Predicting Eyes' Fixations in Movie Videos: Visual Saliency Experiments on a New Eye-Tracking Database

Petros Koutras, Athanasios Katsamanis, and Petros Maragos

School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece 15773
{pkoutras,nkatsam,maragos}@cs.ntua.gr

Abstract. In this paper we describe the newly created eye tracking annotated database *Eye-Tracking Movie Database ETMD* and give some preliminary experimental results on this dataset using our new visual saliency frontend. We have developed a database with eye-tracking human annotation that comprises video clips from Hollywood movies, which are longer in duration than the existing databases' videos and include more complex semantics. Our proposed visual saliency frontend is based on both low-level features, such as intensity, color and spatio-temporal energy, and face detection results and provides a single saliency volume map. The described new eye-tracking database can become useful in many applications while our computational frontend shows to be promising as it gave good results on predicting the eye's fixation according to certain metrics.

Keywords: Eye-tracking Database, Visual Saliency, Spatio-Temporal Visual Frontend, 3D Gabor Filters, Lab Color Space.

1 Introduction

Visual attention is a mechanism employed by biological vision systems for selecting the most salient spatio-temporal regions from a visual stimuli. Attention may have two modes, a top-down task-driven, and a bottom-up data-driven, and so there is often a confusion between attention and visual saliency, which is a bottom-up process and is based on low level sensory cues of a given stimulus. On the other hand, visual attention includes many high level topics, such as semantics, memory, object searching, task demands or expectations.

The development of computational frameworks that model visual attention is critical for designing human-computer interaction systems, as they can select only the most important regions from a large amount of visual data and then perform more complex and demanding processes. Attention models can be directly used for movie summarization, by producing video skims, or by constituting a visual frontend for many other applications, such as object and action recognition. Eyes' fixation prediction over different stimulus appeared to be a

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 183–194, 2014.

widely used way for analyzing and evaluating visual attention models. Although many databases with eye tracking data are available [1], most of them contain only static images, as the first saliency models were based only on static cues.

The purpose of this paper is to describe the newly created eye tracking annotated database *Eye-Tracking Movie Database ETMD* and give some preliminary experimental results on this dataset using our new visual attention frontend, that is based on both low level streams and mid-level cues (i.e. face detection). The existing eye-tracked video databases, in most cases contain very short videos with simple semantic contain. In our effort to deal with more complex problems, such as movie summarization [2], we have developed a database with eye-tracking human annotation, which comprises video clips from Hollywood movies, which are longer in duration and include more complex semantics.

In the second part of the paper, we describe ways for predicting the eye's fixations in movie videos and give preliminary results from our computational framework for visual saliency estimation. Our proposed visual saliency frontend is based on both low-level features [3–5], such as intensity, color and motion, and face detection results and provides a single saliency volume map. We quantitatively evaluate our results according to 3 evaluation scores, as they are described in [6]: Correlation Coefficient, Normalized Scanpath Saliency, Shuffled Area Under Curve. The described new eye-tracking database can become useful in many applications while our computational frontend shows to be promising as it gave good results on predicting the eye's fixation according to all three employed metrics.

2 Eye-Tracking Movie Database (ETMD)

We have developed a new database comprising video clips from Hollywood movies which we have enriched with eye-tracking human annotation: the Eye-Tracking Movie Database (ETMD). Specifically, we cut 2 short video clips (about 3-3.5 minutes) from each one of six Oscar-winning movies of various genres: Chicago (CHI), Crash (CRA), Departed (DEP), Finding Nemo (FNE), Gladiator (GLA) , Lord of the Rings - The return of the King (LOR). We have tried to include scenes with high motion and action as well as dialogues. These clips were annotated with eye-tracking data by 10 different people (annotation data from at least 8 people were collected for each clip). The volunteers viewed the videos both in grayscale and in color, while an eye-tracking system recorded their eyes fixations on the screen.

Specifically, we have used the commercial Eye Tracking System TM3 provided by EyeTechDS. This device uses a camera with infrared light and provides a real time continuous gaze estimation, defined as fixation points on the screen. The tracker's rate has been limited by the video frame rate in order to have one fixation point pair per frame. For a visual attention problem a weighted average between two eye fixations is provided, which is defined either by the mean, if both eyes are found by the eye-tracker, or only by the detected eye's fixation. If neither eye is detected or the fixations lie out of screen boundaries, fixation gets a

(a) CHI Clip 1 (b) CHI Clip 2 (c) CRA Clip 1 (d) CRA Clip 1

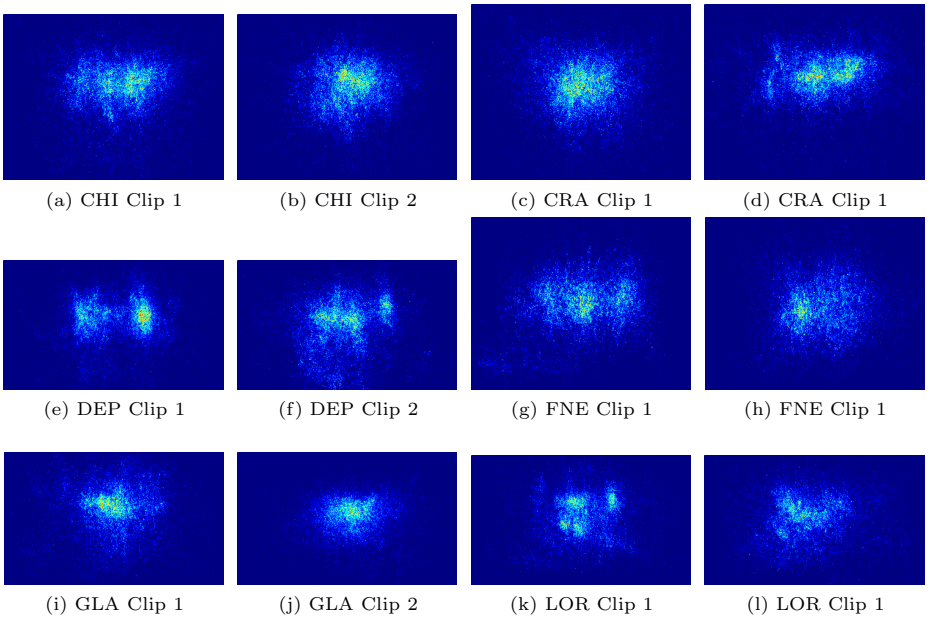
(e) DEP Clip 1 (f) DEP Clip 2 (g) FNE Clip 1 (h) FNE Clip 1

(i) GLA Clip 1 (j) GLA Clip 2 (k) LOR Clip 1 (l) LOR Clip 1

Fig. 1. Examples of the fixation points at frame no. 500 for each of the 12 movie clips. With green + are the fixations points over the color version of each clip, while with red ∗ are the points for the grayscale version. Best viewed in color.

zero value. The eye-tracking system also provides some additional measurements, such as pupil and glints positions and pupil diameter.

Figure 1 shows examples of the fixation points at frame no. 500 for each of the 12 movie clips. We see that in most cases the fixation points of all viewers lie in general close to each other. The fixations for the grayscale version of each clip are highly correlated with the fixation points over the color video as well. Figure 2 shows heatmaps of all fixations over each movie clip for the ETMD database. We see that the most points are clustered at the center of the image which shows that movie clips are highly center-biased. Analyzing the eye-tracking data we provide in Table 1 useful statistics for the database, such as frames number, total duration and valid fixation points per frame, and find correlations among the different viewers and between the color and grayscale version of each movie clip. We see that the fixations are generally correlated both between the different users and the version (color or grayscale) of each movie clip. However, in some movies, such as CHI, the fixations data are highly correlated while other clips (FNE Clip 2, LOR Clip 2) have lower correlation values.



(a) CHI Clip 1 (b) CHI Clip 2 (c) CRA Clip 1 (d) CRA Clip 1

(e) DEP Clip 1 (f) DEP Clip 2 (g) FNE Clip 1 (h) FNE Clip 1

(i) GLA Clip 1 (j) GLA Clip 2 (k) LOR Clip 1 (l) LOR Clip 1

Fig. 2. Heatmaps of all fixations over each movie clip for the ETMD database. We can see that the most points are clustered at the center of the image which shows that movie clips are highly center-biased. Best viewed in color.

Table 1. Statistics for the Eye Tracking Movie Database

Video Clip Name	Number of Frames	Duration (Minutes)	Number of Viewers	Valid Fixations Number per Frame	Average Correlation between Viewers	Average Correlation between Color and Grayscale version
CHI Clip 1	5075	03:22	10	9.50	0.506	0.495
CHI Clip 2	5241	03:29	9	8.63	0.430	0.484
CRA Clip 1	5221	03:28	10	9.47	0.335	0.310
CRA Clip 2	5079	03:23	9	8.47	0.406	0.467
DEP Clip 1	4828	03:13	10	9.45	0.520	0.548
DEP Clip 2	5495	03:39	9	8.25	0.473	0.534
FNE Clip 1	5069	03:22	9	8.45	0.372	0.371
FNE Clip 2	5083	03:23	8	7.50	0.292	0.294
GLA Clip 1	5290	03:31	9	8.18	0.423	0.407
GLA Clip 2	4995	03:19	8	7.61	0.354	0.443
LOR Clip 1	5116	03:24	9	8.38	0.452	0.431
LOR Clip 2	5152	03:26	8	7.56	0.294	0.283

3 Spatio-Temporal Framework for Visual Saliency

3.1 Overall Process

Our proposed visual saliency frontend is based on both low-level features [3–5], such as intensity, color and motion, and face detection results and provides a single saliency volume map. The overall process is shown in Fig. 3. In the first phase the initial RGB video volume is transformed into Lab space [7] and split into two streams: luminance and color contrast. For the luminance channel

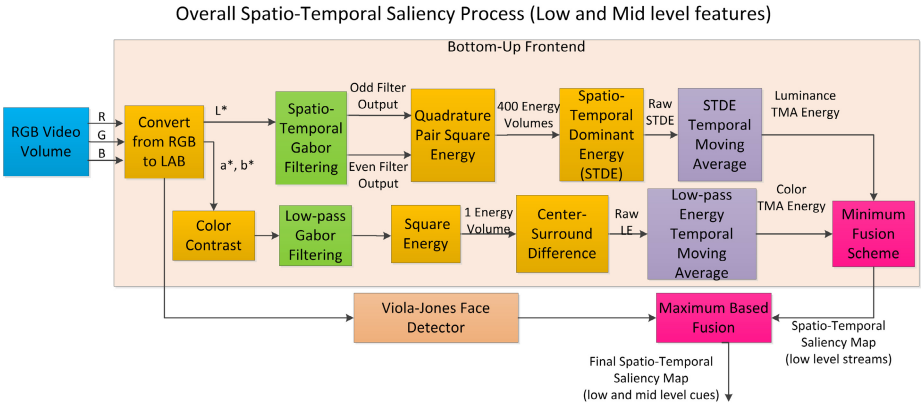


Fig. 3. Overall process for spatio-temporal saliency estimation, which includes both low-level features (i.e. intensity, color and motion) and face detection results and provides a single saliency volume map

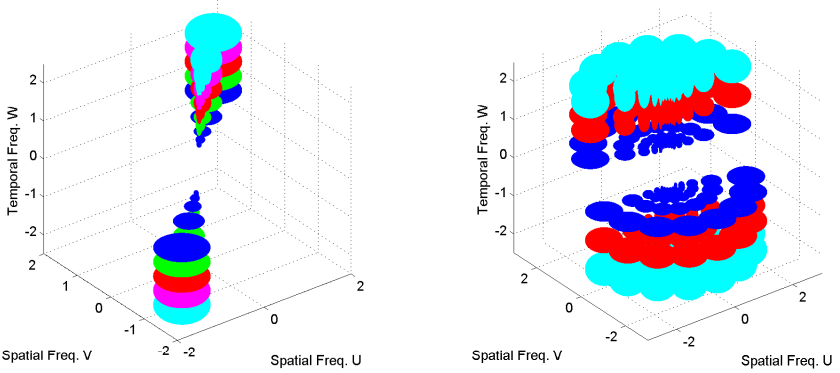
we apply spatio-temporal Gabor filtering [8, 9], followed by Dominant Energy Selection [10, 11], while for the color contrast stream we apply a simple lowpass 3D Gaussian filter followed by a center-surround difference. For integrating the results from the Viola-Jones face detector [12] to a final visual attention map we can use either a maximum or use the face detector's estimation only for the frames that contain faces.

3.2 Spatio-Temporal Bottom-Up Frontend

Preprocessing and Color Modeling. For the color modeling we use the CIE-Lab color space because in this space luminance and chromaticity components can be well separated while it has the additional property to be perceptually uniform. The CIE-Lab space is created from a nonlinear transformation on CIE-XYZ color space [13]. Then the three CIE-Lab components (L^*, a^*, b^*) can be computed by a non-linear transformation of the CIE tristimulus values (X, Y, Z). In the resulting video volume $\mathbf{I}_{Lab}(x, y, t)$ the L^* component expresses the perceptual response to luminance, while a^*, b^* describe differences between red-green and yellow-blue colors respectively. So, the CIE-Lab space includes ideas from the color-opponent theory, which was widely used in visual saliency models and was usually implemented in the RGB color space [5]. In order to describe the color changes in videos by a single measure with positive values, we use the following color contrast operator based on the chromaticity components (a^*, b^*):

$$C_{ab}(x, y, t) = |a^*(x, y, t)| + |b^*(x, y, t)| \tag{1}$$

3D Gabor Filtering. For the filtering process of the video's luminance we choose to use oriented Gabor filters in a spatio-temporal version, due to their



(a) Spatio-Temporal Filterbank at 5 different spatial scales, 1 of 8 orientation and 5 temporal frequencies

(b) Spatio-Temporal Filterbank at 5 different spatial scales, 8 spatial orientations and 3 of 5 temporal frequencies

Fig. 4. Isosurfaces of the 3D Spatio-Temporal Filterbank. Isosurfaces correspond at 70%-peak bandwidth magnitude while different colors are used for different temporal frequencies. We can see that the symmetric lobe of each filter appeared at the plane defined by the temporal frequency $-\omega_{t_0}$ in contrast with the 2D case. We also note that the bandwidth of each filter changes depending on the spatial scale and temporal frequency.

biological plausibility and their uncertainty-based optimality [14,15]. Specifically, we apply quadrature pairs of 3D (spatio-temporal) Gabor filters with identical central frequencies and bandwidth. These filters can arise from 1D Gabor filters [14] in a similar way as Daugman proposed 2D Oriented Gabor Filters [16]. An 1D complex Gabor filter consists of a complex sine wave modulated by a Gaussian window. Its impulse response with unity norm has the form:

$$g(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t^2}{2\sigma^2}\right) \exp(j\omega_{t_0} t) = g_c(t) + j g_s(t) \tag{2}$$

The above complex filter can be split into one odd(sin)-phase ($g_s(t)$) and one even(cos)-phase ($g_c(t)$) filters, which form a quadrature pair filter.

The 3D Gabor extension (as for example used for optical flow in [9]) yields an *even (cos)* 3D Gabor filter whose impulse response is:

$$g_c(x,y,t) = \frac{1}{(2\pi)^{3/2}\sigma_x\sigma_y\sigma_t} \exp\left[-\left(\frac{x^2}{2\sigma_x^2} + \frac{y^2}{2\sigma_y^2} + \frac{t^2}{2\sigma_t^2}\right)\right] \cdot \cos(\omega_{x_0} x + \omega_{y_0} y + \omega_{t_0} t) \tag{3}$$

where $\omega_{x_0}, \omega_{y_0}, \omega_{t_0}$ are the spatial and temporal angular center frequencies and $\sigma_x, \sigma_y, \sigma_t$ are the standard deviations of the 3D Gaussian envelope. Similarly for the impulse response of *odd (sin)* filter which we denote by $g_s(x,y,t)$.

The frequency response of the even (cos) 3D Gabor Filter will have the form:

$$\begin{aligned} G_c(\omega_x, \omega_y, \omega_t) = & \frac{1}{2} \exp[-(\sigma_x^2(\omega_x - \omega_{x_0})^2/2 \\ & + \sigma_y^2(\omega_y - \omega_{y_0})^2/2 + \sigma_t^2(\omega_t - \omega_{t_0})^2/2)] \\ & + \frac{1}{2} \exp[-(\sigma_x^2(\omega_x + \omega_{x_0})^2/2 \\ & + \sigma_y^2(\omega_y + \omega_{y_0})^2/2 + \sigma_t^2(\omega_t + \omega_{t_0})^2/2)] \end{aligned} \quad (4)$$

Thus, the frequency response of an even (cos) Gabor filter consists of two Gaussian ellipsoids symmetrically placed at frequencies $(\omega_{x_0}, \omega_{y_0}, \omega_{t_0})$ and $(-\omega_{x_0}, -\omega_{y_0}, -\omega_{t_0})$. Figure 4 shows isosurfaces of the 3D spatio-temporal filterbank. Note that the symmetric lobes of each filter appear at the plane defined by the temporal frequency $-\omega_{t_0}$ in contrast with the 2D case. So, if we want to cover the spatial frequency plane at each temporal frequency we must include in our filterbank both positive and negative temporal frequencies. Further, the bandwidth of each filter varies with the spatial scale and temporal frequency.

The 3D filtering is a time consuming process due to the complexity of all required 3D convolutions. However, Gabor filters are separable [9], which means that we can filter each dimension separately using an impulse response having the form (2). In this way, we apply only 1D convolutions instead of 3D, which increases the efficiency of the computations. For an image of size $n \times n \times n$ and a convolution kernel of $m \times m \times m$ the complexity is reduced from $\mathcal{O}(n^3 \cdot m^3)$ that is required for 3D convolutions to $\mathcal{O}(3n^3 \cdot m)$ that is required for three separable 1D convolutions.

For the spatio-temporal filterbank we used $K = 400$ Gabor filters (isotropic in the spatial components) which are arranged in five spatial scales, eight spatial orientations and ten temporal frequencies. The spatial scales and orientations are selected to cover a squared 2D frequency plane in a similar way to the design by Havlicek et al. [11]. We also use ten temporal Gabor filters, five at positive and five at negative center frequencies due to the 3D spectrum symmetries. Figure 4 shows spatio-temporal views of our design of this 3D filterbank.

Finally, for the low-pass color filtering we use both spatial and temporal zero frequencies which makes the Gabor filter gaussian.

Postprocessing. After the filtering process, for each filter i we obtain a quadrature pair output $(y_s^{3D}(x, y, t), y_c^{3D}(x, y, t))$ which corresponds to the even- and odd-phase 3D filter outputs. We can compute the total Gabor energy, which is invariant to the phase of the input, by taking the sum of the squared energy of these two outputs:

$$STE_i(x, y, t) = \left(y_s^{3D}(x, y, t)\right)^2 + \left(y_c^{3D}(x, y, t)\right)^2 \quad (5)$$

After this step we have 400 *energy volumes* for the spatio-temporal part (STE_i) and one for the lowpass color filter (LE_0). In order to form one volume for the luminance modality we apply the first step of *Dominant Component Analysis* to

spatio-temporal volumes. Specifically, for each voxel (x, y, t) we keep its maximum value between all existing energy volumes: $STDE = \max_{1 \leq i \leq K} STE_i$ For the lowpass color energy we apply a simple center-surround difference in order to enhance regions which have significantly different values from their background. At each voxel of the video segment we subtract from its lowpass energy value $(LE_0(x, y, t))$ the mean value of the entire energy volume:

$$LE(x, y, t) = \left| LE_0(x, y, t) - \overline{LE}_0(x, y, t) \right| \tag{6}$$

Finally, these energy volumes can become further smoothed by applying a *temporal moving average* (TMA). Thus, each frame energy is computed as the mean inside a temporal window which includes N successive frames whose total duration is 1 second. In this way, we integrate visual events which take place close in time, in a similar way that humans are believed to do. A spatial smoothing with a dilation operator can also by applied, in order to find more compact and dense energy regions.

4 Evaluation on ETMD

4.1 Evaluation Measures

We have tried to keep the same evaluation framework as in [6]. We compared our results according to the three evaluation scores, as they are described in [6]: Correlation Coefficient, Normalized Scanpath Saliency, Area Under Curve. Despite the spatio-temporal character of our method these three measures are computed at each frame separately.

Correlation Coefficient (CC) expresses the relationship between the model's saliency map and the saliency map created by centering a 2D gaussian, with standard deviation 10 pixels, at each viewer's eye fixation.

Normalized Scanpath Saliency (NSS) is computed on the model's saliency map, after zero mean normalization and unit standardization, and shows how many times over the whole map's average is the model's saliency value at each human fixation. For NSS computation we subtract from the saliency map its average value and then divide with its standard deviation. Then the values of this normalized saliency map at each viewer fixation position consist the NSS values. As final NSS value we take the mean over all viewers fixations, while a negative NSS shows that the model cannot predict saliency region better than random selection.

Area Under Curve (AUC) is defined by the area under the Receiver Operating Characteristic (ROC) curve [17]. For our evaluation we consider saliency as a binary classification problem, in which saliency regions are included in the positive class while non-salient pixels form the negative set and model's saliency values are the single features. After thresholding these values we take an ROC curve and subsequently the AUC measure. Instead of selecting the negative points uniformly from a video frame we use the *Shuffled AUC*, which can be more robust across center-bias issue. According to shuffled AUC, we select the negative points

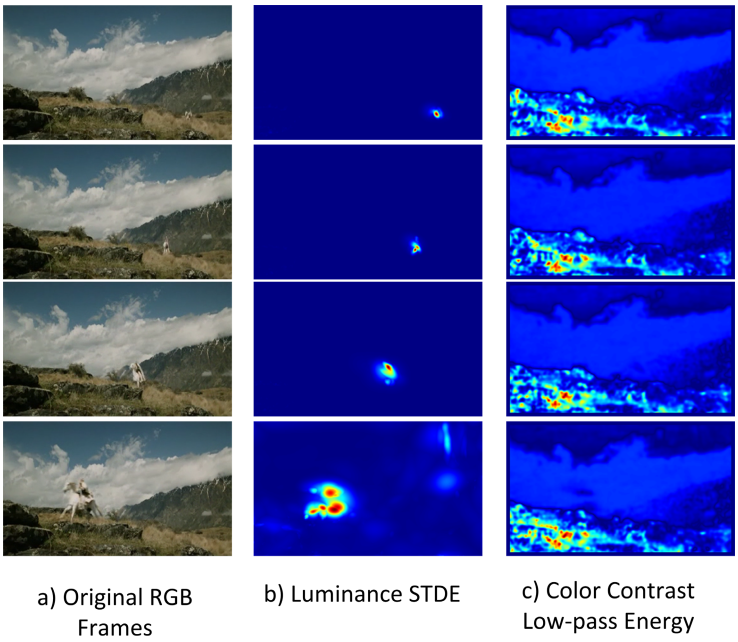


Fig. 5. Example frames of energy volumes computed using our frontend on the *Lord of the Rings* (Clip 1) from our Eye-Tracking Movie Database (ETMD). The galloping horse is perfectly detected by the luminance STDE.

from the union of all viewers' fixations across all other videos except the video for which we compute the AUC. For more details about the above evaluation scores the reader is referred to [1, 6].

4.2 Evaluation Results

We have applied and evaluated our computational model on this novel database. Figure 5 shows example frames of the model's energies computed on the video *Lord of the Rings (LOR)* (Clip 1) from our new Eye-Tracking Movie Database (ETMD). We note that the white galloping horse is detected perfectly by only the luminance $STDE$, since its color information is negligible. The color low-pass energy models static objects or regions in the video sequence, like the rock in the bottom-left and the clouds in the air.

Regarding the feature energy volumes employed we see that luminance $STDE$ and the color low-pass energy using a min fusion performs quite better than using only the $STDE$ energy volume. Moreover, Finally, regarding the grayscale versus color annotation, we saw that the evaluation over color videos yields better results, which indicates that the way color attracts human attention may be predicted more accurately by our model.

Table 2. Evaluation Scores for the Eye-Tracking Movie Database(ETMD) using our **bottom-up frontend**. The employed evaluation measures are Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS) and Shuffled Area Under Curve (AUC). The evaluation of the Luminance $STDE$ was based on Eye-Tracking annotation on both a grayscale and color version of each video.

Evaluation Score	Correlation Coefficient (CC)	Normalized Scanpath Saliency (NSS)	Shuffled Area Under Curve (AUC)
Lum. $STDE$ (Grayscale Annot.)	0.151	0.608	0.611
Lum. $STDE$ (Color Annot.)	0.153	0.632	0.614
MIN(Lum. $STDE$, Color Low-pass)	0.169	0.748	0.635

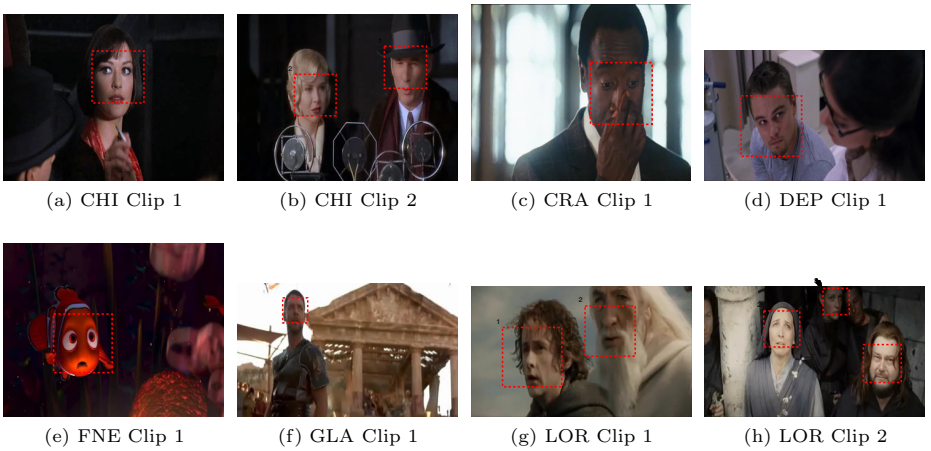
From the database analysis we have seen that in many cases the humans have focused on actors' faces, while their eyes' fixation has also the trend to be center biased. To model these two effect we use two simple methods. The first consists of using a gaussian kernel fixed at the image kernel. The latter provides the use of the Viola-Jones face detector as a saliency estimator only in the frames where people face exist. We have also tried to predict one viewers fixations from the other users eye fixations, as reference results.

4.3 Face Detection

Figure 6 shows examples of the Viola-Jones face detector [12] applied on videos from the new ETMD. We see that in many cases the human faces have been detected accurately while in Fig. 6e the detector find the human-like face of the fish "Nemo". We note that the employed face detector is not very robust with changes in face pose and scale. Thus, it cannot achieve to detect all the faces during a video (low recall) but the obtained results are in most cases true (high precision).

For the fusion of the visual saliency estimation from our bottom-up frontend with the face detection results we can use either a max based method (MAX(Bottom-Up, Face Detection)) or by using the face detector's result for the frames with faces and the BU model for the other frames (Bottom-Up OR Face Detection). We have also applied the face detector and bottom-up model independently only for the frames that contain faces. Table 3 shows the results related with the use of the face detector. We see that face detector improves the final visual saliency result, especially in frames that contain faces, while the two fusion methods' results are very closely.

In this Table are also presented the scores achieved by a Gaussian blob centered at the center of the image as well as the results related with the prediction of each one viewer's fixations from the other users eye fixations. We can see that the Gaussian blob gave just as good results w.r.t. CC and NSS, which confirms the existence of the center-bias effect. Regarding the shuffled AUC it has lower performance since this measure is more suitable for high center-biased database. Regarding the fixation prediction from the other viewers' data, it has achieved high performance w.r.t. CC and NSS measures because even in few occasions humans look at the same direction, this hits very large NN and CC values.



(a) CHI Clip 1 (b) CHI Clip 2 (c) CRA Clip 1 (d) DEP Clip 1

(e) FNE Clip 1 (f) GLA Clip 1 (g) LOR Clip 1 (h) LOR Clip 2

Fig. 6. Examples of Face Detection results over the movie clips from the ETMD database. Best viewed in color.

Table 3. Evaluation Scores for the Eye-Tracking Movie Database(ETMD) using our **bottom-up frontend and the Viola-Jones face detector**. The employed evaluation measures are Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS) and Shuffled Area Under Curve (AUC). There are also presented the scores achieved by the Gaussian blob and the prediction of each one viewer's fixations from the other users eye fixations.

Evaluation Score	Correlation Coefficient (CC)	Normalized Scanpath Saliency (NSS)	Shuffled Area Under Curve (AUC)
Face Detection Only	0.327	1.622	0.807
Bottom-Up (only frames with faces)	0.164	0.719	0.636
Bottom-Up OR Face Detection	0.203	0.933	0.680
MAX(Bottom-Up, Face Detection)	0.201	0.919	0.680
Gaussian Blob	0.197	1.288	0.580
Predict from other viewers' fixations	0.404	2.515	0.617

5 Conclusion

In this paper we presented a new eye-tracking database which comprises video clips from Hollywood movies and eye-tracking data recorded from different viewers. We have also given evaluation results using our proposed spatio-temporal bottom-up frontend for visual saliency estimation. We have also dealt with the problem of "face-biased" movie video by combining the results from our bottom-up saliency frontend with a face detector's estimation. We believe that both the new eye-tracking database and our framework for predicting eye fixations can become useful in many computer applications, such as the producing of movie summaries.

Acknowledgment. The authors wish to thank all the members of the NTUA CVSP Lab who participated in the eye-tracking annotation of the movie database.

This research was supported by the project "COGNIMUSE" which is implemented under the "ARISTEIA" Action of the Operational Program Education and Lifelong Learning and is co-funded by the European Social Fund (ESF) and National Resources.

References

1. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Trans. Pattern Analysis and Machine Intelligence 35(1), 185–207 (2013)
2. Evangelopoulos, G., Zlatintsi, A., Potamianos, A., Maragos, P., Rapantzikos, K., Skoumas, G., Avrithis, Y.: Multimodal saliency and fusion for movie summarization based on aural, visual, textual attention. IEEE Trans. on Multimedia 15(7) (2013)
3. Treisman, A., Gelade, G.: A feature integration theory of attention. Cognit. Psychology 12(1), 97–136 (1980)
4. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology 4(4), 219–227 (1985)
5. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
6. Borji, A., Sihite, D.N., Itti, L.: Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. IEEE Trans. Image Processing 22(1), 55–69 (2013)
7. Poynton, C.: Digital Video and HD: Algorithms and Interfaces, 2nd edn. Morgan Kaufmann (2012)
8. Adelson, E.H., Bergen, J.R.: Spatiotemporal energy models for the perception of motion. J. Opt. Soc. Amer. A 2(2), 284–299 (1985)
9. Heeger, D.J.: Model for the extraction of image flow. J. Opt. Soc. Amer. 4(8), 1455–1471 (1987)
10. Bovik, A.C., Gopal, N., Emmoth, T., Restrepo, A.: Localized Measurement of Emergent Image Frequencies by Gabor Wavelets. IEEE Trans. Information Theory 38, 691–712 (1992)
11. Havlicek, J.P., Harding, D.S., Bovik, A.C.: Multidimensional quasi-eigenfunction approximations and multicomponent am-fm models. IEEE Trans. Image Processing 9(2), 227–242 (2000)
12. Viola, P., Jones, M.J.: Robust real-time face detection. Int'l. J. Comput. Vis. 57(2), 137–154 (2004)
13. Wyszecki, G., Stiles, W.S.: Color Science, 2nd edn. J. Wiley & Sons, NY (1982)
14. Gabor, D.: Theory of Communication. IEE Journal (London) 93, 429–457 (1946)
15. Daugman, J.: Uncertainty Relation for Resolution in Space, Spatial Frequency and Orientation Optimized by Two-Dimensional Visual Cortical Filters. J. Opt. Soc. Amer. A 2(7), 1160–1169 (1985)
16. Daugman, J.G.: Two-dimensional spectral analysis of cortical receptive field profiles. Vision Research 20(10), 847–856 (1980)
17. Green, D.M., Swets, J.A.: Signal detection theory and psychophysics. Wiley, New York (1966)