Question 1

(i) Compare and contrast the problem of edge detection and image segmentation.

edge detection works on locating local gradient changes (1'), image segmentation groups pixels into clusters each being a segment (1'). Edges are essentially what separates different regions (1'), so edges and segmentations are somehow similar.

(ii) Describe a scheme whereby hierarchical segmentations are equivalent to edge maps.

there is no single correct segmentation (1'), probabilistic edgemap can represent hierarchical segmentations (1'), whereby thresholding the edgemap will produce a unique segmentation.

(iii) Explain two common approaches to assess the quality of automatically generated segmentation maps.

subjective (1') and objective (1'), the former employs a list of criteria for evaluating segmentation quality (1'), such as whether each region if of uniform colour, the latter compares with human ground truth data.

(iv) Sketch segmentation is different to image segmentation, working with mere black and white lines other than coloured and textured images. Briefly explain how Gestalt principles commonly used by image segmentation algorithms can be employed to tackle sketch segmentation.

only geometry information can be used in sketch segmentation (1'), common laws such as proximity and similarity can all be used but in different ways (1'), each Gestalt law can induce a particular grouping (1'), the problem is how to combine different laws to reach a balanced grouping(1').

Question 2

The following questions refer to video segmentation:

(i) Explain the main steps of object video segmentation techniques.

background segmentation, foreground object extraction, foreground object linking/tracking, disambiguate multiple objects.

(ii) You are asked by BBC Sports to design a video segmentation technique for Snooker. Given an input video, the ultimate task is to segment the video into two sub-videos, one video consists of all instances when a red ball is potted, and the other consists of non-red balls potted. Explain important design considerations to undertake and main steps of your algorithm.

one needs object segmentation since the requirement is placed on the object. the system should extract all balls in the scene, detect each time they are potted, identify the start and finish of each successful pot, and group based on colour the balls potted, and finally stitch all sub videos together to form a new video.

Question 3

(a)The following questions refer to image segmentation:

(i) Explain the purpose and principle of chroma keying.

Chroma keying is a technique for mixing two images together, in which a colour from one images is removed, revealing another image behind it. The principal subject is filmed or photographed against a background consisting of a single colour (e.g. a blue screen). The proportions of the video which match the preselected colour are replaced by the alternate background video.

(ii) Explain the segmentation technique used in chroma keying.

colour-based segmentation based on thresholding.

(iii) Describe the lighting setup of chroma keying. Explain why such setup is necessary.

diffuse lighting on the blue screen to make sure colour is uniform, spotlight on the foreground object to ensure colour intensities are kept.

(iv) As it happens, you have to perform chroma keying on a playground shoot where school kids are wearing plain colour t-shirts of all dominant colours. Can you still make use of blue screens? If not, what would you choose to use instead?

No, since otherwise some kids will be missing. One way is to upgrade colour-based thresholding to texture-based segmentation, and use unique texture background.

(b) The following questions refer to video segmentation:

(i) Define in mathematical terms the problem of temporal video segmentation.

Temporal segmentation aims to break a video of N frames into sets of disjoint frames. The formulation for this is the same as image segmentation but the student need to understand that the set is now on the temporal dimension other than spatial for each modified equation.

(ii) Under the scenario of traffic monitoring where the camera position is fixed, explain how object segmentation of moving cars can be achieved. List two conditions where your algorithm might fail, and explain why.

Background relatively simple to extract, the foreground is then extracted by calculating difference to background, you also need to ensure object id remains unchanged throughout video. Lighting condition and camera noise are two failure conditions.

(iii) You are asked to design a volleyball video summarisation technique for BBC Sports. Given an input volleyball video, the task is to output a shorter video consisting of all net-touching fails committed by both teams. Explain important design considerations to undertake and main steps of your algorithm.

This is temporal segmentation but based on object segmentation and activity recognition. One needs to first extract basketball players first by performing object segmentation, then classify each object activity through some form of machine learning technique where each segmented objects gets classified into different activities, finally one can segment the original video by cutting chunks of video that correspond to 3points shooting activities.

Question 4

(a) The following questions refer to MPEG family of compression standards.

(i) In MPEG-1, suggest how the motion estimation task can be made quicker.

We need to reduce the search space for matching macroblocks. Information from previously estimated maroblocks can be used, e.g. the motion vectors can tell us where a macroblock is likely to be found. We could also split a frame into smaller areas and only search the areas where most similarities with the macroblock have been found.

(ii) In MPEG-4, briefly explain the need for a “compositor”.

The compositor is in charge of constructing video frames to be played, it is necessary because individual objects have been encoded and decoded separately. It also allows the creation of various contents reusing existing objects.

(iii) State 3 advantages of the use of the XML language in MPEG-7.

human readable, compact, extensible, self-discrimitive (看不清这个词).

Question 5

(a) The following questions refer to frame-based multimedia.

(i) What is the elementary building primitive in MPEG1 standard?

macroblock and gon (看不清这个词)

(ii) For each type of such primitive, what kind of information is transmitted?

Skipped: nothing needs to be transmitted
Intra: DCT coefficients and quantisation values
Inter: motion vector, DCT coefficients for error term, quantisation values

(iii) In a P frame how many types of these primitives are there, name two of them?

Three. Skipped, intra predicted, inter (backward) predicted.

(iv) Explain why Groups of Pictures (GOPs) that contain a large number of B frames may not be suitable for applications such as video conferencing?

Because B needs information from future frames, and this incurs a coding delay not suitable for real time applications.

(b) The following questions refer to MPEG7.

(i) In the context of an information retrieval task, briefly explain what is meant by the "semantic gap". Give two examples applications where effectively solving the semantic gap would help.

The semantic gap refers to the difference between multimedia content (data) and the way infomation retrieval users think about or describe data. Image retrieval, advanced surveillance.

(ii) Explain how does MPEG7 help bridging the semantic gap?

MPEG7 is a standard for describing the features of multimedia content, i.e. it specifies how to associate semantic descriptions with multimedia content, using Descriptors (D) and Description Schemes (DS).

(iii) Explain the use of MPEG7 for search and retrieval of audiovisual material, especially the types of user query it supports.

Audiovisual material that has MPEG7 annotations associated with it can be indexed. The indexed database can be searched more effectively than the raw data. Low level MPEG7 annotations (e.g. image colour histogram) are useful for example-based queries, where an image or a sound wave is submitted as query. Higher level (e.g. the names of the actors in a movie) are useful for text-based queries.

Question 6

(a) The following questions refer to frame-based and object-based multimedia.

(i) Compare the contrast frame-based and object-based multimedia.

Frame-based deals with the whole image without understanding its contents, object-based asks for semantic parsing of the image/video, the latter narrows the semantic gap.

(ii) List 2 applications where an object-based approach is preferred over a frame-based alternative, explain why.

In film retrieval, where the user wants to retrieve based on actor names; For traffic monitor, where one wants to interact with the video using colour of cars.

(iii) You are asked by Scotland Yard to design a new kind of video surveillance storage solution, where instead of compressing and saving the videos as a whole, the system will be able to search and index events (e.g. running, loitering, etc.), objects (e.g. bags) and faces. Discuss the pros and cons of frame-based and object-based approaches under this scenario.

This is video synopsis, one will need to employ object-based multimedia, essence is performing object segmentation of videos, whereby objects (e.g. people) are extracted from the background to construct to spatial-temporal volume, a classifier will then need to be run on top of these spatial-temporal features to classify activities.

(b) The following questions refer to MPEG4.

(i) Give 2 examples of primitive media objects used in MPEG4.

still images (a fixed background)
video objects (a talking person)
audio objects (the voice associated with that person)

(ii) In MPEG4, what does it mean that media objects can be of natural or synthetic nature?

They can be recorded with a camera or a microphone (natural), or generated with a computer (synthetic).

(c) The following questions refer to MPEG7.

(i) In the context of video encoding, briefly explain what “metadata” is. Give concrete examples.

Data about data. Data refers to the video content, the bit stream. Metadata refers to descriptions of the data, such as the theme of the video, the names of the actors of a movie.

(ii) How does MPEG7 allow for the representation of metadata?

MPEG7 is a standard for describing the features of multimedia content, using Descriptors (D) and Description Scheme (DS). A Description Definition Language (DDL) allows the specification of the Ds and DSs.

Question 7

(a) The following questions refer to image retrieval.

(i) Compare and contrast text-based image retrieval and content-based image retrieval.

TBIR uses text as input query, whereas CBIR uses photo as query. Major difference is TBIR does text-level retrieval, while CBIR looks into the pixels.

(ii) Compare and contrast colour and shape as features to conduct retrieval. Discuss under what circumstances would you use two features simultaneously?

Colour is a very local feature, i.e. per pixel, shape is more global but more difficult to compute. You would need both if each on its own is not discriminative enough, e.g. distinguish a red apple from read tomato.

(iii) Describe a colour-based feature scheme where spatial arrangements of image content can be encoded. Discuss how such encoding can assist user relevance feedback.

Start with colour histogram, which is a global measure of an image region. To encode structure, one could divide the image into equal-sized parts and compute colour histogram for each sub-region. It follows that you can stack up the colour histograms for each sub-region to form a longer feature vector.

(b) The following questions refer to sketch-based image retrieval (SBIR).

(i) Discuss why sketches are the most appropriate form of input modality for fine-grained image retrieval.

Sketches are more expressive than text, and more importantly a sketch speaks for a hundred words, moreover it is much easier to sketch other than typing in long sentences.

(ii) Define the problem of fine-grained SBIR. Compare and contrast FG-SBIR with TBIR.

FG-SBIR only retrieves into a single object category, TBIR often deals with multiple categories and text is often ambiguous if trying to distinguish objects of the same category.

(iii) You are asked by Microsoft Bing to design a sketch-based image retrieval system for online retrieval of chairs. The system should work iteratively, where the user has the option to directly sketch on retrieved photos to refine their search results. Discuss important design considerations to undertake and explain main stages of your SBIR system.

The system would need to be able to perform FG-SBIR since only one category, one will also need to allow users to interact by sketching directly on images, therefore on the next round of retrieval, the search will become multi-modal since both sketches and photos can be used together to further constrain the search space. Main stages are FG-SBIR, user interaction, and multi-modal retrieval.

Question 8

(a) The following questions refer to image retrieval.

(i) Explain the pros and cons of TBIR.

Typing is convenient, and intuitive to humans, yet it is also vague and ambiguous.

(ii) Compare and contrast colour-based and shape-based image retrieval. Using examples, explain under what circumstances is one technique better than another, and vice versa.

Colour is global so search results can be imprecise. Shape is relatively local but can be hard to compute. When differentiating a banana with apple, colour is probably sufficient, yet shape might be better if one is comparing a banana with honey melon.

(iii) Imagine you would like to search for a particular pair of shoes for Christmas on Taobao, you have a very good idea of what it looks like but do not have a picture to hand. Describe two common approaches that you could use for locating such pair of shoes online. Discuss how one approach might be better than the other if the shoe is full of visual details (i.e. complicated looking).

You can use either text-based or sketch-based, the latter might be better since we are performing fine-grained retrieval. This is because sketches tend to offer more descriptive details, yet the longer your type the more ambiguous the search would become. Sketch is also in the visual domain so can be better matched to photos without first converting visual impression to text.

(b) The following questions refer to sketch-based image retrieval (SBIR).

(i) Name 3 deficits of using text as an input modality to conduct IR.

inaccurate, ambiguous, language-dependent.

(ii) Define the problem of FG-IR. Discuss why sketches might be a better input modality than text for FG-IR.

FG-SBIR does intra-category retrieval, sketches are better since it offers a detailed visual representation, text can be ambiguous when it comes to detailed descriptions.

(iii) You are asked by Bing to design an image retrieval system for online retrieval of cars and handbags. The system should be able to work both text and sketch as input. Discuss important design considerations to undertake (e.g. should the same input modality be used for each object category, or would a combination of both be better) and explain main stages of your sketch-based image retrieval system.

This is multi-modal retrieval, it is important to figure out which input modality is better for different object categories. For cars, there is less variations than handbags, the latter is also deformable. So text is probably better suited for cars, and sketches play a more important role on handbags.

Question 9

(a) The following questions refer to image retrieval.

(i) Explain the pros and cons of content-based image retrieval.

Taking a photo is convenient nowadays, more accurate than text, but not flexible, i.e. cannot modify search results and search becomes over-constraining ofter resulting in images that are very similar to the input.

(ii) List three types of features commonly used in CBIR. Which two would you use in combination to retrieve into a dataset of fruits (e.g. apple, banana, pear, etc.), and why?

colour, shape, texture, one can use shape and colour, texture is less important since most fruit are of uniform colour.

(iii) Explain what is colour histogram. Describe a scheme where mid-level image structural information can be embedded into histogram-based feature representations.

Colour histogram is a bin count of different colour values. It is a global image feature, structure can be introduced by dividing the image into different regions, and extracting histograms for each sub region, then stacking up each sub histogram and produce a longer feature vector.

(b) The following questions refer to SBIR.

(i) Compared with text-based image retrieval, what is the most significant drawback of photo-based image retrieval.

Photo-based search is overly constraining, it also does not offer a degree of flexibility that is often describe, one often does not have a photo to hand as well.

(ii) Define the problem of FG-IR. Discuss why a photo is not a suitable input modality for FG-IR.

Photo does have details, however some are more important than others, it is not easy to introduce user preferences in photo-based retrieval.

(iii) You are asked by Yahoo to design an image retrieval system for online retrieval of product logos (e.g. Nike logo, Adidas logo, etc.). Discuss which input modality would you choose for such systems, by comparing with two other alternatives.

Detailed shapes. Shapes are hard to describe in text, therefore sketch is the more suitable input here. Photo-based approaches might be too sensitive to changes in colour and texture in logos, companies often modify logos from time to time changing its colour etc. but not the overall shape/ structure.

EBU706U Solution A 2015/16

Question 1

(a) The following questions refer to image segmentation.

(i) Define in mathematical terms the problem of image segmentation.

Segmentation divides an image I into a set of disjoint regions R.
式子见课件

(ii) Compare and contrast top-down and bottom-up segmentation techniques.

Top-down segmentation starts with the whole image and divides it into disjoint regions, you have to know when to stop splitting. Bottom-up segmentation starts with seeding pixels and groups neighbouring pixels at each iteration, knowing positions of seeding pixels is key.

(iii) In the setting of medical imaging, e.g. analysing images of brain tumour, which segmentation technique would you use? Justify your choice in detail.

Bottom-up segmentation is best. For medical imaging, objects are often of irregular shape and normally the doctors can specify starting positions of objects. Top-down segmentation on the other hand might not be the optimal since it would otherwise be hard to decide a stopping criteria. Therefore offering less control over the segmentation quality of the object of interest.

(b) The following questions refer to video segmentation.

(i) Define in mathematical terms the problem of temporal video segmentation.

Temporal segmentation aims to break a video of N frames into sets of disjoint frames. The formulation for this is the same as image segmentation but the student need to understand that the set is now on the temporal dimension other than spatial for each modified equation.

(ii) Compare and contrast temporal and object video segmentation techniques.

Temporal segmentation breaks the video into smaller videos which when stitched together will have the original video back. Object video segmentation aims to make a new video where only the object of interest is captured.

(iii) You are asked to design a video summarisation technique for Scotland Yard, that given an input video of X minutes, the task is to output a video of Y minutes where X >> Y and the output video must not lose objects of large temporal longevity. Which video segmentation technique would you use? Explain in detail important design considerations to undertake and main steps of your algorithm.

One needs object video segmentation since the requirement is placed on the object, the system should extract all objects in the scene, rank them in terms of length of temporal activity, the amount of compression will depend on the actual requirement, in the most extreme case the shortest summarisation would contain only the object having the longest activity.

Question 2

(a) The following questions refer to image retrieval.

(i) Explain in detail the problem of content-based image retrieval.

CBIR extracts semantic info, colour, shape, attributes, etc. from query image and matches to a gallery of images.

(ii) Compare and contrast colour-based, shape-based and texture-based image retrieval techniques. Provide an applicatio scenario where each is best used for.

Colour-based uses colour info as major cue towards retrieval, best used for applications where colour is the dominant cue such as plain cloth retrieval. Shape-based uses shape as major cue and is best when shape the most salient cue such as furniture retrieval. Texture-based uses texture as dominant feature and best when retrieving building materials for example.

(iii) In content-based image retrieval systems, discuss what is meant by relevance feedback. Why it is important and how it can be implement in practice.

Relevant feedback means the user shall provide feedback to initial retrieval results, making consequent retrievals more tailored towards user needs. It is important because most times it is hard to choose the right feature that matches to user needs. It works by first perform automatic retrieval, then ask the users to perform feedback on retrieved images, by marking them as yes/no, specifying particular regions, drawing rough contours, etc.

(b) The following questions refer to SBIR.

(i) Motivate the problem of SBIR and provide an application scenario where SBIR is best used for.

SBIR is best used if the object is hard to describe in text and you don't have an image to hand as query. For example, retrieving fashion items such as shoes and bags.

(ii) Define the problem of FG-SBIR. Compare and contrast FG-SBIR with TB and CB IR.

FG-SBIR means we are not interested in objects of one category, but specific instance of that category there the input sketch matches exactly to the input sketch. TB means user would have to describe the visual object in detail therefore producing rather complicated text query which is complicate retrieval systems. CBIR require input image which are harder to parse than sketches.

(iii) You are asked by Baidu to design a SBIR system for online retrieval of shoes. The system should be able to differentiate different styles of shoes (sneakers, flats, boots, etc) and as well as their make. Discuss in detail important design considerations to undertake and explain main stages of your SBIR system.

We need to first think about detailed features that might be needed such as logos, styles, etc. Then have means of detecting them. Once that's done a FG matching paradigm needs to be implemented. SBIR is the best here, since shoes are otherwise hard to describe. A machine learning framework should be employed which involves collecting of dataset of training and testing pairs of sketches and images of shoes.

Question 3

(a) Consider the 2 following content delivery methods: (1) information pyramid and (2) scalable coding:

(i) Both of these methods serve the same purpose. What is it?

Both methods adapt the content to be delivered so it can be transmitted on a variety of networks to different kinds of devices, and to serve different applications. With both methods, a receiver will receive only the data it needs or can afford.

(ii) Explain the differences between the two methods.

Information pyramid generates different versions of the data, and the server decides which version should be transmitted to the receiver. Scalable coding generates only one data stream. However, the data stream can be truncated, and the receiver decides when to stop receiving the data.

(iii) Give 4 examples of scalability in scalable coding with brief explanations for each.

没拍到……

(b) The followings refer to MPEG7.

(i) In the context of an information retrieval task, briefly explain what is meant by the “semantic gap”. Give concrete examples.

The semantic gap refers to the difference between multimedia content (data) and the way information retrieval users think about or describe the data. Example of data: an image (bitmap of pixels), example of user query: a flower (i.e. meaningful/semantic object)

(ii) How does MPEG7 help bridging the semantic gap?

MPEG7 is a standard for describing the features of multimedia content, i.e. it specifies how to associate semantic descriptions with multimedia content, using Descriptors (D) and Description Schemes (DS).

(iii) Explain the use of MPEG7 for search and retrieval of audiovisual material, especially the types of user query it supports.

Audiovisual material that has MPEG7 annotations associated with it can be indexed. The indexed database can be searched more effectively than the raw data. Low level MPEG7 annotations (e.g. image colour histogram) are useful for example-based queries, where an image or a sound wave is submitted as query. Higher level (e.g. the names of the actors in a movie) are useful for text-based queries.

Question 4

(a) The followings refer to multimedia delivery.

(i) What is meant by “transcoding”? How can it make multimedia delivery to a small device more efficient.

Transcoding means that the content is converted into another signal with different properties. For example, an image is converted into a textual description, which might be all a small device can display or is capable of receiving.

(ii) Describe 4 types of transcoding.

Content blind transcoding, i.e. independent of the semantics, e.g. lower spatial or temporal resolution.
Content aware transcoding, e.g. bit rate budget or frame skip.

Intra-media transcoding, i.e. do not change the nature of the media, e.g. a lower resolution video (content blind) is intra-frame transcoded.
Inter-media transcoding, i.e. process of transforming one media in another media format, also called trans-moding, e.g. speech to text or video to text.

(b) The followings refer to 3D multimedia.

(i) Explain the difference between model-based rendering and image-based rendering.

Model-based: A 3D model of the scene is generated using geometric information and camera parameters. Virtual views are produced rendering the texture and lighting in the 3D model and projecting it back to a virtual image plane.
Image-based: A scene is regarded as a set of images (reference frames) taken from predefined reference viewpoints. The goal of image-based rendering techniques is to generate intermediate views of the object using only the reference frames, i.e. without generation of a complete 3D model.
The computational cost depends only on the number of pixels in the reference frames and not on the geometric complexity of the object. This is a fundamental difference between direct 3D modelling and image-based rendering.

(ii) In image-based rendering, what is disparity map? Explain in detail how disparity map can be calculated given a pair of images.

It is a set of disparity vectors, which show the displacement of feature points between two reference frames. It can be calculated by performing pixel matching between two images, then calculate a global displacement field.

(c) The followings refer to data protection.

(i) In terms of data protection, what is the main difference between analogue media and digital media? Which of the two is easier to protect.

Copying analogue media involves some degradation of the data, whereas digital media can be copied an infinite number of times without occurring degradation. The degradation of copied analogue media is a built-in protection, hence it is easier to protect analogue media than digital media. Digital media is also easier to distribute, for example via the Internet.

(ii) Give 3 traditional types of media protection (other than watermarking) and their respective drawbacks.

Access control headers, can easily be removed and altered.
Encryption, once decrypted, the data is unprotected.
Copy protection, is susceptible to hacking.

(iii) Explain in detail what is a digital watermarking?

It is a signal embedded in another signal.