[13] R. B. Fisher and D. K. Naidu, "A comparison of algorithms for subpixel peak detection," in *Image Technology, Advances in Image Processing, Multimedia and Machine Vision*. New York: Springer-Verlag, 1996, pp. 385–404.
[14] D. G. Bailey, "Sub-pixel estimation of local extrema," in *Proc. Image and Vision Computing,* 2003, pp. 408–413.
[15] M. Shimizu and M. Okutomi, "Precise sub-pixel estimation on area-based matching," in *Proc. 8th IEEE ICCV*, Vancouver, BC, Canada, 2001, pp. 90–97.
[16] J. I. Woodfill, G. Gordon, D. Jurasek, T. Brown, and R. Buck, "The Tyzx DeepSea G2 vision system, a taskable, embedded stereo camera," in *Proc. Embedded Comput. Vis. Workshop*, 2006, pp. 126–132.
[17] I. Haller, C. Pantilie, F. Oniga, and S. Nedevschi, "Real-time semi-global dense stereo solution with improved sub-pixel accuracy," in *Proc. IEEE IV Symp.*, Jun. 2010, pp. 369–376.
[18] S. Hermann, R. Klette, and E. Destefanis, "Inclusion of a second-order prior into semi-global matching," in *Proc. 3rd Pacific Rim Symp. Adv. Image Vid. Technol.*, Jan. 2009, vol. 5414, Lecture Notes in Computer Science, pp. 633–644.
[19] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Proc. IEEE CVPR*, Jun. 2007, pp. 1–8.
[20] H. Hirschmuller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1582–1599, Sep. 2009.
[21] R. Szeliski and D. Scharstein, "Sampling the disparity space image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 3, pp. 419–425, Mar. 2004.
[22] S. Gehrig and U. Franke, "Improving stereo sub-pixel accuracy for long range stereo," in *Proc. IEEE 11th ICCV*, Rio de Janeiro, Brazil, 2007, pp. 1–7.
[23] I. Haller, C. Pantilie, M. Tiberiu, and S. Nedevschi, "Statistical method for sub-pixel interpolation function estimation," in *Proc. IEEE ITSC*, Sep. 2010, pp. 1098–1103.
[24] F. Oniga and S. Nedevschi, "Processing dense stereo data using elevation maps: Road surface, traffic isle and obstacle detection," *IEEE Trans. Veh. Technol.*, vol. 59, no. 3, pp. 1172–1182, Mar. 2010.

Eye-Tracking Database for a Set of Standard Video Sequences

Hadi Hadizadeh, *Student Member, IEEE*, Mario J. Enriquez, and Ivan V. Bajić, *Member, IEEE*

***Abstract*—This correspondence describes a publicly available database of eye-tracking data, collected on a set of standard video sequences that are frequently used in video compression, processing, and transmission simulations. A unique feature of this database is that it contains eye-tracking data for both the first and second viewings of the sequence. We have made available the uncompressed video sequences and the raw eye-tracking data for each sequence, along with different visualizations of the data and a preliminary analysis based on two well-known visual attention models.**

***Index Terms*—Gaze tracking, video compression.**

I. Introduction

The perceptual coding of video using computational models of visual attention (VA) has been recently recognized as a promising approach to achieve high-performance video compression [1], [2]. The idea behind most of the existing VA-based video coding methods is to encode a small area around the gaze locations with higher quality compared with other less visually important regions [1]. Such a spatial prioritization is supported by the fact that only a small region of several degrees of visual angle (i.e., the fovea) around the center of gaze is perceived with high spatial resolution due to the highly nonuniform distribution of photoreceptors on the human retina [1], [3].

In recent years, various video quality assessment approaches have been proposed based on psychophysical properties of the human visual system [4], [5]. The performance of many video quality assessment methods, however, can be improved by incorporating VA information. The reason is that visual artifacts are more disturbing to a human observer in regions with higher saliency than in other nonsalient regions [6].

In the literature, several computational models of VA have been developed to predict gaze locations in digital images and video [7]–[9]. Although the current VA models provide an easy and cost-effective way for gaze prediction, they are still imperfect. One must consider that human attention prediction is still an open and challenging problem. Ideally, the most accurate approach to find actual gaze locations is to use a gaze-tracking (eye-tracking) device. In a typical gaze-tracking session, the gaze locations of a human observer are recorded when watching a given video clip using a remote screen- or head-mounted eye-tracking system. However, eye trackers are still fairly expensive and are not easily accessible to most researchers.

Manuscript received October 14, 2010; revised February 08, 2011, May 31, 2011 and August 03, 2011; accepted August 03, 2011. Date of publication August 18, 2011; date of current version January 18, 2012. This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) under Grant RGPIN 327249 and in part by the NSERC/Canada Council for the Arts New Media Initiative under Grant STPGP 350740. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ali Bilgin.

The authors are with the School of Engineering Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada (e-mail: hha54@sfu.ca; mario_enriquez@sfu.ca; ibajic@sfu.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2011.2165292

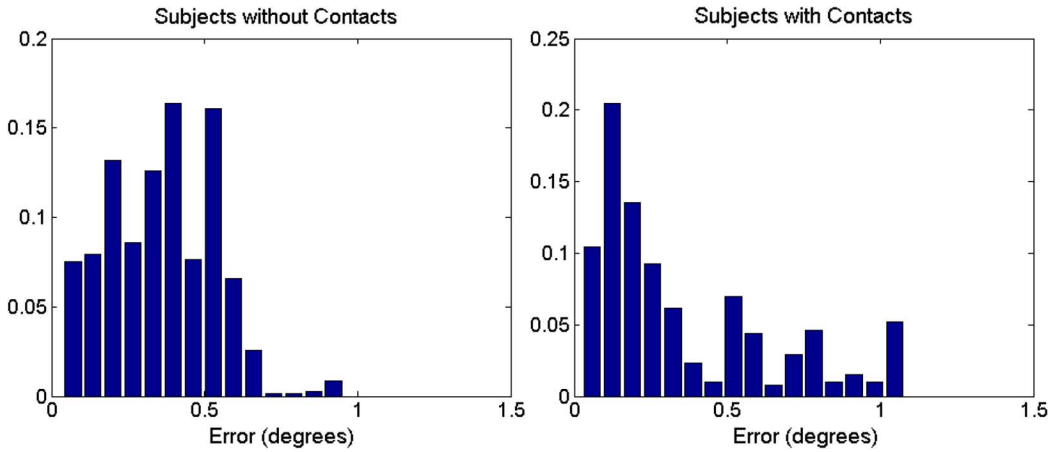

Fig. 1. Normalized histograms of measurement errors (in degrees of visual angle) for subjects (left) without and (right) with contacts.

Over the past two decades, a set of common standard video sequences (for example, *Foreman*, *Flower Garden*, etc.) have been frequently used by many researchers in the field of video compression and quality assessment. Given the growing popularity of VA-based video compression and quality assessment methods, the need for an eye-tracking database for these "standard" sequences is becoming apparent. Recently, a database of nonstandard high-definition video clips alongside their eye-tracking data was presented in [10]. However, to the best of our knowledge, there is no publicly available database of eye-tracking data for the standard sequences previously mentioned.

In this correspondence, we present a publicly available free online database of gaze-tracking data collected on a set of standard video sequences. The database includes 12 standard uncompressed YUV (one luma channel Y and two chroma channels U and V) video sequences in Common Intermediate Format (CIF) resolution with their corresponding eye-tracking data. To generate the eye-tracking data, the sequences were presented to 15 nonexpert subjects two times, and their gaze fixation points were recorded for each frame of each of the 12 selected video sequences using a head-mounted eye-tracking device. The recorded visual gaze location points provide subjects' gaze shifts caused by subjects' overt VA in both the first and second viewings. We present a preliminary analysis of the congruency of the first and second viewings for each sequence. We also compare the accuracy of two popular VA models, i.e., the Itti–Koch–Niebur (IKN) [7] and Itti–Baldi (IB) models [9], on the obtained eye-tracking data. The data set can be utilized for various applications including psychovisual video compression, perceptual video quality assessment, and attention prediction purposes.

The correspondence is organized as follows: Section II describes the video sequences available in the database, the procedure followed for eye-tracking data collection and analysis, and the structure of the database. Results are presented in Section III, and conclusions are drawn in Section IV.

II. Database

A. Video Sequences

To generate the eye-tracking data, we used the following 12 standard video sequences: *Foreman* (300 frames), *Bus* (150 frames), *City* (300 frames), *Crew* (300 frames), *Flower Garden* (250 frames), *Mother and Daughter* (300 frames), *Soccer* (300 frames), *Stefan* (90 frames), *Mobile Calendar* (300 frames), *Harbor* (300 frames), *Hall Monitor* (300 frames), and *Tempete* (260 frames). The sequences were stored in YUV 4:2:0 format at CIF (352×288) resolution and 30 frames per second (fps). These sequences were selected based on the fact that they are frequently used to test video compression, processing, and transmission algorithms. We believe that eye-tracking data for these sequences will facilitate the development and the testing of novel perceptually motivated algorithms and models of VA.

B. Eye Tracker

In our experiments, we used a Locarna "Pt-Mini" head-mounted eye tracker [11]. The advertised accuracy of the Locarna eye tracker is 1 ° or better in the field of view, which is the same as the advertized accuracy of other eye trackers on the market (e.g., Tobii, faceLAB, etc.). This eye tracker is head mounted (using lightweight eyeglasses) and allows subjects to move their head naturally. The eye tracker has two cameras, i.e., one pointing toward the subject's eye ("eye camera" of resolution 320 × 240) and the other pointing forward ("scene camera" of resolution 720 × 480). Both cameras operate at 30 fps. The pupil images captured by the eye camera are analyzed (in real time) by specific image processing techniques implemented in the eye tracker's software in order to find the exact location of the pupil center.

In order to verify the accuracy of the Locarna eye tracker, we performed a simple experiment with ten subjects different from the 15 subjects who participated in the eye-tracking data collection study in Section II-C. Out of these ten subjects, four persons were wearing contact lenses, and the other six persons had normal vision. In the experiment, we first calibrated the eye tracker using nine calibration dots displayed on the same monitor used in the eye-tracking study. The calibration procedure and the experimental setup are described in greater detail in the next section. We then asked the subjects to fixate on 12 test dots of radius 32 pixels and measured their gaze locations. Following that, we displayed a video (*Stefan*) at the center of the screen for about 7 s and finally repeated the gaze point measurement test on the 12 test dots. From the recorded data, we isolated the fixation groups of frames, i.e., those frames where the point of gaze does not move by more than 50 pixels in seven consecutive frames. We then computed the Euclidean distance between the center of each test dot and the gaze location provided by the eye tracker in each frame of the corresponding fixation group. The computed distances were considered as the measurement errors in estimating the gaze location by the eye tracker. The results showed that the mean measurement error was around 0.5 ° of visual angle (with standard deviation up to 0.45 °), confirming the advertised accuracy of the Locarna eye tracker for both subjects with and without contact lenses. We found that the accuracy for subjects with contact lenses was slightly lower than those without contact lenses, but the difference was very small, i.e., less than 0.1 °. Similarly, the accuracy on the second test was somewhat lower than the accuracy on the first test (just after the calibration), but again, the difference was very small, i.e., less than 0.1 °. Fig. 1 shows the normalized histograms of all the obtained measurement errors (in degrees of visual angle) for subjects with

and without contacts. A more detailed description of this experiment is provided in a separate document with the database.

C. Eye-Tracking Data Collection

A total of 15 nonexpert participants (2 women and 13 men) took part in the eye-tracking data collection study. They were recruited by a mass e-mail invitation and were paid $15 for their participation. All of them had normal or corrected-to-normal vision and were asked to wear the Locarna "Pt-Mini" head-mounted eye tracker to determine their gaze direction. The participants consisted of undergraduate and graduate Simon Fraser University students aged between 18 and 30. None of the participants wore spectacles.

To track the movement of the head relative to the screen, two red dots of radius 1 cm were placed in the left and right bottom corners of the screen. Tracking these two dots in the scene camera view made it possible to compensate for the head movement and map the gaze locations onto the screen using a homographic transformation, without a head tracker. Given our experimental setup, subjects did not need to move their head much, and furthermore, they remained at a fixed distance from the screen (80 cm), which allowed for a more precise mapping of the gaze data back onto the screen plane.

In order to map the location of the pupil to the real-world scene (i.e., scene camera view), a calibration matrix is used. In the calibration procedure, the subject is instructed to successively fixate on the center of nine separate dots (of radius 1 cm) represented on a 3×3 grid mounted on a wall at a typical viewing distance of about 80 cm. Each dot has a unique label displayed at the center of it. Each fixation was triggered by a vocal sound instructing the subject to look at the next dot of a specific number. After each fixation, the two images captured by the scene and eye cameras were recorded for further processing. At the end of the calibration procedure, we obtained nine sets of coordinates in the real-world scene (the centers of the nine black dots) and nine pupil locations. A manual inspection was then performed to make sure that the obtained center locations are correct and accurate. Finally, the obtained coordinates were used to compute the calibration matrix using a typical quadratic transformation and singular value decomposition.

In order to verify that the eye tracker remained calibrated throughout the duration of the experiment run, a small crosshair was displayed on a blank screen after presenting each video clip, and the subjects were asked to fixate on the center of the crosshair. Any severe deviation (more than one degree of visual angle) from the true location was used as an out-of-calibration indicator. This allowed us to recalibrate the system in case of any miscalibration.

The study was performed one participant at a time over a period of two days in June 2010. The experiment was run in a quiet room with an ambient light of 200 lx, as recommended in [12] to simulate a "home environment." Each participant was seated in front of a 19-in Samsung SyncMaster 915N color monitor (maximum brightness of 200 cd/m^2) at a distance of 80 cm, and each watched a video with prerecorded instructions on how to complete the experiment before getting started. The monitor resolution was set to 800×600, with a vertical frequency of 75 Hz and a horizontal frequency of 46.875 kHz. Other options were set to their factory default values. The video clips were shown on the screen at twice their normal size so that they would occupy approximately 84% of the screen. The actual size of the video frames was about 25° of diagonal visual angle. The video resolution was increased using nearest-neighbor interpolation. This did not create visible or salient artifacts at the viewing distance of 80 cm, based on our subjective assessment.

The 12 short video sequences were sequentially presented in a fixed order with a 3-s pause in between. During this pause and before the beginning of each video, a small crosshair (centered on the video display area) was presented, and the participants asked to fixate on it. After the 12 videos had been presented, participants then had a 2-min break after which the 12 videos were again presented. The participants were asked to naturally look at the videos and were not given any instructions as to what to look for in the sequences.

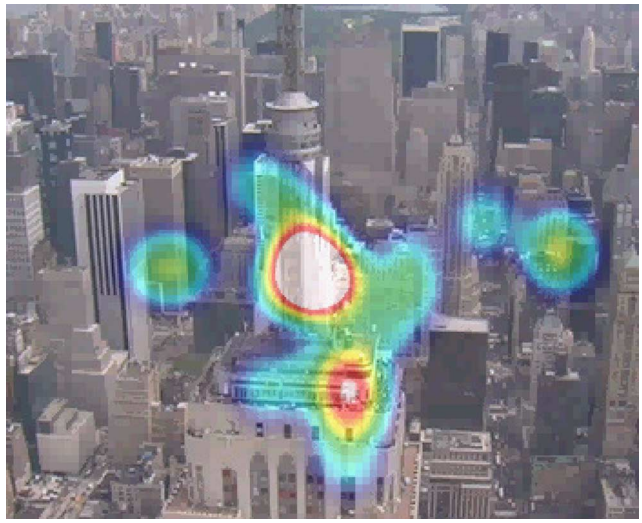

Fig. 2. Heat map visualization of *City*.

D. Gaze Data Visualization

The collected raw gaze data were analyzed, mapped from the head-mounted eye tracker onto the video plane and stored in a comma-separated value (CSV) file format. This file contains the frame-by-frame pixelwise x- and y-coordinates (measured from the bottom left corner) of the gaze location for each participant and each of the video sequences. All the obtained gaze data were both manually and automatically inspected to ensure that they are fairly reliable. Each gaze location stored in the mentioned CSV files was flagged as either correct $(\text{flag} = 1)$ or incorrect $(\text{flag} = 0)$ in a separate CSV mask file. We flagged as invalid those gaze locations that satisfied the following conditions:

1) The gaze location is out of frame boundaries.
2) The gaze location is at the frame boundaries or very close to the frame boundaries (within 5 pixels).
3) The gaze location remains constant for N consecutive frames. The value of N can be arbitrarily set by the user through the MATLAB code of the flagging procedure, which is provided with the database. The default value of N was set to 30 frames (i.e., 1 s).
4) The gaze location remains constant in all frames.

The gaze data were also represented in two different visualizations for each video sequence, i.e., a moving heat map and a gaze plot comparing participants' first and second viewings of the sequences. In the heat map visualization (see Fig. 2), the areas of the video that received the most VA are presented in white, followed by red, yellow green, and blue as VA dropped. The heat maps were generated from the valid raw gaze location points collected for all participants based on the characteristics of the fovea. In each frame, we create a circular area with values following a Gaussian distribution around the gaze location of each participant. This Gaussian models the nonuniform distribution of the photoreceptors on the retina (i.e., the eccentricity of the fovea). The width of the Gaussian was set to 2° of visual angle, which translates to 64 pixels in our case. The accumulation of the obtained Gaussian values resulted in the heat map for that frame. In the gaze plot visualization (see Fig. 3), a pair of connected circles represent where each participant looked at the sequences the first and second times they were presented to them. Each participant's gaze location for the first and second views

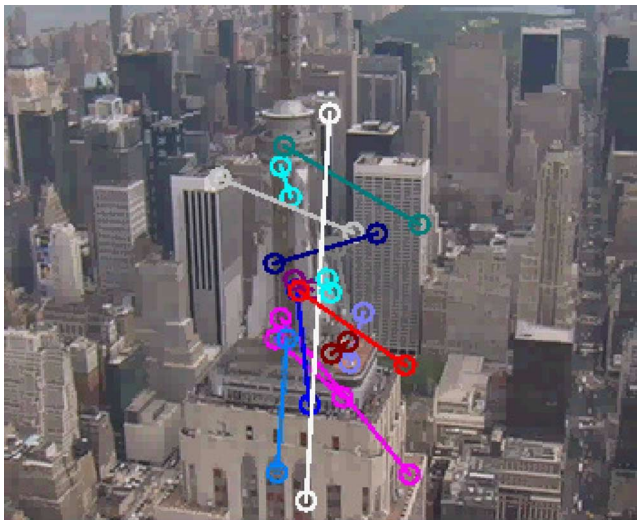

Fig. 3. Gaze plot visualization comparing the first and second viewings of *City*.

is represented in a different color. These data were collected in an effort to determine if a person who had just seen a particular video and was thus familiar with it would look at the same locations when viewing it a second time.

E. Database Location, Structure, and Accessibility

The database is available online at www.sfu.ca/~ibajic/datasets.html. Each of the 12 video sequences is stored in a separate folder that contains the following.

1) Original uncompressed sequences in YUV 4:2:0 format.
2) Heat map visualization (-heatmap) video clips in compressed Audio Video Interleave (AVI) format, similar to Fig. 2.
3) First- and second-view visualization (-1vs2) video clips in compressed AVI format, similar to Fig. 3.
4) A CSV file containing the x- and y-coordinates for each participant's first and second viewings for each frame of each video sequence.
5) A CSV file containing the binary flag matrix (-Mask) for each frame of each video sequence.
6) A number of MATLAB functions to generate and visualize the heat maps and gaze data, as well as a user manual for the code.
7) A brochure for the employed eye tracker (Pt-Mini) and a number of white papers and technical papers, which are also accessible at http://www.locarna.com/docs/. We also provided our data for measuring the actual accuracy of Locarna's eye tracker in a separate document with the database.

III. Results

A. Congruency of the First Viewing Versus the Second Viewing

It is natural to ask whether people who view a particular video multiple times look at the same locations each time they view it. We hypothesized that this would not always be the case. In other words, we expect that, in many cases, people would tend to shift their gaze to different locations each time they view a particular video clip. We thus collected gaze location data for two sequential viewings of each sequence in our database in order to corroborate this hypothesis.

The gaze tracking data allowed us to compare where the participants' gaze was directed for each of the sequences the first time participants saw them, as well as when they were viewed a second time. In each frame, there is a gaze location for the first and second viewings for each participant. Visualizations of the gaze locations for the first and second viewings (similar to Fig. 3) are also made available in the database. As anticipated, there was a notable difference in the locations of the participants' gaze for the first and second viewings. We computed the Euclidean distance between participants' gaze location on the first and second viewings, and then averaged those distances across different participants. The average distance for each of the video sequences is presented in Table I, both in terms of pixels and of the percentage of the size of the CIF frame diagonal, which is $\sqrt{352^2 + 288^2} \approx 454.8$. As shown in Table I, the average distance between the gaze locations could be as large as a quarter of the frame. Note however that the variability between the first and second viewings is likely to be influenced by the amount of time elapsed between the two viewings. Our main goal here is to raise awareness among the readers that such variability may exist, rather than provide an accurate model for such variability.

TABLE I
Average Distance Between Gaze Locations in the First and Second Viewings

Sequence	Average distance	
	Pixels	% of diagonal
Bus	91.52	20.12
City	72.35	15.91
Crew	99.23	21.82
Foreman	46.18	10.15
Flower Garden	92.74	20.39
Hall Monitor	67.62	14.87
Harbor	78.27	17.21
Mobile Calendar	145.95	32.09
Mother & Daughter	70.03	15.40
Soccer	82.62	18.17
Stefan	41.38	9.10
Tempete	65.60	14.42

The shift in gaze locations was particularly evident for sequences such as *Crew*, *Flower Garden*, and *Mobile Calendar*, where there are numerous objects (none of which are strongly dominant) that compete for the viewer's attention. Here, the word "dominant" refers to our subjective impression of what was dominant in a particular sequence or set of frames (e.g., the face in the initial part of *Foreman*). In cases where there was no single dominant object, the viewers tended to shift their gaze to a different object in the second viewing. On the other hand, in sequences with a single dominant object of interest, such as *City* and *Stefan*, the differences in gaze locations were related to the size of the object; the small object (the tennis player) in *Stefan* gave rise to a small difference, whereas the large object (the central building) in *City* gave rise to a large difference. Bear in mind that, in these sequences, as in those with multiple objects of interest, the gaze location did change between the first and second viewings but usually remained within the dominant object of interest, as illustrated in Fig. 3.

Certain sequences presented interesting patterns when comparing the first and second viewings. An example is *Foreman*, where the distance between gaze locations of the first and second viewings varied as the sequence progressed (see Fig. 4). In the beginning of the sequence, when there is a face present in the video, the gaze was concentrated on this face in both viewings. Hence, the gaze location difference in this part of the sequence was relatively small. As the sequence progresses, the camera pans to show a construction site, and there was a larger disparity between participants' gaze locations for the first and second viewings. In Fig. 4, we can see a definite trend; the gaze distance between the first and second viewings increases as the sequence

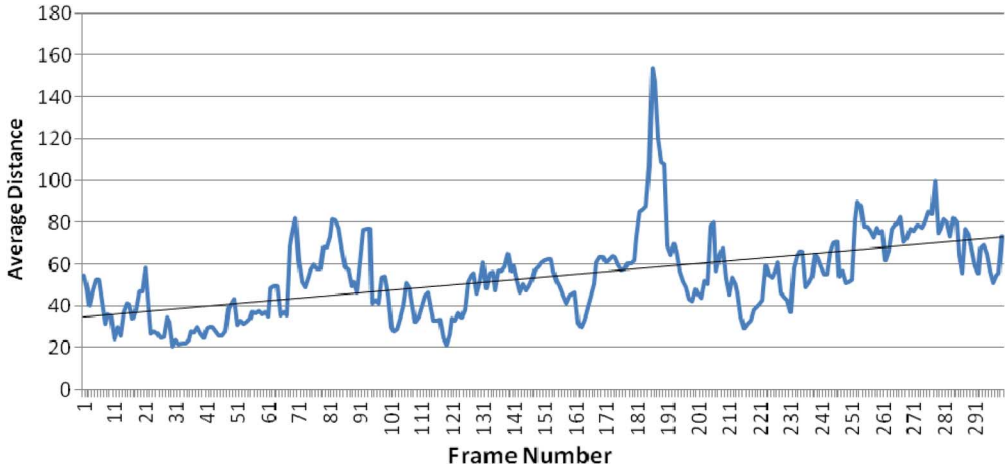

Fig. 4. Average distance (in pixels) between gaze locations in the first and second viewings for *Foreman*, presented frame by frame.

progresses and peaks between frames 180 and 190 when the camera starts to pan to the right.

B. Accuracy of Two Popular Attention Models

One of the possible uses of this database is in testing prediction models of human attention. To show how this can be done, we utilized the gaze location data to determine the accuracy of two well-known VA prediction models, i.e., the IKN [7], and IB [8], [9] models. Using the gaze location data, we were able to determine how well these two attention prediction models perform on each of the sequences in the database.

For each frame, both models produce a saliency map $s(x, y)$ that contains a predicted attention potential value (ranging from 0 to 255) for each pixel. However, they do not produce the same total saliency in each frame. In other words, $\Sigma s(x, y)$ is, in general, different for the two models. In order to have a fair comparison between the two models, we normalized saliency values as follows:

$$s'(x, y) = \frac{s(x, y)}{\sum s(x, y)} N_{\text{pixel}} \tag{1}$$

where N_{pixel} is the number of pixels in the frame. In our case (CIF resolution), $N_{\text{pixel}} = 352 \times 288 = 101,376$. After this normalization, both models produce the same total normalized saliency per frame, i.e., $\Sigma s'(x, y) = N_{\text{pixel}}$ for both models.

Using the normalized saliency maps, we proceeded to calculate the accuracy of the models by adding the normalized values of every pixel where a gaze was directed. If (x_i, y_i) is the pixel where the ith viewer's gaze was directed in a particular frame, the accuracy score of a model for that frame was computed as

$$\text{Score} = \sum_{i=1}^{15} \sum_{x,y} w_{x_i,y_i}(x, y) s'(x, y) \tag{2}$$

where i goes from 1 to 15 because there were 15 viewers in our study and $w_{x_i,y_i}(x, y)$ is a 2-D Gaussian function of width δ centered at the ith gaze location (x_i, y_i), defined as

$$w_{x_i,y_i}(x, y) = \frac{1}{2\pi\delta^2} \exp\left(-\left(\frac{(x - x_i)^2 + (y - y_i)^2}{2\delta^2}\right)\right). \tag{3}$$

In our calculations, we set $\delta = 64$ pixels, which corresponds to 2° of visual angle. The average accuracy scores (over all frames) for each sequence are presented in Table II for both the first and second viewings. To examine whether the difference in the average scores between the IKN model and the IB model is statistically significant, we performed a paired t-test [13] on the frame-by-frame scores for each sequence and each viewing. The null hypothesis was that the scores of both models come from the distributions with the same mean. Based on the results, the null hypothesis was rejected (at the 5% significance level) in both viewings for all sequences. The obtained p-values (probability that the null hypothesis holds) were less than 10^{-6}, except for *Foreman* for the first viewing ($p = 0.000055$) and *Mother and Daughter* for the first viewing ($p = 0.002416$) and the second viewing ($p = 0.000027$). Therefore, based on our data, the difference in the average accuracy scores of two models was highly statistically significant in each case, and the model with the higher average score on a particular sequence can be considered more accurate on that sequence.

TABLE II
AVERAGE ACCURACY SCORE FOR PREDICTING GAZE LOCATION IN THE FIRST AND SECOND VIEWINGS

Sequence	IKN model [7]		IB model [9]	
	View-1	View-2	View-1	View-2
Bus	23.50	20.83	15.77	14.94
City	9.12	10.67	13.39	14.61
Crew	19.60	18.95	16.16	16.53
Foreman	28.99	30.01	25.15	24.50
Flower Garden	51.31	48.93	19.35	20.48
Hall Monitor	81.62	83.71	59.35	59.00
Harbor	31.12	36.81	20.73	23.71
Mobile Calendar	44.74	40.21	21.40	21.17
Mother & Daughter	32.63	35.13	29.48	30.96
Soccer	31.19	29.12	23.04	22.62
Stefan	67.42	66.91	51.26	48.69
Tempete	34.33	34.05	28.02	28.80

Several observations can be made from the data in Table II. First, the IKN model [7] showed better accuracy than the IB model [9] in 11 out of 12 sequences, whereas the IB model was more accurate in just one case (*City*). This finding is somewhat surprising, given that the IB model is more recent [9] and claimed to be an improvement over the IKN model.

We also ran the t-test to determine if there is any statistical basis for claiming that a particular model had better accuracy on the first or second viewing. The results were mixed. At the 5% significance level, the IKN model showed better accuracy on the first viewing for three sequences (*Bus*, *Mobile Calendar*, and *Soccer*) and on the second viewing for four sequences (*City*, *Foreman*, *Harbor*, and *Mother and Daughter*), whereas for the remaining sequences, the difference was

TABLE III
AVERAGE ACCURACY SCORE FOR UNIFORMLY SPREAD SALIENCY IN THE FIRST AND SECOND VIEWINGS

Sequence	View-1	View-2
Bus	14.55	14.63
City	14.57	14.44
Crew	13.21	13.62
Foreman	14.67	14.66
Flower Garden	14.48	14.55
Hall Monitor	13.48	14.65
Harbor	14.50	13.86
Mobile Calendar	14.26	14.41
Mother & Daughter	14.47	14.73
Soccer	12.91	14.26
Stefan	14.82	13.37
Tempete	14.42	13.79

not statistically significant. The IB model showed better accuracy on the first viewing for three sequences (*Bus*, *Foreman*, and *Stefan*) and on the second viewing for five sequences (*City*, *Flower Garden*, *Harbor*, *Mother and Daughter*, and *Tempete*), whereas there was no statistically significant difference on other sequences. Overall, according to our data, both models seem to be roughly equally suitable for the first and second viewings.

Finally, while Table II provides the data to compare the relative accuracy of the two models, it is natural to ask how accurate these models are in absolute terms. One way to tackle this question is to compare these models with uniformly spread saliency. Suppose we assign the same saliency to each pixel, i.e., $s_u(x, y) = 1$ for all (x, y). With such uniformly spread saliency, the total normalized saliency is the same as for the two aforementioned models ($\Sigma s_u(x, y) = N_{\text{pixel}}$); thus, a fair comparison is possible. The average accuracy scores for such uniformly spread saliency computed using (2) are listed in Table III. One could argue that, if a particular model does not produce a score significantly above that listed in Table III, it is really not any more accurate than uniformly spread saliency. Again, we used the t-test to assess whether a particular model's score was significantly better (or worse) than that produced by uniform saliency. The IKN model's score on *City* was significantly lower than that produced by uniform saliency on both views of *City* and was significantly better in all other cases. Meanwhile, the IB model's score was significantly lower than that produced by uniform saliency on the first view of *City*, whereas there was no significant difference on the second view of *Bus* and *City*. In all other cases, the IB model had a significantly higher accuracy than uniform saliency. Overall, the scores were the highest (and the models were most accurate) on sequences with few dominant moving objects, such as *Stefan* and *Hall Monitor*, whereas both models showed lower accuracy on sequences where there were multiple objects competing for viewers' attention. One surprising finding was perhaps that both models had a problem with the sequence *City*, which contains a single large dominant object (the central building). A possible reason may be that this dominant object has a similar color and texture distribution as the background and appears relatively static relative to the background as the camera revolves around it; thus, it is not being picked up by the contrast analysis modules employed by both models.

IV. CONCLUSION

As video compression and processing algorithms evolve to incorporate models of human perception and attention, it becomes imperative to have the tools to test them. In this correspondence, we have presented an eye-tracking database for a set of 12 standard CIF video sequences commonly used in the literature to compare video compression and processing algorithms. The database itself is available for public download at www.sfu.ca/~ibajic/datasets.html. We have described the procedure followed in order to produce this database and have also presented a preliminary analysis of the obtained data. An interesting finding stemming from the data is that gaze locations tend to be different in different viewings of the same video, which may have implications in the design of compression algorithms intended for one-time viewing (e.g., videoconference), as compared with those intended for multiple viewings (e.g., DVD and Blu-ray). We have also shown how the data can be used to compare models of VA in terms of their accuracy in predicting gaze locations.

Some of the limitations of the database include the accuracy of the data, which is limited to about 1 ° in the field of view by the eye-tracking equipment and setup, and the number of video sequences and participants, both of which should ideally be as high as possible. Furthermore, the distances between participants' gaze locations in the first and second viewings should be taken with a grain of salt since they likely depend on the amount of time elapsed between the viewings. Our data is mainly intended to raise awareness that such variability may exist. Nonetheless, despite these limitations, we hope the data will be useful to the research community.

REFERENCES

[1] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.
[2] Z. Chen, W. Lin, and N. K. Ngan, "Perceptual video coding: Challenges and approaches," in *Proc. IEEE ICME*, Singapore, Jul. 2010, pp. 784–789.
[3] B. Wandell, *Foundations of Vision*. Sunderland, MA: Sinauer, 1995.
[4] *Digital Video Image Quality and Perceptual Coding*, H. R. Wu and K. R. Rao, Eds. New York: CRC Press, 2006.
[5] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[6] U. Engelke, M. Barkowsky, P. Le Callet, and H.-J. Zepernick, "Modelling saliency awareness for objective video quality assessment," in *Proc. 2nd Int. Workshop QoMEX*, 2010, pp. 212–217.
[7] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[8] L. Itti and P. F. Baldi, "A principled approach to detecting surprising events in video," in *Proc. IEEE Conf. CVPR*, San Diego, CA, Jun. 2005, pp. 631–637.
[9] L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," *Vis. Res.*, vol. 49, no. 10, pp. 1295–1306, Jun. 2009.
[10] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *Image Vis. Comput.*, vol. 29, no. 1, pp. 1–14, Jan. 2011.
[11] Locarna Systems [Online]. Available: http://www.locarna.com
[12] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, ITU-R BT.500-11, 2002.
[13] J. T. McClave and T. Sincich, *Statistics*, 9th ed. Upper Saddle River, NJ: Prentice-Hall, 2003.