|  |  |  |
| --- | --- | --- |
| No. | Questions asked in Zoom Chat | Tian Jing’s answer |
| VSE course (day 4) | | |
| 1 | What are the use case for doing analysis in frequent rather than spatial domain for computer vision? There is method to analyse image in spatial and freq domain? How to do that in freq domain? | There are many frequency-domain image processing. For example, apply wavelet transform on the input image; manipulate wavelet coefficients (e.g., high frequency coefficients) to remove noise. In addition, there are some works to perform image augmentation in frequency domain. |
| 2 | Color histogram doesn’t have spatial data, will it be accurate to just use it in neural network for background comparison, or do we need to have other preprocessing step? | Color histogram is a global-based (whole image-based) feature. Similarity, the ‘fc’ layers of the CNN model also lose spatial information from the image. However, they are effective to perform image classification, where we care more about image content (cat or dog) than the spatial information of objects (cat under table or cat on the table). |
| 3 | Some models/algorithms are scale or rotation invariant...what is the exact meaning of this? Are all deep learning model scale or rotation invariant as long as there is enough image augmentation. Tradition method? | If the method claims to be X-invariant, that means, if the input image is X-transformed, then the method will obtain the same result. Some methods have explicit operation to handle X-invariant, such as pyramid (scale-invariant) or rotation calibration (rotation-invariant) in SIFT. Some methods have implicit operation, such as max-pooling (translation-invariant) or pooling (scale-invariant) in CNN. |
| 4 | Is it possible to detect smoothness in texture? | Smaller gradient or intensity variance indicate smooth texture. |
| 5 | texture = pattern = change in colour. Can there be texture in grayscale image? how do you calculate gradient for a color image? isit an avg of each color channel? | The methods, which are defined on gray-scale images, can also be applied on each color component of color image. |
| 6 | How about one object has two shape, for example, dumbell, there is both long and round shape, if the HoG histogram can still be used to identify shapes? | In this case, we might need to consider combine HoG with other types of features. |
| 7 | If I want to draw the boundary based on the gradient directions, can I just draw a line perpendicular to the arrows | Yes, correct. |
| 8 | Can you explain about padding, why and how do we apply it for HOG? | Padding is used for boundary pixels of images in all image processing methods, where we don’t have sufficient number of pixels to perform certain operation (e.g., filtering or convolution). |
| 9 | In HoG, will the histogram be taken for each pixel cell and then averaged over all pixel cells histograms? Same as in LBP? | Yes. Certainly, you can modify it by changing ‘averaging over all cells’ to be ‘concatenating histograms obtained from different cells’. |
| 10 | So I can feed the output feature of hog to a NN for classification task? | Yes. HoG is just feature extractions. We will further use such features in other machine learning models. |
| 11 | How do we assess the quality of features extracted by CNN? Do we need to test it in another classification/detection model? | Yes, it depends on your analytics tasks, classification/detection/segmentation. All of them have different performance metrics. |
| 12 | Are the residual in CNN and the laplacian pyramid the same? is it related to the res block in ResNet? | From signal processing aspect, they have similar concepts. The Laplacian pyramid considers the difference between an input and its filtered version, while ResNet considers the combination (via trainable parameters) between an input and its filtered version. However, more importantly, from machine learning aspect, ResNet uses the skip connection to help back propagation. |
| 13 | Is the similarity measure not scale invariant? It means that two images of a person might be mis-labelled as dissimilar simply because they appear at different positions? If that's the case, should there be a localization process before evaluating the similarity? | For this scenario, we usually localize the object first, and then feed the localized object region (e.g., face region, not the whole image) to the Siamese network. |
| 14 | What is the effect of changing the margin value in Contrastive Loss? | The margin means that we hope the model is able to produce greater distance than some margin value for negative pairs. That means, as long as the model output produced for a negative pair is distant enough (larger than a margin), no efforts are wasted on enlarging that distance. A smaller margin indicates an easy training task, while a larger margin indicates a difficult training task. |
| 15 | What should be the output size of the last layer of the CNN in Siamese network? FCN or dense layer? | Either way is okay, it depends on your design of Siamese model. |
| 16 | The margin in SVM is calculated by algorithm instead of set as expectation, can this Siamese margin be calculated as well? | No, the margin is the user-defined parameter. |
| 17 | What is difference between contrastive loss vs binary log loss. | They are fundamentally different. Contrastive loss is used for distance learning, while binary log loss is used for classification. |
| 18 | If face verification will be affected by fake image (generated by GAN or something)? Will there be a possibility for common similarities in the fake image group, and common similarity in the real image group? | Yes, it is a challenging task for fake image. |
| 19 | We used FaceNet in our previous module, I think it is an example of triplet network, right? Also they have an embedding process, can you briefly share how the embedding works? if possible | Yes, you are correct. The face embedding can be considered as 'CNN features (before the loss)' (in slide 35). For the face dataset, you train the siamese network. For another face dataset, you apply the pre-trained Siamese network to get the CNN features (mentioned above), then you use this features to train another classifier model. So the CNN features are considered as 'embedding'. |
| 20 | Models which has more trainable parameters are generally "heavier" and therefore unsuitable for edge deployment? For edge deployment have to use a light model which generally means less parameters? | Yes. We have to design simplified model (e.g., MobileNet) or use model optimization tool (e.g., TensorRT, OpenVINO). |
| 21 | During the random generation of pos/neg pair in Siamese model training, what if a person called Norman, none of his photos are being used, can we still predict him? | It doesn’t matter. The Siamese network model learns similarity between faces. We don’t care whose faces are used in training. |
| 22 | Assuming we use face detection to localize faces in input images, what should we do to make the bounding boxes' shape the same as what the network's input layer? Do we simply resize? If so, won't that squash or stretch the input and affect the result? | Yes, resizing it according to the requirement of the model input shape. |
| 23 | The contrastive loss in the colab and the lecture notes.. are they different? it looks different for the y=0 part colab squares the max but lecture notes get max of the individual squares? | They are same. Please be careful in formula whether it is defined in ‘distance’ or ‘similarity’. Larger distance means smaller similarity. |
| SRSD course (days 1, 2, 3) | | |
| 24 | Why Gaussian filter again in SIFT? | SIFT uses pyramid to detect keypoints in multiple scales. You can change Gaussian filter to other smoothing filters. |
| 25 | So there will be more than one key point at different scales in SIFT? | Yes, possible. |
| 26 | Each pixel need to scan all the layers? if the same pixel has a local maximum in different layers, which s to choose? | Yes, we need to scan all pixel locations in all scales. It is not likely to have local maximum at the same pixel locations (x, y) at different scales (s). If it happens, there might be ambiguity for this keypoint. |
| 27 | So potentially matched keypoints can come from different scale? | Yes, possible if two images are captured using different scales (e.g., zoom in factors of phone camera). |
| 28 | If all the X-invariant features have been found out by scientist before they switch to CNN? are they still looking for new X-invariant features? | We cannot say all X-invariant features are found in the literature. We are still looking for better X-invariant features, or new X-invariant (e.g., cross modality, infrared image vs cctv image) |
| 29 | In pure CNN, we use pooling to scale down the input image...This means we can capture features at various scales? Why do we say that features capture at the previous levels are lost as we go deeper into the network (unless we implement pyramid or concatenation like unet)? Isn't all the information actually embedded in the subsequent activation map? | Yes, pooling can help to generate feature maps with various scales.  With deeper layers, we do more convolutions, we lost information in the previous convolution layers; however, we also create new information. |
| 30 | Does deep learning network produces invariant features? Unless there is enough image data or augmentation (eg image in different orientation)? How do we tell the level of invariance in features produced neural network? | CNN model is fairly invariant to translation (max pooling), or scale (pooling). We will use experiment to evaluate its invariant performance (e.g., can a CNN model recognize rotated traffic sign, or flipped human face). |
| 31 | In ORB method, check all 16 pixels and check if 12 contiguous pixels fall in the criterion -> means all 12 pt must be side by side with no break in between? Why must be contiguous ? | ORB is a faster (simplified) method to find keypoint via checking 12 continuous neighbourhood pixels (3 quarters). |
| 32 | Is there a chance that the test will have feature not in dictionary in feature encoding? | Yes, possible. However, we also assign such features into its closest cluster in our dictionary. |
| 33 | In VLAD method, the residuals are with respect to the cluster centers? | Yes, it is calculated as the difference between the input feature and the cluster center. |
| 34 | In BoW method, can dispersion within clusters be used as a feature also? | It is not in the original BoW and VLAD method. You can explore to use such intra-cluster dispersion features. |
| 35 | In VLAD method, we use the residual to measure the probability or accuracy to the specific cluster? | No, the residual indicates how the feature is close to the cluster centre. They will be used as additional features for feature matching. |
| 36 | In VLAD method, for image with multiple features of the same cluster, is it average of the vectors or vector addition? | For a single cluster, we need to use summation (not average) to combine residual vectors.  After we obtain aggregated vectors (from different cells), we need to normalize them. |
| 37 | How do u combine the encoded feature and the residual feature to make final comparison? | A simple way is to just concatenate them. |
| 38 | Could we use silhouette score to determine to optimum cluster? Or there is other way to do so? | Yes, possible. This is a separate task to try to determine the optimal # of clusters used in BoW method. |
| 39 | Can I say that feature encoding is an optional step to increase the speed of computation? It does not improve the accuracy at all? | Feature encoding is additional step to improve accuracy, particularly for large reference image database; however, it introduces additional computational cost. |
| 40 | For page 37, could you explain more about the pooling process? from my understanding the pooling process generate new centroids by concatenating the Vi for each clusters, then we can make use of the new centroids to assign new clusters for the input vector again? | Pooling is a fancy term to combine multiple input to be single one. |
| 41 | How is structure light different from TOF since all measure time of flight? | ToF sensors flood the entire scene with light and calculate depth by the time it takes each photon to come back to the sensor. Structured light, on the other hand, entails a known pattern projected by the IF sensor. |
| 42 | Is the convention for depth map to be white as closer to camera? | No, the color (lighter/darker color for smaller/larger distance) is just visualization. |
| 43 | How is Projections method different from Multiview method? I suppose in Multiview, the various views (View 1...View 2...View N) are also kind of projections of from 3D shape? | Projection is able to simulate data in the view, which is not available in the multiple-view acquisition. |
| 44 | If the scene only have breath and no depth...e.g. different objects directly placed in-front of a wall....Does using an additional depth image helps in the classification/detection task? | In this scenario (a plain wall or document), the depth might not be helpful. |
| 45 | In object detection, how should you handle cases where a bounding box overlaps more than 1 class? | It is a trade-off between whether you trust box classification score (two objects, since both boxes return high classification scores) or box overlapping (one object, since boxes should not overlap too much). |
| 46 | During training phase, should we generate patches with different aspect ratios? then do resize before model training? | Yes, similar to image augmentation in model training. |
| 47 | Is box regression stage part of training step? | Yes, box regression is trainable. |
| 48 | Since region proposal depends on the feature maps, does that mean the CNN filters should already be trained to a certain extend before the Region Proposal can begin? | You can treat it as pre-trained CNN or re-train it. |
| 49 | Can I ask how to evaluate the box regressor and know it is a good selection box? As every time the items are in different location. | It is regression loss between predicted (x,y,w,h) with ground truth (x,y,w,h). |
| 50 | Will there be a feedback from Classifier to Regression? | No, there is no feedback between these two. |
| 51 | How are the anchor points determined? | Anchor locations are determined by users. Usually, they are located at each positions in various feature maps. |
| 52 | In object detection, is the size of the input image fixed? Could it be adopted to other input image size without performing resizing? | Yes, because we have box classification model, which usually have fully-connected layers. |
| 53 | So the backbone CNN is set to un-trainable, then we do anchor points on backbone CNN's output (feature map). | Yes, usually, we use pre-train CNN as backbone. |
| 54 | Will we define our anchor points in shape and ratio in terms of size of feature map or raw image? (then model will somehow auto map anchors coordinate on image to coordinate on feature map?) | Yes, users need to configure the anchor box size, aspect ratio, etc. |
| 55 | Why is it called 2 stage? We can't train both RPN and per region network together at the same time? | In the literature, they are trained separately. |
| 56 | Shared layers saves computation time and resources? any other benefits? | From machine learning aspect, shared weights can reduce the number of trainable parameters. |
| 57 | How do we annotate the depth image? | We don’t need to annotate the depth image, which has same resolution with the RGB image. |
| 58 | Can I ask how to get geometrical features? | I think you mean HHA image, which stands for horizontal disparity, Height above ground, and Angle with gravity, defined in the ECCV 2014 paper in the slides. |
| 59 | Dimensionality of RGB and geometric features must be the same? | Resolution must be same, as they use shared conv filters. |
| 60 | Semantic segmentation looks like cartoonization of image, could you explain how the cartoonization is usually achieved? | Similar concept. |
| 61 | In practise, if I want to train with my own dataset, how do I obtain the label map? | You need to either label yourself or pay vendor to label for you. |
| 62 | Why full convolution layer can take input of any size while CNN+fully connected layer can only accept certain input size? | The fully connected layer reshape 2D feature map into 1D vectors, where the number of nodes depend on the resolution (width, height) of 2D feature map. If the input image has different solution, it will generate feature maps with different resolution. |
| RTAVS course (days 1, 2, 3AM) | | |
| 63 | For training the talking face model, why we need to use next frame's video image to train? how about testing, do we generate the video image for current frame or next frame when playing the voice? | Given an audio signal at the certain time index, I think these two different models (in terms of problem formulation) to ‘generate current’ frame or ‘predict next’ frame. |
| 64 | When we identify two pixels as the same one, if we identified wrongly due to a very similar pixel in another far away place in the image, the motion vector will be larger than usual, do we supress this motion vector? | We don’t search for the best matched ‘pixel’ for every pixel. We use math of optical flow to calculate such pixel-based motion vector. |
| 65 | When you mention motion vector has row and col, is it expressed in "change in pixel location" in row and col? | Yes, motion vector is defined as ‘displacement’ (so-called change in pixel locations) in row and column. |
| 66 | Are there heuristics for which of 4 types of motion representation to apply for specific scenarios? | * Frame-based: global motion for the whole frame such as the camera motion. * Block-based: fast, detect scene change, etc. * Pixel-based: rich motion information for action recognition, etc. * Region-based: requires to segment the scene into various objects, video conferencing, etc. |
| 67 | For block-based, are you saying each block is one dim, or all the blocks is 1 dim? | For every block, it has 1 motion vector (2 dimentions). |
| 68 | Can I say pixel based motion vector is similar to disparity ? | Both pixel-based motion vector and disparity can be expressed in terms of displacement at each pixel location. However, their physical meaning is different. Motion vector is due to motion, while disparity is due to view point change. |
| 69 | If there are 4 blocks, then you have to match each block (in current frame) to another block (in previous frame)? Then there are 16 possible combination? How to find the best match combination? | No, we don’t consider such one-to-one combination. |
| 70 | If the origin position of the motion vector is each block center or each pixel from the previous frame, and destination position will be the corresponding point in next frame? | Yes. |
| 71 | I have seen an alternative constraint imposed on LK equations, which is expressed in variational form. The idea behind that is to impose smoothness on the motion vector field. Is this formulation equivalent to the one in the slide? | Yes, this is the same spatial coherence constraint. It could be expressed in variational form (as in your question) or even same (in my slide). |
| 72 | If losing information between frames is a problem, will it be easier to just increase sampling rate? | Yes, you can increase frame rate to analyse frames with shorter interval, subject to your computational power. |
| 73 | If the brightness used in Optical flow calculation has to use RGB color space? | It does not matter, either color or gray scale are okay. We can simply calculate gradient for each color component and then average them. |
| 74 | Will optical flow fail if the texture is very constant/similar e.g. a grid of squares, since its harder to find a unique match? | Yes, it is a challenging task. |
| 75 | Is there a way to decompose the resultant frame-to-frame movement into sum of global camera movement and instances movement? | We usually do not ‘decompose’ the calculated optical flow into two components (one for camera motion, one for object motion). We usually compensate/calibrate camera motion BEFORE we apply optical flow. |
| 76 | For those matching method, the size of the slide window depends on the template, right? so zoom out or zoom in will influence the detection? | Yes, the same block size. Since these two input frames have very short interval (e.g., 1/20 second), the object is assumed to be same size. Therefore, we use same block size in block matching. |
| 77 | Am I right to say that pixel level optical flow can be derived by below two main methods?  -       mathematical equation (Lucas-Kanade equations)  -        neural network (e.g FlowNetSimple, FlowNetCorr)  So what is the main diff in the pixel flow obtained via these two methods? Training vs non-training etc... | Yes, correct. These two methods obtain the same results (optical flow), but they are different in the sense of training-based/non-training based. |
| 78 | Correlation filter is on-train training while pos/neg is off-line training? Why can't Correlation filter be trained off-line? Can't i generate various response off-line to determine the best correlation filter? | These are ‘on-line’ training, because we don’t know what is ‘target object’. |
| 79 | Detection task in video has to take temporal info into consideration else the bounding box will shift (not smooth) from frame to frame. Can explain again? i suppose as the target moves across frame, the BB will also shift...So what do u mean by BB shifting? | I refer to bounding box (x, y, w, h) parameters might be noisy (not smooth) for neighbouring frames. |
| 80 | For real time model over a sequence or video, if not all frames are discriminative enough, is there guideline about the intervals to set when fetching the frames? e.g. every 5 seconds or 10 seconds? | We still extract frames, but we analyse motion BEFORE we perform action recognition. For example, if there is no motion in these 5 frames, we don’t need to perform action recognition. |
| 81 | Shouldn't the LSTM/3D-CNN be trained to be intelligent enough to pick up the specific motion? What is the difference between sequent info and motion info? Since both are sequential? Why can't LSTM gives us motion info? | In summary, the motion can be obtained from either of following ways   * Explicit motion estimation, optical flow. * Sequential modelling, such as LSTM. * 3D-CNN model |
| 82 | But in practice, other body parts will also have slight movements, the combined image will be fuzzy? | Yes, so this motion energy or motion history method is useful when there is only 1 object in the video. |
| 83 | there is also limitation for the motion history, only up to 255 frames. Can we overlap the optical flow with the stacked image instead? | No, we don’t have such limitation ‘255’. |
| 84 | If we don't stack different frames in the same position, instead, we put each frame in a different block to within a single image for learning, will this method help? | No, there is no correction among these segments (in a single frames) |
| 85 | regardless of motion energy/history, its assumed the object to be analyzed is already segmented correct? | Yes, it requires object mask to be obtained. |
| 86 | Can you share how can we use this special image in further application? For example, can we use them as input to a neutral net? In this case, the feature will be special image, the label/target will be the specific action? Any other possible application examples? | This 1 special image (e.g., motion energy image or motion history image) represents information from the whole sequence. You can apply whatever we have learned from static image on such special image. |
| 87 | One last question, why can't we use optical flow from two consecutive special image? | In motion energy image and motion history image method, there is only 1 special image obtained. |
| 88 | Yes, there is only 1 special image but if you want to use CNN, you ahve to create many special image? | For every sequence, we obtain 1 special image.  So 1000 sequences have 1000 special images, which are used to train the model. |
| 89 | The cuboids is 3 x 3 x 2...How to understand this dimension? | It splits the cuboid into 3\*3\*2 (3 in row direction, 3 in column direction, 2 in time direction). |
| 90 | do we have to understand each method's concept (will these be tested) or these are for our info? | We test concepts, not the individual paper. It is important to understand these innovative ideas from various papers so that we can build our own model. |
| 91 | If we don't split into Horizontal and vertical, directly do OF+HOG. able to get Object motion? | We can get motion, which is a combination of camera motion and object motion. |
| 92 | How about the camera motion is not constant? | Camera motion is considered global (same motion is applied on the whole image) and constant (within 1/20 second) |
| 93 | If there is no camera motion, what will be the values of the MBH? | No change to MBH, as MBH only depends on object motion. |
| 94 | What is the meaning "Detect human body to remove spurious trajectories."? | Apply human detection on each frame, set the motion vector (outside of the detected human bounding box) to be zero. |
| 95 | How to understand why single frame (randomly chosen) CNN works? How is prediction possible when the frames are just randomly chosen? | In 2014, It was a good model to extend the advanced image CNN models to be multiple-frame video sequence. |
| 96 | what is LRCN used for? classification or predict what activity will happen next? | It is for classification. |
| 97 | why is LRCN in the slide has so many outputs? | It is sequential model. Also, we will use Many-to-One LSTM for this classification task. |
| 98 | Can clarify on page 24 and page 25. the late fusion in page 24 show only two photos. but page 25, late fusion uses the 6 photos | Yes, it places two separate single-frame networks for two consecutive or sampled frames. The point is that it applies single frame CNN. |
| 99 | So concatenation: (1,2) and (3,4) becomes (13,24)? | No, it changes the dimension of arrays, not the element values in the arrays. Please refer to various concatenate functions provided in Numpy. |
| 100 | The H & W in the slides refers to the resolution at the fusing layer, not the res of the frame correct? | Yes, here H and W refers to height and width of feature map, so called tensor. |
| 101 | Is it possible to do sum fusion for the network shown in pg. 28? The 2 branches have different no of frames | In slide 28, it is multiple-channel 2D-CNN. In this method, after the convolution layer, the feature map dimension depends on # of convolution filters, not the # of frames in the input. |
| 102 | Can I say late fusion is frame skipping? | No, these are two different concepts. |
| 103 | Can also clarify what does the D stand for which we have been using in the slides? is it the number of different possible features? Because from the paper, they say channels of the respective feature maps. | Yes, you can call D is # of channels of the feature map. |
| 104 | All models learned today are for classification. How can we modify some of them to do prediction? Like prediction of customer response to video... | I personally think we need to predict based on the structure data (e.g., number of humans, facial expression,) extracted from the video, not predict ‘image content’. |
| 105 | Could I treat the RGB channels as the 3 input channels of 3D-CNN instead? | Yes, but in this case, your input is only a 3\*T (T=1, so totally 3) array, you don’t need to do weight sharing. That means, the 3D-CNN is more useful when your 3\*T is large (i.e., T is a large number). |
| 106 | For the model summary on slide 32, does it mean that the input frame counts (T) need to be fixed? | Yes, it is fixed. Otherwise, the feature map might have different size N (# of conv) \* T (# frames) \* W (# width) \* H (# height). If we further flatten it into fully-connected layer, it will have different # parameters. |
| 107 | Can we upsample to make up the lost fidelity to make the sounder clearer like from sample 1 to sample 2? | Yes, we can use signal processing method (interpolation) to increase # of samples in the audio sequence. |
| 108 | how does quantization affects music quality? | It will affect the data quality in terms of precision or details. For example, if only 1 bit is used in quantization, it only can present silent audio and non-silent audio. |
| 109 | Is D and G the same signal? same values at each point of the curve? | Yes, they have the same values. |
| 110 | Any particular parameter we need to pay attention to for quantization to prevent loss of vital information? | This is in audio acquisition step. We don’t need to consider this in audio analytics step. |
| 111 | What's difference between zero cross and signal frequency? Seem similar... | They are similar in the concept to evaluate how the signal change over the time. But they have different mathematical formula. |
| 112 | why not 3d feature as there is frwq and amplitude and time dimesnion? Spectrogram gives info of freq and amplitude info over time? Y axis is frequency. X-axis is index of window (time). Amplitude is given by colour intensity. Since there are 3 dimensions, why feature is not 3d but 2d? | The spectrogram is 2D, where the # of rows is determined by # of segments, # of columns is determined by Fourier coefficients.  We don’t treat amplitude (color intensity illustrated in spectrogram) as the 3rd component. Otherwise, a gray-scale image would be 3D (row, column, intensity). |
| 113 | For language classification, which type of model will have better accuracy? audio based or text based? | I don’t have experience for this specific application, but I think audio features would be good feature, as different language have different phenome. |