Interactive User Oriented Visual Attention Based Video Summarization and Exploration Framework

Yiming Qian
Ryerson University
Department of Electrical and Computer Engineering
350 Victoria St, Toronto, Canada
yqian@ryerson.ca

Matthew Kyan
Ryerson University
Department of Electrical and Computer Engineering
350 Victoria St, Toronto, Canada
mkyan@ee.ryerson.ca

***Abstract*— an interactive user oriented high definition visual attention based video summarization and exploration framework is proposed to extract feature frames from a video collection and allow users to interactively explore those feature frames. It is based on previous work [1] that applies high definition visual attention algorithm mapping and multivariate mutual information to select a feature frames to represent each shot, then uses a self-organizing map to remove the redundant frames. After the video summary process, the extracted feature frames are connected into a network structure. Each node contains the information of the feature frame and the relation to other nodes. The relation between nodes are defined by clustering algorithms (self-organizing map, k-means, support vector machine, etc), expert systems (look-up table, fuzzy logic statement, etc) or any algorithm that defines similarity (sift, surf, etc). When a user select one node, depending on the user setting, the related nodes will be displayed onto a 2D canvas. In this way user is be able to interactively to browse through the whole video collection.**

I. INTRODUCTION

As video recordings become part of people's everyday activities, the question of how to access and manage their recorded video increasingly becomes a challenge. In this work, we consider the problem of presenting a reasonable summary of the video to allow users to interactively access video segments via those summary results. The framework contains two parts. The first part is a video summary algorithm that extracts feature frames from a video to create a summary. The second part is an interactive image exploration framework that allows users to explore the feature frames in the collection. The video summarization algorithm is based on previous work [1]. It uses colour histogram shot detection to separate the video into shots, and then applies a novel high definition visual attention algorithm to construct a saliency map for each frame. The saliency map is constructed based on a hybrid between Itti's visual attention theory [2] and colour theory[3]. The frame is first processed by a Gaussian pyramid algorithm to create an array of low resolution feature images, then those low resolution feature images are compared with the original image to construct the array of saliency maps. The comparison algorithm is based on CIE Delta E, a standard developed from psychological studies of human vision identifying the difference between two colours proposed by the *International Commission on Illumination* (abbreviated CIE for its French name, *Commission Internationale de l'éclairage*) [3]. The array of saliency maps are fused together to form a final Saliency map of the image. A multivariate mutual information algorithm is then employed to select a feature frame to represent each shot based on the saliency information. The selected feature frames are then processed by a self-organizing map to remove any redundant frames.

The summary results are stored into an image database. Organizing this image database becomes a challenge. Strong and Dong [4] proposed a similarity based multi-resolution algorithm that takes real time GPU processing power using self-organizing map to cluster images. Images are first grouped together by self-organizing map and display onto a 2D canvas. When user selects an image or an image group, further clustering processing will be conducted around the selected target to display the most relevant set of images. Worrying, Rooij and Rijn [5] proposed an algorithm using K-nearest neighbour network and graph theory to construct a graph network which allows user to browse through the image collections in 3D space. Each image is connected with K most similar other images based on a similarity distance function.

In this paper, a novel graph based image collection exploration method is proposed. Similar to Worrying's graph structure, the image collection is constructed into a network structure and then projected to a 2D canvas. Each image is a node that contains its own set of information. Each node is connected to other nodes in different relations where the relations are defined by clustering algorithms, expert systems or any algorithm that defines similarity. The degree of node similarity to start a connection is defined by user changing the number of clusters or defining thresholds. The colour histogram group (clustered by self-organizing map) is first displayed on a 2D canvas; one image with median weight is selected to represent the group. When user selects an image, the images that share a connection with it will be displayed on the canvas. The user can filter out some connection nodes by adjusting the framework setting.

II. VIDEO SUMMARIZATION

A. High Definition Human Attention Model

The High Definition Human Attention model is inspired by anatomical studies of the human vision system. Image colour features are fed into a centre surround algorithm to construct multiple saliency maps in different scales. The final saliency map is the fusion of all the saliency maps [6].

978-1-4799-3010-9/14/$31.00 ©2014 IEEE

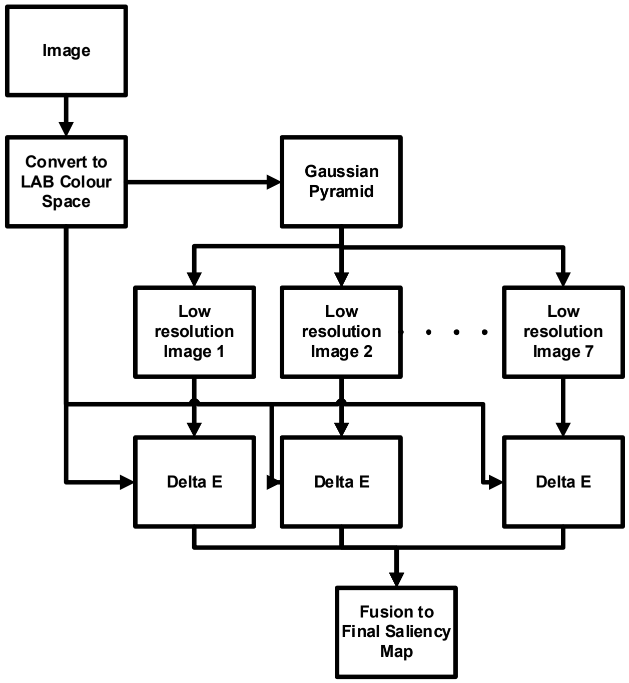


Figure 1 high definition visual attention algorithm flowchart

The centre surround algorithm that proposed by Itti [2] is used to define differences between a small centre region and its close surrounding. It is based on the idea that colour differences at different scales trigger neural responses in the human visual system [7]. It is implemented by decomposing an image into lower scale versions using Gaussian image pyramids. The low resolution version images are then resized by Bicubic interpolation algorithm to its original image size. In this work a series of 7 low resolution images are constructed and resized back to the original size. The saliency maps are constructed by taking the colour features in LAB colour space from the original image and comparing with the resized low resolution image features.

$$I_{c,s}(x,y) = \Delta E_{00}(I_c(x,y), I_s(x,y)) \tag{1}$$

Where
I_c is the original image features
I_s is the resized low resolution image features
ΔE_{00} is the colour difference calculation

The colour difference calculation that based on vision theory is implemented. When humans observe a colour, they will react to hue difference first, Chroma difference second and lightness differences last [8]. This phenomenon was observed by International Commission on Illumination (CIE) and it is been used to measure the visual difference between two colours which is known as the Delta E standard. The Delta E 2000 standard is used in the proposed algorithm. The Delta E 2000 colour space is an ellipsoid space which is more accurate than Delta 1976. Furthermore Delta E 2000 corrected the assumption that made in Delta E 1994 which made the lightness weighting varied. Those improvements help Delta E 2000 quantify small perceived colour difference more accurately than other methods [9]. The Delta E 2000 standard calculation in Lab colour space is following [10]:

$$\Delta E_{00}(L_1^*, a_1^*, b_1^*; L_2^*, a_2^*, b_2^*) = \Delta E_{00}^{12} \tag{2}$$

$$\Delta E_{00}^{12} = \sqrt{\begin{matrix}(\frac{\Delta L^{'}}{k_L S_L})^2 + (\frac{\Delta C^{''}}{k_C S_C})^2 + \\ (\frac{\Delta H^{'}}{k_H S_H})^2 + R_T(\frac{\Delta C^{''}}{k_C S_C})(\frac{\Delta H^{'}}{k_H S_H})\end{matrix}} \tag{3}$$

Where
L_1, L_2, a_1, a_2, b_1, b_2 are the two colours value in LAB colour space

$$\overline{L}' = (L_1 + L_2)/2 \tag{4}$$

$$\overline{C} = (\sqrt{a_1^2 + b_1^2} + \sqrt{a_2^2 + b_2^2})/2 \tag{5}$$

$$G = (1 - \sqrt{\frac{\overline{C}^7}{\overline{C}^7 + 25^7}})/2 \tag{6}$$

$$a_1' = a_1(1+G) \tag{7}$$

$$a_2' = a_2(1+G) \tag{8}$$

$$\overline{C'} = \left(\sqrt{a_1'^2 + b_1^2} + \sqrt{a_2'^2 + b_2^2}\right)/2 \tag{9}$$

$$h_1' = \begin{cases} \tan^{-1}(b_1/a_1') & \tan^{-1}(b_1/a_1') \geq 0 \\ \tan^{-1}(b_1/a_1') + 360^o & \tan^{-1}(b_1/a_1') < 0 \end{cases} \tag{10}$$

$$h_2' = \begin{cases} \tan^{-1}(b_2/a_2') & \tan^{-1}(b_2/a_2') \geq 0 \\ \tan^{-1}(b_2/a_2') + 360^o & \tan^{-1}(b_2/a_2') < 0 \end{cases} \tag{11}$$

$$\overline{H}' = \begin{cases} (h_1' + h_2' + 360^o)/2 & |h_1' - h_2'| > 180^o \\ (h_1' + h_2')/2 & |h_1' - h_2'| \leq 180^o \end{cases} \tag{12}$$

$$\begin{aligned} T &= 1 - 0.17\cos(\overline{h}^{'} - 30^o) \\ &+ 0.24\cos(2\overline{h}^{'}) + 0.32\cos(3\overline{h}^{'} + 6^o) \\ &- 0.20\cos(4\overline{h}^{'} - 63^o) \end{aligned} \tag{13}$$

$$\Delta h' = \begin{cases} h_2' - h_1' & |h_2' - h_1'| \leq 180^o \\ h_2' - h_1' + 360^o & |h_2' - h_1'| > 180^o; h_2' \leq h_1' \\ h_2' - h_1' - 360^o & |h_2' - h_1'| > 180^o; h_2' > h_1' \end{cases} \tag{14}$$

$$\Delta L' = L_2 - L_1 \tag{15}$$

$$\Delta C' = C_2 - C_1 \tag{16}$$

$$\Delta H' = 2\sqrt{C_1'C_2'}\sin(\Delta h'/2) \tag{17}$$

$$S_L = 1 + \frac{K_2(\overline{L}' - 50)^2}{\sqrt{20 + (\overline{L}' - 50)^2}} \tag{18}$$

$$S_C = 1 + K_1\overline{C}' \tag{19}$$

$$S_H = 1 + K_2\overline{C}'T \tag{20}$$

$$\Delta\theta = 30\exp\left\{-\left(\frac{\overline{H}' - 275^o}{25}\right)^2\right\} \tag{21}$$

$$R_T = -2\sin(2\Delta\theta)\sqrt{\frac{\overline{C}'^7}{\overline{C}'^7 + 25^7}} \qquad (22)$$

K_C and K_H are usually both unity and the weighting factors K_L, K_1 and K_2 depend on the application

Table 1 Delta E 2000 Constant Table

	Graphic Arts	Textiles
K_L, K_C, K_H	1	2
K_1	0.045	0.048
K_2	0.015	0.014

The proposed algorithm creates a series of 7 saliency maps. Those saliency maps are normalized (equation 23) and fused together (equation 24) to form a final saliency map (N_0).

$$N_i(x,y) = \{D_i(x,y) - d_{\min}\} / \{d_{\max} - d_{\min}\} \qquad (23)$$

$$N_0 = \sum_{i=1}^{7} N_i \qquad (24)$$

Where
N_i is the normalized saliency map
D_i is the saliency map before normalization
d_{max} is the maximum value of the saliency map
d_{min} is the minimum value of the saliency map

B. *Creating Video Summary Story Boards*

The saliency map obtained by the proposed method indicates a high resolution map of attention areas. An attention curve is constructed from it based on an assumption that people tend to choose frames that contain more information with respect to adjacent frames. This assumption was modeled by calculating the multivariate mutual information [11] within a shot. The multivariate mutual information calculates the similarity of a frame against all the frames in a shot. When a frame has the highest multivariate mutual information value, it means that frame contains higher information (relatively) in that shot. The high definition saliency map is used as a special grayscale version of image. The advantage of using high definition saliency map against regular grayscale image is the saliency map emphasized the human attention region and filtered out unimportant information. After obtaining a feature frame for each shot, a self-organizing map is applied to remove the redundant frames. The self-organizing maps algorithm processes the colour histogram information for each feature frames and categorize them into different groups. One frame with median weight is selected from each group to form the final feature frame summary.

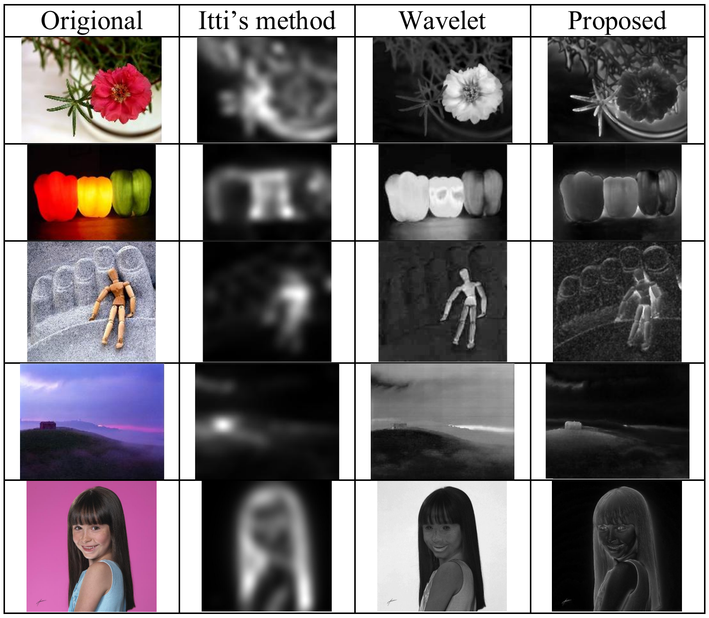


Figure 3 comparisons of visual attention methods

III. Exploration Framework

The proposed exploration framework is based on graph theory which is similar to the social network structure. On the social network, people have their own profiles and they are connected by different relations (friend, family, colleague, etc.), common interests (hiking, cheese burger, cats, etc.) or common groups (MADD, CAA, IEEE, etc.). The proposed framework applied the same concepts. Each frame is an individual and contains its own information (profile) such as colour histogram (CH), gray-level co-occurrence matrix (GLCM), histogram of oriented gradients (HOG), timestamp in the video, and so forth. Different frames are connected by relations that defined by clustering algorithms (self-organizing map, support vector machine, k means), expert systems (lookup table, fuzzy logic statements, etc.) or any algorithm that define similarity (sift, surf, etc.). One frame could have multiple connections with another frames, or it could only have

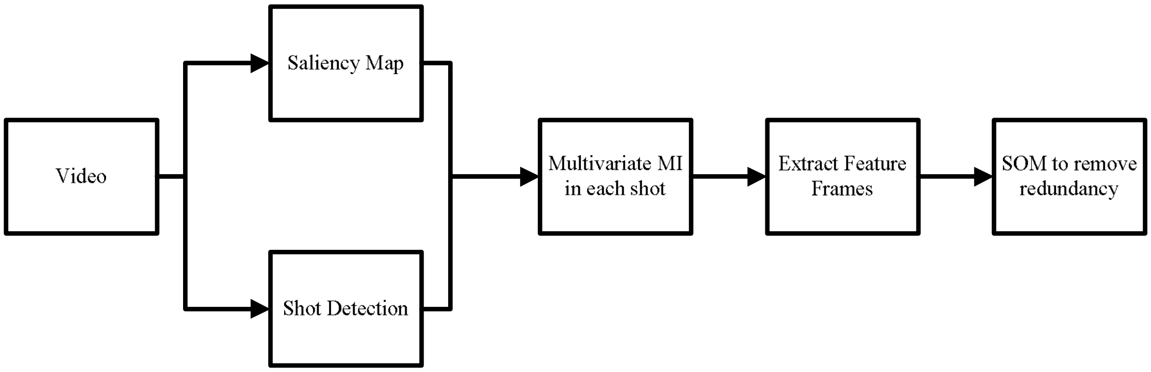


Figure 2 high definition video summarization flowchart one connection (see figure 4). During the clustering process, only the frames with direct relation will be connected. In this way, the number of connection level is able to indicate the strength of the relation for example if object A needs 3 connections to connect object B but object C just need 2 connections which indicates object C has stronger connection to object B. When user selects a node, the nodes that share a connection with will be displayed on a 2D canvas. User can filter out some connection nodes by adjusting the framework setting. User could browse through those connections to explore the whole video collections.

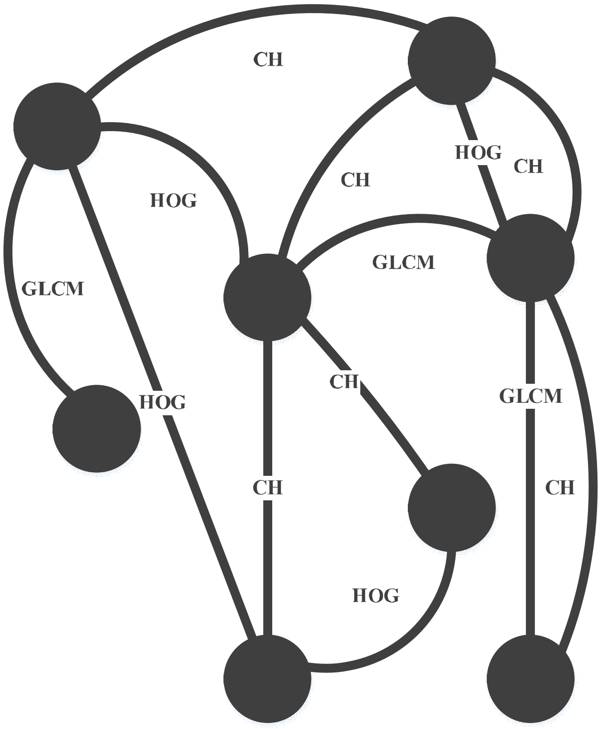


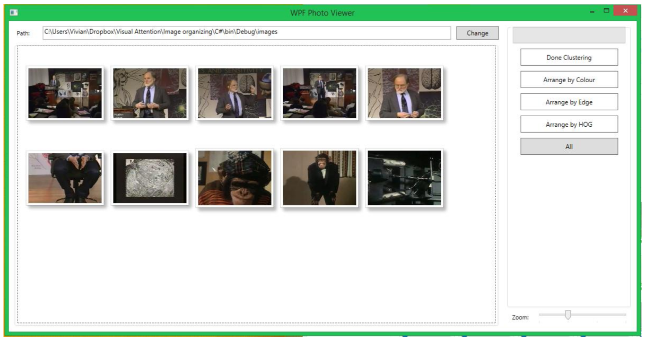
Figure 4 example of proposed network structure

This structure has a few advantages:

1. User interaction: user is be able to interactively explore the whole frame collection.
2. Adjustable sensitivity: the degree of similarity limitation is adjustable through control the number of clusters or a threshold.
3. Flexibility: clustering algorithms, expert system, or any algorithms that define similarities could be used in the framework to define connections.
4. Mobile platform friendly: the process does not require large real time processing power, so it could run on low computational power mobile platforms.
5. Upgradability: new clustering features or new videos could be added to the system without large structure modification.

IV. Results and Discussions

In the experiment, the framework was implemented in C# WPF and self-organizing maps were implemented to perform the clustering process. Three features were extracted from images which were HSV colour histogram, gray level co-occurrence matrix (GLCM), and histogram of oriented gradients (HOG). Those three features represent three kinds of connections from the selected image. Initially users have the options to start with arrange all the images in the database by one of those three features into an 8 by 8 cluster grid (One image with median weight was selected to represent its own cluster) or user can select to display all the images. Once an image was selected, users have the options to select what kind of relation they want to view that is relevant to the selected image. The relation could be based on sharing the same colour distribution, the same texture or similar edge information. From there, users are able to explore the image collections as they navigate from node to node in the image network which is similar to social network experience: i.e. people browse through others profiles to find interesting people. Once user found an interest frame, user have an option to playback the shot where that frame is located or play the entire video.


WPF Photo Viewer

Path: C:\Users\Vivian\Dropbox\Visual Attention\Image organizing\C#\bin\Debug\images

Figure 5 demonstration GUI

V. Conclusions

The proposed framework provides the user with a new way to access and manage their home video collections. The framework consists of two parts: feature frame extraction and frame exploration. The feature frame extraction creates summary story boards from videos, depends on the user setting, a 30-minute video could be summarized into a 25-frame story board. The frame exploration provides user a new way to explore their video collections in a network structure. Users are able to navigate through the frames from one node to another node. The future work is to implement more feature similarity comparison algorithms, such as sift or surf, into the framework. In this way the frames could be connected by location-based scene without any GPS location information [12]. The framework is implemented in relational database. Due to its graph structure it could be implemented in a graph database (such as Neo4J [13] or OrientDB [14]) to more efficiently handle large information.

References

[1] Y. Qian and M. Kyan, “High Definition Visual Attention Based Video Summarization”, the 9th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), Lisbon, Portugal, 2014.

[2] L. Itti, C. Koch, and E. Niebur, A model of saliency-based visual attention for rapid scene analysis. IEEE PAMI, 20(11):1254-1259, 1998

[3] S. Millward, "Color Difference Equations and Their Assessment." Test Targets(2009): 19.

[4] G. Strong, and M. Gong, "Similarity-based image organization and browsing using multi-resolution self-organizing map." Image and Vision Computing 29, no. 11 (2011): 774-786.

[5] M. Worring, O. Rooij, and T. Rijn, "Browsing visual collections using graphs." In Proceedings of the international workshop on Workshop on multimedia information retrieval, pp. 307-312. ACM, 2007.

[6] S. Frintrop, "Computational visual attention." In Computer Analysis of Human Behavior, pp. 69-101. Springer London, 2011.

[7] Y. Saber and M. Kyan. "High resolution biologically inspired salient region detection." In Image Processing (ICIP), 2011 18th IEEE International Conference on, pp. 649-652. IEEE, 2011.

[8] X-Rite, Incorporated , “A Guide to Understanding Color Communication.”, 2007

[9] G. Sharma, W. Wu, E. N. Dalal, and M. U. Celik, "Mathematical discontinuities in CIEDE2000 color difference computations." In Color and Imaging Conference, vol. 2004, no. 1, pp. 334-339. Society for Imaging Science and Technology, 2004.

[10] B. J. Lindbloom, "Delta E (CIE 2000)." Bruce Lindbloom.com. [online] http://www.brucelindbloom.com/index.html?Eqn_DeltaE_CIE2000.html , Feb. 2009. (Accessed: 12 Jan. 2014)

[11] Z. Z. Tabrizi, B. M. Bidgoli, and M. Fathi, "Video summarization using genetic algorithm and information theory." In Computer Conference, 2009. CSICC 2009. 14th International CSI, pp. 158-163. IEEE, 2009.

[12] X. Chen, M. Das, and A. Loui, "An efficient framework for location-based scene matching in image databases." International Journal of Multimedia Information Retrieval 1, no. 2 (2012): 103-114.

[13] P. Neubauer. "Graph databases, NOSQL and Neo4j." , 2010.

[14] Tesoriero, Claudio. “Getting Started with OrientDB.” Packt Publishing Ltd, 2013.