THE IMPACT OF COLOUR VISUAL ATTENTION FOR VIDEO SUMMARIZATION

By

Yiming Qian, B.Eng.
Bachelor of Engineering (B.Eng.), Ryerson University, Toronto, 2012

A thesis
presented to Ryerson University

in partial fulfillment of the
requirement for the degree of
Master of Applied Science in the Program of
Electrical and Computer Engineering.

TORONTO, ONTARIO, CANADA 2014

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

Yiming Qian

Abstract

The Impact of Colour Visual Attention for Video Summarization

Master of Science 2014

Yiming Qian

Electrical and Computer Engineering

Ryerson University

A High Definition visual attention based video summarization algorithm is proposed to extract feature frames and create a video summary. Specifically, the proposed framework is used as the basis for establishing whether or not there is a measurable impact on summaries constructed when choosing to incorporate visual attention mechanisms into the processing pipeline. The algorithm was assessed against manual human generated key-frame summaries presented with tested datasets from the Open Video Dataset (www.open-video.org). Of the frames selected by the algorithm, up to 68.1% were in agreement with the manual frame summaries depending on the category and length of the video. Specifically, a clear impact of agreement rate with the ground truth is demonstrated when including colour-attention models (in general) into the summarization framework, with the proposed colour-attention model achieving stronger agreement with human selected summaries, than other models from the literature.

Acknowledgments

First of all I would like to thank my parents for all those years to support my education and then I would like to thank my supervisor Dr. Matthew Kyan. He gave this research opportunity and guides me throughout my master thesis. Also I would like to thank my friends Ting Hao, Nawar Mahfooth and Mario Garingo without them I would have a hard time to finish my thesis.

Table of Contents

List of Tables and Figures

Chapter 1 : Introduction

A video summary is an abstract version of a video that is significantly shorter while still remaining informative – it is able to provide a sufficient idea of the main content in the video, without requiring the observer to physically watch the entire piece. Video can be considered to fall into two primary forms: structured or scripted video and unstructured raw footage. Structured video relates to video that has been edited or organized. It could be a highlights package, a film or news report – for which the footage has been structured to reflect some desired narrative. On the other hand, the unstructured video relates to raw video footage that has not been edited or organized, and thus no narrative exists. It may consist of repeated shots, long still scenes, or extremely short surprise scenes. Given the ubiquity of digital video in today's world, easily acquired through an array of mobile devices, tablets, DSLR cameras, camcorders, webcams, surveillance cameras and the next wave of wearable smart cameras and devices (e.g. Go-Pro and Google Glass), individuals are rapidly amassing larger and larger collections of such unstructured video through their everyday interactions. The problem of how to organize and archive such collections is increasingly becoming a problem. With video this is exacerbated particularly when attempting to locate parts of a video where a particular event may have occurred, where typically the user is forced to tediously experience the content in a linear fashion, watching hours of footage in order to locate content of interest.

In order to organize such unstructured collections for more non-linear access, some form of annotation is needed. The annotation could be knowledge based automatic annotation (where there is some prior information about the nature of the content to be annotated) or it could be purely manual. For example iPhoto is a hybrid system that uses domain knowledge (face detection) with manual annotation in order to organize photo collections in different ways based on the people that might exist in the photos. The goal of such an approach is to provide alternative entry points when browsing: the collection may be re-organized based on different people of interest. In order to automate such annotation (either fully or semi)

a detector may be defined that specifically locates particular content, however we must know what the video might be comprised of – as an example, genre classification of sports video has been used to label sections that observers might be interested in viewing at another stage for example a ‘goal’ event or particular play in a hockey or soccer match. Other types of labels may be associated with view types (e.g. far, close up, etc). For these, domain specific detectors must be designed, which are only suitable for organizing video of a particular type. There are severe limitations in organizing unstructured video because there is no clear model for what content may be considered as having the potential to be 'interesting' to the user.

In the absence of a useful model defining ‘interesting content’, this thesis considers the use of models based on pre-attentive mechanisms known to exist in the human visual system [1] to suggest events in the video that could potentially be of interest - for further consideration or processing. Such models reflect the involuntary response we tend to have when some form of anomaly is present in the visual field of view (FOV) – for example, a region of the FOV that undergoes unusual relative motion or colour difference. In particular, the models aim to simulate a property encoded in the visual cortex known as ‘centre-surround’, which detects 'salient' (inconspicuous) local regions as a differential between visual properties in the local vicinity of a pixel, versus a more general surrounding region. This could be considered for properties relating to colour, texture, brightness or motion. Attention has also been considered with respect to audio properties, in order to define possible events of interest in time within a video. The principle is analogous to visual, with the exception that locations in time that are salient with respect to temporally nearby audio in the video stream are detected and highlighted as 'interesting'. In identifying possible sites of interest, it is then ultimately left to the user to decide whether or not to dig further into the video artifact or collection, thus the summary acts merely as a mechanism for enabling alternative entry point to the data, as a preprocessing step for search, annotation or other related tasks.

The goal of this thesis is to consider the impact of attention models as a specific indicator of regions and frames that might be emphasized in terms of their 'interest' factor, and their subsequent inclusion into a first stage of unstructured video processing: the construction of a 'video summary'. The purpose of a video summary is to provide a snapshot or efficient and fast way of establishing the 'gist' of what might exist in a particular video clip, or more generally, in a collection of unstructured videos. The video summary allows the browser to do much more than simply get an idea of what content may be present, it may serve as the basis for some interaction with the content itself - such as allowing the user to direct where to focus and apply annotations (used to further facilitate search/organization); it may provide sample frames to query or launch into a browsing mode that enables the user to jump directly to clips or sections of a clip that they wish to view, etc. Some applications are elaborated upon in next section. Typically; the goal of a video summary is first and foremost, to construct an abstract version of the video/set of videos which facilitate entry points directly to specific locations in the content. With an initial summary, it is expected that the user can more efficiently navigate to areas of the content based on their own interpretation of what has been summarized.

1.1 Types of Video Summarizations

There are two main approaches to video summarization, first: *video skimming* which provides a fast forwarded version of the video; second: *key frame extraction*, which extracts feature frames and presents them to the users as a storyboard. The advantage of the video skimming summary is it reduces the video play time and kept the content in a sequential order which is perfect to process videos such as surveillance video and sport video. Surveillance video and sport video consist of long, monotonous, repeated scenes, it is common for audience to manually fast-forward the video but often some important detail may be missed. The video summary processes the video and provides a variable speed; fast forward playback which shortens the redundant parts of the video at the same time allows important segments play at regular speed or even slow motions. The playback speed is determined by a frame weighting matrix,

where the shot with higher order of importance has slower playback speed; on the contrary, shots with lower order of importance has higher playback speed. As shown in Figure 1 this is typically achieved through the calculation of an importance curve that is used to control the sampling rate of frames to be included into the video skim. The area that the curve covered indicates the sampling rate in the shot. The larger area means more important content appears in the shot which needs higher sampling rate.

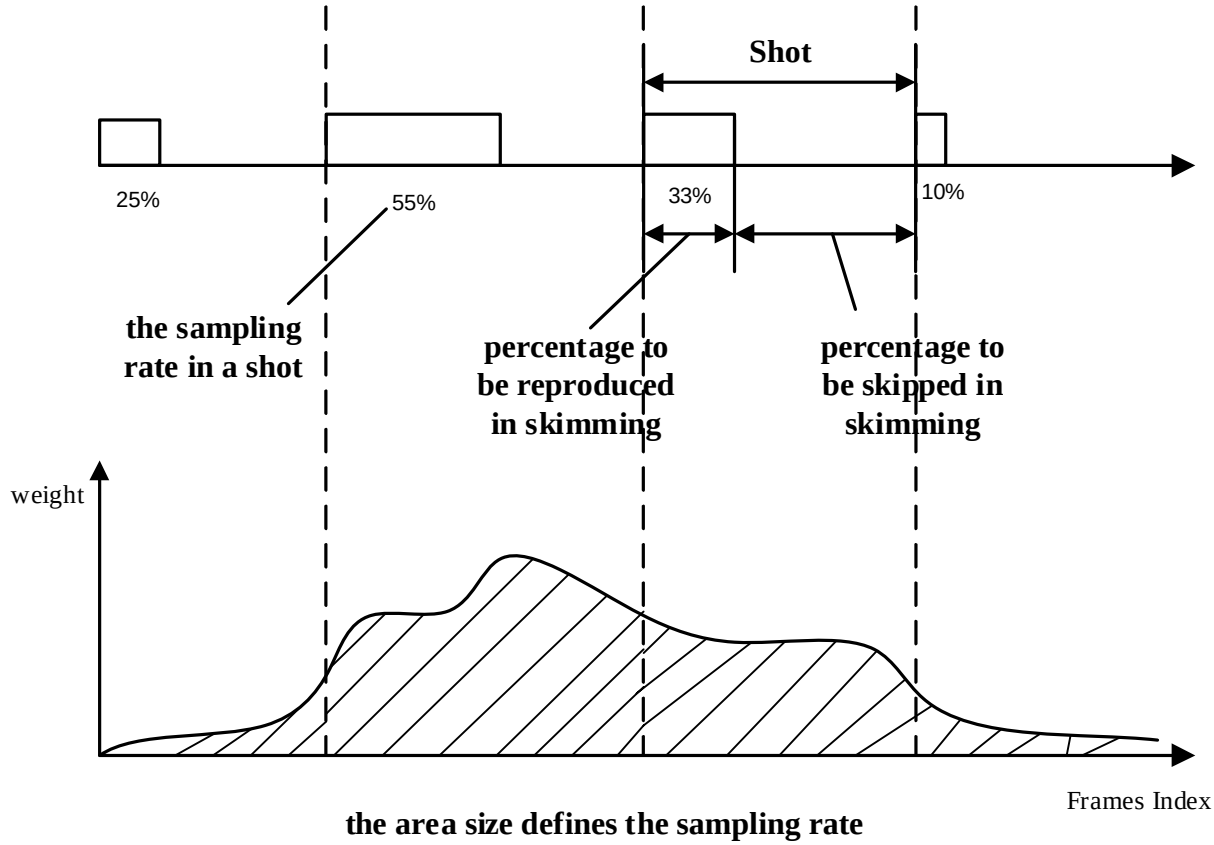


Figure 1: Video skimming diagram

Unlike the video skimming type, the video storyboard type provides an instant view of content from throughout the video sequence which could be interactively defined / scaled so as to include larger granularity in the visual summary in order to give the user an overall set of possible entry points that could be used to further delve into the collection. It extracts key frames that give a maximum amount of the information for the target video. The most common form of the storyboard is a comic book (e.g.

manga books) which uses a limited number of pictures to describe a story. The following example (Figures 3) shows an example storyboard for “the dumbest soccer player”, while the typical process for construction is highlighted in Figure 2.

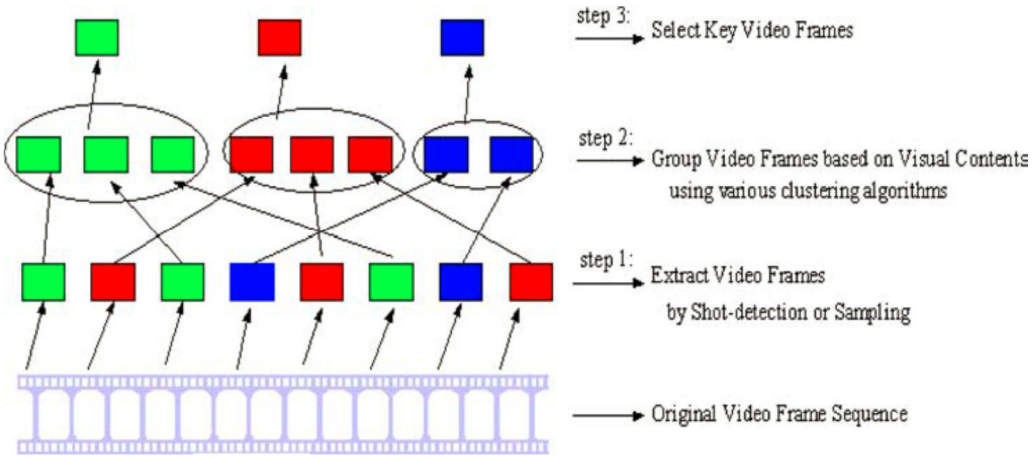


Figure 2: Clustering based key frame video summary [2]

Figure 3: An example of a storyboard video summary

1.2 Applications

The main advantage of video summaries lay in their ability to convert the video to reduced length. Such that, to perform search, annotation, or analysis tasks on a video will be made significantly easier. In essence, the summary can be thought of as a mechanism for extracting context from the video which can be interpreted quickly by the user when deciding which clips should be considered further. For instance, in a search task, the video summary could provide potential queries that could be used to extract or navigate directly to particular scenes of interest. There are several of areas of application for the video summary algorithm: rushes summarization and video file organizing, surveillance video.

1.2.1 Rushes Summarization

Rushes are raw material used to produce a video. It often contains as much 20 to 40 times as finished product [3]. Often the raw frames or sequence of frames are highly repetitive. Manually extracting the stock footage (reusable shots of people, objects, events, locations) from the raw material often takes tremendous amount of time. The proposed video summary algorithm provides a good solution for rushes summarization. It processes raw uncut video and extracts the key frames to create a storyboard and at the same time removes redundant video shots. Editors then will be able to select important video segments to edit based on the key frames which dramatically reduce the video editing time.

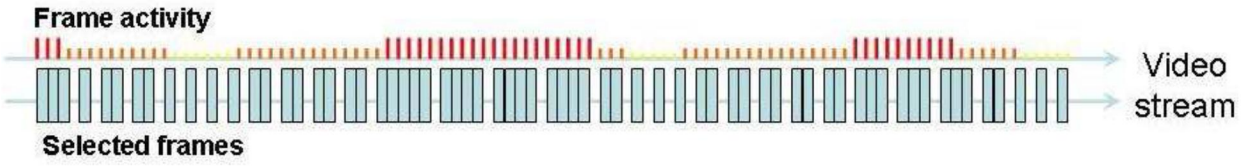


Figure 4: An example of algorithm selects key frames from a video [4]

1.2.2 Video Search

If a user has a query image and wants to find all the relevant frames in a video sequence, then the depth-first-search with pruning at each node can carry out the desired functionality much faster than the serial

search [5]. The key frames provide access points for the video searching algorithm to search relevant frames which significantly reduce the number of frames need to be searched. The same technology can also be used in the image file organizing software to process a collection of images. The following screen shot is Apple's iPhoto. It has a hybrid system that is able to semi-automatically annotate images (based on face detection), while at the same time accepting user manual annotation in order to organize photos in different ways. This same principle could be utilized in video; however there are significantly more objects/events (apart from faces) that might be of interest in a video. If a user were to return from a diving trip in the Caribbean for instance, and had a large collection of Go-Pro videos from a diving expedition, it would not be trivial to find an event that occurred in the video where the user was able to follow an exotic colorful fish that appeared for only a brief moment. Again without a suitable model for detecting such an event, it would be a tedious exercise in attempting to relocate that within the set of videos collected.

Figure 5: screenshot of iPhoto 11

1.2.3 Surveillance Video Processing

Surveillance video often consists of hundred hours of archived video. Currently, security professionals go through the time-consuming process of human review of archived video in order to reconstruct or revise events relating to possible infractions [6]. Video summarization provides an easy solution to reduce the time required to analyze the video by removing unnecessary video segments and provides a quick access to rapidly search for past incidents from archived video. In most scenes, the surveillance video contains a relatively static background and some moving objects. The video summaries can automatically respond to moving stimuli or objects (people, cars, etc.) of interest and present them to the security professionals for further consideration.

1.3 Fundamentals of Visual Attention

As indicated earlier, the fundamental mechanism in visual attention relates to how differentials are detected in highly localized spatial regions in the field of view. To understand this mechanism, it is important to first review some basic physiology of the human eye. The human eyes have two important optical functions: to receive light from the surrounding environment, to focus on certain objects and project a clear image onto the back of the eye. There are multiple parts that working together to perform those optical functions. First, the light enters the *cornea*, a transparent bulge on the front of the eye behind which is a cavity filled with clear liquid, called the *aqueous humor*. Next, light travel through the *pupil*, a variably sized opening in the opaque *iris*, which gives the eye its external colour. Behind the iris, light passes through the *lens*, whose shaped is controlled by *ciliary muscle*s. The len's optical properties can be altered by changing its shape, a process called accommodation. The photon then travels through the clear *vitreous humor* that fills the central chamber of the eye. Finally, it reaches the *retina*, the curved surface at the back of the eye. The retina is covered by over 100 million light-sensitive cells, *photoreceptors* [7].

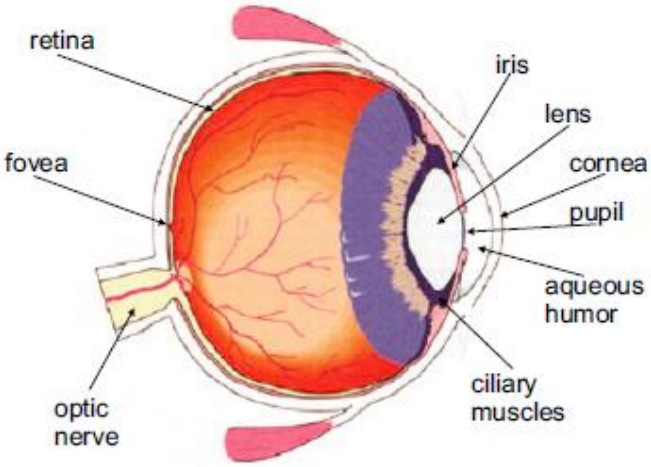


Figure 6: A cross section of the human eye

When light reaches retina, it is converted into neural signal which is fed to the human brain for further processing. In the visual system, this function is carried out by two types of *photoreceptors*: rods and cones, in the retina. Cones (about 8 million cells) concentrated in the centre of the retina (*fovea*) which only covers 2 degree of visual angle, less sensitive to light than rods. They are responsible for normal lighting condition (*photopic conditions*) and colour sensing. On the contrary, rods (about 120 million cells) are extremely sensitive to light locates all over the retina except the centre part. They are used for vision at low lighting conditions (scotopic conditions).

Besides the individual stimulation of rods/cones due to incident light through the pupil, there is physiological evidence that nerves further down the pipeline (within the visual cortex) respond to the collective stimulus of groups of rods/cones in a location versus those in the immediate vicinity. One such collective stimulus is known as the *centre-surround* response, and has been related to the control of eye-gaze during pre-attention [1]. Essentially this means that when some of the rods/cones are stimulated locally, and their surrounding cells are not, this event is detected at later stages in the HVS. Similar spatial patterns of stimulation are known to reflect oriented stimulus such as vertical vs horizontal lines, etc. This principle is at work in generating a response to isolated differences in light patterns that occur in the visual field – they are also at play when considering localized colour-based differences.

The Figure 7 [8] shows the centre-surround mechanism, which highlights measured neural response at later levels in the HVS due to patterns of centre or surround only stimulation on the retina. Its relationship to eye-gaze means that we tend to look and focus more on regions of the visual field that contain patterns that may reflect these salient or unusual anomalies. Our interest in this type of model is that such an attention mechanism could serve as a useful alternative for suggesting content that could potentially be of interest to users when surveying an unknown scene.

Based on the biological eye structure, Itti [1] proposed a centre-surround model simulating the neural activities to find the attention region of an image (predominantly as a means of explaining eye gaze movements). The human eye centre-surround behavior is modeled as an array of Gaussian pyramid images in different sets of colour opponencies (shown as opposing sides of the colour circle depicted in Figure 8).

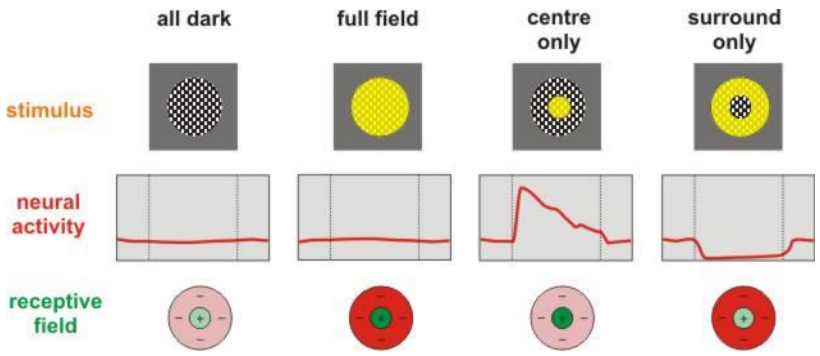


Figure 7: Neural centre surround response [8]

By comparing the colour opponencies in different Gaussian pyramid levels, an array of saliency 'maps' (a probabilistic distribution across the FOV that represents regions that are of more or less ('inconspicuity') is constructed. Additional mappings may be considered for visual properties other than colour – for example motion or texture information. Fusing the array of saliency maps together will form the final

saliency map giving an overall indication of potential regions of interest, which may be exploited in the construction of visual summaries of the content.

Figure 8: LUV colour circle

1.4 Contributions of this Thesis

The primary contribution of this thesis is to establish and evaluate whether or not there is a measurable impact for including a visual attention mechanism within the video summarization process. In order to assess this impact, a novel high definition visual attention based self-organizing map video summary algorithm is proposed and then evaluated with and without various visual attention mechanisms embedded. The test data used in the evaluation was taken from a popular video dataset online [9], for which a number of ground truth summaries (key-frame storyboards), have been manually pre-selected by humans. The video summarization mechanism is kept consistent across all experiments, while different attention mechanisms are evaluated for their ability in selecting key-frames that echo human selections. The attention mechanism explored is restricted to colour visual attention models, in an effort to justify whether or not any improvement is achieved, while the use of more elaborate attention models such as those that incorporate prior knowledge, or a combination of audio and visual information, is left for future

studies. In the process of this evaluation, a new model for colour-attention was also developed and compared to existing models within the video summarization framework.

Contributions thus fall into two parts: a novel high definition visual colour-attention algorithm and a clustering based video summary algorithm. The novel high definition visual colour-attention algorithm is based on a hybrid between Itti's visual attention theory [1] and colour theory [10]. Instead of taking opponency of fixed colour channels within the centre-surround calculation, the CIE Delta 2000 standard is applied to perform the comparison calculation. The result provides more colour uniformity and consideration of more subtle opponencies across the colour spectrum. The CIE Delta E 2000 standard is 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*). The CIE Delta E 2000 standard is based on calculations performed in the La*b* colour space (Figure 9), in which the proximal relationships between different colours is more uniformly representative of their perceptual differences: e.g. the difference between two points in this space can be equated to how different we 'perceive' the colours to be.

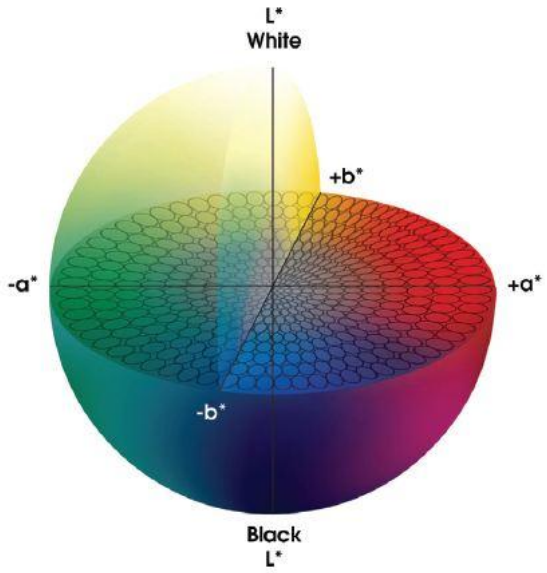


Figure 9: Delta E colour measurement in LAB colour space

The video summary algorithm takes the visual attention algorithm and uses this to enhance frames that may be more important in the context of any given shot, for consideration when extracting key frames. Figure 10 shows a brief example of the type of salient mapping that can be generated from the visual attention algorithm prior to frame selection – the column on the right showing the proposed attention mapping and its boosted resolution.

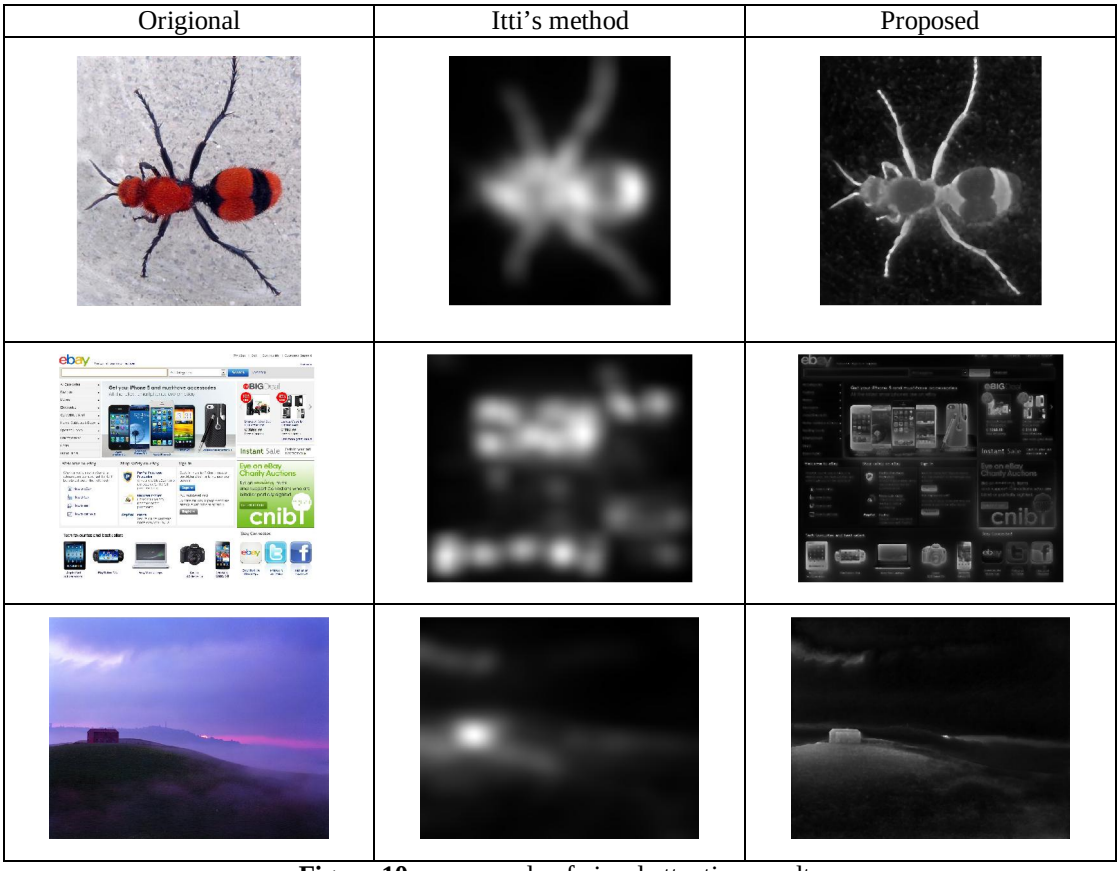


Figure 10: an example of visual attention results

Key frames are then selected by comparing the multivariate mutual information within the frame, where mutual information is based on salient content (as opposed to the raw image). The frame with the highest multivariate mutual information will be selected as the key frame to represent the shot. The key frames are then processed by self-organizing map to remove redundant frames. The size of the self-organizing map is user defined, thereby reduce or increasing the amount of images created to summarize the video. As will be discussed in the thesis, the self-organizing map was chosen for its potential in supporting post-

processing activities (such as facilitating interactive search or annotation processes that may be invoked from the summary).

1.5 Thesis Organization

The objective of this thesis is to propose a key-frame based video summarization algorithm that uses visual attention algorithms to preprocess frames in order to filter out the less important information, and then evaluate the impact of including visual attention as a key step in the summarization process. A novel visual attention algorithm is proposed to perform the saliency mapping process. The summarization results between different visual attention algorithms and without visual attention process were compared with ground truth. The video summary with visual attention enhanced summarization achieved better agreement rate with ground truth. The remainder of this thesis is organized as follows:

In Chapter 2 (Past Video Summary Approaches) is a literature review of past key-frame based and attention-based approaches to video summarization. In this chapter, we establish the main motivation for a hybrid approach.

In Chapter 3 (Proposed Visual Attention Model), detailed information is given regarding the proposed visual attention algorithm itself, as a mechanism for generating highly resolved saliency maps for individual video frames. Some preliminary results are then provided drawing comparison between the proposed, and various popular visual colour-attention methods.

In Chapter 4 (Proposed Video Summarization Algorithm) introduces a novel colour-attention, enhanced video summarization algorithm with a detailed test on 5 different videos. A comparison was made both with and without colour-attention driven frame selection, and with different types of visual attention

algorithm (Itti, Wavelet, and Proposed Visual Attention Algorithm). Evaluations are made against ground truth data.

Finally, Chapter 5 (Conclusion and Future Work) provides an overview of the proposed algorithm and future improvements, including the proposal of an interactive mechanism for search applications.

Chapter 2 : Past Video Summary Approaches

There are two main approaches to video summarization; the first being video skimming which provides a fast forwarded version of the video and the second, key frame extraction, which extracts feature frames and presents them to the users as a key-frame based storyboard.

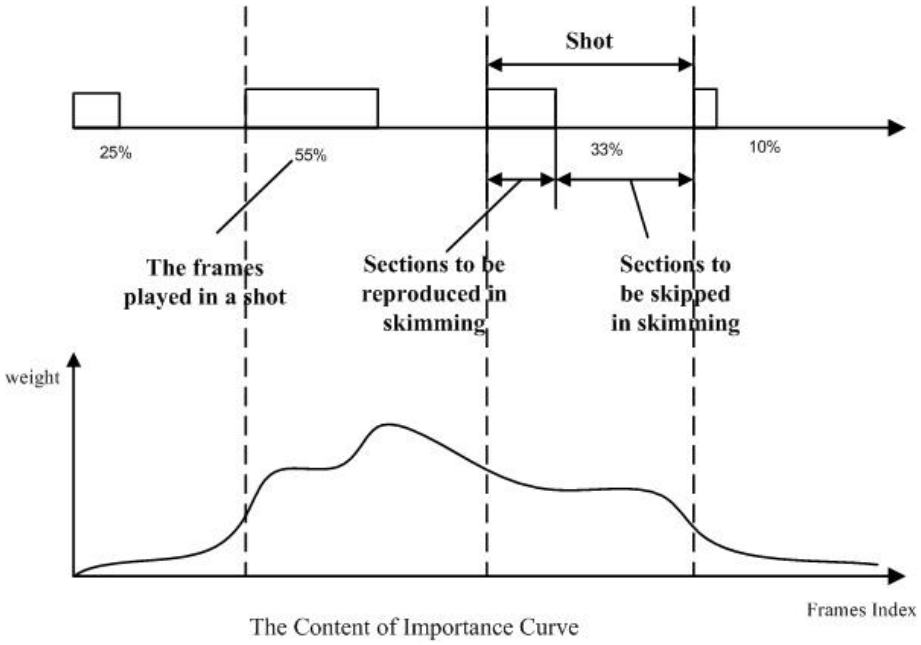


Figure 11: Video skimming diagram

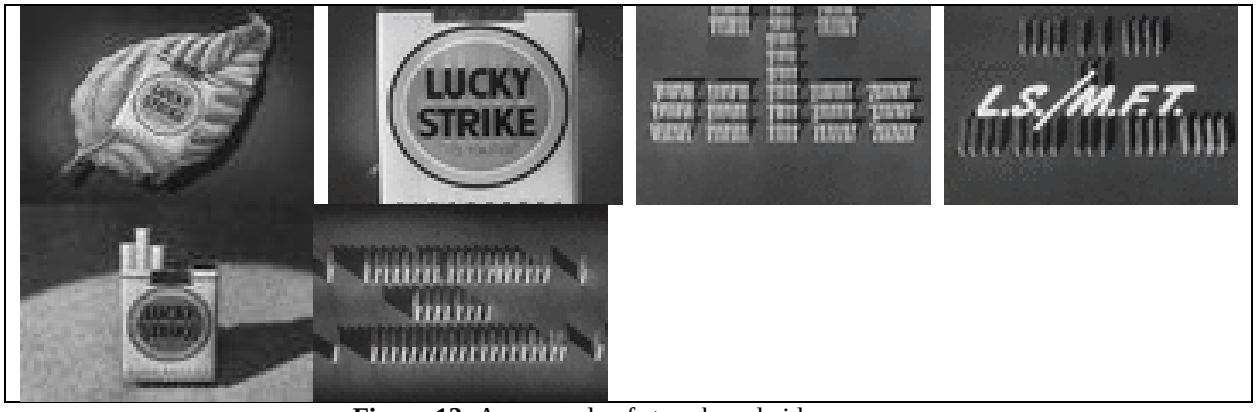


Figure 12: An example of story board video summary

This chapter will focus on presenting the popular approaches for key-frame type of video summary. The most popular type of video summarization methods are key-frame based video summarization and attention based video summarization. Key-frame based methods group frames into different clusters and

one frame is selected to represent each cluster. Attention-based methods process different video/audio features to create attention curves. Multiple attention curves are then fused together to create a final attention curve used to control and vary the sampling of key frames.

2.1 Key-frame Based Video Summarization

In key-frame based approaches, machine learning is used to extract different parameters from the raw frames, and then use those parameters as a basis of measuring inter-frame similarity. In this similarity space, clustering or classification algorithms group frames that are close to one another (or close to a known class label) in terms of some distance metric. One frame from each group will be selected as a feature frame. The most popular methods for grouping/labeling in video summary are K-means clustering, Support Vector Machine (SVM), Self-Organizing Maps (SOM), and Fuzzy C Means.

The general key frame video summary process is shown in the figure below. First, the video frames are extracted from the original video. Video Frame extraction selects frames to represent each shot or samples frames directly on a predefined time interval. Following this, different features are extracted and fed into a clustering algorithm to group frames. Finally one frame will be selected to represent each cluster.

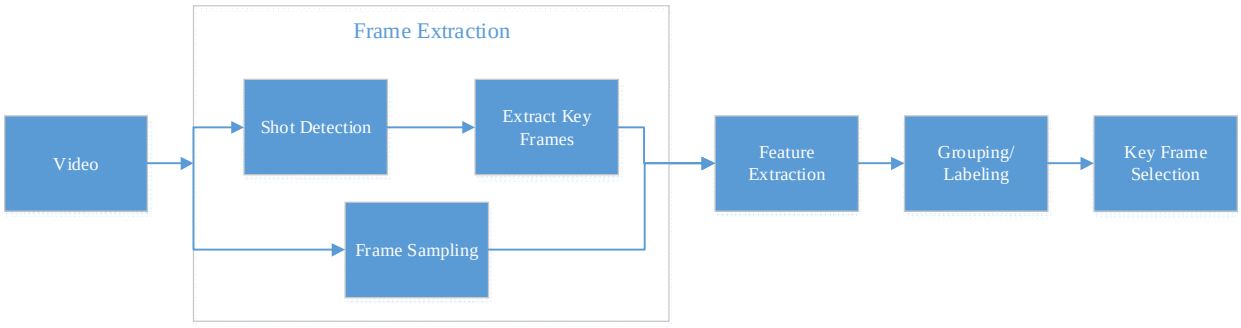


Figure 13: Clustering based key frame video summary

2.1.1 Frame Extraction

The key frames in a shot are commonly extracted by certain measuring metric or sampling frames at a fixed rate. Chasanis, Likas and Galatsanos [11] deploy k-means clustering method within a shot to cluster frames into different groups. One frame from each group is selected as the key frame for the shot. In Zhong, Zhang and Chang's work [12] the key frames in a shot are extracted by direct subsampling. Cvetkovic, Jelenkovic, Nikolic [13] selected a middle frame of a shot as the key frame.

2.1.2 Feature Extraction

In the Feature extraction stage, low level features are computed and used to describe the visual content in an image. Common image features are HSV colour histogram, histogram of oriented gradients (HOG), Colour layout descriptor (CLD), colour edge co-occurrence histogram and codebook. The HSV colour histogram is the simplest colour feature from the image; it represents a statistical pixel colour distribution, which can be used to describe the spatial layout/shape of the image throughout the frame. The histogram of oriented gradients is another statistical histogram to record the edge gradient orientation distribution which can be used to describe the spatial layout/shape of the image. The colour layout descriptor is also designed to capture the spatial information using a combination of colour selection and discrete cosine transforms (DCT). The colour edge co-occurrence histogram adds geometric information to the normal colour histogram, which is good for describing both shape and colour spatial distribution within an image. Among these feature extraction methods, a codebook is generated from the collection of feature frames that contains a reduced set of frames from across the whole video.

2.1.2.1 HSV Colour Histogram

HSV colour histogram is the most common feature extraction algorithm which is based on measuring histogram value on each Hue, saturation and lightness channel. The advantage of using HSV colour space over RGB colour space is that HSV colour space is more uniform throughout the colour space. As such, distance metrics comparing two frames described with this feature associate frames with similar global colour content. The HSV colour histogram has some limitations however, in that spatial layout of colour is not considered, it is thus possible that two frames with a different spatial arrangement of the same set of colour pixels could be considered similar. It never-the-less serves as a common, yet basic form for grouping frames.

2.1.2.2 Histogram of Oriented Gradients

Histogram of oriented gradients (*HOG*) [14] is a descriptor that is based on evaluating well-normalized local histograms of image gradient orientations in a dense grid where local object appearance and shape can often be characterized by the distribution of local intensity gradients or edge directions. The image is divided into small spatial regions (*cells*) where each region accumulating a local 1-D histogram of gradient directions or edge orientations over the pixels of the cell. The combinations of the all the histogram in the cell constructs a general histogram representation. The advantages of the HOG representation is that captures edge or gradient structure that is very characteristic of local shape. Also those extracted characters are insensitive to rotational variations. The image is first processed by colour normalization before computing gradient information. Then each pixel calculates a weighted vote for an edge orientation histogram channel based on the orientation of the gradient element centred on it. Those votes are accumulated into orientation bins over local spatial regions (cells). This algorithm extracts local edge orientation information and normalized to form the global information for the frame. This descriptor

attempts to consider the arrangement of shape and texture across the image and is more exacting in assessing similarly located objects and regions between two frames.

2.1.2.3 Colour Layout Descriptor

The colour layout descriptor [15] captures the spatial layout of the representative colours on a grid in an image as opposed to edge information. An 8 by 8 grid followed by a discrete cosine transform is used to represent the colours. First, an image is converting into YUV colour space then divides into small block with block size of 8 by 8. Then a representative colour is selected from each block which is obtained by computing the average of pixel value in the block. In the third part of this process, each block is processed by DCT to compute three set of 64 DCT coefficients. At last each set of coefficient is zigzag-scanned to form the final colour layout descriptor. This descriptor provides a good alternative for colour histograms to provide more localized colour texture information for frames.

2.1.2.4 Colour Edge Co-occurrence Histogram

The colour edge co-occurrence histogram [16] considered the geometric relationships between pixels and is widely used in image retrieval and object detection. It extracts texture information from the image then constructs that information into a histogram. The image is first constructed into a gradient map. Then it is normalized and filtered by a threshold into a binary edge map. The gradient is calculated by taking the first order partial derivate at each pixel. It is approximated by applying a Sobel Operator to the image and extracts the magnitude and phase value. After obtaining the edge map, each edge point p at location (x, y) is processed to obtain the edge pair p_1 and p_2 locations (x_1, y_1) and (x_2, y_2):

$$\begin{cases} x_1 = x - d \times \cos\theta \\ y_1 = y + d \times \sin\theta \end{cases} \tag{1}$$

$$\begin{cases} x_2 = x + d \times \cos\theta \\ y_2 = y - d \times \sin\theta \end{cases} \tag{2}$$

Where

d denotes as a fix distance between p_1 and p_2

The size of the colour edge co-occurrence histogram is calculated as:

$$CECH(c_1, c_2, d) = \\ size = \left(\left\{ (p_1, p_2) \middle| \begin{matrix} p_1, p_2 \in F, and \\ \|p_1 - p\| = \|p_2 - p\| = d \end{matrix} \right\}, c_1, c_2 \in C \right) \tag{3}$$

Where

c_1 and c_2 are the colour values at points p_1 and p_2

C is the colour set of the image

Function size measures the number of elements in a set

The edge points in the image are scanned and the frequency of occurrence of the same colour pairs are recorded to construct the colour edge co-occurrence histogram. This feature provides the best of both colour and edge spatial layout and is often used when trying to identify exact matches between frames. It is often used in content-based copy detection tasks (e.g. when attempting to identify plagiarized images or video clips)

2.1.3 Grouping, Labeling and Selection of Key Frames

After extracting different features, those features are fed into a clustering algorithm for clustering into groups. This progress aims to remove the redundant frames from the key frame collection. The unsupervised clustering algorithm automatically finds the correlation between the frames (based on features extracted in the previous stage), and divides those frames into a predefined number of groups. One frame will be selected from each group as the key frame of that group.

2.1.3.1 Unsupervised Frame Clustering

Unsupervised clustering algorithms are a good option to process the unstructured video without any prior knowledge on the content of the video. They automatically cluster frames into different groups without human interaction. The most popular unsupervised algorithms for video summarization are K-means, and Self-Organizing Map. These approaches are however, quite dependent on the feature(s) used to describe the frames and the associated similarity metric (distance) used to compare them. For instance, using an HSV colour histogram would formulate groups based on global colour distribution, while spatial layout based features would form groups that share similarities in terms of regions or objects contained within.

The K-means clustering [11, 17, 18] is one of the fast and simplest unsupervised clustering algorithms that classify a given set of data into a certain number of clusters (*k*). Each cluster has a centroid that changes its location according to the input the data set. After certain round of the clustering process, the related frames will group together around a common centroid. Fuzzy C-means [19] is similar to K-means clustering but it allows one data set belongs to two or more clusters, and is better able to deal with overlapping clusters.

A Self-Organizing Map [20, 21] is an abstract mathematical model of topographic mapping from the visual sensors to the cerebral cortex. When presented with a stimulus, neurons compete among themselves for possession or ownership of this input. The winners then strengthen their weights or their relationships with this input (Yin, 2008). The advantage of using the self-organizing map is that is not only cluster the input into different groups at the same time defines the similarities between different clusters where the direct neighbour clusters shares more similarities. Such topological arrangement can be particularly useful in an ensuing search task, which may require the user to switch to nearby/similar content as their information need requires. After initially noticing a frame that might be relevant (used as an early query), a more appropriate frame may be later selected when digging into the dataset further.

2.1.3.2 Supervised Frame Classification

Supervised classification algorithms require human labeling and a training process. This approach is only applicable to specific kind of video or scene: for example, distinguishing between soccer/tennis video or identifying the location of a goal event in soccer video. The labeling process requires definition of exacting criteria of what user is looking for in the video such as score board, the type of shots (long, medium. Close-up and out of field), score scenes and so forth. Visual features are extracted to describe the frame are then associated with a particular label, and a training mechanism is employed such that the label can be predicted from the occurrence of similar features. Support Vector Machine [22, 23, 24, 25] is one of the most commonly used supervised classification algorithms for video summarization. The advantage of the support vector machine is that works well on high dimension data. In supervised classification, the trained model is used to group/label the target video before the key frame selection.

Figure 14: Labelling model training flow chart

2.1.3.3 Evolutionary-Computing Methods

Video summarization has also be defined as an optimization problem in which the goal becomes finding a fixed number of frames that contain the most information about the video. In this case the genetic algorithm has been used. Genetic Algorithm on video summarization [26] is an optimization algorithm to maximize the differences between the selected key frames. First the number of frames in the video summary result is defined. The algorithm automatically selects the frames from the frame extracted in the first stage to form the summary. Genetic algorithm is applied to perform the selection and measurement process. The fitness function of the genetic algorithm is the sum of the mutual information between each

frame pairs in the population. The groups with the minimum mutual information will be selected as “the fittest”, which are then retained to populate the next generation of solutions. When the algorithm reaches a predefined stop criterion, the algorithm picks the frame group with the minimum mutual information value as the final video summary result.

2.2 Attention Based Video Summarization

Visual Attention models are based on human perception, using an algorithm to rank the video frames from different visual parameters. The summarized frames are selected based on the ranking value as calculated throughout the video. According to the human attention mechanism, the attention model is defined from two intrinsic stimulus driven attributes: appearance and motion, which can be considered using either a static or motion model respectively. Additional to these two models, other knowledge based models have been used. Face attention models simulate the fact that humans may pay more attention to human faces. Alternatively, as in commercial for instance, video for instance, video producer implicitly may utilise components that direct the viewers to certain events in the video for example professional cameraman will tend to centre objects of interest (frame) can be detected using camera motion models. Alternatively, speech, music and other sound effects that may shift audience’s attention to certain events have also been considered. For example, comedy videos may include artificial laugh tracks to give audiences hints for the funny part, and can be detected through an aural saliency model. [27]. The majority of attention models from the literature (outlined in this section), produce an attention curve (much like that used in video skimming), from which frames are selected for the summary. Figure 15 summarizes the techniques employed in the literature.

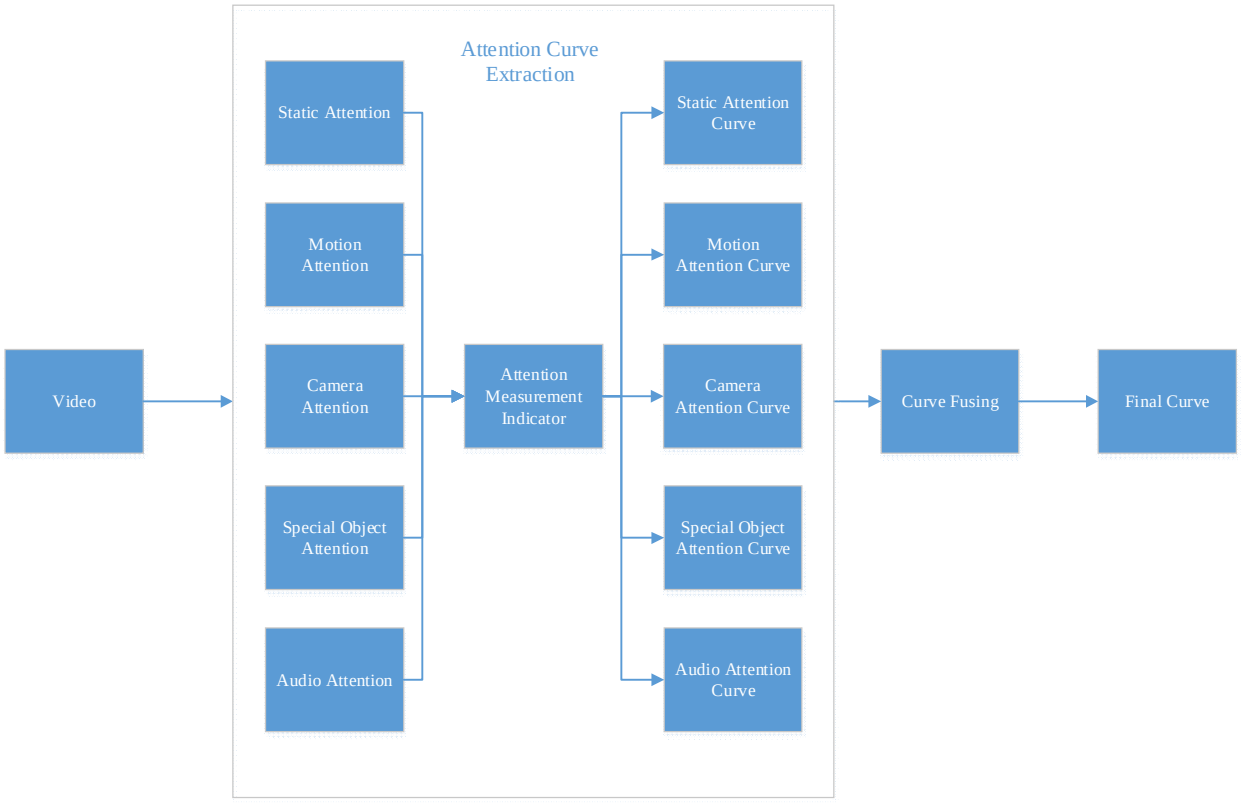


Figure 15: Attention based video summary flow chart

2.2.1 Static Attention

Static attention models are generated by only appearance information in the image, which compares one area with its surroundings to measure its distinctiveness. Peng [28] proposed a contrast based model to compute a static attention feature which takes an intensity feature (I) and colour features (H) to calculate the centre-surround difference (C_i).

There are many other types of the static attention models available. Itti [1] proposed the first visual attention algorithm which applying centre-surround operation to calculate the primary colour oppoenecies (e.g. red-green and blue-yellow), intensity and orientations.

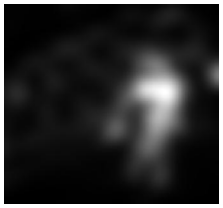
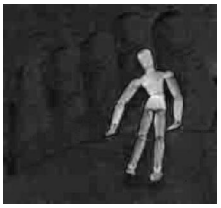


Figure 16: an example of Itti's saliency map left is original, right is saliency map

The Wavelet based method or the Definition Human Visual System (HDHVS) Method [29], [30] wavelets are used to decompose the primary colour opponencies into different scales, perform centre-surround differentials, then inverse transform those image back to the original scales. The HDHVS employs similar concepts as the Itti-Koch algorithm such as centre-surround, colour feature extraction, and normalization to obtain a saliency map. But the key difference is using wavelet to replace the Gaussian pyramids. Wavelet was used to compare local areas to global areas in an image because wavelet enables for approximations to be extracted while details are left untouched. These details can be added back to the final result helping maintain resolution in the saliency map generated. This multi-resolution feature is coupled with lossless resizing capabilities to upsize and downsize the image without loss of information. In terms of detecting colour opponencies that occur close to black/white – i.e. the saliency is not necessarily consistent across all colour opponencies (opposing colours in Figure 8), rather it emphasizes opponencies between blue/yellow and red/green which can be computed as a transformation of the standard red/green/blue (RGB) channels from the raw image.

Origional

Wavelet based



Frequency Tuned

Figure 17: An example saliency maps

The frequency tuned approach applies a band pass filter in frequency domain to blur the image without reducing the resolution of the image. In this way, the algorithm keeps the component detail, but it tends to emphasize the boundaries of the attention regions. Achanta [31] proposed a Difference of Gaussians (DoG) filter to perform the band pass filtering. The DoG filter is widely used in edge detection because of its high efficient on approximating the Laplacian of Gaussian (LoG) filter and it has high sensitivity on intensity changes.

The final saliency value (M_{static}) also depends on the number of pixels and their position, size and saliency map pixel value (C_i). The vertical axis of Figure 18 shows M_{static} as a function of frame number within a video sequence.

$$M_{static} = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot C_i \tag{4}$$

$$w_i = \exp\left(\frac{|p_i - p_{center}|}{2\delta_w^2}\right) \tag{5}$$

A sliding window is used to compare the local average with each value to pick out the key frames from the attention curve. For example, in figure 18, the dots are the key frames locations.

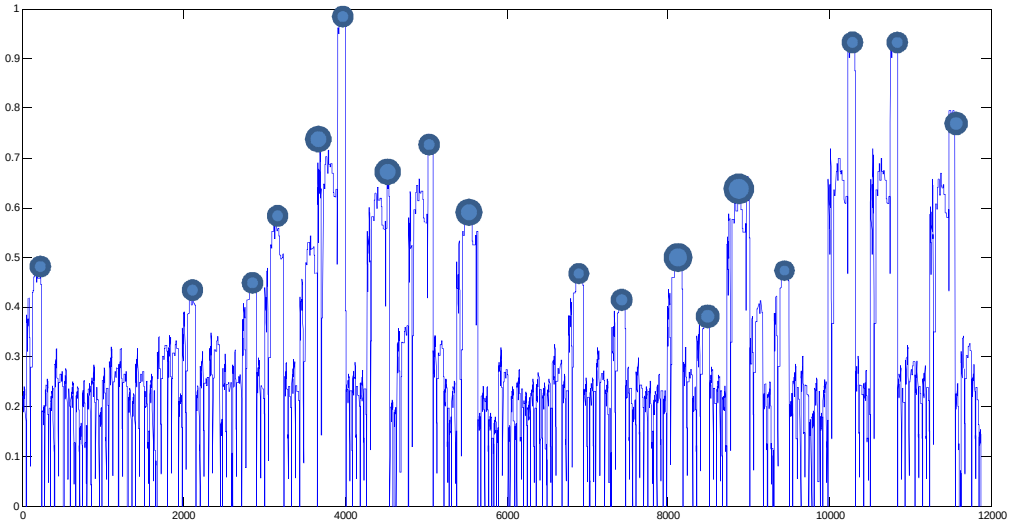


Figure 18: examples of Static Attention Curve, the dots are key frame locations

2.2.2 Motion Attention

The Motion attention models are based on the assumption that people pay more attention to moving objects rather than static objects [27]. The object motion is estimated by a motion vector field (MVF), which divides the image into different blocks with size of *M* by *N* and then tracks the moving direction of the blocks. The MVF has three inductors: intensity, spatial coherence and temporal coherence - which are constructed into three motion maps. The intensity inductor *I* at each block $MB_{i,j}$ ($0 \le i < M$, $0 \le j < N$) is expressed as follows:

$$I(x, y) = \frac{\sqrt{dx_{i,j}^2 + dy_{i,j}^2}}{MasMag} \tag{6}$$

Where

$dx_{i,j}$, $dy_{i,j}$ are two components of motion vector along the *x* axis and *y* axis.

MaxMag is a normalization factor

The spatial coherence inductor induces the spatial phase consistency of motion vectors which is measured using an entropy based method. The phase histogram with a spatial window size of *w* by *w* is sampled in each block. The entropy is calculated as follows:

$$Cs(x, y) = -\sum_{t=1}^{n} p_s(t) Log(p_s(t)) \tag{7}$$

$$p_s(t) = \frac{SH_{i,j}^{w}(t)}{\sum_{k=1}^{n} SH_{i,j}^{w}(k)} \tag{8}$$

Where

$SH_{i,j}^{w}(t)$ is spatial phase histogram

$p_s(t)$ is the corresponding probability distribution function

n is the number of histogram bins

The temporal coherence inductor C_t is calculated by a sliding window with the size of *L*

$$Ct(x, y) = -\sum_{t=1}^{n} p_t(t) Log(p_t(t)) \tag{9}$$

$$p_t(t) = \frac{TH_{i,j}^{L}(t)}{\sum_{k=1}^{n} TH_{i,j}^{L}(k)} \tag{10}$$

Those three motion inductors are fused together as

$$B = I \times C_t \times (1 - I \times C_s) \tag{11}$$

The attended area motions are computed as the motion value of the MVF as the average value in the motion saliency map.

$$M_{motion} = \frac{\sum_{r \in \Lambda} \sum_{q \in \Omega} B_q}{N_{MB}} \tag{12}$$

Where

M_{motion} is the attended motion value

B_q is the motion value of the macroblock in saliency map

Ω is the set of macroblock in the attended area

Λ is the set of attended areas caused by motion activities

2.2.3 Face Attention

When people observe an image with humans in it, people will subconsciously pay more attention to human face [32]. Therefore, a face attention model aims at improving the attention result on frames that include human objects. While this type of model represents the incorporation of top-down prior knowledge, and can be quite effective in producing a summary, the restriction is made on pre-attentive mechanisms that are not domain-based, in order to gather an unbiased sense of any improvement gained by such pre-attentive mechanisms when producing a summary.

2.2.4 Camera Motion Attention

The camera is operated by cameraman to guide audiences to the attention objects [32]. There are six types of Camera motions:

1. Panning and tilting
2. Rolling
3. Tracking and booming
4. Dollying
5. Zooming
6. Still

The first four motions can be modeled in Cartesian coordinates where panning and tilting is camera

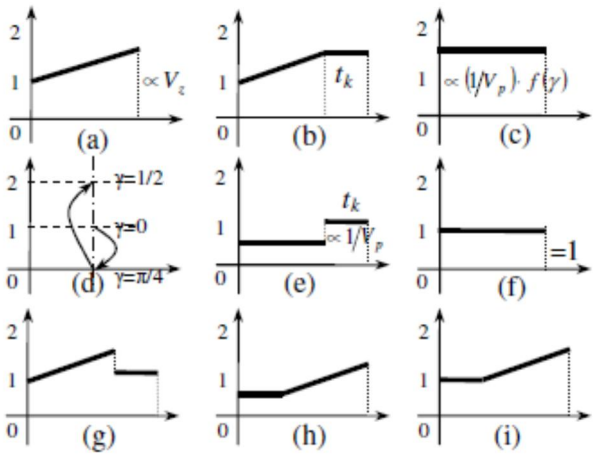


Figure 19: Camera attention modeling (a) Zooming, (b) Zooming followed by still, (c) Panning, (d) Direction mapping function of panning, (e) Panning followed by still, (f) Still and other types of camera motion, (g) Zooming followed by panning, (h) Panning followed by zooming, (i) Still followed by zooming.

rotating around x and y axis; rolling is camera rotating around the z axis; tracking and booming are camera moving along x and y axis; dollying is camera moving along z axis. The zooming is a lens focusing adjustment while still is the camera standing still without any motion.

Based on the above assumptions, the camera motion model could be simplified into 9 attention curves. When a motion attracts human attention it will be labeled as value 2 otherwise it will be labeled as 1. Those motion values will be combined to other saliency maps to create a new saliency map.

2.2.5 Audio Saliency Attention

The audio signal can shift human attention in the same way as visual signals. For example speech, music, or other special sound such as shistle, applause, laughing and expolsion will always appear as louder or sundden sound to attract people's attention which can be defined by sound energy [27]. A sliding windows is used to compute audio saliency along an audio segment. Such that the audio saliency model can be defined as following:

$$M_{as} = \overline{E}_a \cdot \overline{E}_p \tag{13}$$

$$\overline{E}_a = E_{avr} / MaxE_{avr} \tag{14}$$

$$\overline{E}_p = E_{peak} / MaxE_{peak} \tag{15}$$

Where

E_{avr} is the average energy in the sliding windows

E_{peak} is the energy peak in the sliding windows

$MaxE_{avr}$ is the maximum average energy of the audio segment

$MaxE_{peak}$ is the maximum peak energy of the audio segment

Similar to sound energy, speech and music attention is specifically focused on human speech and music other than special sound effects [27]. The assumption behind this model is that music is used to emphasize the atmosphere of scenes in the video and naturally draws an audience's aural attention.. The saliency of speech or music can be measured by the ratio of speech or music to other sounds in an audio segment. In order to measure the ratio a K-Nearest Neighbour is used to classify the speech and music segments (with feature extraction based on common features based on the MPEG7 standard).

Chapter 3 : Proposed Visual Attention Model

As discussed in Chapter 1, Itti-Koch [1] proposed a visual attention model to simulate the way human eyes detect attentional regions. Itti's algorithm extracts three types of features from the input image which are; intensity, colour, and orientation. Those features are processed by a centre-surround operator to generate low-level saliency maps. The centre-surround operator applies a Gaussian image pyramid to blur and down sample the image multiple times then up sample to a designated size and performs a pixel to pixel subtraction to simulate the centre-surround differential operation producing the saliency map. All saliency maps are normalized then added together to generate a final saliency map.

3.1 High Definition Colour 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 [33]. The centre surround algorithm proposed by Itti [1] which is based on the idea that colour differences at different scales trigger neural responses in the human visual system [30]. It is implemented by decomposing an image into lower scale versions in a factor of 2 using Gaussian image pyramids. The low resolution version images are then resized by Bicubic interpolation algorithm to its original image size.

In this work, we propose a new model that takes a series of 7 low resolution images - each 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.

$$I_{c,s}(x,y) = \Delta E_{00}(I_c(x,y), I_s(x,y)) \tag{16}$$

Where

I_c is the original image features

I_s is the resized low resolution image features

ΔE_{00} is the colour difference calculation

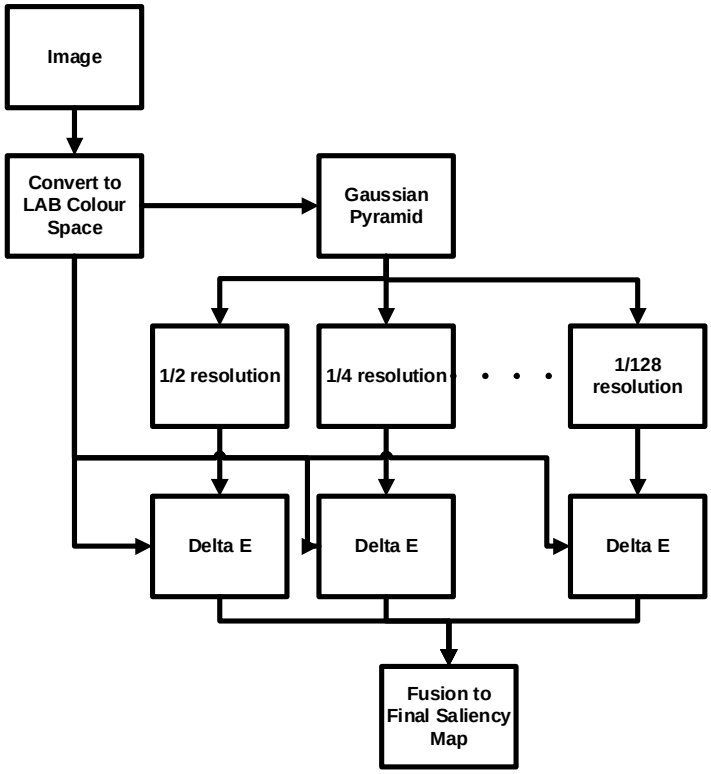


Figure 20: High definition visual attention algorithm flowchart

The colour difference calculation based on colour theory is implemented. When humans observe a colour, they will react to hue difference first, Chroma difference second and lightness differences last [34]. 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.

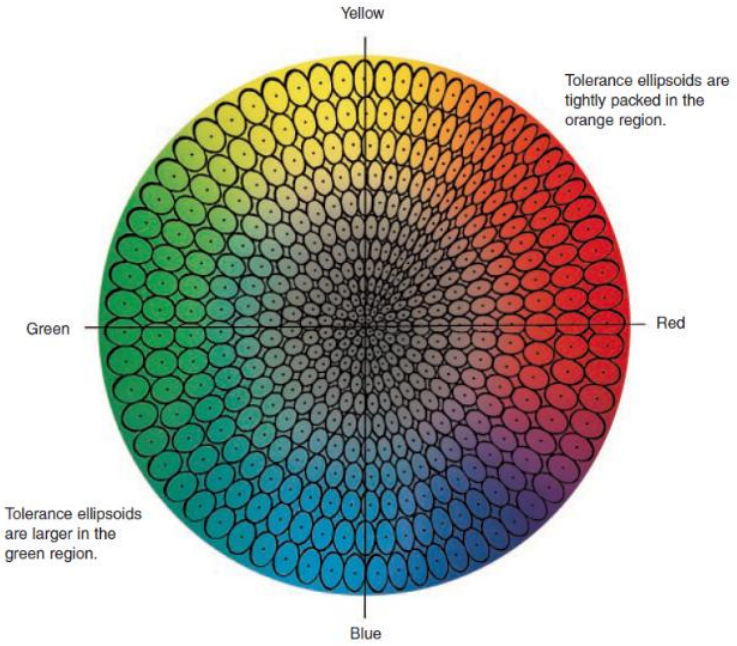


Figure 21: Tolerance ellipsoids in colour space

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 [35]. For more detail on Delta E 2000 standard calculation please refers to appendix A:

$$\Delta E_{00}(L_1^*, a_1^*, b_1^*; L_2^*, a_2^*, b_2^*) = \Delta E_{00}^{12} \tag{17}$$

$$\Delta E_{00}^{12} = \sqrt{\left(\frac{\Delta L'}{k_L S_L}\right)^2 + \left(\frac{\Delta C''}{k_C S_C}\right)^2 + \left(\frac{\Delta H'}{k_H S_H}\right)^2 + R_T\left(\frac{\Delta C''}{k_C S_C}\right)\left(\frac{\Delta H'}{k_H S_H}\right)} \tag{18}$$

Where

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

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

$$N_i(x,y) = \{D_i(x,y) - d_{\min}\} / \{d_{\max} - d_{\min}\} \tag{19}$$

$$N_0 = \sum_{i=1}^{7} N_i \tag{20}$$

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

3.2 Visual Attention Algorithm Results

Before comparing various saliency algorithms it is important to recognize that there are currently no quantifiable methods in which to measure the quality of the saliency map. The comparison in this section is purely subjective and is application driven, meaning different applications will dictate which object is more salient.

Three types of visual attention algorithms: Itti's [1] , wavelet [36], frequency tuned [31] are compared with the proposed algorithm. Itti's algorithm gives a good approximation about the visual attention object's location. The frequency tuned solution provides a fast high resolution solution against the Itti's method, which applies a DoG (difference of Gaussian) operator to perform the centre surround algorithm operation. Instead of Gaussian pyramid to create an array of low resolution image, the DoG operator functions as a band pass filter to remove the high frequency and low frequency components while retaining the same resolution, thereby resulting a high definition version of the saliency map. The wavelet based algorithm process images in the frequency domain, in different scales to give low

resolution, and directional information about the image. The Gaussian pyramid is replaced by the wavelet algorithm and after the centre surrounds operation the low resolution image is inverse-transformed with the previous decomposed frequency components. The wavelet approach improves the visual attention algorithm process and at the same time preserving the image in high resolution. The results from each algorithm will be compared in this section.

From the results, the Itt's method shows the attention locations but without any indications about the objects under the attention region. Itti's method takes the colour opponencies, light intensity and object orientation to perform the centre-surround operation but during those operations, the detail in the image is lost. On the other hand since it approximates the shape of the attention region it can be used for coarse grain object detection and as a seed for semi-supervised image segmentation.

The frequency tuned method provides high definition solution for the visual attention algorithm. It applies a DoG filter to simulate the centre-surround operation which significantly increases the image processing speed and the saliency map resolution. The DoG filter acts like a band-pass filter which removes the low and high frequency information and preserves the middle frequency. From the test results shown later in this section the frequency tuned approach gives a good high resolution detail about the attention object but it tends to ignore less important details. For example the second image of figure 22, the frequency tuned method gives more weight on the persons skin and tends to give significantly lower values to the persons clothing as well as the tree in the background.

The wavelet method provides another alternative high definition approach. It takes advantage of the wavelet decomposition of the image into a low resolution image and 3 coefficients in horizontal, vertical and diagonal direction. Similar to the Itti's approach, a pyramid of image arrays is constructed but after the centre-surround operation, the coefficients are adding back to the low resolution image through wavelet inverse transformation to form high resolution images. The wavelet approach is very sensitive to

bright colours making it ideal for when the object has large colour contrast with the background, such as a bright object on a black background seen in the third and fourth images on figure 22.

On the other hand, it reveals a disadvantage which is shown in figure 23 the first row. When the background is in purple and the person has dark hair. The wavelet method tend to only pick the background and the blue dress that girl is wearing. Furthermore, when the image is black and white the wavelet method seems to indicate that the background is more important. For example in the figure 25 the third image, the wavelet method does not highlight the person in the image.

The proposed method is aiming to provide a high resolution solution for the visual attention process. It is based on Itti's approach. Instead of decomposing the colour data into different opponency pairs, it applies the Delta E standard in the LAB colour space to perform the colour comparisons. Gaussian pyramid low resolution images are compared with the original image to construct saliency maps in different scale which helps the final saliency map stay in high resolution while at the same time capturing the attention object in different levels. This approach has a better ability to capture small attention details than the previous other methods. For example the first image on figure 23, only the proposed method detected both the girl and the signature. In figure 25, the algorithm detected both the man in a black suit and the black balls in the background. Compared to the wavelet algorithm the proposed algorithm does not do a very good job in detecting salient objects when they have high contrasting colours with the background, such as the results seen on the third image on figure 22.

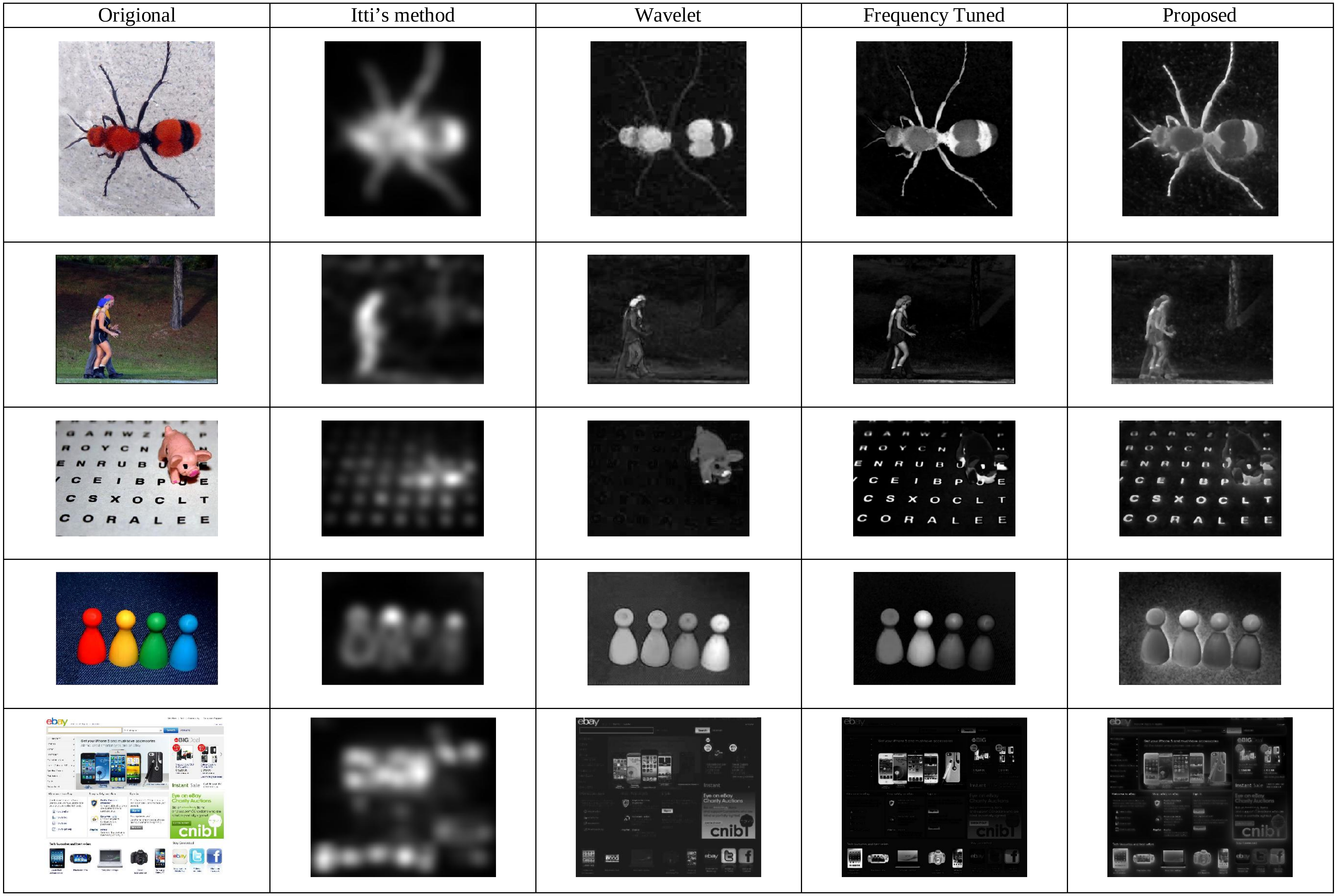


Figure 22: Visual Attention Method Comparison 1

Origional	Itti's method	Wavelet	Frequency Tuned	Proposed

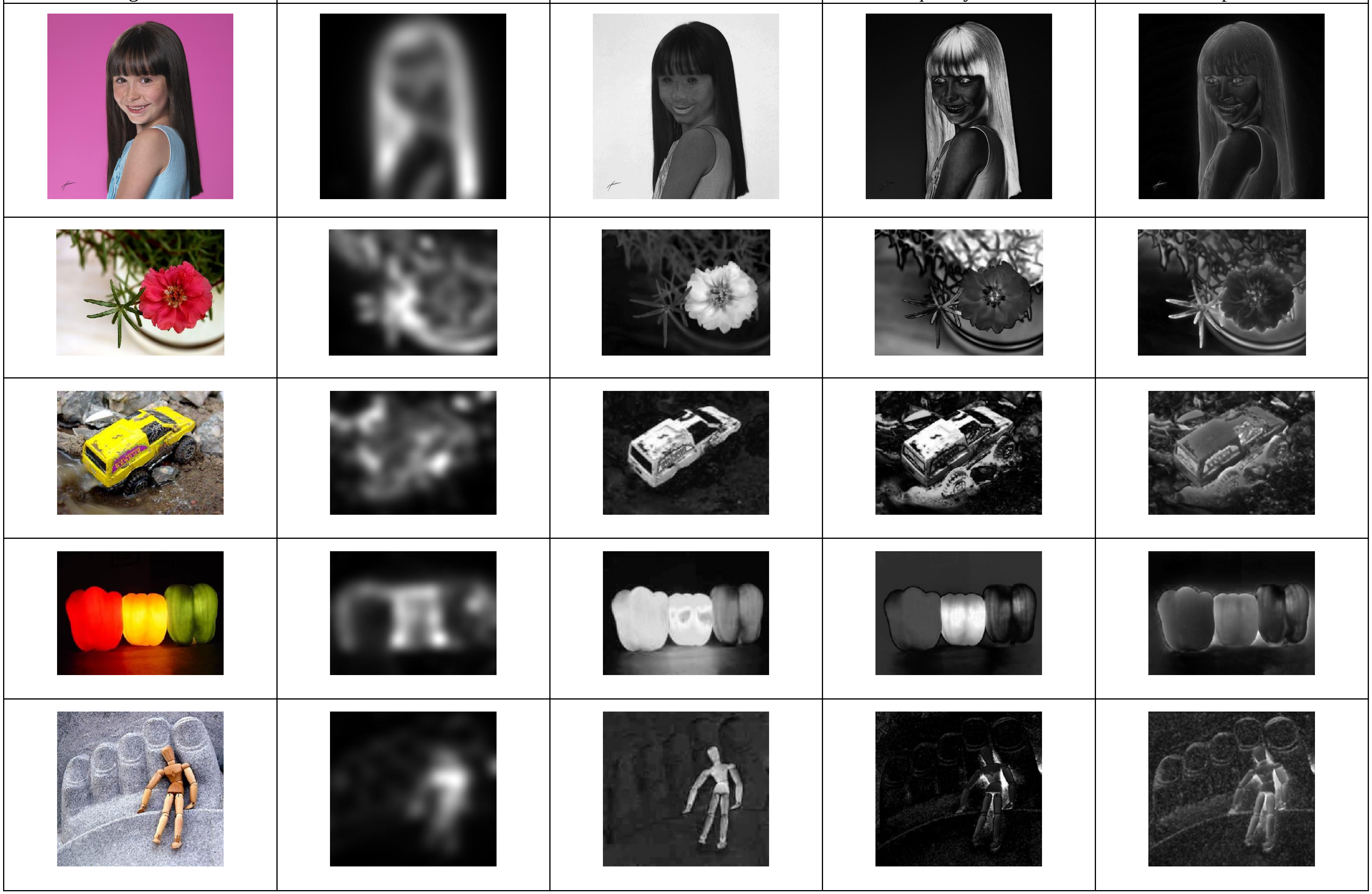


Figure 23: Visual Attention Method Comparison 2

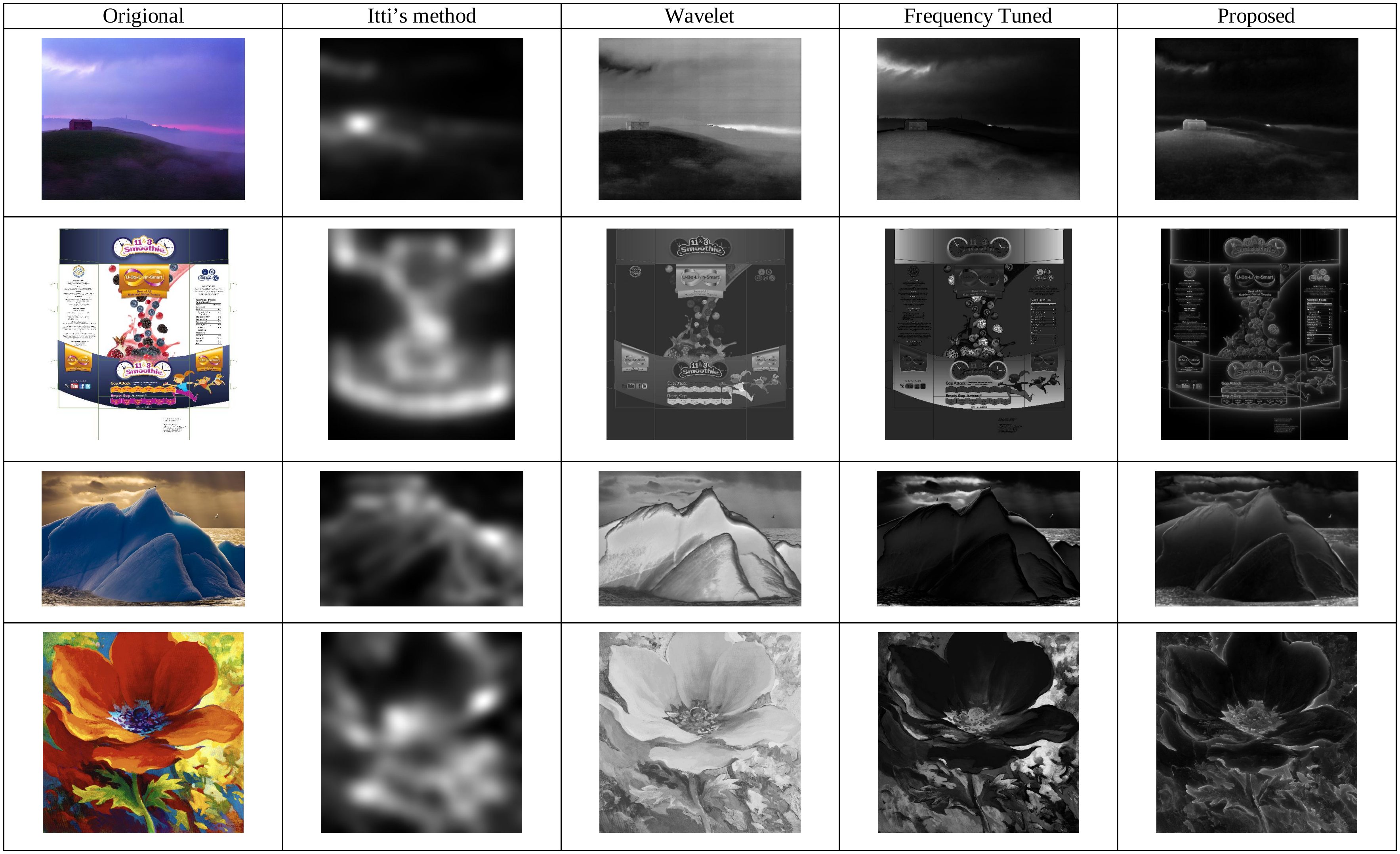


Figure 24: Visual Attention Method Comparison 3

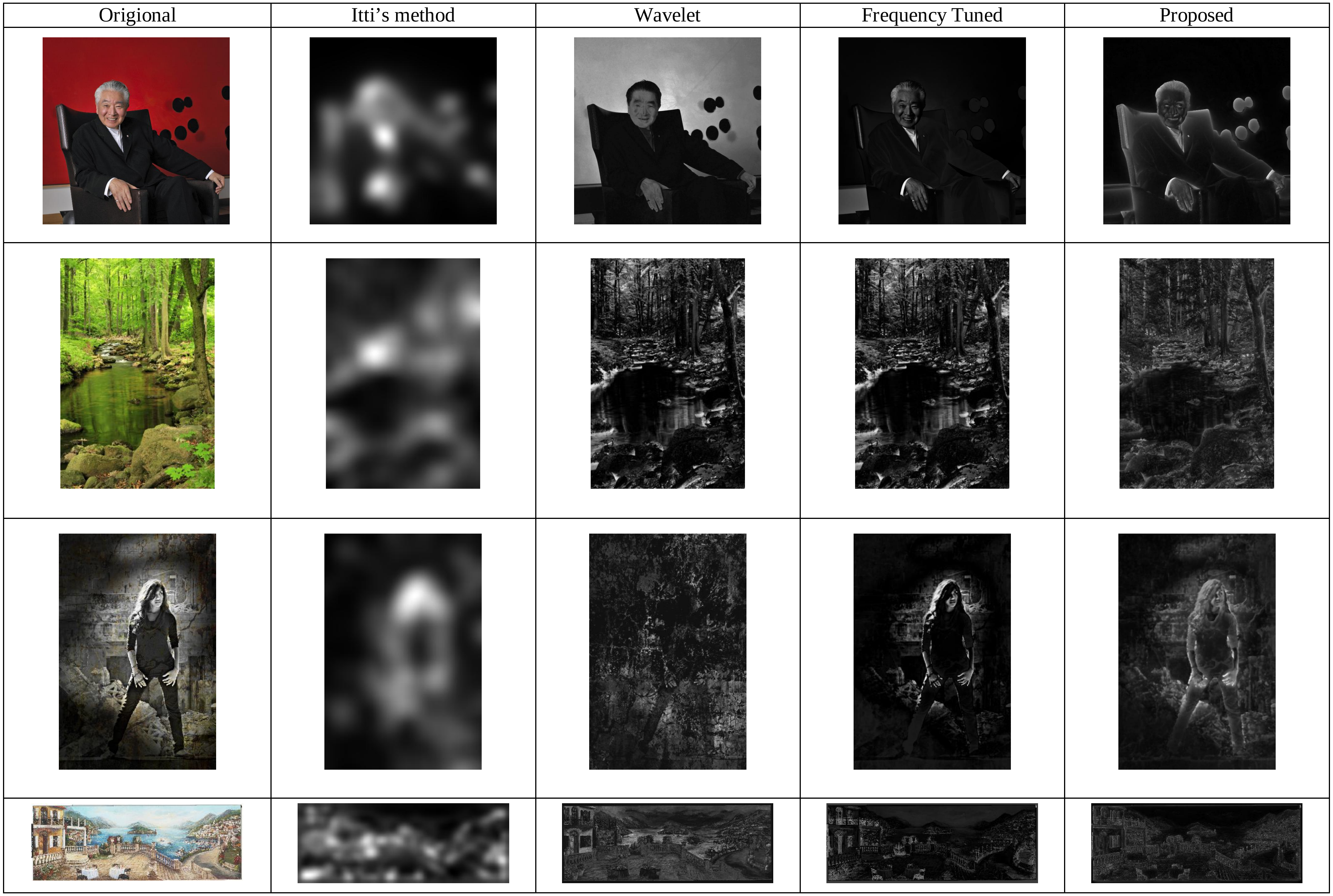


Figure 25: Visual Attention Method Comparison 4

Chapter 4 : Proposed Video Summarization Algorithm

A high definition visual attention based self-organizing map video summary algorithm is proposed. 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. 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. From the experiment visual attention algorithm improved the agreement rate between the ground truth and predicted results.

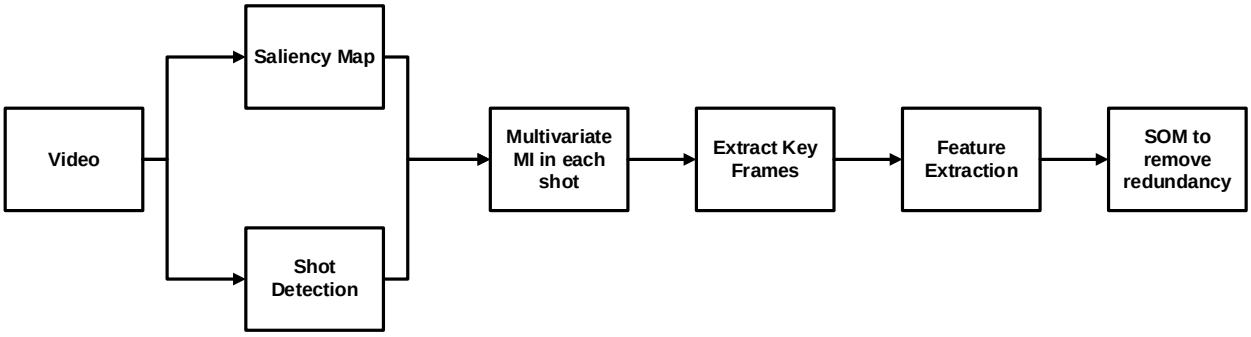


Figure 26: High definition video summarization flowchart

4.1 Shot Detection

An HSV histogram based adaptive threshold shot boundary detection algorithm is implemented. The frames are first converted from RGB to HSV colour space. Three separate 512 bin histogram are constructed on H, S, and V channel. The Euclidean distances between adjacent frames are calculated as a parameter to construct a curve determining the shot boundary. The threshold of this shot boundary curve is adaptively determined by a sliding window [37, 38]. In this experiment, the windows size is set as 40. The threshold in the window is calculated by following equation:

$$Threshold = \mu + Td\sqrt{\sigma} \tag{21}$$

Where

Td is a constant, in the experiment *Td* is set to 5

μ is the local mean

σ is the local variance

4.2 Extract Attention Curve

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 within a shot. The multivariate mutual information calculates the similarity of a frame against all the frames in the 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 that is generated by proposed Visual Attention Model 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 potentially less important information.

The mutual information is a measure of the amount of information one random variable contains about another which also could be seen as a measure of the distance between two probability distributions [39, 26]. Let χ be a finite set and X be a random variable taking values x in χ with distribution $p(x) = Pr[X=x]$. Similarly, Y is a random variable taking values y in Υ. The Shannon entropy $H(X)$ of a random variable X is defined by

$$H(X) = -\sum_{x \in \chi} p(x) \log p(x) \tag{22}$$

The joint entropy of X, Y was defined as

$$H(X,Y) = -\sum_{x \in \chi} \sum_{Y \in \gamma} p(x,y) \log p(x,y) \tag{23}$$

The mutual information of the X and Y was expressed as

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{24}$$

$$I(X,Y) = \sum_{x \in \chi} p(x) \sum_{Y \in \gamma} \log p(y \mid x) \log \frac{p(y \mid x)}{p(y)} \tag{25}$$

Instead of only calculating the mutual information between two frames, the multivariate mutual information is calculated within a single shot.

$$M(k) = \sum_{v=1}^{S} I(k,v) \tag{26}$$

Where

M is the multivariate mutual information for the kth frame

S is the number of frames in the shot

4.3 Feature Frame Extraction

One frame from each shot is selected to represent the whole shot. The selection algorithm is based on select the frame with highest the multivariate mutual information value with in that shot.

$$VAI = Max(M) \tag{27}$$

Where

M is the multivariate mutual information value

VAI is the frame index that selected as a feature frame

4.4 Self-Organized Mapping

A self-Organizing Map (SOM) is an abstract mathematical model of topographic mapping from the visual sensors to the cerebral cortex. When presented with a stimulus, neurons compete among themselves for

possession or ownership of this input. The winners then strengthen their weights or their relationships with this input [40]. The self-organizing map learning process is following:

1. Initialize each node's weights
2. Choose vector from training data and input into SOM
3. Find the best Matching Unit (BMU) by calculating the distance between the input vector and the weight of each node

$$Dist = \sqrt{\sum_k \|V_k - W_k\|^2} \tag{28}$$

4. The radius of the neighbourhood around the BMU is calculated. The size of the neighbourhood decreases with each iteration.

$$\sigma(t) = \sigma_0 \exp(-\frac{t}{\lambda}) \tag{29}$$

Where

t is the number count of iteration loops

$\sigma(t)$ is the neighbourhood size at t^{th} loop

σ_0 is the initial radius

λ is the time constant

V_k is the input vector value

W_k is the node weighting vector

5. Each node in the BMU's neighbourhood has its weights adjusted to become more like the BMU. Nodes closest to the BMU are altered more than nodes furthest away in the neighbourhood.

$$\eta(t) = \exp[-\frac{Dist^2}{2\sigma(t)^2}] \tag{30}$$

$$L(t) = L_0 \exp[-\frac{t}{\lambda}] \tag{31}$$

$$W(t+1) = W(t) + \eta(t)L(t)(V(t) - W(t)) \tag{32}$$

6. Repeat from step 2 to 5 till reached the stopping iterations number

The frames with the median weight in its group will be selected as feature frames

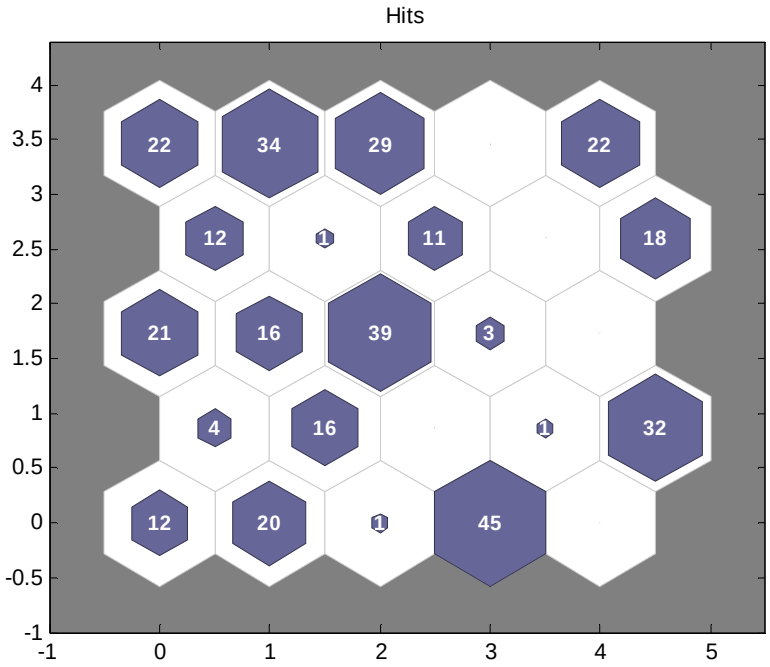


Figure 27: An example of Self-Organizing Map results

The HSV colour histogram and Histogram of Gradient features were used in the clustering. After the self-organizing maps algorithm progresses, the features frames were categorized into different groups. One frame with median weight was selected from each group to form the final feature frame summary. The advantage of SOM over other clustering method is that the SOM clusters frames of similar content and arrange them spatially such that frames close to one another are more similar, which facilitates search processes that might follow. The framework can facilitate the act of 'drilling down' into the video collection by selecting and re-displaying related sets of frames.

4.5 Video Summary Algorithm Results

The test videos for this project are from the Open Video Project (http://www.open-video.org). In the website, it provides both the original video files and a ground truth summary. The ground truth summary is created by a hybrid machine-human process which a colour histogram based frame selection algorithm generates hundreds of 'candidate' key frames then human viewer selects the key frames from those candidates [9]. The proposed method was tested to 5 videos with different length and types. Moreover one of the videos was a black and white video. The self-organizing map size for this experiment was set to 5x5 solely because of the ground truth frames were on the order of 20 frames. The computer used for this experiment was an i7 due core laptop with 16G of RAM.

The videos were first processed by the proposed algorithm to generate a storyboard. Due to the size of the Self-organizing map, the maximum number of feature frames that the proposed algorithm selected from a video was 25 frames. The storyboard generated by the visual attention enhanced video summary and without visual attention enhanced video summary are then manually compared with the ground truth from the open video library. The agreement rate was recorded in the following table.

Table 1: Video Summary results

	Calvin Workshop	Hurricanes	Seamless Media Design	Senses and Sensitivity, Lecture	Lucky Strike
Length (min)	6:35	3:54	5:57	27:14	1:00
Type	Comedy	Documentary	Educational	Lecture	Commercial
Colour	Coloured	Coloured	Coloured	Coloured	Black/White
Itti's Method	9/18(50%)	17/27(62.9%)	4/18(22.2%)	12/22(54.5%)	5/6(83.3%)
Wavelet Method	7/18(38.9%)	16/27(59.3%)	3/18(16.7%)	13/22(59.1%)	5/6(83.3%)
Proposed Method	10/18 (55.6%)	17/27(62.9%)	5/18(27.8%)	15/22(68.1%)	6/6(100%)
Method without Visual Attention	7/18(38.9%)	14/27(51.9%)	2/18(11.1%)	12/22(54.5%)	5/6(83.3%)

As shown in the experimental result above, the proposed method shows reasonable agreement with frames chosen in the ground truth in all those 5 videos. Depends on the length and type of the video, the result is range from 16.7% to 100% agreement for visual attention enhanced video summary and 11.1% to 83.3% agreement for without visual attention enhanced video summary. It is important to note that this does not represent “accuracy” percentage, but rather a tendency for the algorithm to automatically select summary frames that correspond to human choices. The visual attention algorithm generally could provide about 15% increases in agreement rate. This impact is from the visual attention algorithm filtering out the unnecessary background information and enhances the attention area details that are used when comparing frames.

The first video is a gorilla that acts like a human to perform some tasks. This video contains multiple scenes which were detected correctly by the shot detection algorithm. The main attention object that is picked up by the proposed algorithm are the gorilla, work station and camera which gives out a general idea about the video. The visual attention summary picked up more attention object than without visual attention. Viewers can guess the video is about a gorilla working in a film studio to produce movies. The second video is a cigarette commercial in black/white where, due to its simple and repeated video structure, the difference between with and without visual attention enhancement is minor. The video summary algorithm correctly picked up most of the important object from the commercial. The third video is a documentary about the destruction power of the hurricanes. The documentary type video has well-structured scenes so that the key frames are easier to be identified. This documentary shown the power of hurricane flooded the city and even blow a ship onto the land. The fourth video demonstrates the potential markets of the seamless media design technology. Due to its complex structure and large number of attention objects, the video summary algorithm gives a lower rate of agreement with the ground truth. But again, the visual attention provides a strong impact on this video and increases the agreement rate.The final video is a 30 minutes lecture, which is structurally simpler with large number of still and repeated scenes. The video summary algorithm compressed the lecture into a 25 frames

storyboard. There are three main groups that draw viewers' attention, first professor, second student in the class, and third the board. The video summary correction picked up all these objects without any prior knowledge. The detailed test results are in next sections.

4.5.1 Calvin Workshop

With Proposed Visual Attention

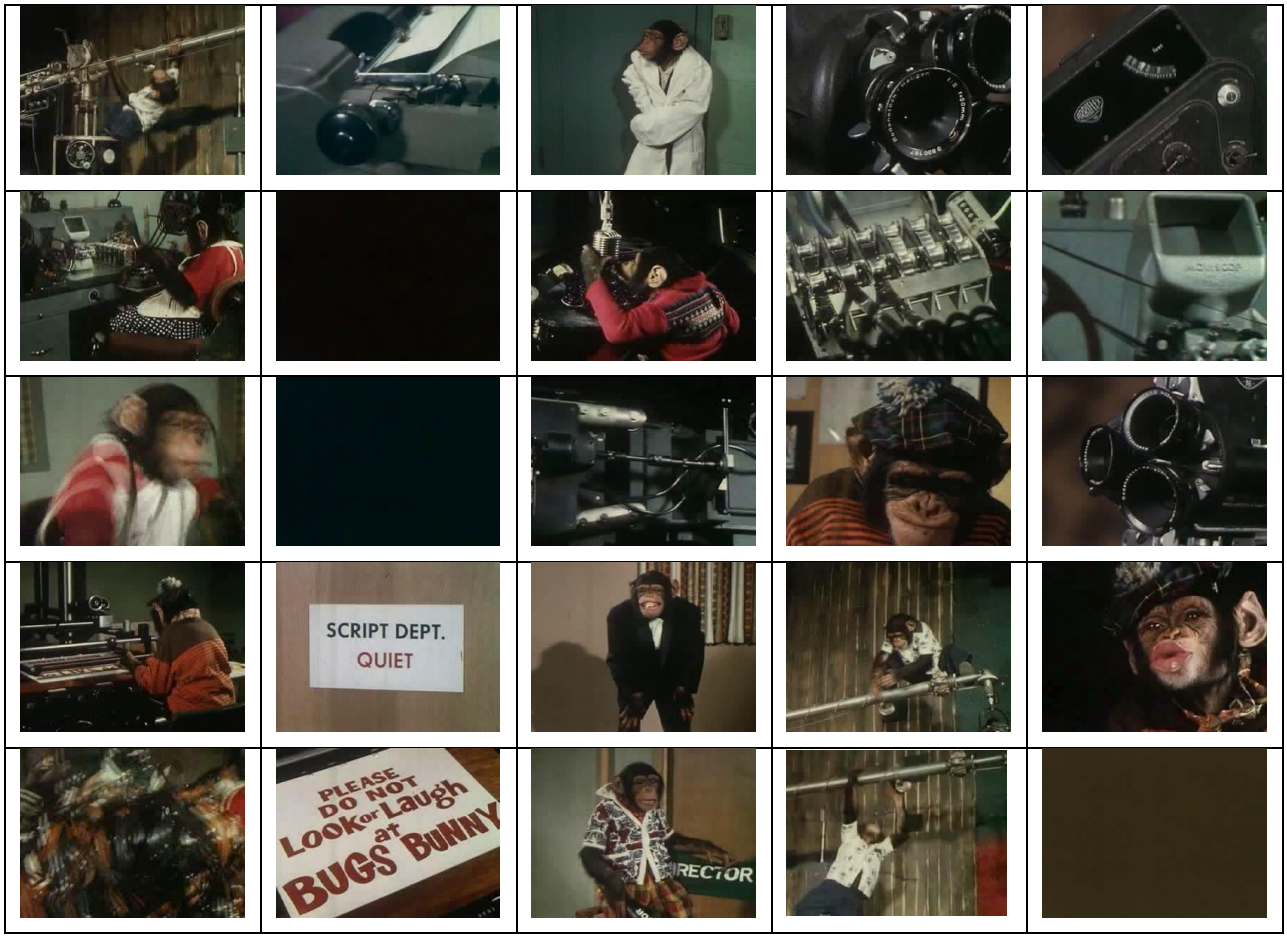


Figure 28: Proposed Algorithm with Proposed Visual Attention Process Calvin Workshop Video Test Results

With Itti's Method

Figure 29: Proposed Algorithm with Itti's Visual Attention Process Calvin Workshop Video Test Results

With Wavelet Method

Figure 30: Proposed Algorithm with Wavelet Visual Attention Process Calvin Workshop Video Test Results

Without Visual Attention

Figure 31: Proposed Algorithm without Visual Attention Image Process Calvin Workshop Video Test Results

Ground Truth

Figure 32: Ground Truth of Calvin Workshop Video

4.5.2 Lucky Strike Cigarette Commercial

With Proposed Visual Attention

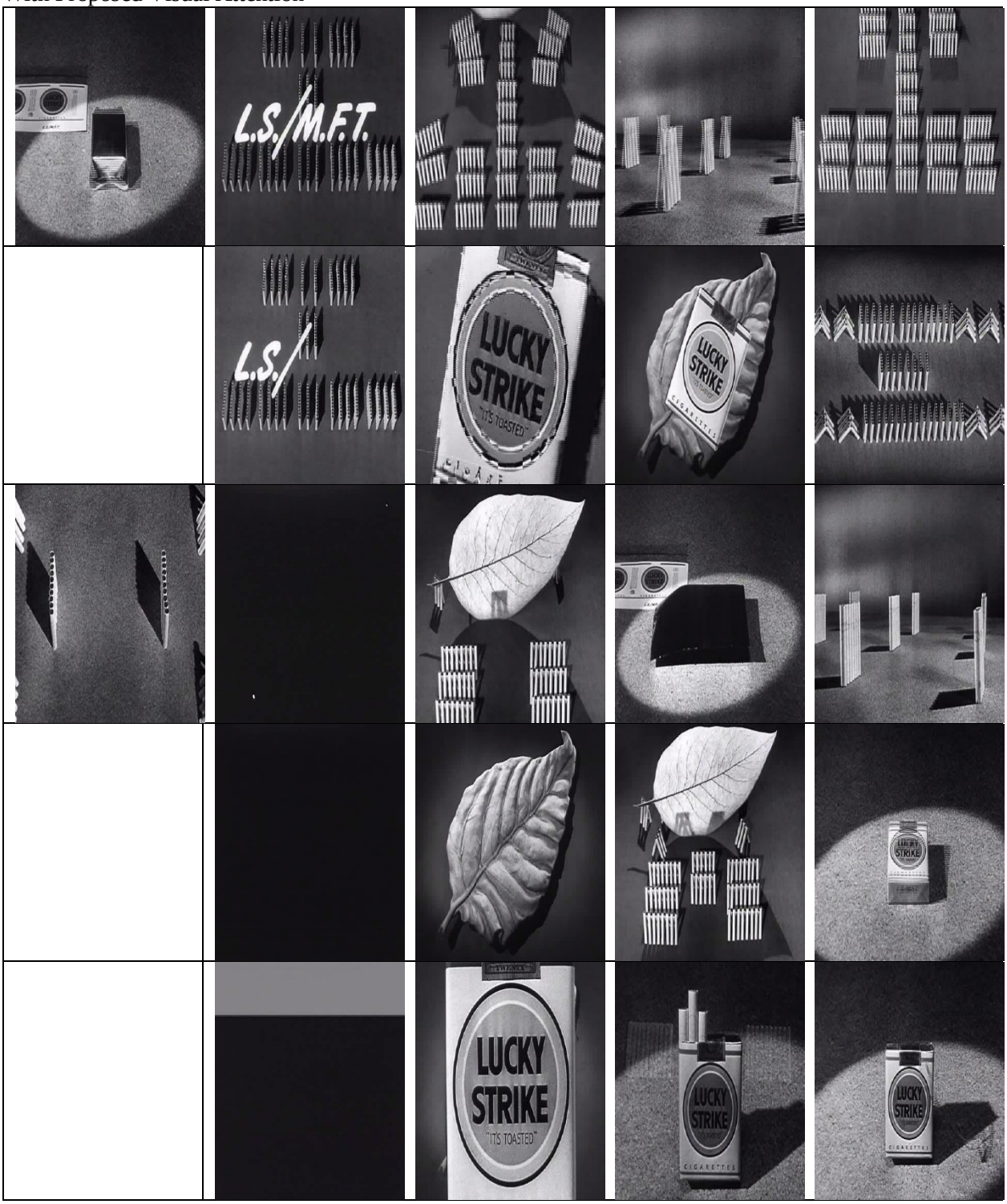


Figure 33: Proposed Algorithm with Proposed Visual Attention Process Lucky Strike Cigarette Commercial Video Test Results

With Itti's Method

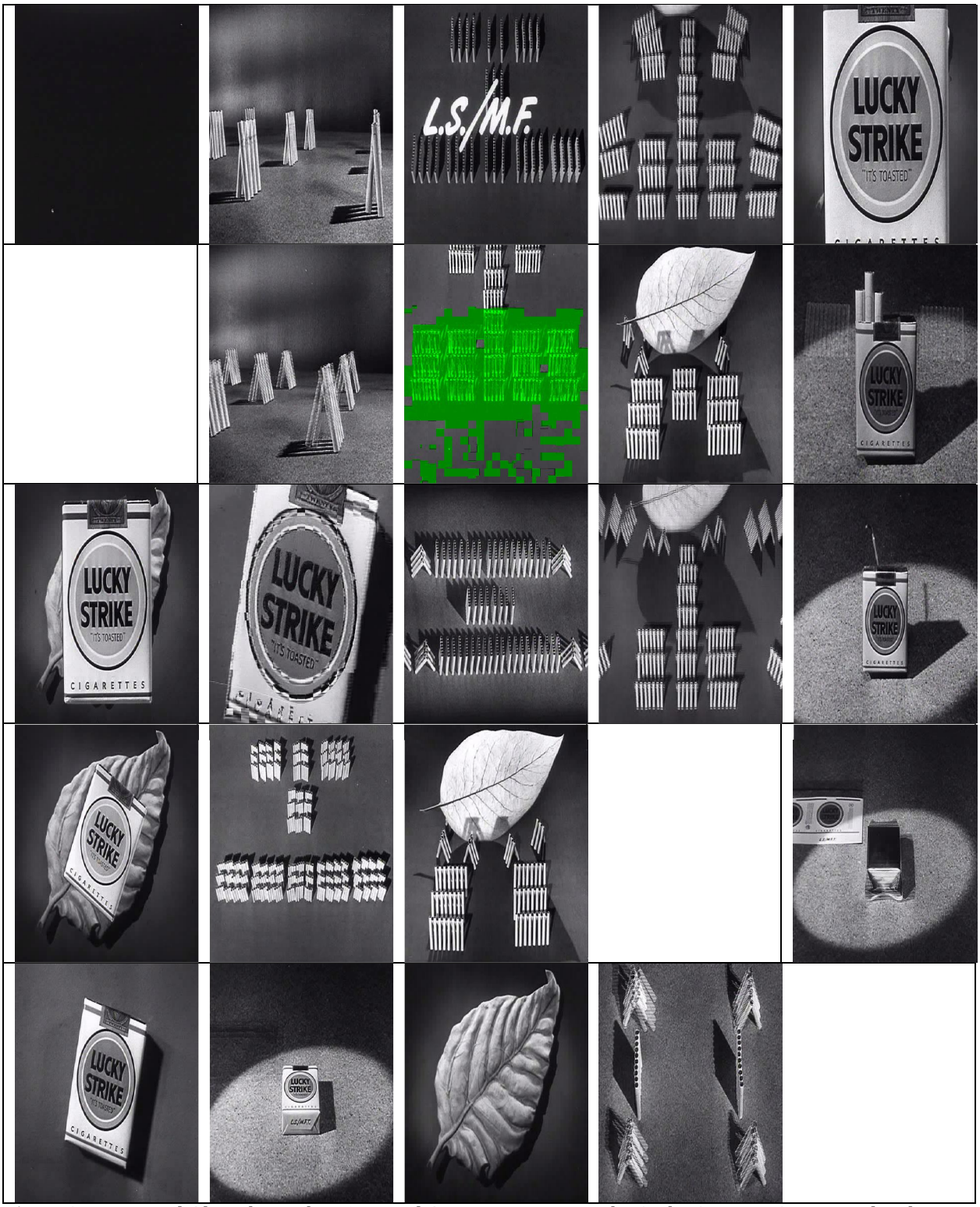


Figure 34: Proposed Algorithm with Itti's Visual Attention Process Lucky Strike Cigarette Commercial Video Test Results

With Wavelet Method

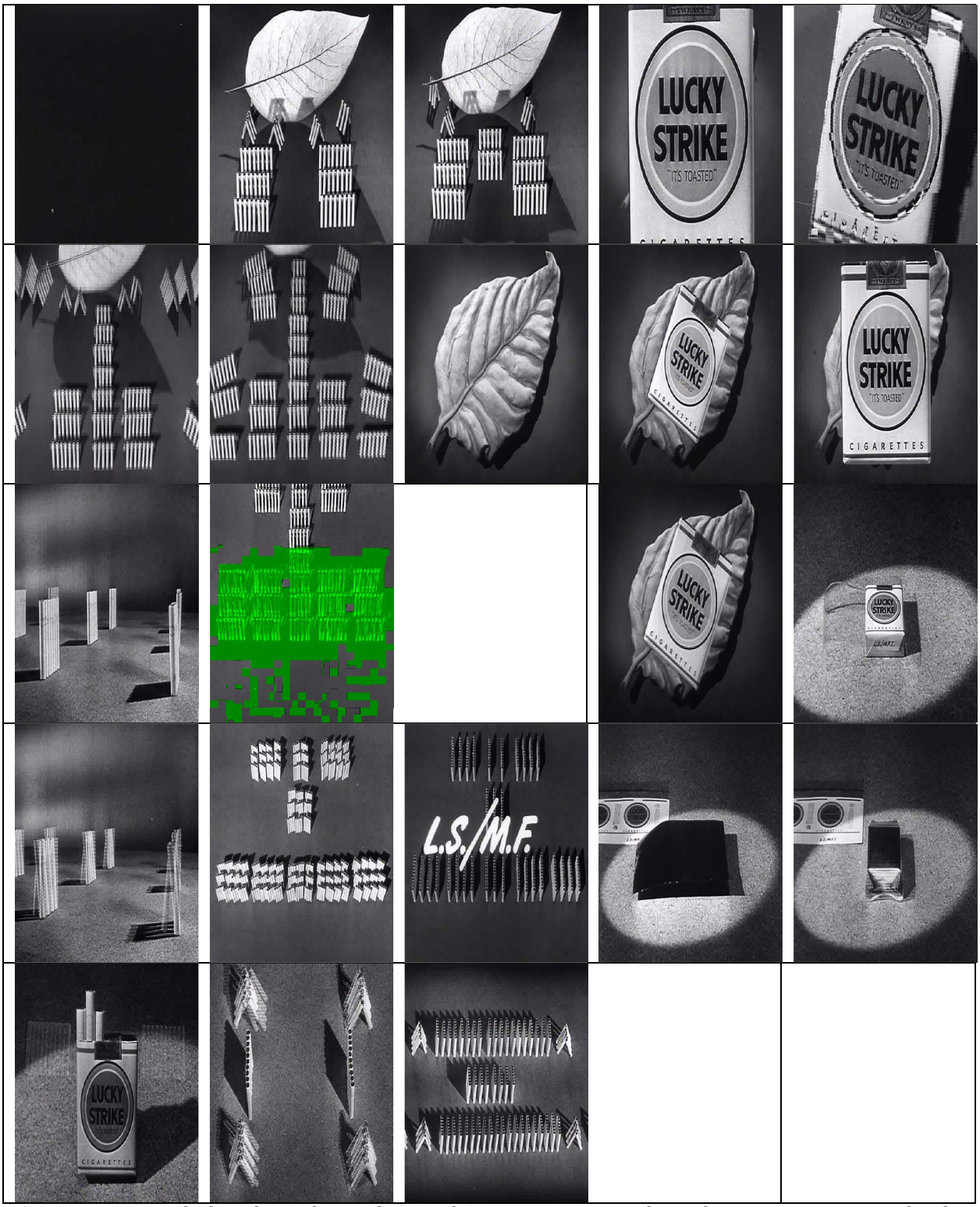


Figure 35: Proposed Algorithm with Wavelet Visual Attention Process Lucky Strike Cigarette Commercial Video Test Results

Without Visual Attention

Figure 36: Proposed Algorithm without Visual Attention Process Lucky Strike Cigarette Commercial Video Test Results

Ground Truth

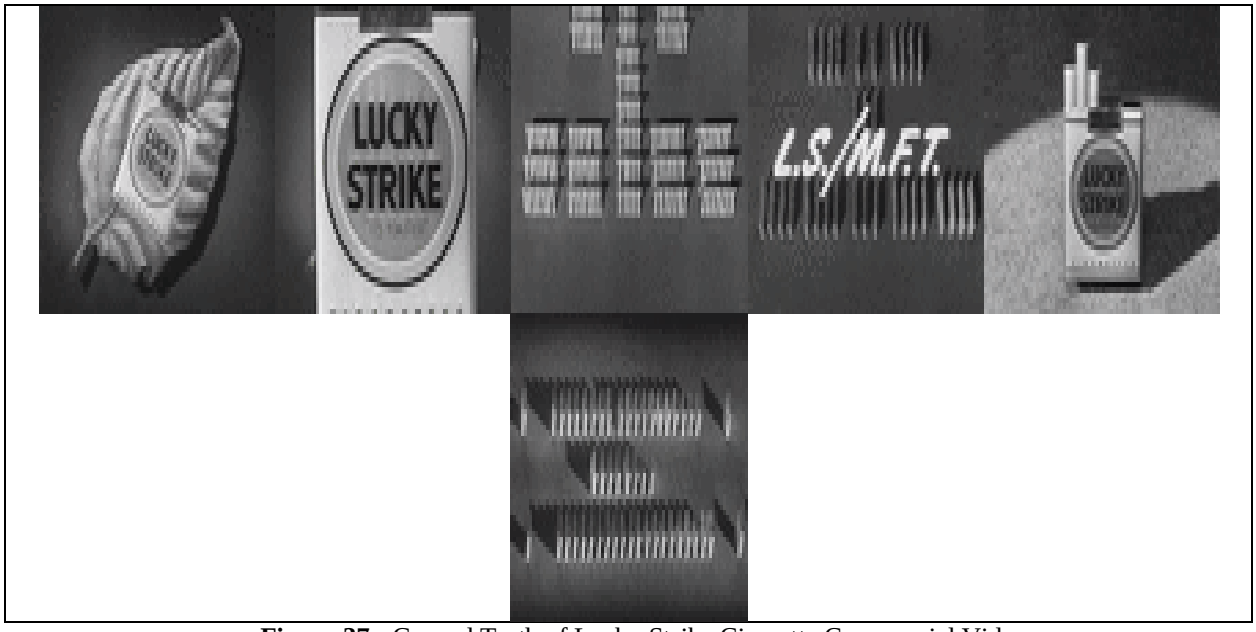


Figure 37: Ground Truth of Lucky Strike Cigarette Commercial Video

4.5.3 Hurricanes

With Proposed Visual Attention

Figure 38: Proposed Algorithm with Proposed Visual Attention Process Hurricanes Video Test Results

With Itti's Method

Figure 39: Proposed Algorithm with Itti's Visual Attention Process Hurricanes Video Test Results

Without Wavelet Method

Figure 40: Proposed Algorithm without Visual Attention Process Hurricanes Video Test Results

Without Visual Attention

Figure 41: Proposed Algorithm without Visual Attention Process Hurricanes Video Test Results

Ground Truth

Figure 42: Ground Truth of Hurricane Video

4.5.4 Seamless Media Design

With Visual Attention

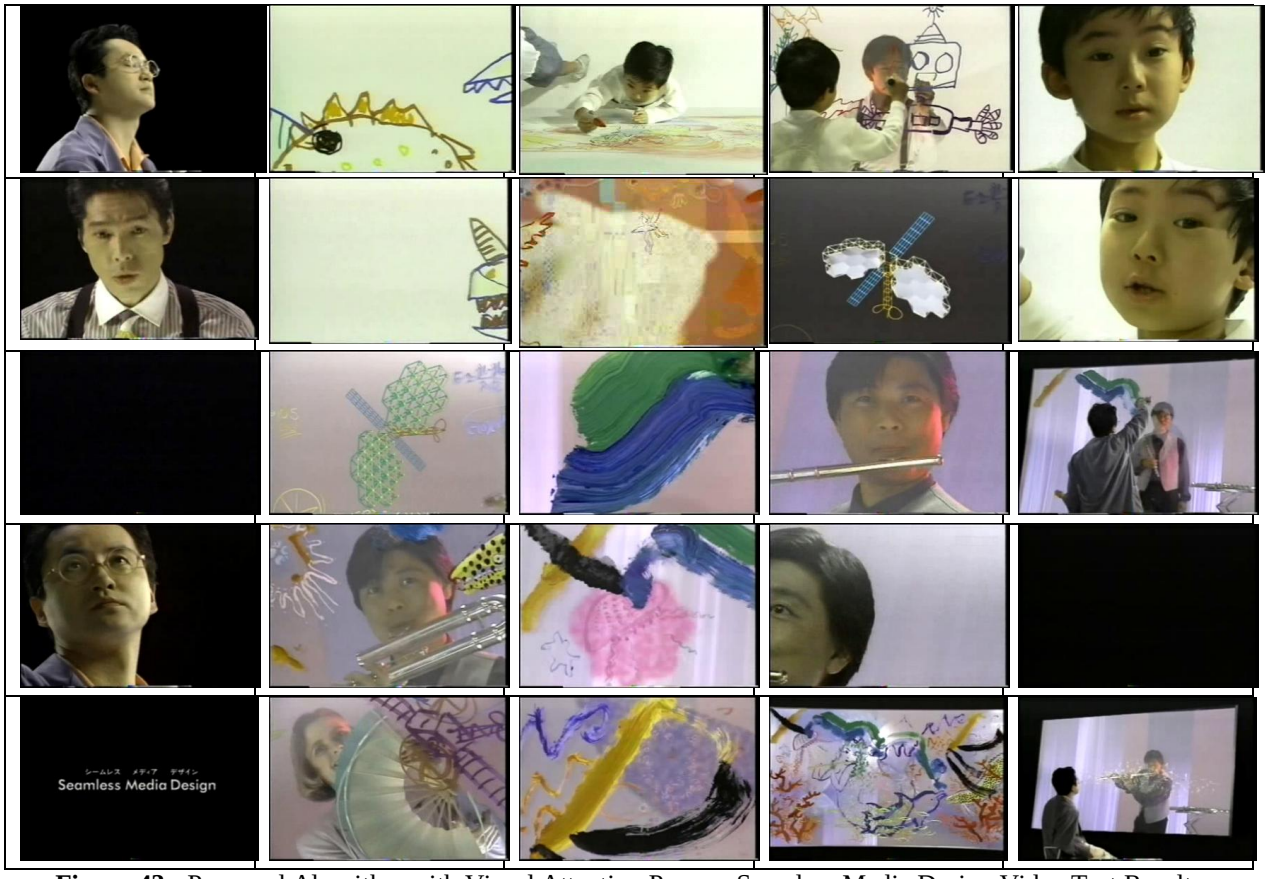


Figure 43: Proposed Algorithm with Visual Attention Process Seamless Media Design Video Test Results

With Itti's Method

Figure 44: Proposed Algorithm with Itti's Visual Attention Process Seamless Media Design Video Test Results

With Wavelet Method

Figure 45: Proposed Algorithm with Wavelet Visual Attention Process Seamless Media Design Video Test Results

Without Visual Attention

Figure 46: Proposed Algorithm without Visual Attention Process Seamless Media Design Video Test Results

Ground Truth

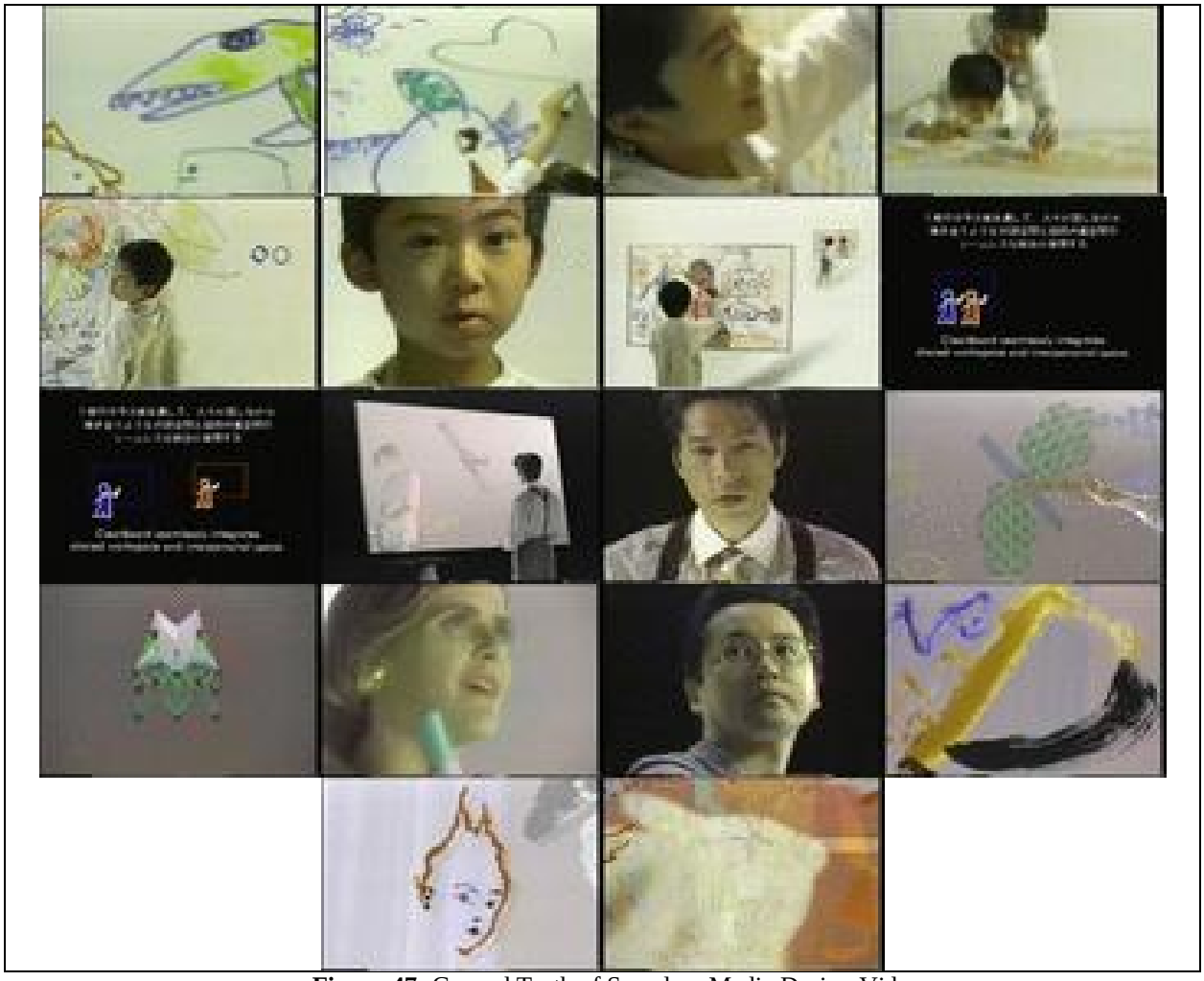


Figure 47: Ground Truth of Seamless Media Design Video

4.5.5 Lecture

With Visual Attention

Figure 48: Proposed Algorithm with Visual Attention Process Lecture Video Test Results

With Itti's Method

Figure 49: Proposed Algorithm with Itti's Visual Attention Process Lecture Video Test Results

With Wavelet Method

Figure 50: Proposed Algorithm with Wavelet Visual Attention Process Lecture Video Test Results

Without Visual Attention

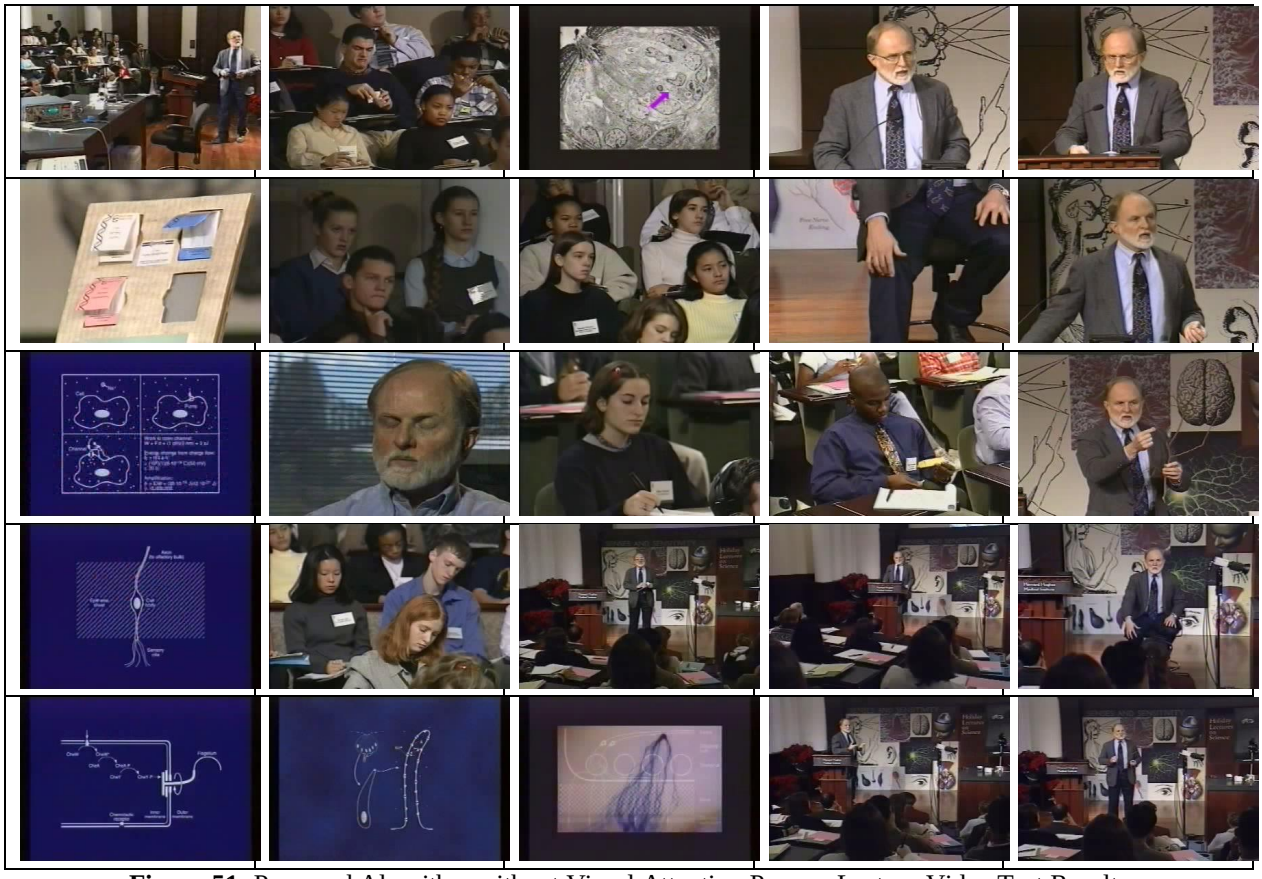


Figure 51: Proposed Algorithm without Visual Attention Process Lecture Video Test Results

Ground Truth

Figure 52: Ground Truth of Lecture Video

Chapter 5 : Conclusions and Future work

The proposed video summarization algorithm applies a novel high definition visual attention algorithm and a multivariate mutual information algorithm to select a series of feature frames from a video. Then a self-organizing map is applied to those feature frames to remove redundant frames. The size of the self-organizing map can be interactively defined by user, thereby reduce or increasing the amount of images created to summarize the video. The advantage of this method is it simulates the human visual system's sensitivity to colour opponency by using colour theory to extract a detailed attention region from the background. The proposed approach works on both colour videos and black-white films in a manner that produces summaries that are more aligned with human choices with respect to the use of previous colour attention models from the literature.

The storyboard generated by the visual attention enhanced video summary and without visual attention enhanced video summary are compared with the ground truth from the open video library (http://www.open-video.org). From the comparison the visual attention enhanced video summaries deliver higher agreement rate with the ground truth compare with the video summary without visual attention. It is worth to mention that across the board, for all attention models tested the summarization showed stronger alignment to the human constructed storyboards. It is expected that the incorporation of some top down knowledge (such as face detection or presence of humans may further improve summaries generated).

The proposed method provides another alternative method to produce saliency maps, which is further extended in its use in the video summarization algorithm. It provides a different perspective of the representation of the video in the form of a story board, which can be used as access entry point for video search, indexing and annotation purposes.

5.1 Merging Summarization with Video Search

The future work for this algorithm is a user oriented exploration framework for video searching based on graph theory. In social network graphs, 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 applies similar 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 a chosen clustering algorithm. One frame could have multiple connections with another frames (based on different features considered), or it could only have only one connection (see figure 54). During the clustering process, the frames within the same group will be connected.

In this way, the number of connection levels can be used to indicate the strength of the relation for example if object A is connected to object B in terms of 3 different feature descriptors, while only connected to object C in terms of one feature descriptor, then object A might be considered to have a stronger relationship with object B. When user selects a node, the nodes that share a connection with will be displayed on a 2D canvas, which can be considered in terms of one or more features (colour, edge histogram, or some hybrid/combination). User can filter out some connection nodes by adjusting the framework setting (i.e. switching a feature type on/off). The user could then browse through those connections to explore the whole video collection.

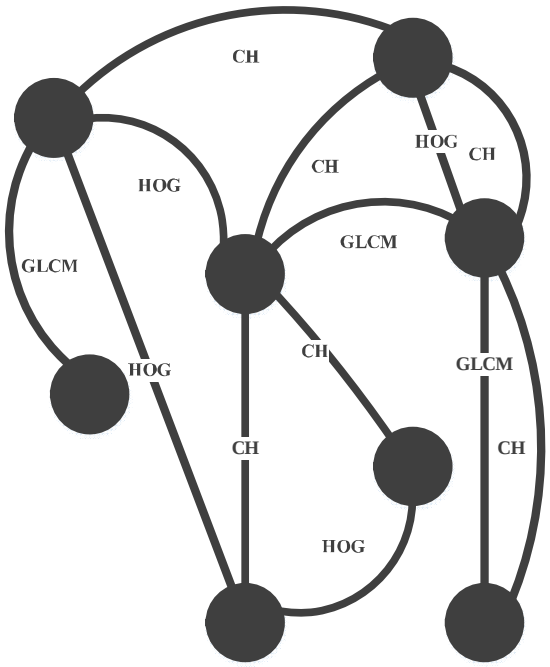


Figure 53: An example of proposed network structure

This structure has a few advantages:

1. User interaction: user is being 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 clustering process can run on the cloud server, accessible through low computational power mobile platforms.

5. Upgradability: new cluster features or new videos could be added to the system without large structure modification.

This framework will provide 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 storyboards from videos, depending on the user setting, a 30-minute video could be summarized into an N-frame storyboard, which could be adjusted by the user interactively. The frame exploration provides the user with 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. While such frameworks could be implemented in a relational database, extracting such interconnections would lend themselves to a graph database implementation, to more efficiently handle large information.

This network structure could enable more interactive experience as user could use a video summary as a starting point and browse through the feature frame collection by defining different bias weight toward to different features. Although a video summary varies from person to person and may not satisfy all users' informational need as initially generated, it does offer a rapid mechanism for establishing context for the user as to what might exist in their clip/collection. Moreover, a video summary can serve as a mechanism for suggesting possible queries or entry points for non-linear access into the resource, which can greatly facilitate tasks such as search, annotation and editing.

Appendix A:

The Delta E 2000 standard

The Delta E 2000 standard calculation in Lab colour space is following [41]:

$$\Delta E_{00}(L_1^*, a_1^*, b_1^*; L_2^*, a_2^*, b_2^*) = \Delta E_{00}^{12} \tag{33}$$

$$\Delta E_{00}^{12} = \sqrt{\begin{array}{l}(\frac{\Delta L^{'}}{k_L S_L})^2 + (\frac{\Delta C^{"}}{k_C S_C})^2 + \\ (\frac{\Delta H^{'}}{k_H S_H})^2 + R_T(\frac{\Delta C^{"}}{k_C S_C})(\frac{\Delta H^{'}}{k_H S_H})\end{array}} \tag{34}$$

Where

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

$$\overline{L}' = (L_1 + L_2)/2 \tag{35}$$

$$\overline{C} = (\sqrt{a_1^2 + b_1^2} + \sqrt{a_2^2 + b_2^2})/2 \tag{36}$$

$$G = (1 - \sqrt{\frac{\overline{C}^7}{\overline{C}^7 + 25^7}})/2 \tag{37}$$

$$a_1' = a_1(1+G) \tag{38}$$

$$a_2' = a_2(1+G) \tag{39}$$

$$\overline{C'} = \left(\sqrt{a_1'^2 + b_1^2} + \sqrt{a_2'^2 + b_2^2}\right)/2 \tag{40}$$

$$h_1' = \begin{cases} \tan^{-1}(b_1/a_1') & \tan^{-1}(b_1/a_1') \geq 0 \\ \tan^{-1}(b_1/a_1') + 360^o & \tan^{-1}(b_1/a_1') < 0 \end{cases} \tag{41}$$

$$h_2' = \begin{cases} \tan^{-1}(b_2/a_2') & \tan^{-1}(b_2/a_2') \geq 0 \\ \tan^{-1}(b_2/a_2') + 360^o & \tan^{-1}(b_2/a_2') < 0 \end{cases} \tag{42}$$

$$\overline{H}' = \begin{cases} (h_1' + h_2' + 360^o)/2 & |h_1' - h_2'| > 180^o \\ (h_1' + h_2')/2 & |h_1' - h_2'| \leq 180^o \end{cases} \tag{43}$$

$$T = 1 - 0.17\cos(\bar{h}' - 30^o) + 0.24\cos(2\bar{h}') + 0.32\cos(3\bar{h}' + 6^o) - 0.20\cos(4\bar{h}' - 63^o) \quad (44)$$

$$\Delta h' = \begin{cases} h_2' - h_1' & |h_2' - h_1'| \le 180^o \\ h_2' - h_1' + 360^o & |h_2' - h_1'| > 180^o; h_2' \le h_1' \\ h_2' - h_1' - 360^o & |h_2' - h_1'| > 180^o; h_2' > h_1' \end{cases} \quad (45)$$

$$\Delta L' = L_2 - L_1 \quad (46)$$

$$\Delta C' = C_2 - C_1 \quad (47)$$

$$\Delta H' = 2\sqrt{C_1' C_2'}\sin(\Delta h'/2) \quad (48)$$

$$S_L = 1 + \frac{K_2(\bar{L}' - 50)^2}{\sqrt{20 + (\bar{L}' - 50)^2}} \quad (49)$$

$$S_C = 1 + K_1\bar{C}' \quad (50)$$

$$S_H = 1 + K_2\bar{C}'T \quad (51)$$

$$\Delta\theta = 30\exp\left\{-\left(\frac{\bar{H}' - 275^o}{25}\right)^2\right\} \quad (52)$$

$$R_T = -2\sin(2\Delta\theta)\sqrt{\frac{\bar{C}'^7}{\bar{C}'^7 + 25^7}} \quad (53)$$

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 2: 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

Gaussian Pyramid

The Gaussian pyramid is an algorithm to reduce image resolution by using a low pass filter (Gaussian blur filter) and sub sampling process by the factor of 2 [42]. The process is defined recursively as follows,

$$G_0(x, y) = I(x, y), \quad \text{for level, } l = 0 \tag{54}$$

$$G_l(x, y) = \sum_{m=-2}^{2} \sum_{n=-2}^{2} w(m, n) G_{l-1}(2x + m, 2y + n), \quad \text{otherwise} \tag{55}$$

Where

w(m,n) is a weight function an example weight for the impulse response from binomial weight is $\frac{1}{16}\begin{bmatrix}1 & 4 & 6 & 4 & 1\end{bmatrix}$.

l is the level index

The reason to downsize the image by Gaussian pyramid is that keeps the maximum feature from the original image. In image processing theory, the low resolution image is a blurred then down sampled version of high resolution image [43]. The regular interpolation method downsizes the image and same time it also removes the noise from the original image which will lead the image lose some features.

Gabor Filter

The Gabor filter, named after Dennis Gabor [44], is a band pass filter for orientation sensitive edge detection. It is convolute an image with different sets of Gabor kernels to detect different orientations. In the spatial domain, a 2D Gabor filter is a Gaussian kernel function modulated by a sinusoidal plane wave.

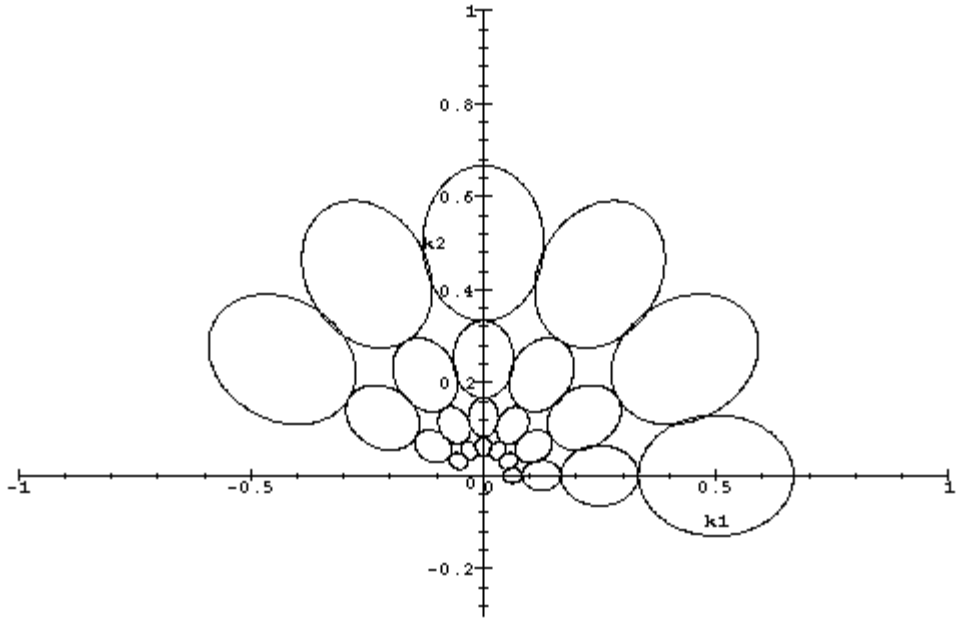


Figure 54: Half-value plot of the Gabor filters in frequency plane tuned to different frequencies and orientations (30 degree resolution)

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp(-\frac{x'^2 + \lambda^2 y'^2}{2\sigma^2}) \cos(2\pi \frac{x'}{\lambda} + \psi) \tag{56}$$

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp(-\frac{x'^2 + \lambda^2 y'^2}{2\sigma^2}) \sin(2\pi \frac{x'}{\lambda} + \psi) \tag{57}$$

$$x' = x\cos\theta + y\sin\theta \tag{58}$$

$$y' = -x\sin\theta + y\cos\theta \tag{59}$$

A number of parameters are required to perform the filtering process which including wavelength of sinusoid (λ), the orientation of the filter (θ), the phase offset (ψ), and the spatial aspect ratio.

Appendix B: Thesis Related Publications

Qian, Y., Kyan, M. High Definition Visual Attention Based Video Summarization. In The 9th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), 2014

Qian, Y., Kyan, M. Interactive User Oriented Visual Attention Based Video Summarization and Exploration Framework, in IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 2014

References:

[1] L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions,* pp. 1254-1259, 1998.

[2] P. Mundur, Y. Rao and Y. Yesha, "Keyframe-based video summarization using Delaunay clustering," *International Journal on Digital Libraries,* pp. 219-232, 2006.

[3] Trecvid2008, "Guidelines for the TRECVID 2008 Evaluation," 5 January 2009. [Online]. Available: http://www-nlpir.nist.gov/projects/tv2008/tv2008.html#4.4.

[4] E. Dumont and B. Mérialdo, "Sequence Alignment for Redundancy Removal in Video," in *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, 2008.

[5] S. Sull, J.-R. Kim, Y. Kim, H. S. Chang and S. U. Lee, "Scalable Hierarchical Video Summary and Search," in *Photonics West 2001-Electronic Imaging*, 2001.

[6] IntelliVision, "Video Summary," 5 April 2014. [Online]. Available: http://www.intelli-vision.com/products/video-search/video-summary.

[7] M. Čadík, "Human perception and computer graphics," Czech Technical University Postgraduate Study Report, 2004.

[8] S. Durant, "Lecture 2: Learning to Read the Neural Code," in *PS1061: Sensation and Perception*, Nova Southeastern University, 2014.

[9] G. Marchionini, B. M. Wildemuth and G. Geisler, "The open video digital library: A möbius strip of research and practice," *Journal of the American Society for Information Science and Technology,* pp. 1629-1643, 2006.

[10] S. Millward, "Color Difference Equations and Their Assessment," *Test Targets,* pp. 19-26, 2009.

[11] V. Chasanis, A. Likas and N. Galatsanos, "Video rushes summarization using spectral clustering and sequence alignment," in *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop,* 2008.

[12] D. Zhong, H. Zhang and S.-F. Chang, "Clustering methods for video browsing and annotation," in *Electronic Imaging: Science & Technology*, 1996.

[13] S. M. J. a. S. V. N. Cvetkovic, "Video summarization using color features and efficient adaptive," *PRZEGLĄD ELEKTROTECHNICZNY,* pp. 247-250, 2013.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference*, 2005.

[15] C. V. Royo, "Image-Based Query by Example Using," Universitat Politecnica de Catalunya, Barcelona, 2010.

[16] W. Jia, H. Zhang, X. He and Q. Wu, "Image Matching Using Colour Edge Cooccurrence Histograms," in *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference*, 2006.

[17] J. Calic, D. P. Gibson and N. W. Campbell, "Efficient layout of comic-like video summaries," *Circuits and Systems for Video Technology, IEEE Transactions on 17, no. 7,* pp. 931-936, 2007.

[18] A. Amiri and M. Fathy, "Hierarchical keyframe-based video summarization using QR-decomposition and modified k-means clustering," in *EURASIP Journal on Advances in Signal Processing 2010*, 2010.

[19] E. J. Y. Cayllahua-Cahuina and D. M. G. Camara-Chavez, "Static Video Summarization Approach With Automatic Shot Detection Using Color Histograms," UFOP - Federal University of Ouro Preto,

Ouro Preto, 2012.

[20] M. M. S. Koskela, J. Laaksonen, V. Viitaniemi and H. Muurinen, "Rushes summarization with self-organizing maps," in *Proceedings of the international workshop on TRECVID video summarization*, 2007.

[21] T. Ayadi, M. Ellouze, T. M. Hamdani and A. M. Alimi, "Movie scenes detection with MIGSOM based on shots semi-supervised clustering," in *Neural Computing and Applications*, 2013.

[22] H. Jiang and M. Zhang, "Tennis video shot classification based on support vector machine," in *Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference*, 2011.

[23] L. Li, X. Zhang, W. Hu, W. Li and P. Zhu, "Soccer video shot classification based on color characterization using dominant sets clustering," in *Advances in Multimedia Information Processing-PCM*, 2009.

[24] L. Li, K. Zhou, G.-R. Xue, H. Zha and Y. Yu, "Video summarization via transferrable structured learning," in *Proceedings of the 20th international conference*, 2011.

[25] H. M. Zawbaa, N. El-Bendary, A. E. Hassanien and A. Abraham, "SVM-based soccer video summarization system," in *Nature and Biologically Inspired Computing (NaBIC), 2011 Third World Congress*, 2011.

[26] 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*, 2009.

[27] Y.-F. Ma, X.-S. Hua, L. Lu and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *Multimedia, IEEE Transactions*, pp. 907-919, 2005.

[28] J. Peng and Q. Xiao-Lin, "Keyframe-based video summary using visual attention clues," *IEEE Multimedia*, pp. 64-73, 2010.

[29] Y. Saber, "High-definition human visual attention mapping using wavelets," Ryerson University, Toronto, 2011.

[30] Y. Saber and M. Kyan, "Frequency tuned salient edge detection," in *Electrical and Computer Engineering (CCECE), 2011 24th Canadian Conference* , 2011.

[31] R. Achanta, S. Hemami, F. Estrada and S. Susstrunk, "Frequency-tuned salient region detection," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference*, 2009.

[32] Y.-F. Ma, L. Lu, H.-J. Zhang and M. Li, "A user attention model for video summarization," in *Proceedings of the tenth ACM international conference on Multimedia*, 2002.

[33] S. Frintrop, "Computational visual attention," *Computer Analysis of Human Behavior*, pp. 69-101, 2011.

[34] X-Rite-Incorporated, A Guide to Understanding Color Communication, X-Rite Incorporated, 2007.

[35] G. Sharma, W. Wu, E. N. Dalal and M. U. Celik., "Mathematical discontinuities in CIEDE2000 color difference computations," in *Color and Imaging Conference*, 2004.

[36] Y. Saber and M. Kyan, "High resolution biologically inspired salient region detection," in *Image Processing (ICIP), 2011 18th IEEE International Conference*, 2011.

[37] W. C. Y. Yusoff and J. Kittler, "Video Shot Cut Detection using Adaptive Thresholding," *BMVC2000*, pp. 1-10, 2000.

[38] L. Krulikovská and J. Polec, "Shot Detection using Modified Dugad Model," *World Academy of Science, Engineering and Technology*, pp. 123-126, 2012.

[39] T. M. Cover and J. A. Thomas, Elements of information theory, John Wiley & Sons, 2012.

[40] H. Yin, "The self-organizing maps: Background, theories, extensions and applications," *Computational intelligence: a compendium*, pp. 715-762, 2008.

[41] B. J. Lindbloom, "Delta E (CIE 2000)," 12 January 2014. [Online]. Available:

http://www.brucelindbloom.com/index.html?Eqn_DeltaE_CIE2000.html .

[42] K. Derpanis, "The Gaussian Pyramid," 2005. [Online]. Available: http://www.cse.yorku.ca/~kosta/CompVis_Notes/gaussian_pyramid.pdf.

[43] J. Tian and K.-K. Ma, "A survey on super-resolution imaging," *Signal, Image and Video Processing*, p. Springer, 2011.

[44] D. Gabor, "Theory of communication," *Journal of the Institute of Electrical Engineers*, pp. 429-457, 1946.

[45] X. Chen, M. Das and A. Loui, "An efficient framework for location-based scene matching in image databases," *International Journal of Multimedia Information Retrieval*, pp. 103-114, 2012.

[46] S. Benini, A. Bianchetti, R. Leonardi and P. Migliorati, "Hierarchical Summarization of Videos by Tree-Structured Vector Quantization," in *Multimedia and Expo, 2006 IEEE International Conference*, 2006.