Unsupervised Identification of Multiple Objects of Interest from Multiple Images: dISCOVER

Devi Parikh and Tsuhan Chen

Carnegie Mellon University
{dparikh,tsuhan}@cmu.edu

Abstract. Given a collection of images of offices, what would we say we see in the images? The objects of interest are likely to be monitors, keyboards, phones, etc. Such identification of the foreground in a scene is important to avoid distractions caused by background clutter and facilitates better understanding of the scene. It is crucial for such an identification to be unsupervised to avoid extensive human labeling as well as biases induced by human intervention. Most interesting scenes contain multiple objects of interest. Hence, it would be useful to separate the foreground into the multiple objects it contains. We propose dISCOVER, an unsupervised approach to identifying the multiple objects of interest in a scene from a collection of images. In order to achieve this, it exploits the consistency in foreground objects - in terms of occurrence and geometry - across the multiple images of the scene.

1 Introduction

Given a collection of images of a scene, in order to better understand the scene, it would be helpful to be able to identify the foreground separate from the background clutter. We interpret foreground to be the objects of interest, the objects that are found frequently across the images of the scene. In a collection of images of offices, for instance, we may find a candy box in some office image. However, we would perceive it to be part of the background clutter because most office scenes don't have candy boxes. Most interesting scenes contain multiple objects of interest. Office scenes contain monitors, keyboards, chairs, desks, phones, etc. It would be useful if, given a collection of images of offices, we can identify the foreground region from the background clutter/objects and further more, separate the identified foreground into the different objects. This can then be used to study the interactions among the multiple objects of interest in the scene, learn models for these objects for object detection, track multiple objects in a video, etc. It is crucial to approach this problem in an unsupervised manner. First, it is extremely time consuming to annotate images containing multiple objects. Second, human annotation could introduce subjective biases as to which objects are the foreground objects. Unsupervised approaches on the other hand require no hand annotations, truly capture the properties of the data, and let the objects of interest emerge from the collection of images.

In our approach we focus on rigid objects. We exploit two intuitive notions. First, the parts of the images that occur frequently across images are likely to belong to the foreground. And second, only those parts of the foreground that are found at geometrically consistent relative locations are likely to belong to the same rigid object.

Y. Yagi et al. (Eds.): ACCV 2007, Part II, LNCS 4844, pp. 487–496, 2007.

Several approaches in literature address the problem of foreground identification. First of, we differentiate our work from image segmentation approaches. These approaches are based on low level cues and aim to separate a given image into several regions with pixel level accuracies. Our goal is higher level, where we wish to separate the local-parts of the images that belong to the objects of interest from those that lie on background clutter using cues from multiple images. To re-iterate, several image segmentation approaches aim at finding regions that are consistent within a single image in color, texture, etc. We are however interested in finding objects in the scene that are consistent across multiple images in occurrence and geometry.

Several approaches for discovering the *topic* of interest have been proposed such as discovering main characters [1] or objects and scenes [2] in movies or celebrities in collections of news clippings [3]. Recently, statistical text analysis tools such as probabilistic Latent Semantic Analysis (pLSA) [4] and Latent Dirichlet Allocation (LDA) [5] have been applied to images for discovering object and scene categories [6,7,8]. These use unordered *bag-of-words* [9] representation of documents to automatically (unsupervised) discover topics in a large corpus of documents/images. However these approaches, which we loosely refer to as *popularity* based approaches, do not incorporate any spatial information. Hence, while they can identify the foreground separate from the background, they can not further separate the foreground into multiple objects. Hence, these methods have been applied to images that contain only one foreground object. We further illustrate this point in our results. These popularity based approaches can separate the multiple objects of interest only if they are provided images that contain different number of these objects. For the office setting, in order to discover the monitor and keyboard separately, plSA, for instance, would require several images with just the monitor, and just the keyboard (and also a specified number of topics of interest). This is not a natural setting for images of office scenes. Leordeanu, *et al.* [10] propose an approach to unsupervised learning of the object model from its low resolution video. However, this approach is also based on co-occurrence and hence can not separate out multiple objects in the foreground.

Several approaches have been proposed to incorporate spatial information in the popularity based approaches [11,12,13,14], however, only with the purpose of robustly identifying the single foreground object in the image, and not for separation of the foreground into multiple objects. Russell, *et al.* [15], through their approach of breaking an image down into multiple segments and treating each segment individually, can deal with multiple objects as a byproduct. However, although from multiple segmentations, they rely on consistent segmentations of the foreground objects.

Further on the object detection/recognition front instead of object discovery, object localization approaches could be considered, with a stretch of argument, to provide rough foreground/background separation. Part-based approaches, as is ours, however towards this goal of object localization, have been proposed such as [16,17] which use spatial statistics of parts to obtain objects masks. However, these are supervised approaches for single objects. Unsupervised part-based approaches for learning the object models for recognition have also been proposed, such as [18,19]. However they deal with single objects.

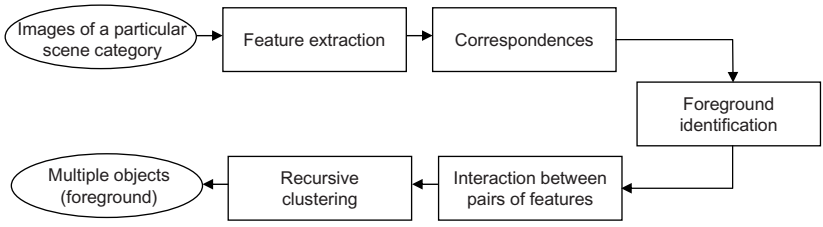


Fig. 1. Flow of dISCOVER for unsupervised identification of multiple objects of interest

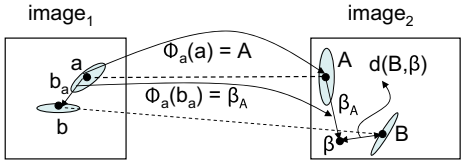


Fig. 2. An illustration of the geometric consistency metric used to retain *good* correspondences

The rest of the paper is organized as follows. Section 2 describes our algorithm dISCOVER, followed by experimental results in Section 3 and conclusion in Section 4.

2 dISCOVER

Our approach, dISCOVER, is summarized in Fig. 1. The input to dISCOVER is a collection of images taken from a particular scene, and the desired output is the identified foreground separated into the multiple objects it contains.

2.1 Feature Extraction

Given the collection of images taken from a particular scene, local features describing interest points/parts are extracted in all the images. These features may be appearance based features such as SIFT [20], shape based features such as shape context [21], geometric blur [22], or any such discriminative local descriptors as may be suitable for the objects under consideration. In our current implementation, we use the Derivative of Gaussian interest point detector, and SIFT features as our local descriptors.

2.2 Correspondences

Having extracted features from all images, correspondences between these local parts are to be identified across images. For a given pair of images, potential correspondences are identified by finding k nearest neighbors of each feature point from one image in the other image. We use Euclidean distance between the SIFT descriptors to determine the nearest neighbors. The geometric consistency between every pair of correspondences is computed to build a geometric consistency adjacency matrix.

Suppose we wish to compute the geometric consistency between a pair of correspondences shown in Fig. 2 involving interest regions *a* and *b* in $image_1$ and *A* and *B* in $image_2$. All interest regions have a scale and orientation associated with them. Let ϕ_a be the similarity transform that transforms *a* to *A*. β_A is the transformed b_a, the relative location of *b* with respect to *a* in $image_1$, using ϕ_a. β is thus the estimated location of *B* in the $image_2$ based on ϕ_a. If *a* and *A*, as well as *b* and *B* are geometrically consistent, the distance between β and *B*, $d(B, \beta)$ would be small. A score that decreases exponentially with increasing $d(B, \beta)$ is used to quantify the geometric consistency of the pair of correspondences. To make the score symmetric, *a* is similarly mapped to α using the transform ϕ_b that maps *b* to *B*, and the score is based on $max(d(B, \beta), d(A, \alpha))$. This metric provides us with invariance only to scale and rotation, the assumption being that the distortion due to affine transformation in realistic scenarios is minimal among local features that are closely located on the same object.

Having computed the geometric consistency score between all possible pairs of correspondences, a spectral technique is applied to the geometric consistency adjacency matrix to retain only the geometrically consistent correspondences [23]. This helps eliminate most of the background clutter. This also enables us to deal with incorrect low-level correspondences among the SIFT features that can not be reliably matched, for instance at various corners and edges found in an office setting. To deal with multiple objects in the scene, an iterative form of [23] is used. However, it should be noted that due to noise, affine and perspective transformations of objects, etc. correspondences of all parts even on a single object do not always form one strong cluster and hence are not entirely obtained in a single iteration, instead they are obtained over several iterations.

2.3 Foreground Identification

Only the feature points that find geometrically consistent correspondences in most other images are retained. This is in accordance to our perception that the objects of interest are those that occur frequently across the image collection. Also, this post processing step helps to eliminate the remaining background features that may have found geometrically consistent correspondences in another image by chance. Using multiple images gives us the power to be able to eliminate these random errors which would not be consistent across images. However, we do not require features to be present in all images either in order to be retained. This allows us to handle occlusions, severe view point changes, etc. Since these affect different parts of the objects across images, it is unlikely that a significant portion of the object will not be matched in many images, and hence be eliminated by this step. Also, this enables us to deal with different number of objects in the scene across images, again, the assumption being that the objects that are present in most images are the objects of interest (foreground), while those that are present in a few images are part of the background clutter. This proportion can be varied to suit the scenario at hand.

We now have a reliable set of *foreground* feature points and a set of correspondences among all images. An illustration can be seen in Fig. 3 where only a subset of the detected features and their correspondences are retained. It should be noted that the approach being unsupervised, there is no notion of an object yet. We only have a cloud of features in each image which have all been identified as foreground and

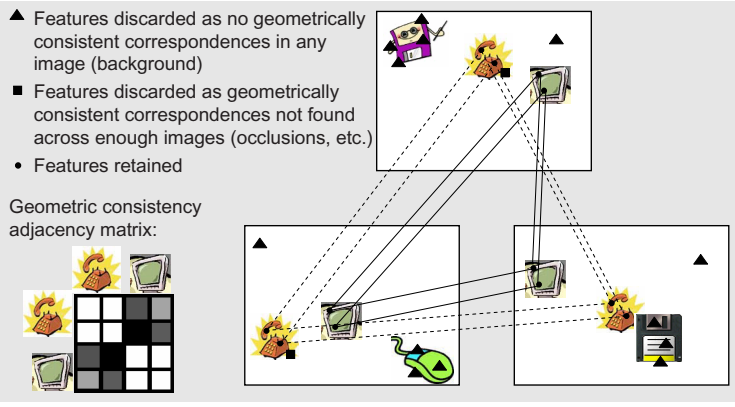


Fig. 3. An illustration of the correspondences and features retained during feature selection. The images contain two foreground objects, and some background. An illustration of the geometric consistency adjacency matrix of the graph that would be built for this set up is also shown.

correspondences among them. The goal is to now separate these features into different groups, where each group corresponds to a foreground object in the scene.

2.4 Interaction Between Pairs of Features

In order to separate the cloud of retained feature points into clusters, a graph is built over the feature points, where the weights on the edge between the nodes represents the interaction between the pair of features across the images. The metric used to capture the interaction between the pairs of features is the same geometric consistency as computed in Section 2.2, except now averaged across all pairs of images that contain these features. While the geometric consistency could contain errors for a particular pair of images due to errors in correspondences, etc. averaging across all pairs suppress the contribution of these erroneous matchings and amplifies the true interaction among the pairs of features.

If the geometric consistency between two feature points is high, they are likely to belong to the same rigid object. On the other hand, features that belong to different objects would be geometrically inconsistent because the different objects are likely to be found in different configurations across images. An illustration of the geometric consistency adjacency matrix can be seen in Fig. 3. Again, there is no concept of an object yet. The features in Fig. 3 are arranged in an order that correspond to the objects, and each object is shown to have only two features, only for illustration purposes.

2.5 Recursive Clustering

Having built the graph capturing the interaction between all pairs of features across images, recursive clustering is performed on this graph. At each step, the graph is clustered into two clusters. The properties of each cluster are analyzed, and one or both of the clusters are further separated into two clusters, and so on. If the variance in the

(a) (b)

Fig. 4. (a) A subset of the synthetic images used as input to dISCOVER (b) Background suppressed for visualization purposes

adjacency matrix corresponding to a certain cluster (subgraph) is very low but with a high mean, it is assumed to contain parts from a single object, and is hence not divided further. Since the statistics of each of the clusters formed are analyzed to determine if it should be further clustered or not, the number of foreground objects need not be known *a priori*. This is an advantage as compared to plSA or parametric methods such as fitting a mixture Guassians to the foreground features spatial distribution. dISCOVER is non-parametric. We use normalized cuts [24] to perform the clustering. The code provided at [25] was used.

3 Results

3.1 Synthetic Images

dISCOVER uses two aspects: popularity and geometric consistency. These can be loosely thought of as first order as well as second order statistics. In the first set of experiments, we use synthetic images to demonstrate the inadequacy of any of these alone.

To illustrate our point - we consider 50×50 synthetic images as shown in Fig. 4(a). The images contain 2500 distinct intensity values, of which 128, randomly selected from the 2500, always lie on the foreground objects and the rest is background. We consider each pixel in the image to be an interest point, and the descriptor of each pixel is the intensity value of the pixel. To make visualization clearer, we display only the foreground pixels of these images in Fig. 4(b). It is evident from these that there are two foreground objects of interest. We assume that the objects undergo pure translation only.

We now demonstrate the use of pLSA, as an example of an unsupervised popularity based foreground identification algorithm, on 50 such images. Since pLSA requires negative images without the foreground objects we also input 50 random negative images to pLSA, which dISCOVER does not need. If we specify pLSA to discover 2 topics, the result obtained is shown in Fig 5. It can be seen that it can identify the foreground from the background, but is unable to further separate the foreground into multiple objects. One may argue that we could further process these results and fit a mixture of Gaussians (for instance) to further separate the foreground into multiple objects. However this would require us to know the number of foreground objects *a priori* and also the distribution of features on the objects need not be Gaussian as in these images. If we specify pLSA to discover 3 topics instead, with the hope that it might separate the foreground into 2 objects, we find that it randomly splits the background into 2 topics, while still maintaining a single foreground topic, as seen in Fig. 5. This is because pLSA simply incorporates occurrence (popularity) and no spatial information.

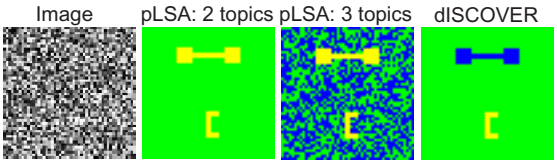


Fig. 5. Comparison of results obtained using pLSA with those obtained using dISCOVER

Hence, pLSA is inherently missing the information required to perceive the features on one of the foreground objects any different than those on the second object and hence separate them.

On the other hand, dISCOVER does incorporate this spatial/geometric information and hence can separate the foreground objects. Since the input images are assumed to allow only translation of the foreground objects and the descriptor is simply the intensity value, we alter the notion of geometric consistency than that described in Section 2.2. In order to compute the geometric consistency between a pair of correspondences, we compute the distance between the pairs of features in both images. The geometric consistency decreases exponentially as the discrepancy in the distances increases. The result obtained by dISCOVER is shown in Fig. 5. We successfully identify the foreground from the background and further separate the foreground into multiple objects. Also, dISCOVER does not require any parameters to be specified, such as number of topics or foreground objects in the images. The inability of a popularity based approach to obtain the desired results illustrates the need for geometric consistency in addition to popularity.

In order to illustrate the need for considering popularity and not just geometric consistency, let us consider the following analysis. If we consider all pairs of images such as those shown in Fig. 4 and keep all features that find correspondences that are geometrically consistent with at least one other feature in atleast one other image, we would retain approximately 2300 of the background features. This is because even for background, it is possible to find at least some geometrically consistent correspondences. However the background being random, this would not be consistent across several images. Hence, instead of retaining features that have geometrically consistent correspondences in one other image, if we now retain only those that have geometrically consistent correspondences in at least two other images, only about 50 of the background features are retained. As we use more images, we can eliminate the background features entirely. dISCOVER being an unsupervised approach, the use of multiple images to prune out background clutter is crucial. Hence, this demonstrates the need for considering popularity in addition to geometric consistency.

3.2 Real Images

In the following experiments with real images, while we present results on specific objects, it is important to note that the recent advances in object recognition that deal with object categories complement the proposed work. Since any particular features are not an integral part of dISCOVER, it can be applied to object categories by using appropriate features. However, the focus of our work is to identify the multiple objects

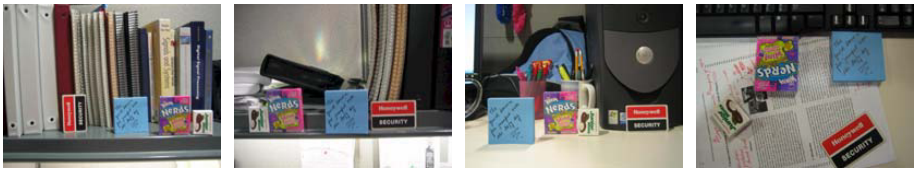
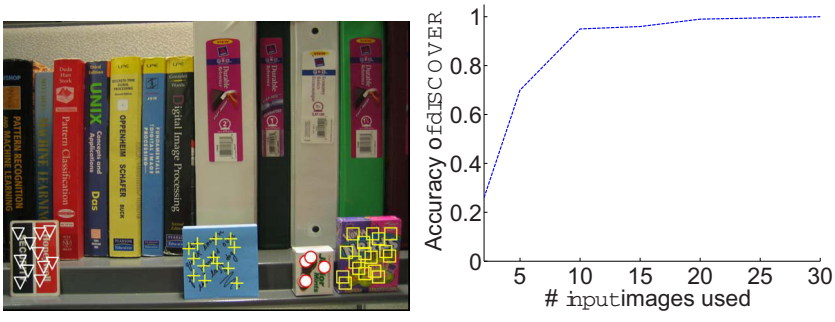


Fig. 6. A subset of images provided as input to dISCOVER



(a) (b)

Fig. 7. (a) Visual results obtained by dISCOVER. The cloud of features retained as foreground and further clustered into groups. Each group corresponds to an object in the foreground. (b) Quantitative results obtained using dISCOVER.

of interest in the scene, and not object categorization. Hence, to illustrate our algorithm, we show results on specific objects (however with considerable variations) using SIFT.

We first illustrate dISCOVER on a collection of 30 real images as shown in Fig. 6. Note the variation in orientation, scale and view-point of objects as well as in lighting conditions along with the highly cluttered backgrounds. We now use the descriptors as well as geometric consistency notions as described in our approach in Section 2. The results obtained are shown in Fig. 7(a). All background features have been successfully eliminated and the foreground features have been accurately clustered into multiple objects. In order to quantify the results obtained, we hand labeled the images with the foreground objects. This being a staged scenario where the objects were intentionally placed, the ground truth foreground objects of interest were known and hence such an analysis is possible. The portion of features that were assigned to their appropriate cluster in the foreground was computed as the accuracy of dISCOVER. The accuracy is shown in Fig. 7(b) with varying number of images used as input. It can be seen that while we need multiple images for accurate unsupervised multiple-object foreground identification, our accuracy reaches its optimum with a fairly small number of images.

Let us now consider a real scene where the objects are not staged. Consider a collection of 30 images such as those shown in Fig. 8. These are images of an office taken at different times. Note the change in view-points, scale of objects and varying lighting conditions. We run dISCOVER on these images, and the result obtained is as shown in Fig. 9. The monitor, keyboard and CPU are identified to be the foreground objects.

Fig. 8. A subset of images provided as input to dISCOVER

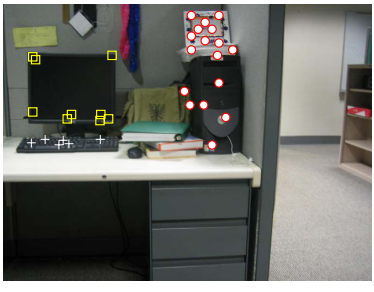


Fig. 9. Results obtained by dISCOVER. The cloud of features retained as foreground and further clustered into groups. Each group corresponds to an object in the foreground.

This seems reasonable. The mouse is not identified to be the foreground object because very few features were detected on the mouse, which were not stable across images mainly due to the lighting variation and pose changes. The photo frame and the CPU are clustered together. This is because these objects are stationary in all the input images and hence are found at identical locations with respect to each other (whenever present) across images, and are hence perceived to be one object. This is an artifact of dISCOVER being an unsupervised algorithm. Also, the bag next to the CPU is not retained. This is because the bag is occluded in most images, and hence is considered to be background. Overall, the foreground is successfully separated from the background, and is further clustered into the different objects of interest it contains.

4 Conclusion

We propose dISCOVER, which, given a collection of images of a scene, identifies the foreground and further separates the foreground into the multiple objects of interest it contains - all in an unsupervised manner. It relies on occurrence based popularity cues as well as geometry based consistency cues to achieve this. Future work includes loosening the geometric consistency notion to deal with non-rigid objects, learning models for the identified objects of interest for detection and studying interactions among the multiple objects in the scene to provide context for robust object detection.

Acknowledgments

We thank Andrew Stein and Dhruv Batra for code to compute geometrically compatible correspondences among images.

References

1. Fitzgibbon, A., Zisserman, A.: On affine invariant clustering and automatic cast listing in movies. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, Springer, Heidelberg (2002)
2. Sivic, J., Zisserman, A.: Video data mining using configurations of viewpoint invariant regions. CVPR (2004)
3. Berg, T., Berg, A., Edwards, J., White, R., Teh, Y., Learned-Miller, E., Forsyth, D.: Names and faces in the news. CVPR (2004)
4. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning (2001)
5. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research (2003)
6. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. CVPR (2005)
7. Quelhas, P., Monay, F., Odobez, J., Gatica, D., Tuytelaars, T., Van Gool, L.: Modeling scenes with local descriptors and latent aspects. ICCV (2005)
8. Sivic, J., Russell, B., Efros, A., Zisserman, A., Freeman, W.: Discovering objects and their location in images. ICCV (2005)
9. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, Springer, Heidelberg (2004)
10. Leordeanu, M., Collins, M.: Unsupervised learning of object models from video sequences. CVPR (2005)
11. Liu, D., Chen, T.: Semantic-shift for unsupervised object detection. In: CVPR. Workshop on Beyond Patches (2006)
12. Li, Y., Wang, W., Gao, W.: A robust approach for object recognition. In: PCM (2006)
13. Fergus, R., FeiFei, L., Perona, P., Zisserman, A.: Learning object categories from Googles image search. In: ICCV (2005)
14. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. CVPR (2006)
15. Russell, B., Efros, A., Sivic, J., Freeman, W., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. CVPR (2006)
16. Marszałek, M., Schmid, C.: Spatial weighting for bag-of-features. CVPR (2006)
17. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, Springer, Heidelberg (2004)
18. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. CVPR (2003)
19. Weber, M., Welling, M., Perona, P.: Unsupervised learning of models for recognition. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1842, Springer, Heidelberg (2000)
20. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
21. Belongie, S., Malik, J., Puzicha, J.: Shape context: a new descriptor for shape matching and object recognition. In: NIPS (2000)
22. Berg, A., Malik, J.: Geometric blur for template matching. CVPR (2001)
23. Leordeanu, M., Hebert, M.: A spectral technique for correspondence problems using pairwise constraints. In: ICCV (2005)
24. Shi, J., Malik, J.: Normalized cuts and image segmentation. In: PAMI (2000)
25. Shi, J.: `http://www.cis.upenn.edu/jshi/software/`