Detecting Object with Content Images

Dr A J Patil
Principal, Gulabrao Deokar, COE, Jalgaon

Kiran Patil
Student Gulabrao Deokar, COE, Jalgaon

Abstract
The advances in the data set in the form of image are available on the internet. The paper proposes the system with the retrieval and detection of images from the image. Detecting particular objects in image is an important step for understanding the visual imaginary semantically. In content based retrieval, the ability to detect people any specified object such as animals, cycles and automobiles gives the option of advanced queries such as “Find an image clip which contains a crowded area or a fast moving car.”, or we detect the various contents present in the image for which the application is trained for. The application can be designed and trained for the same from the dataset.

KEYWORDS
Content based image retrieval, semantic information, and colour histogram, segmentation

INTRODUCTION
Content based image retrieval may be defined as an approach in which the images are retrieved from the large database based upon their visual contents[1][6][7]. Content based image retrieval is desirable because most image or image search engines rely purely on metadata and this produces a lot of wrong results. Also these keywords are annotated manually and are completely based upon human perception, so this procedure will not be sufficient enough to capture every keyword that describes the image or image. Thus a system that can filter images on their content would provide better indexing and return more accurate results. The content of image or image are colour, shape, texture etc. Image content can be grouped into two levels: low level visual features and high-level semantic content.

CONTENT BASED IMAGE RETRIEVAL

As image data is very complex; understanding of its unique characteristics is essential to develop techniques for managing it. There are some significant characteristics that distinguish image from other classes of data.[2]

1. Image is stored as binary; therefore, in contrast to alphanumeric data, image has higher resolution, larger data volume, larger set of data that can be originated, higher interpretation ambiguity, and needs more interpretation efforts.

2. Image has spatial and temporal dimension, whereas text is only non-spatial static and image is spatial static. Moreover, image semantic is unstructured and generally contains complex relationships.
semantic information of image can be expressed either implicitly or explicitly. Figure 2 depicts a high-level concept on the contents of images data.

Figure 3 illustrates a very simple example on how semantic contents can be conveyed through the spatio-temporal AV presentation of image data. Since shots are merely depicted as a sequence of image frames in the second diagram, thus the sequence of three shots convey semantic information, which indicates that a car has moved from right to left (spatial-temporal information), and a tower is located in the middle of the background (spatial-static information).[8]

It is important to note that the AV components of image data are not always equally important in conveying the semantic content. Instead, it depends on the purpose and use of the image data. For example, in analysing the game strategy and techniques during a soccer match, the information about the motion and position of the players are most important. Along with this, depending on how the image was produced, many different AV features can represent the same semantic content and vice versa, the same AV features can represent different semantic contents due to the subjectivity of the annotators. Hence, similar to the human perception of image document, the semantic content in image will be more accurately interpreted when more channels are perceived.

A general structure of CBVR components is depicted in Figure 4 and can be described as follows.

User/application requirements determine the retrieval and browsing. The success of retrieval depends on the completeness and effectiveness of the indexes. Indexing techniques are determined by the extractable information through automatic or semi-automatic content extraction. Since image contains rich and multidimensional information, it needs to be modelled and summarized to get the most compact and effective representation of image data. Prior to image content analysis, the structure of image sequence needs to be analysed and separated into different layers such as shot and scene. Since the design of each component is affected by other components, it is generally difficult to discuss each component separately. For example, if the retrieval is based on high-level semantic features such as specific sport highlight segments like goal, the indexing can be based on the hierarchy of summarized (key) events. Each event can be abstracted using the face of the actors who participate during that event whereas the event itself can be described using some statistical measures like excitement ratio (e.g. the higher excitement, the more significant an event should be). As a result, the content extraction process aims to automatically identify and classify the event(s) that is contained within each play-break segment.

**IMAGE SEGMENTATION**

In general, indexing could be performed on the whole image stream but it would be too coarse [11]. On the other hand, if the indexing is based on each frame in the clip, it would be too dense as a frame often does not contain any important information. Researchers have commonly indexed on a group of sequential frames with similar characteristics. Figure 5 shows how images are divided into frames.

It is important to note that the AV components of image data are not always equally important in conveying the semantic content. Instead, it depends on the purpose and use of the image data. For example, in analysing the game strategy and techniques during a soccer match, the information about the motion and position of the players are most important. Along with this, depending on how the image was produced, many different AV features can represent the same semantic content and vice versa, the same AV features can represent different semantic contents due to the subjectivity of the annotators. Hence, similar to the human perception of image document, the semantic content in image will be more accurately interpreted when more channels are perceived.

A general structure of CBVR components is depicted in Figure 4 and can be described as follows.

User/application requirements determine the retrieval and browsing. The success of retrieval depends on the completeness and effectiveness of the indexes. Indexing techniques are determined by the extractable information through automatic or semi-automatic content extraction. Since image contains rich and multidimensional information, it needs to be modelled and summarized to get the most compact and effective representation of image data. Prior to image content analysis, the structure of image sequence needs to be analysed and separated into different layers such as shot and scene. Since the design of each component is affected by other components, it is generally difficult to discuss each component separately. For example, if the retrieval is based on high-level semantic features such as specific sport highlight segments like goal, the indexing can be based on the hierarchy of summarized (key) events. Each event can be abstracted using the face of the actors who participate during that event whereas the event itself can be described using some statistical measures like excitement ratio (e.g. the higher excitement, the more significant an event should be). As a result, the content extraction process aims to automatically identify and classify the event(s) that is contained within each play-break segment.

**IMAGE SEGMENTATION**

In general, indexing could be performed on the whole image stream but it would be too coarse [11]. On the other hand, if the indexing is based on each frame in the clip, it would be too dense as a frame often does not contain any important information. Researchers have commonly indexed on a group of sequential frames with similar characteristics. Figure 5 shows how images are divided into frames.

**SHOT BOUNDARY DETECTION** [8]

Shot is a sequence of image frames which have similar characteristics. Shots extraction requires the computation of an appropriate metric (algorithm) to characterize the change of image content between two frames and a threshold to determine whether the change is important enough to be defined as a shot boundary.

Three main methods which can be used for shot boundaries detection are described below:

- Pixel-wise frame difference.
- Histogram comparison.
- Audio assisted.
Figure 6: Shot Boundary Detection

Pixel-wise frame difference technique detects shot boundaries by measuring a qualitative change between two frames by simply comparing the spatial corresponding pixels in the two frames and determining the amount of the pixels that have changed. Thus it is also called pair-wise pixel comparison algorithm shown in Figure 7.

Figure 7: Pixelating the Image from the Image

While pixel-wise comparison is easy to be done, it is too sensitive against noise, illumination changes and camera motion. Alternatively, color histogram can be used due to its effectiveness in characterizing the global distribution of an image without knowing the component objects that make up the image. Color histogram shows statistically the intensities of the three-color channels in the image, such RGB (Red Green Blue) or HSV (Hue Saturation Value).

Color histogram comparison should also be less sensitive to object motion than pixel difference technique [1]. Histogram is used for shot boundary detection by assuming that two frames which have unchanging backgrounds and objects will show little difference in the histograms. This method can be summarized as follows: histogram of a frame is computed by measuring probability distributions of pixel values in the entire image on a frame; therefore each frame will produce a distinct histogram. To identify shot boundaries, the histogram difference between a frame and its consecutive frames is compared against a threshold. If the difference is bigger than the threshold, a shot boundary is found.

SCENE DETECTION

Although shot-based indexing can be fully automated, the major problem is the lack of semantic information [12]. As mentioned, a shot is only a group of sequential frames with similar characteristics; therefore, it does not actually correspond directly to semantic content. For example, the semantic content often does not change during shot boundary. To overcome this limitation, scenes need to be extracted as a sequence of shots which represent and can be described by a semantic content description. Scene detection is generally more difficult than shots detection due to a need for understanding the image contents. One approach is by measuring the semantic correlation of consecutive shots based on dominant colour grouping and tracking. For example, a shot grouping method called expanding window has been designed to cluster correlated consecutive shots into one scene. Similarly, scene can be detected via continuous coherence in which related shots are grouped into scenes which are defined as a single dramatic event taken by a small number of related cameras. Alternatively, a scene can be formed by grouping a sequence of shots which depict a particular object activity or event. Figure 8 shows frame retrieval.

Figure 8: Frame retrieval

IMAGE ANALYSIS[2]

General image or visual features include colour, texture, and shape. The focus of this section is to review the very useful image features that can be used for image content analysis, namely,
1) Colour feature,
2) Shape feature and
3) Texture feature.

Figure 9: Image Analysis
Colour Features:

Colour is the most intensively used feature for image content management due to its robustness to complex background (occlusion), scaling (image size), orientation, and perspective. Colour histogram and colour moments are some of the fundamental features of colour. The colour feature detection is shown in Figure 10.

Colour Histogram:

Colour histogram is the most common feature representation as it can characterize the global distribution of an image effectively without knowing the components' objects that made up the image [10]. It shows the intensities of the three-color channels in the image statistically. For example, Figure 11 depicts the histogram of two typical views in soccer images; where the frame with large playing field shows dominant intensities in colour index of 0.4 to 0.6 while the frame with player close-up has a more equal distribution of colour intensities.

Colour Moments:

Colour moments of an image are chosen due to their effectiveness and simplicity. Any colour distribution can be represented by its moments. In particular, first (moment) order captures mean colour, while second and third order capture colour variance and skewness respectively.

Shape features:

Shape is another important image feature as any good shape representation should be unique, robust against translation, scaling and rotation. To extract shape features, image segmentation techniques are required to segment an image into its individual objects. In general, shape representations can be categorized into boundary- (outer shape) and region- (entire shape) based. The Figure 12 shows the shape feature detection.

Texture Features:

Texture is an important image feature as it describes visual patterns which are homogeneous and not produced from single colour or intensity. These visual patterns distinguish structural arrangement of surfaces from the surrounding environment; therefore, texture is a natural property of all object surfaces, such as clouds, woods and bricks. However, it is generally difficult to describe texture and its perception can be subjective to a certain degree. For this reason, texture specification techniques often need to mimic human perception on texture. Figure 13 shows the texture feature detection in the images from the images.

CONCLUSION

The content-based image retrieval system can be best applied to any sports where the team wants to track the record and to improve the game of a particular player. The CBVR can be improved and combined with histogram of oriented gradient for tracking the humans and also detecting various objects like cycles or animals like cat or dog.

This can be implemented in a single application with a multiple detectors becomes a difficult approach when comes to the application level of the approach. Detecting humans and its application implementation for pedestrian detection and setting up an alarm as the pedestrian is detected. The various classifiers with scale invariant feature transform (SIFT) can be used to improve the detection of the object in the content-based image retrieval system.

REFERENCES


[21]. TRECVID, 2006. An overview of up-to-date methods in content-based image retrieval -- by examining top performances in TREC image retrieval evaluation. TRECVID.