

Computer Vision
Prof. Jayanta Mukhopadhyay
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 28
Feature Detection and Description Part – V

We are discussing about descriptors, different kinds of featured descriptors in an image. In our previous lectures we discussed how key points could be described, then we also considered description of patches or regions. And after that we have also discussed about similar kind of descriptors of regions, where textures are described. And, today we will discuss another kinds of description of images. When the whole image is required to be described, the whole visual content in an image needs to be represented. And, in that scenario let us see what kind of techniques are there for image description. Typically, we will discuss two such popular techniques. One of them is known as bag of visual words and then we will consider vector of locally aggregated descriptions.

(Refer Slide Time: 01:21)




Image Descriptor

- Bag of visual words
 - Compute key-point based feature descriptors over a library of images.
 - Quantize them (clustering) to form a finite set of representative descriptors (visual words). K-means clustering
 - For an image assign the nearest visual word corresponding to the feature descriptor of a key point.
 - Represent by each image by a histogram of visual words.

Sivic et. al., Discovering objects and their location in images, ICCV' 2005.

So, we will see that kind of descriptors. So, bag of visual words in this representation what we do. First we compute the key point based feature descriptors over a library of images. So, in this case we need to construct visual words from a set of images and use those visual words. So, it is a kind of dictionary and dictionary in terms of visual words and then

describe any image using those visual words whether they are present in the image or not or in what number they are present in that image.

So, the analogy of this kind of approach came from natural language processing or document retrieval if I say particularly. So, in the textual documents, there are words which are present in the documents and some of those words which play their distinctive roles on the nature of the documents. So, those are the words which are textual words and a document can be represented by some feature vector where the presence of those words are counted for.

Now, the similar kind of concept has been extended when we are trying to describe the visual content of an image, but the problem here is that, we do not have any precisely defined set of words, as we can get it from the documents. Because, dictionaries are independently created by linguists, by any language there are dictionary of words. So, we can use those dictionaries to represent any word and those words could be used for representing the document.

But, in the images, the first task could be to create a dictionary of words and which should be dependent on the visual content or local visual features of an image. So, this locality is first addressed by considering the key points of an image because, key points as we understand they play a very important role in defining transformation invariant features in an image. So, they are the locations, they are kind of landmark locations which are easily detectable even after the image is transformed. So, key points are those positions in an image where you are considering those landmarks. then the description of those key points are possible candidate of visual words.

So, that is the philosophy when we are considering these step. You consider a set of images for each image you compute the key points by using different kinds of detectors what we discussed earlier. It could be a SIFT detector, it could be a detector based on SURF, it could be other detectors like FAST and those will be giving you those locations. And, also you understand that even the scaling variance is also considered for detection of key points in some of these techniques. And, then around these locations, you are deriving a descriptor which also should have the property of transformational invariance.

But the problem here is that you get so many varieties of descriptors that there would be a large number of descriptors in that case and it is very difficult to represent an image with

so much of variations. So, what we need to do next is that all these key points we need to put them under some quantization schemes. That means, we would like to get some representatives of these visual words. So, it is not key point, it is a descriptor which we are considering here. A descriptor corresponds to a visual word in our candidate visual word dictionary, but as there could be lot of variations.

So, we would like to put those visual words into buckets which are very similar and choose one representative out of them. So, this task in the vector space also it is called quantization, it is also called clustering when you are bucketizing them or when you are grouping them into a similar setup in similar visual words or similar feature vectors and then choose a representative out of it. So, there are different algorithms for clustering and there is a technique called K means clustering technique by which you can easily perform this kind of grouping and we can choose the representatives from that group.

So, if I consider a set of feature vectors from a group, then the mean of that feature vector is used as a representative for that set. So, the parameter K, it denotes that how many such representatives or how many such groups you are going to create. And in fact, that would give you the number of visual words in your dictionary and using that fixed number of visual words we are going to describe an image. So, it has certain advantages that dimension of your feature description is determined by this number K and it remains the same for every kind of image under your consideration.

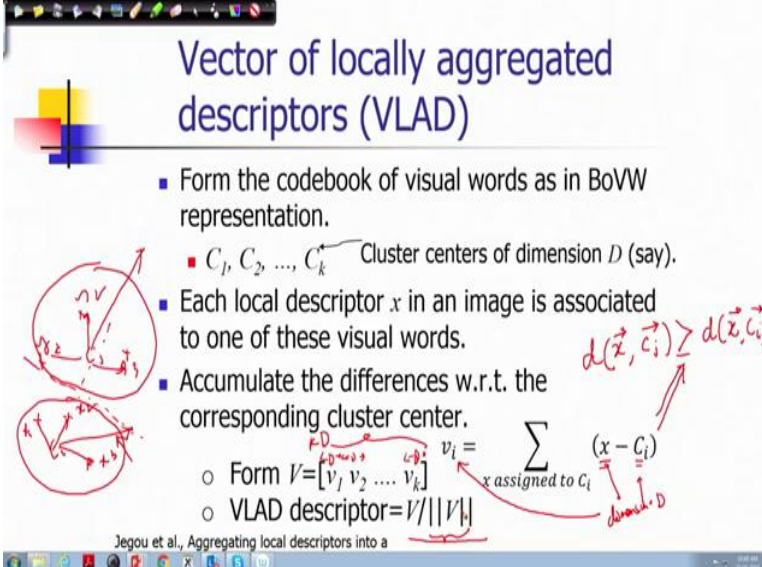
So, once again just to summarize these steps: you take a set of images which you can call a library of images from where you are trying to form a dictionary of visual words. For that what you are doing? For each image you are computing the key points, at each key point you are computing its feature descriptors. Then you get a collection of feature descriptors from all such images and use that collection in clustering algorithm, apply a clustering algorithm to that collection of feature descriptors. So, group them into a finite number of clusters like using K means clustering algorithm which I will discuss in later on in some other topic of clustering and classifications in this particular course.

Right now, let us understand what it does. It simply partitions the descriptors into K groups or K partitions. So, from each partition you take the mean as a representative of that cluster and they form visual words. So, you have K visual words out of this process. Next, what you are doing? That, now your dictionary is prepared, now you consider any image. And,

for any image you again compute its key words, key points and consider a descriptor of that key point and then you find out that which is the nearest visual word corresponding to that feature descriptor and then that description is associated with that word.

So, here you are counting how many such representative visual words, how many times they occur in an image. It is the similar strategy where you count how many key words occur in a document from a dictionary, similarly how many key visual words they occur in your dictionary in a particular image that we are doing. And, that can be represented by a histogram because histogram gives you frequency distribution of these visual words of your dictionary and the number of bins in this histogram is fixed by the number K . So, you have a K dimensional feature representation for an image and this is what the representation by bag of visual words is.

(Refer Slide Time: 10:35)



Vector of locally aggregated descriptors (VLAD)

- Form the codebook of visual words as in BoVW representation.
 - C_1, C_2, \dots, C_k Cluster centers of dimension D (say).
- Each local descriptor x in an image is associated to one of these visual words.
- Accumulate the differences w.r.t. the corresponding cluster center.
 - Form $V = [v_1 \ v_2 \ \dots \ v_k]$

$$v_i = \sum_{x \text{ assigned to } C_i} (x - C_i)$$
 - VLAD descriptor = $V / \|V\|$

Handwritten notes on the slide:
 - A diagram on the left shows a 2D space with points x_1, x_2, \dots, x_k and cluster centers C_1, C_2, \dots, C_k . Arrows indicate the assignment of points to cluster centers.
 - A red arrow points from the text "Accumulate the differences" to the formula for v_i .
 - A red arrow points from the text "VLAD descriptor" to the formula $V / \|V\|$.
 - A red arrow points from the text "x assigned to C_i " to the summation in the formula for v_i .
 - A red arrow points from the text "dimension D" to the variable D in the formula for v_i .
 - A red arrow points from the text "distance" to the formula $d(x, C_i) = \|x - C_i\|$ shown in red.

Jegou et al., Aggregating local descriptors into a

Next we will consider another kind of descriptor which is an extension of bag of visual words description we will see, the concepts of visual words are used here also. But, the summarization of the visual content or the nature of description is bit different and that we would like to see what kind of variation is there in this description. So, this technique is known as Vector of Locally Aggregated Descriptors or in an acronym we call it VLAD. So, the first step is similar, you have to form the codebook of visual words as you did in the bag of visual word representation.

So, which means that, once again you should take a set of images, you should extract the key points for each image and at each key point again you form the feature descriptors; you will get a set of feature descriptors. And, then perform K means clustering on those feature descriptors and get K visual words or that is what your bag of K visual words is, that you derive in the first step form the dictionary. Suppose, these representative visual words are denoted by the symbols like you have K such visual words so, each code we can call it also the codebook. So, each is a feature vector and they are denoted by say C_1 C_2 to C_k .

So, they are just the cluster centers of dimension D in this case. So, this D dimension is determined by the dimension of the feature descriptor. For example, we know that for SIFT feature vector the dimension is 128, for SURF feature vector the dimension is 64; likewise there are different other feature descriptors from where these dimensions are defined. These are all conventions as you understand, you can have your own feature descriptor also and accordingly you can determine this dimension. Then after that we consider that aggregation operations with respect to these centers. Let us understand this mathematical operation. So, consider a local descriptor 'x' in an image and that is associated to one of these visual words.

So, then we are accumulating the differences with respect to the corresponding cluster center. So, this operation can be described by this particular summation operation.

$$v_i = \sum_{x \text{ assigned to } C_i} (x - C_i)$$

We can see in this representation x is the unknown local descriptor and C_i is a cluster center. And, such that the distance from x to C_i is the minimum among all cluster centers that is how the x is assigned to C_i . So, if I compute the distance, suppose you use some distance function let me represent the distance function as say 'd'. And, distance between feature vector x and any cluster center C_j that is denoted in this case like this.

$$d(x, C_i)$$

$$d(x, C_i) > d(x, C_j)$$

In this case, you can see equality also in a degenerate case. So, at least all the distances should be either greater or equal to this distance then you can assign x to any such cluster center; if there are multiple centers which are of equal distances and which are also minimum you have to choose one of them. So, then what you are doing? You are simply aggregating the differences of that cluster center and with respect to a cluster center C_i . So, if I geometrically look at it say C_i and then suppose there is another cluster center C_j and there are feature vectors of this document say this is (x_1, x_2, x_3) and say (y_1, y_2, y_3) .

So, these are all feature vectors of the same document, I could have written $x \ 4 \times 4$; also let us follow that and what we are doing? You are considering say the Euclidean distances. So, you know that if I draw the perpendicular distance connecting these two feature vectors a perpendicular then this hyper plane will separate into two. So, all these feature vectors that would be close to C_i and all these feature vectors would be close to say C_j and then what you are doing? See this is the vector, this is vector, and this is a vector. So, you are just adding this vector; so, it is a resultant of all these vectors with respect to C_i . Similarly, so if I take the resultant, it would be some resultant direction say this is some resultant direction, this is the resultant direction.

Similarly, for C_j also you consider the resultant directions, some resultant directions would be this. So, so that is what is your feature descriptors v_i , accumulate the differences with respect to the corresponding cluster center; you are basically concatenating all these descriptors. So, if they are so the dimension is quite large as you can see, that if you have C_i as a dimension say D ; so, dimension of v_i would be also D . So, what we are doing here so, that if I consider the dimension of C_i is D dimension, the dimension of v_i would be also D and there are concatenation of all D dimensional feature vectors. So finally, you will have a feature dimension of K into D . So, you note that your dimension of feature representation, it increases in VLAD considering the number of bag of visual words. But, in reality what happens that in this representation your number of K is kept small and that is how it matches with the bag of visual word; it can compete with the bag of visual words representation also in terms of the length of a feature vector. It has been found this is also efficiently discriminating an image.

So, VLAD descriptor is the normalized vector of this representation of aggregated differences or concatenated aggregated differences. So, that is how the VLAD descriptor is defined. You see that this vector is divided by its magnitude. So, this particular operation is the magnitude of the vector which is L2 norm of that vector and if you divide it, then you will get the normalized representation. So, this is how another kind of representation is obtained. Let us try to understand that what could be the motivation of this kind of representation. So, what are the applications?

(Refer Slide Time: 18:57)

Vector of locally aggregated descriptors (VLAD)

- Form the codebook of visual words as in BoVW representation.
 - C_1, C_2, \dots, C_k Cluster centers of dimension D (say).
- Each local descriptor x in an image is associated to one of these visual words.
- Accumulate the differences w.r.t. the corresponding cluster center.
 - Form $V = [v_1 \ v_2 \ \dots \ v_k]$

$$v_i = \sum_{x \text{ assigned to } C_i} (x - C_i)$$
 - VLAD descriptor = $V / \|V\|$ ← Dimension: $k \cdot D$

Jegou et al., Aggregating local descriptors into a compact image representation, CVPR, 2010.

(Refer Slide Time: 19:01)

Application of global image descriptor

- Content based image retrieval
 - Image search based on visual content



So, one of the application of this kind of global image descriptor is content based image retrieval. So, let us try to understand what is meant by this content based image retrieval. It is the image search that you can perform using any search engine where your query would be an image and it would search similar images in your library of images, from your database or from your library from different repositories scattered over the web.

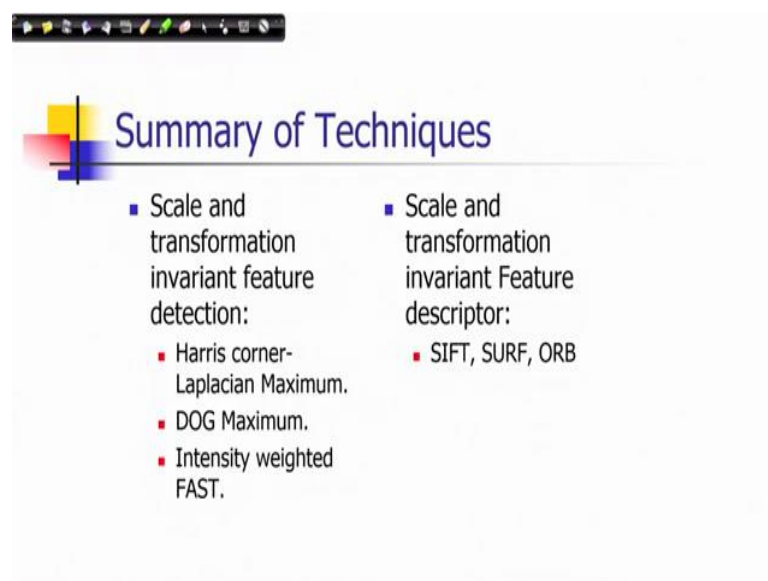
So, content based image retrieval is meant for that so, for that what you need to do? You need to represent all the candidate images or images of your library using some of these descriptors and given any query image you also convert it to its descriptor. And, then try to match between the query descriptor and the descriptors of your library images and which one is closest or rather a set of images which are closer than the others; so, you can even rank them; that means, stop match, next stop matching etcetera. So, in this way you can rank those images say top 50 images, top 5 images you can find out that list and you can report them, which is the way content based image retrieval works.

So, it is an image search based on visual content, there is a thing and one such example of these operations can be shown here. Suppose, this is a query image and you have a library of images and as I mentioned all those images in that library they are stored to they are descriptors of stored in your database. And, then your objective would be to match the descriptor of query image to those descriptors of images in your database. So, if I perform this kind of operations and if I report few top ranking matches we can get some results like this. So, this is the operations I am showing from one of the processings, one kind of one of the systems of content based image retrieval and, you can see the image of a chariot with wheel, it is a very famous image from on the heritage site Hampi and there is a temple called Vithala temple where this image has been captured. And, in your library also it contains several such instances taken from different views and some of those instances are also captured by the search and those are shown here.

And, the idea is that when you create an image database with the image, images can also keep various other metadata information, its descriptions. Suppose you have captured newly this image while traveling and you would like to know what this particular object is called. So, you give a search and associated images should be shown including their description.

So, this is one kind of very useful applications which are presently available in different systems. I can give you another example. So, these are the retrieved images from a database. So, you see that one of them is not really a match, it happens because these are all represented in a feature space and there would be ambiguities in representation. So, many of the images which are different, but they can have similar feature vectors. This is another example, this is from a place called Bishnupur in West Bengal and there are terracotta temples which are very famous. This is image of one such temple and using our image search technique, we could get this kind of representations.

(Refer Slide Time: 23:19)

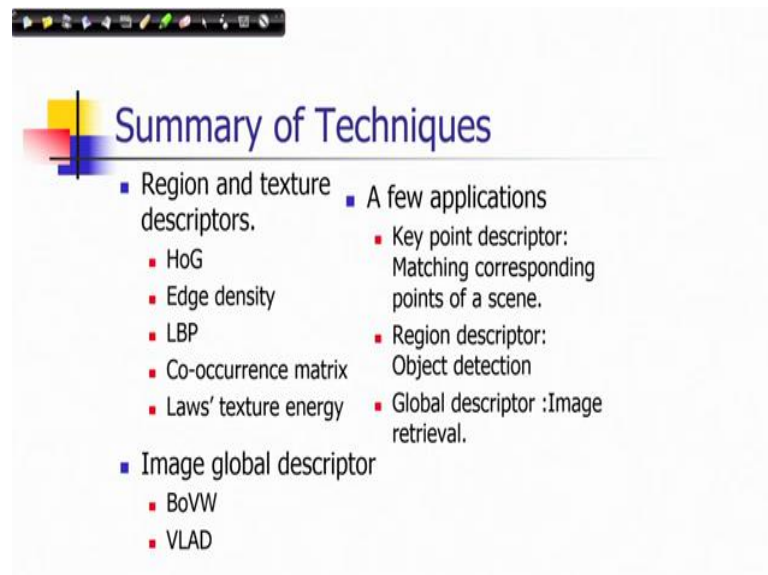


So, in this way we have seen that there are different kinds of descriptions of an image including its descriptors for the key points or for the points or descriptors for regions even descriptors for the image globally. So, in this topic we discussed all such techniques of descriptions and also detections of the point key points of the images particularly. So, we will summarize this all this topics, all these different issues what we discussed, different techniques what we discussed in this in this lectures on this topic.

And so, we have discussed about scale and transformation invariant feature detection. And, some of them are like Harris corner or Laplacian maximum; those are the methods of feature detections. You can use also difference of Gaussian maximum or extremum, it could be minimum also and then we have used intensity weighted first which is also gives a detection in terms of rotational invariant detections in that way.

And, we have also discussed different kinds of feature descriptors like they are transformation invariant descriptions; typically the descriptors like you know scale invariant feature transform, SIFT or SURF or ORB. So, these are different kinds of descriptors that we discussed.

(Refer Slide Time: 25:13)



And, we discussed also the region and texture descriptors that is another kind of description related to images. Here you consider a patch or a region instead of a point and usually those patches or regions are there of rectangular shape and in that there are different kinds of descriptors that we discuss like histogram of gradient. And, there are different texture descriptors, edge density based descriptors, local binary patterns and its variations where you can make it rotational invariant. You can also handle noise in this and make it more robust in that way. That is not only rotational invariant there is a concept called uniform pattern that has been used to make it more robust, but for handling noises. And, then there are co-occurrence matrix based description. So, local binary pattern descriptors, you have to create histograms in that particular region of these patterns that gives you a descriptor.

In co-occurrence matrix based descriptors you have to form co-occurrence matrix depending upon the coupling of pixel values; that means, pairs of pixel values with respect to predefined relationships of spatial relationships among the paired pixels. So, in this way you can define a frequency distribution of different pairs of values and that would give you the co-occurrence matrix. And, from co-occurrence matrix again different kinds of

features, different kinds of measurements and features you can define and which can be used for the description of a region with certain texture.

Laws' texture energy is also another technique which we discussed and finally, we discussed about the global descriptor of an image. And, we have seen its application or content based image retrieval or image based searches and two techniques are particularly discussed. One is bag of visual word representation, another one is vector of locally aggregated differences VLAD technique. So, these are the two techniques we discussed. So, with this of course, you know there are few applications that also we considered, let us look at them also.

Like for key point descriptors we know that they could be used for matching corresponding to get corresponding points of a scene in a multi view imaging system, particularly stereo imaging systems. And, they could be used for obtaining different kinds of characteristics matrices of those imaging systems like fundamental matrix of its history imaging system. Then for region descriptors object detection could be one applications, there are using HoG, there are applications of pedestrian detection, there are applications of character recognition etcetera.

And, for global descriptors we have discussed about a technique an application of image retrieval. So, with this let me stop this particular lecture and also we conclude this topic with this summarizations. We will move over to our next topic of matching and we will be discussing also matching and model fitting in our next topic. Some parts of matching of feature descriptors we have discussed in this topic also, but we will have more elaborations in the next topic.

Thank you very much, for your listening.

Keywords: VLAD, visual words, image descriptor, co-occurrence matrix, local binary pattern.