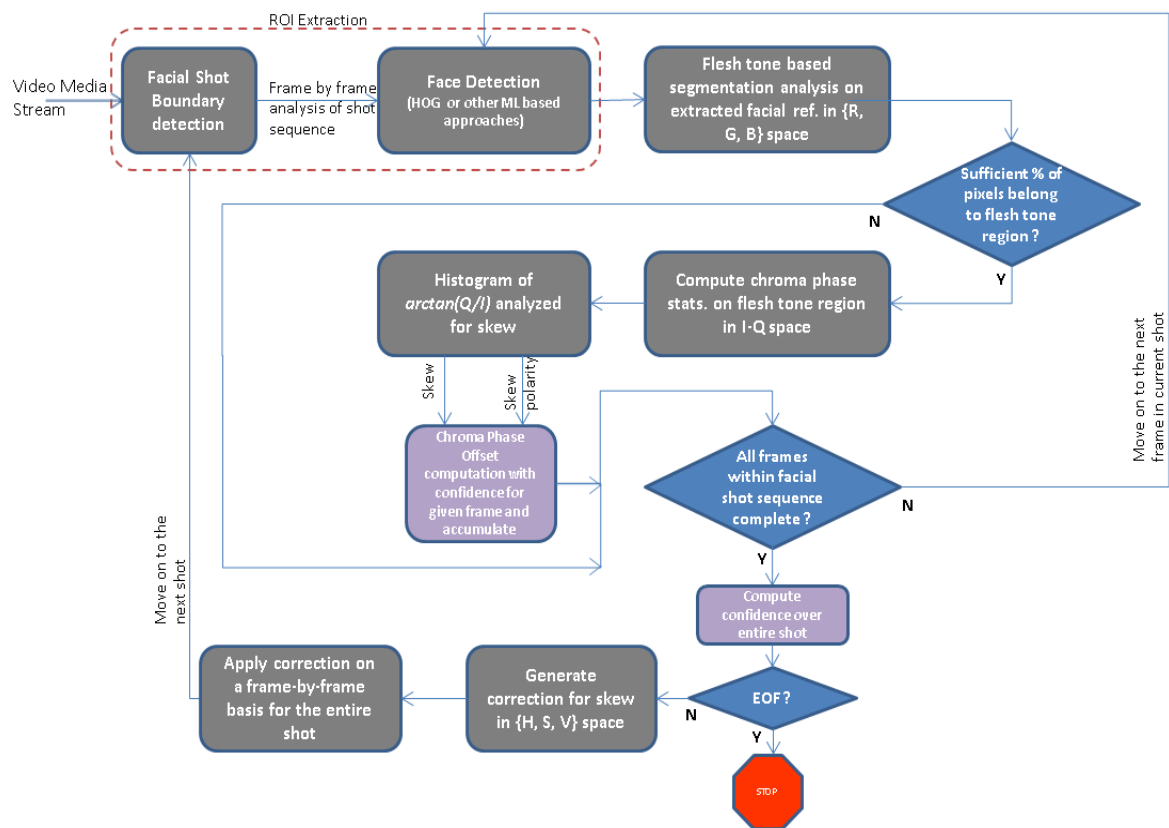


Figure 3. Chroma phase detection and correction flow chart

3.1.1. Self-reference ROI extraction

First obtain the self-reference if available, i.e. facial shot(s) from the video clip that is to be analysed; thereby precluding the non-essential portions of the image sequence. This is the ML based pre-classification stage that extracts the facial RoIs. Here, we note that it is not necessary that every shot within the given video content needs to have human subjects-in theory, one shot is sufficient (assuming that the chroma offset anomaly is prevalent throughout the entire video clip, which in practical scenarios is generally a valid assumption). The pre-classification has been achieved using the CNN developed by the visual geometry group (VGG) at Oxford University [20]. Their approach is outlined as follows: i) Implementation of a histogram of oriented gradients (HOG) based approach for face detection [21], [22]; and ii) Subsequent face tracking done using a Kanade, Lucas & Tomasi (KLT) tracker [23], [24].

A couple of examples of this process are shown in Figures 4(a) and 4(b), wherein each figure represents a frame from the respective chroma phase affected content. Here, the content in Figure 4 (a) depicts positive skew or offset, i.e. a reddish tint or hue. Figure 4 (b) on the other hand, shows a negative skew or offset, i.e. a yellowish tint or hue. Thus, if the chroma offset affected video under analysis has human subject content, then this methodology enables us to extract the relevant facial shots, which form the RoIs for our subsequent steps.

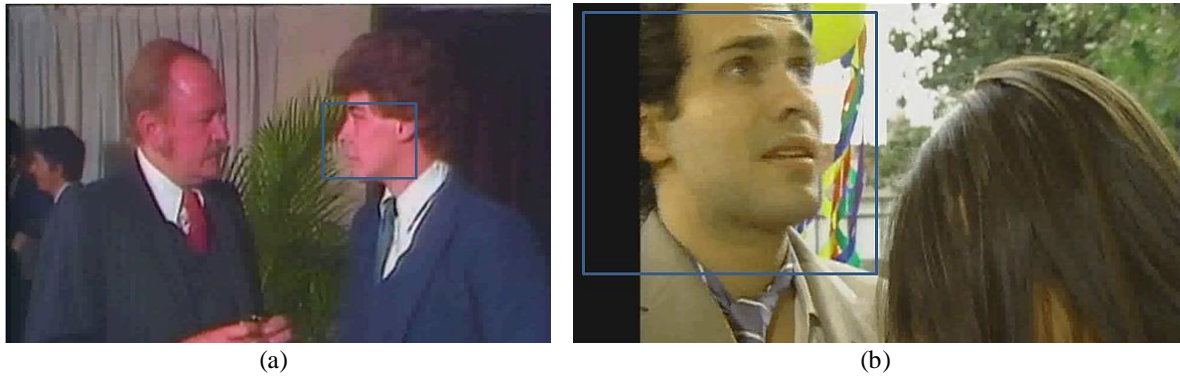


Figure 4. Facial tracking (blue frame) within affected video content (a) reddish hue and (b) yellowish hue

3.1.2. Flesh-tone pixel segmentation

Next, use the above reference RoIs to obtain pixel-wise statistics that further indicate the presence of skin tone, as well as inherent color phase deviation (positive or negative), if any. This is done by referring to the skin classifier rules described in (1). However, in order to mitigate false positives we further augment the $\{R, G, B\}$ component-wise constraints described in (1), to yield (3).

$$\begin{aligned}
 & \text{If } (R > 95 \text{ and } G > 40 \text{ and } B > 20 \text{ and} \\
 & \max\{R, G, B\} - \min\{R, G, B\} > 15 \text{ and} \\
 & |R - G| > 15 \text{ and } R > G \text{ and } R > B \text{ and} \\
 & R - B < 255 \text{ and } G > B) \text{ is 'True'} \\
 & \text{Then } (R, G, B) \text{ is a skin-pixel}
 \end{aligned} \tag{3}$$

3.1.3. Chroma phase angle histogram extraction

For those pixels that satisfy (3), they undergo a conversion to the $\{Y, I, Q\}$ space in order to extract the chroma phase angle histogram statistics. Upon doing so, a histogram of the phase angle $\arctan(Q/I)$ is obtained in order to extract the statistics of the proportion of pixels falling outside of the bounds as described by (2). In the following examples, a set of images for each selected example has been provided, where in:

- The first image is an available facial shot (ROI) from an affected video content.
- From the available facial shot (ROI), the second image is the coarse flesh tone segmentation obtained using the rule provided by (3).
- The third plot is the I versus Q scatter plot of the flesh tone pixels tagged within the segmented image.
- The fourth plot is the histogram of $\arctan(Q/I)$ for the flesh tone pixels. The red vertical lines denote $\pm \pi/6$ around the I axis.

Example 1 Analyzing content with a reddish hue. If the skin pixels have a reddish hue, (i.e. phase angle is positive towards the color red), then more of the phase angle histogram bins closer to or $> \pi/6$ will be populated and the histogram will have a positive skew. The facial tracking indicated by the blue bordered frame shown in Figure 4(a) is used as an ROI to perform skin tone segmentation followed by histogram extraction, as shown in Figure 5.

Example 2 Analyzing content with a reddish yellowish hue. On the other hand, if the video content has a yellowish hue, (i.e. phase angle is negative towards the color Yellow), then more of the phase angle histogram bins closer to or $< -\pi/6$ will be populated, and the histogram will have a negative skew. Similarly, the facial tracking indicated by the blue bordered frame shown in Figure 4(b) is used as an ROI to perform skin tone segmentation followed by histogram extraction, as shown in Figure 6.

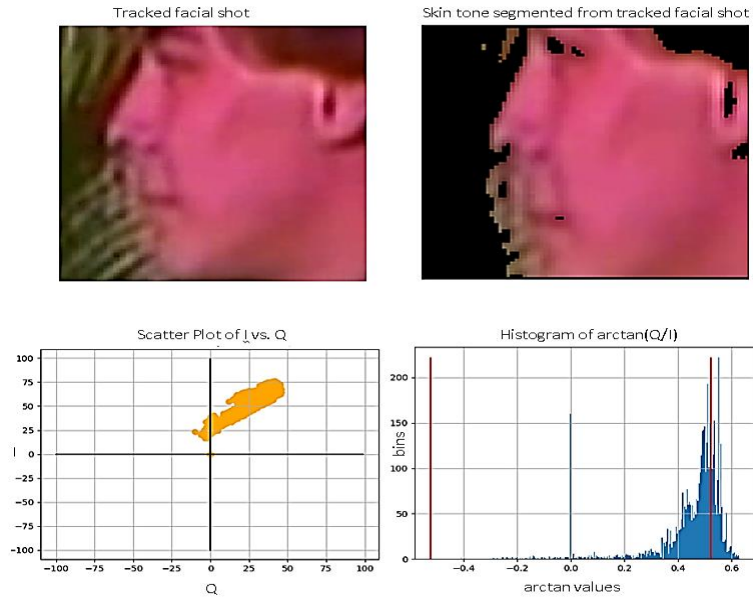


Figure 5. Example of content with positive (reddish) skew

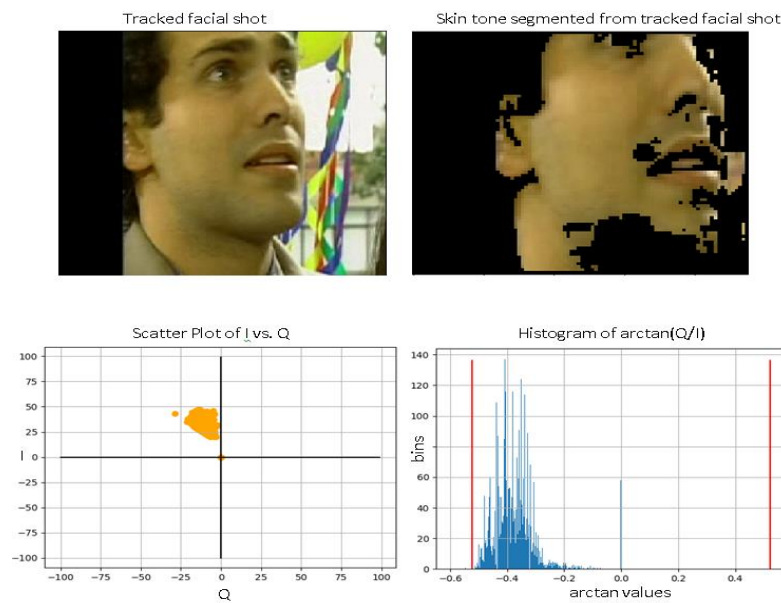


Figure 6. Example of content with negative (yellowish) skew

3.1.4. Histogram analysis and offset confidence metric generation

Conclusions regarding phase deviation, if any, are made only if a “significant” percentage of the image contains flesh tone pixels that satisfy (3). This thresholding is done to avoid unreliable conclusions or false positives. Here, the extracted histogram described in section 3.1.3 then allows us to observe how much deviation there is within the flesh tone pixels from the ‘normal’ gamut and thus deduce if any chroma phase offset is present within the given content. These observations are taken on a frame-by-frame basis and then aggregated over the entire ROI sequence in order to obtain a shot wide average. This parameter is then used to generate a confidence metric whose magnitude is between [0.0, 1.0] and is associated with the chroma phase offset observed. The pseudo code is outlined as follows:

For every frame ‘n’:
 Initialize: $\theta_{\text{aggregate}}$, $\text{red_skew_pixel_count}$ & $\text{yellow_skew_pixel_count}$ to 0; (4)

```

If (skin_pixel_count > K1*h*w) Then
  For each skin-tone pixel converted to (Y, I, Q):
    theta = arctan(Q/I)
    If (arctan(Q/I) > π/6) Then
      increment red_skew_pixel_count by 1
    Else If (arctan(Q/I) < -π/6) Then
      increment yellow_skew_pixel_count by 1
  theta_aggregate = theta_aggregate + theta

```

Where,

H : Height of the input image
w : Width of the input image
theta_aggregate : The accumulate of the skin pixel phase angles within the ROI(s)
skin_pixel_count : The total number of skin-tone pixels detected in the segmented ROI(s)
red_skew_pixel_count : The amount of skin-tone designated pixels within the ROI whose phase exceeds $\pi/6$
yellow_skew_pixel_count : The amount of skin-tone designated pixels within the ROI whose phase is $< -\pi/6$
K1 : A constant chosen typically ≥ 0.1 , (i.e. at least 10% of the entire image)

The theta aggregate and skew counts are then normalized by the *skin_pixel_count* to yield the following parameters for the given frame 'n':

$$norm_red_skew(n) = red_skew_pixel_count / skin_pixel_count \quad (5)$$

$$norm_yellow_skew(n) = yellow_skew_pixel_count / skin_pixel_count \quad (6)$$

$$theta_average(n) = theta_aggregate / skin_pixel_count \quad (7)$$

Note that typically, only the count that reflects the skew, (either towards red, i.e. positive skew, or towards yellow, i.e. negative skew), will be non-zero. On the other hand if there is no chroma skew, then clearly both the counts will be zero. Thus, having extracted the chroma phase histogram skew, we are now in a position to use this parameter to derive a confidence metric associated with the chroma phase offset, outlined as follows:

$$\begin{aligned}
 & \text{if}(theta_average(n) \geq 0) \{ \\
 & \quad chroma_phase_offset_confidence(n) = (1 - e^{-(K2 * norm_red_skew(n))^{K4}} * \\
 & \quad \quad (1 - (e^{-(abs(theta_average(n))^{K3}/K5}))) \\
 & \} \text{ else } \{ \\
 & \quad chroma_phase_offset_confidence(n) = -(1 - e^{-(K2 * norm_yellow_skew(n))^{K4}} * \\
 & \quad \quad (1 - (e^{-(abs(theta_average(n))^{K3}/K5}))) \\
 & \}
 \end{aligned} \quad (8)$$

Where,

chroma_phase_offset_confidence(n) : The phase offset confidence metric derived from the chroma phase analysis for given frame 'n'
K2 : programmable parameter, typically = -30
K3 : programmable parameter, typically = 0.3
K4 : programmable parameter, typically = 0.04
K5 : programmable parameter, typically = 6

The choice of the exponent functions in formulating the *chroma_phase_offset_confidence* metric for a given frame is governed by the fact that we need to continuously and monotonically map the entities *norm_red_skew* (or *norm_yellow_skew*), which are essentially in the domain of [0, 1] and $|theta_average|$, which is in the domain of [0, π], to a single metric function with its range bounded by [0, 1]. One also notes from (8) that the polarity of the *chroma_phase_offset_confidence* metric is defined by the polarity of *theta_average*. Now analysing the chroma phase histogram data from the two (single frame) examples shown in Figure 4a and the corresponding Figure 5; as well as Figure 4b and the corresponding Figure 6, we get: For the chroma phase analysis from the results of Figure 4(a) and Figure 5:

$$norm_red_skew = 0.0288$$

$$\theta_{average} = 0.4502$$

Using (8), we get:

$$\text{chroma_phase_offset_confidence} = 0.9783$$

Similarly, for the chroma phase analysis from the results of Figure 4(b) and Figure 6:

$$\begin{aligned} \text{norm_yellow_skew} &= 0.0003 \\ \theta_{average} &= -0.3747 \end{aligned}$$

Using (8), we get:

$$\text{chroma_phase_offset_confidence} = -0.8095$$

Finally, the confidence extracted from each frame within the shot is aggregated over all the frames within the shot ('N'), in order to produce an overall shot-wide confidence, (see Table 1, section 4, results and discussion):

$$\text{color_phase_metric_average} = \frac{\sum_{n=0}^{N-1} \text{color_phase_offset_confidence}(n)}{N} \quad (9)$$

3.1.5. Chroma phase offset correction

Having obtained the chroma phase offset (if any), from the content analysis as described in Sections 3.1.1 to 3.1.4, we are now in a position to correct for the offset. This is done under the following assumptions:

- Chroma phase 'wraps around' $\pm\pi$ with respect to the I-axis. So it is assumed that we are dealing with phase offsets that are significantly less (in magnitude) than π which in a practical sense is a fair assumption, considering that the normal skin-tone gamut itself exists within a relatively narrow region of $\pm\pi/6$ around the I axis.
- The content is currently corrected for phase (or hue) anomalies only, not for saturation. In this regard, converting from {R, G, B} \rightarrow {H, S, V} space is a viable option, with the {H} component (i.e. hue), being the one that is adjusted by a correction angle proportional to the derived chroma phase offset, aggregated or filtered over the entire shot.

Initialize: $\theta_{comp} = 0, \theta_{filt} = 0$

For every frame 'n':

$$\theta_{filt}(n) = ((1.0 - \text{ALPHA}) * \theta_{filt}(n-1)) + (\text{ALPHA} * \theta_{average}(n)) \quad (10)$$

$$\theta_{comp}(n) = \min(\theta_{filt}(n) / K_scale / \pi, 1.0) \quad (11)$$

For each pixel (i, j), within frame 'n':

$$H_{adjusted}(i, j, n) = \max(H(i, j, n) + \theta_{comp}(n), 0.0) \quad (12)$$

Where:

$\theta_{average}(n)$: From (7), section 3.1.4. The average chroma phase offset derived for the given frame 'n'

$\theta_{filt}(n)$: The filtered value of $\theta_{average}(n)$ over successive 'n+1' frames {0, ..., n} within the given shot, using a first order low pass IIR filter

ALPHA: The parameter of the first order IIR used; chosen between [0.05, 0.5]

K_scale: A programmable parameter used to scale the filtered value, depending upon the desired sensitivity of the correction process; chosen between [1.5, 6.0].

$\theta_{comp}(n)$: The compensation angle for the Hue component, to correct for the chroma phase in frame 'n'

$H_{adjusted}(i, j, n)$: The corrected Hue pixel at (i, j) for frame 'n', after applying $\theta_{comp}(n)$

Each corrected $\{H_{adjusted}, S, V\}$ pixel is then converted back to $\{R, G, B\}$ for subsequent transcoding or display.

3.1.6. Content correction example

So continuing with the affected video stream example represented by the single frame in Figure 4(a), and its subsequent chroma phase analysis as depicted in Figure 5, the steps described in sections 3.1.1 to 3.1.4. are carried out to determine the $\theta_{average}$ for each frame. This parameter is then iteratively used to generate the Hue correction, as described using (10) to (12) in Section 3.1.5, using a sequence of frames from which the ROI is extracted in terms of a facial shot.

Thus we note that while a single frame has been shown for convenience in Figure 7, the extracted ROI is actually a sequence of sub-regions that comprise an extracted facial shot. As described in Section 3.1.5, this leads to the generation of θ_{filt} , the filtered value of $\theta_{average}$ over successive frames with the given shot; which in turn is used to generate θ_{comp} , the compensation angle for the Hue component, to correct for the chroma phase. The correction settings used to generate this result were: $ALPHA=0.5$ and $K_{scale}=1.5$.

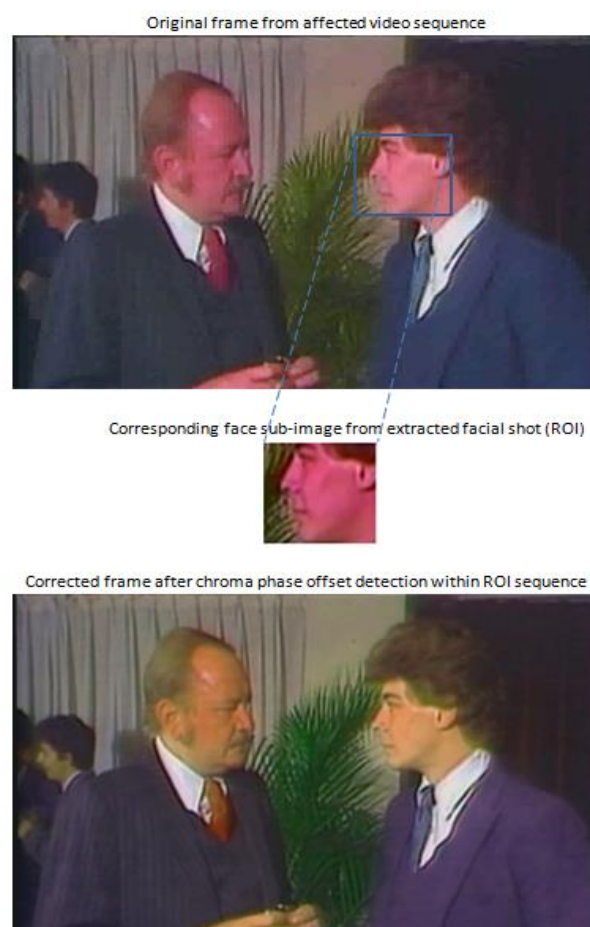


Figure 7. Chroma phase offset correction

4. RESULTS AND DISCUSSION

This section captures and discusses some of the results obtained upon analyzing a variety of motion video content. As a part of the algorithm verification process, around 3000 video streams were analyzed for chroma phase issues. Out of all these, only the ones in which significant chroma phase issues were detected have been tabulated in Table 1:

The results obtained demonstrate the viability of the overall chroma offset detection and correction method in the absence of a known-reference [25], in terms of extracting a reliable ROI based upon ML based facial shot extraction, in order to then be in a position to detect genuine cases of chroma phase offset within

the given content, using known skin-tone properties. To reiterate, this method works when human subjects are present in the content; as the human flesh tone provides for a reliable self-reference, more so when it is a sequence of spatially correlated flesh tone regions, (as in a facial shot). Once a reliable ROI has been extracted, then chroma phase analysis can be done to discern for phase anomalies, which can be used subsequently for hue correction. This has been demonstrated by the example shown in Figure 7 and summarized by the system flow chart in Figure 3.

Table 1. Summary of results over a variety of video media content

Stream Header	Average chroma_phase_metric (over all ROI shots)	Maximum chroma_phase_metric (over all ROI shots)
EMIVideo-0094636667652	0.4438	0.7399
EMIVideo-0094636513454	0.2543	0.5695
BPR01625701-dog	0.1193	0.7009
B000341033S0_1*	0.0523	0.7402
B000341033S0_2*	0.0595	0.7447
HD_D4605111	0.3076	0.7557
AK44ASFMeer	0.1919	0.6814
SDPS8U_See1	0.6327	0.8602
IMX50-video-dropouts-and-block-errors	0.1062	0.7283
Interlaced	0.4904	0.4904
VTS_02_1	0.2212	0.7423
RCOW_FINAL_HD_DELIVERY.1	0.0650	0.7470
2_audio_induction_hissyscratchB	0.4593	0.7408
FASE_00000 (a single shot comprising of 144 frames, with a representative frame shown in Figure 4(a))	0.4532	0.4532
FASE_00006 (a single shot comprising of 200 frames, with a representative frame shown in Figure 4(b))	-0.3920	-0.3920

*These streams have animation content and hence technically, should not be flagged as they are subject to the animator's artistic intent. However, animation detection [26], [27] is a separate component of the overall flow and will be dealt with in the future.

5. CONCLUSION

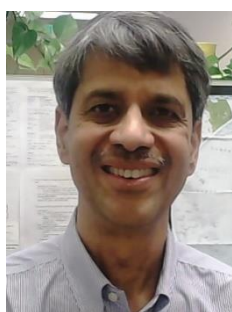
In the realm of no-reference video quality analysis, using flesh tone detection to discern and correct for chroma phase offset anomalies has been shown to be a viable option during the course of this work, provided we have a robust flesh tone detector to generate reference information. The latter has been achieved using a ML based face tracker, which provides a spatio-temporal correlation within the facial regions within successive frames of the extracted facial shot; i.e. the overall Region of Interest, within which the color phase analysis using (3) is to be performed. This method has proven to be robust, as seen by its viability across a variety of content. However, there is scope for further work. In particular, the current methodology applies only if human faces are present in at least one shot, if not more. So other modes of extracting a self-reference, or additionally, a known-reference (specific pattern or template), within the content can also be explored. Furthermore, animation content is subject to the creativity of the animator or colorist; wherein the normal rules of flesh-tone gamut need not necessarily apply. Therefore for future enhancements to this work, animation content has to be flagged during the ML based pre-processing step, and then dealt with separately, (for instance, using known-reference techniques).




REFERENCES

- [1] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A survey on pixel-based skin color detection techniques," *Proceedings of GraphiCon 2003*, vol. 85, no. 0896-6273 SB-IM, pp. 85–92, 2003, [Online]. Available: <http://academic.aua.am/Skhachat/Public/Papers>.
- [2] S. K. Singh, D. S. Chauhan, M. Vatsa, and R. Singh, "A robust skin color based face detection algorithm," *Tamkang Journal of Science and Engineering*, vol. 6, no. 4, pp. 227–234, 2003, doi: 10.6180/jase.2003.6.4.06.
- [3] A. Elgammal, C. Muang, and D. Hu, "Skin detection - a short tutorial," in *Encyclopedia of Biometrics*, Boston, MA: Springer US, 2009, pp. 1218–1224.
- [4] S. Kolkur, D. Kalbande, P. Shimpi, C. Bapat, and J. Jatakia, "Human skin detection using RGB, HSV and YCbCr color models," in *Proceedings of the International Conference on Communication and Signal Processing 2016 (ICCASP 2016)*, 2017, pp. 324–332, doi: 10.2991/iccasp-16.2017.51.
- [5] M. A. Berbar, "Skin colour correction and faces detection techniques based on HSL and R colour components," *International Journal of Signal and Imaging Systems Engineering*, vol. 7, no. 2, pp. 104–115, 2014, doi: 10.1504/IJSISE.2014.060056.
- [6] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1999, vol. 1, pp. 274–280, doi: 10.1109/cvpr.1999.786951.
- [7] M. M. Fleck, D. A. Forsyth, and C. Bregler, "Finding naked people," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 1065, pp. 594–602, 1996, doi: 10.1007/3-540-61123-1_173.




- [8] G. Wei and I. K. Sethi, "Face detection for image annotation," *Pattern Recognition Letters*, vol. 20, no. 11–13, pp. 1313–1321, Nov. 1999, doi: 10.1016/S0167-8655(99)00100-2.
- [9] R. M. Jiang, A. H. Sadka, and H. Zhou, "Automatic human face detection for content-based image annotation," in *2008 International Workshop on Content-Based Multimedia Indexing, CBMI 2008, Conference Proceedings*, Jun. 2008, pp. 66–69, doi: 10.1109/CBBI.2008.4564929.
- [10] K. Jack, "NTSC and PAL digital encoding and decoding," in *Video Demystified (Fourth Edition)*, Elsevier, 2005, pp. 394–471.
- [11] A. Nadian-Ghomsheh, "Color constancy for improving skin detection," *International Journal of Image Processing (IJIP)*, vol. 8, no. 6, pp. 479–496, 2014, doi: 10.1.1.734.9380.
- [12] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, vol. 1, pp. 511–518, doi: 10.1109/cvpr.2001.990517.
- [13] M. H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002, doi: 10.1109/34.982883.
- [14] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1867–1874, doi: 10.1109/CVPR.2014.241.
- [15] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2015, pp. 5325–5334, doi: 10.1109/CVPR.2015.7299170.
- [16] Ujjwalkarn, "An intuitive explanation of convolutional neural networks," *The data science blog*. 2016, [Online]. Available: <https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/>.
- [17] S. Srinivas, R. K. Sarvadevabhatla, K. R. Mopuri, N. Prabhu, S. S. S. Kruthiventi, and R. V. Babu, "A taxonomy of deep convolutional neural nets for computer vision," *Frontiers Robotics AI*, vol. 2, Jan. 2016, doi: 10.3389/frobt.2015.00036.
- [18] P. Sharma and V. Yadav, "Discrimination Between Skin and Non-Skin Pixels in Image Using the Range of HSV Color Space," *International Journal of Computer Science and Technology*, vol. 4, no. 1, pp. 438–439, 2013.
- [19] L. Krouse, "What is skin pigmentation," *verywellhealth*. 2021.
- [20] J. S. Chung and A. Zisserman, "Out of time: Automated lip sync in the wild," in *Computer Vision – ACCV 2016 Workshops. ACCV 2016. Lecture Notes in Computer Science*, vol. 10117, 2017, pp. 251–263.
- [21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, 2005, vol. 1, pp. 886–893, doi: 10.1109/CVPR.2005.177.
- [22] W. T. Freeman and M. Roth, "Orientation Histograms for Hand Gesture Recognition," in *IEEE Intl. Wkshp. on Automatic Face and Gesture Recognition*, 1995, pp. 296–301.
- [23] C. Tomasi and T. Kanade, "Detection and tracking of point features," *International Journal of Computer Vision*, 1991, doi: 10.1.1.45.5770.
- [24] M. S. Sri, "Object detection and tracking using KLT algorithm," *International Journal of Engineering Development and Research, IJEDR*, vol. 7, no. 2, pp. 542–545, 2019.
- [25] S. A. De Araujo and H. Y. Kim, "Ciratefi: An RST-invariant template matching with extension to color images," *Integrated Computer-Aided Engineering*, vol. 18, no. 1, pp. 75–90, 2011, doi: 10.3233/ICA-2011-0358.
- [26] X. Qin, Y. Zhou, Z. He, Y. Wang, and Z. Tang, "A faster R-CNN based method for comic characters face detection," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Nov. 2017, vol. 1, pp. 1074–1080, doi: 10.1109/ICDAR.2017.178.
- [27] Y. G. Ichihara, "Color constancy in Japanese animation," in *Color Imaging XI: Processing, Hardcopy, and Applications*, Jan. 2006, vol. 60580C, p. 6058, doi: 10.1117/12.639253.

BIOGRAPHIES OF AUTHORS



Advait Mogre    is currently a Principal Scientist at Interra Systems since November 2012, with an emphasis on Video Quality assessment. Prior to this assignment, he held a similar position at Broadcom Corporation, with a focus on algorithmic and architectural issues pertaining to picture quality within a Set Top Box receiver. He began his professional career with an involvement in Systems Analysis and modeling of video compression standards at LSI Logic. He has over 30 years of experience in the consumer product industry. He holds Bachelor and Master of Technology degrees from the Indian Institute of Technology-Bombay, and a Doctorate from the University Of Missouri-Columbia with an emphasis in Image Processing, Computer Vision and AI. He is currently a Senior Member of the IEEE. He can be contacted at email: advait@interrasystems.com.



Shekhar Madnani    is a Director of Engineering, Video Quality Research Group at Interra Systems. He is a Gold Medalist in his Bachelor of Technology (Electrical Engineering) graduating class from College of Technology, Pantnagar University, India and Master of Technology in Computer Technology from the Indian Institute of Technology, Delhi, India. He has more than 20 years of experience in the area of Image and Video processing, Computer Vision, Machine Learning, No-reference Video Quality Assessment, Real-time Acquisition, Delivery and Processing of video/audio data. He has research publications in IEEE, SMPTE and other conferences. He has filed more than fifteen patents in the area of Video Quality Analysis. He is currently leading the Video Quality initiatives for Baton™, a Scalable, Enterprise-class, Automated File-based QC Solution. He can be contacted at email: shekhar@interrasystems.com.