Object Co-skeletonization with Co-segmentation

Koteswar Rao Jerripothula[1,2], Jianfei Cai[2], Jiangbo Lu[3,4] and Junsong Yuan[2]
[1]Graphic Era University, India [2]Nanyang Technological University, Singapore
[3]Advanced Digital Sciences Center, Singapore [4]Shenzhen Cloudream Technology, China
krjimp@geu.ac.in, {asjfcai,jsyuan}@ntu.edu.sg, jiangbo.lu@gmail.com

Abstract

Recent advances in the joint processing of images have certainly shown its advantages over the individual processing. Different from the existing works geared towards co-segmentation or co-localization, in this paper, we explore a new joint processing topic: co-skeletonization, which is defined as joint skeleton extraction of common objects in a set of semantically similar images. Object skeletonization in real world images is a challenging problem, because there is no prior knowledge of the object's shape if we consider only a single image. This motivates us to resort to the idea of object co-skeletonization hoping that the commonness prior existing across the similar images may help, just as it does for other joint processing problems such as co-segmentation. Noting that skeleton can provide good scribbles for segmentation, and skeletonization, in turn, needs good segmentation, we propose a coupled framework for co-skeletonization and co-segmentation tasks so that they are well informed by each other, and benefit each other synergistically. Since it is a new problem, we also construct a benchmark dataset for the co-skeletonization task. Extensive experiments demonstrate that proposed method achieves very competitive results.

1. Introduction

Our main objective in this paper is to exploit joint processing [30, 13, 6] to extract objects' skeletons in images of the same category. We call it *object co-skeletonization*. By objects, we mean something which interests the image viewer more compared to the stuff like sky, roads, mountains, sea, etc, in its presence. Automatic skeletonization of such objects has many applications such as image search, image synthesis, generating training data for object detectors, etc. However, it is difficult to solve this problem as a standalone task, because it requires objects shape information as well. Existing methods either need pre-segmentation [3, 21] of the object in the image or groundtruth skeletons for the training images to learn [25, 20] to perform skeletonization on test images. The recent deep learning based method [22] requires not only the skeleton location information but also the skeleton scale information that accounts for shape information. The skeleton scale is basically the distance between a skeleton point and the nearest boundary point of the object.

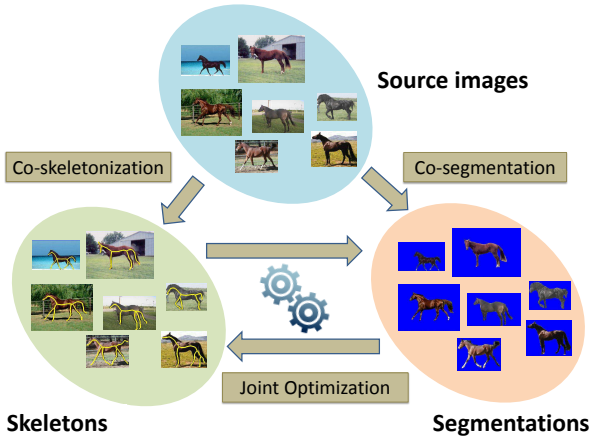

Figure 1. Object co-skeletonization with co-segmentation. Skeletons are in yellow.

In contrast, in this paper we consider the skeletonization problem with weak supervision, i.e. co-skeletonization, which does not need pre-segmentation or groundtruth skeletons of training images. Particularly, we leverage the existing idea of object co-segmentation to help co-skeletonization. It turns out that co-skeletonization can also help co-segmentation in return by providing good scribbles. In this way, both co-skeletonization and co-segmentation benefit each other synergistically. We couple these two tasks to achieve what we call "*Object Co-skeletonization with Co-segmentation*" as shown in Fig. 1.

There are several challenges involved in performing co-skeletonization and the coupling with co-segmentation. First, existing skeletonization algorithms [21, 17, 3, 19] can yield a good skeleton if a good and smooth shape is provided, but they are quite sensitive to the given shape,

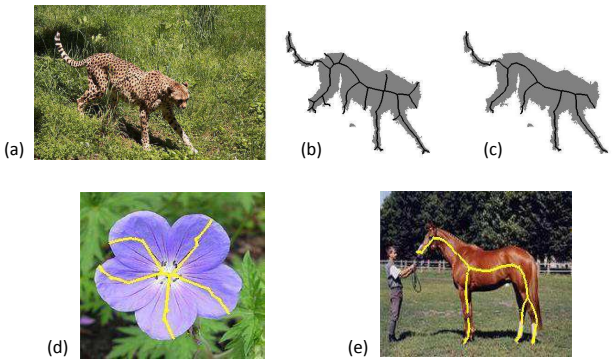

Figure 2. Example challenges of co-skeletonization. The quality of segmentation affects the quality of skeletonization. (b) The result of [21] for (a). (c) Our result. Skeletons lie on homogeneous regions, such as in (d) and (e), which are difficult to be detected and described.

as shown for the image in Fig. 2(a) which has unsmooth segmentation. The skeleton produced by [21] in Fig. 2(a) has too many unnecessary branches, while a more desirable skeleton to represent the cheetah would be the one obtained by our method in Fig. 2(c). Thus, the quality of the provided shape becomes crucial, which is challenging for the conventional co-segmentation methods because their complex way of co-labeling many images may not provide good and smooth shapes. Second, joint processing of skeletons across multiple images is quite tricky. Because most of the skeleton points generally lie on homogeneous regions as shown in Fig. 2(d) and (e), it is not easy to detect and describe them for the purpose of matching. Third, how to couple the two tasks so that they can synergistically assist each other is another challenge.

Our key observation is that we can exploit the inherent interdependencies of two tasks to achieve better results jointly. For example, in Fig. 3, although the initial co-segmentation produces a poor result, most of the skeleton pixels still remain on the horse, which gradually improve the segmentation by providing good seeds for segmentation in the subsequent iterations of joint processing. In turn, co-skeletonization also becomes better as the co-segmentation improves. Our another observation is that we can exploit the structure-preserving quality of dense correspondence to overcome the skeleton matching problem.

To the best of our knowledge, there is only one dataset where co-skeletonization could be performed in a weakly supervised manner, i.e. WH-SYMMAX dataset [20], and it only contains horse images. To extensively evaluate co-skeletonization, we construct a new benchmark dataset called CO-SKEL dataset, which consists of images ranging from animals, birds, flowers to humans with total 26 categories. Extensive experiments show that our approach

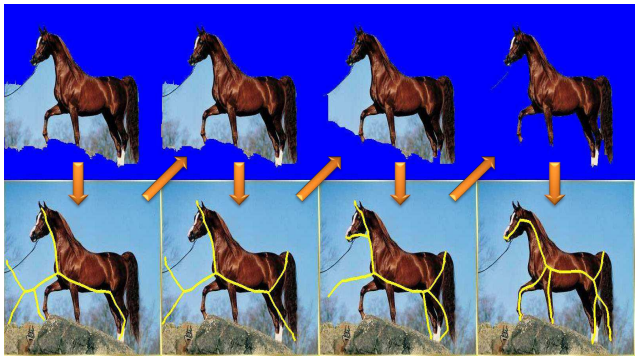

Figure 3. Inherent interdependencies of co-skeletonization and co-segmentation can be exploited to achieve better results through a coupled iterative optimization process.

achieves state-of-the-art co-skeletonization performance in the weakly supervised setting.

2. Related Work

Skeletonization: The research on skeletonization can be divided into three categories. First, there are some algorithms [17, 3, 19] which can perform skeletonization if the segmentation of an object is given. Generally, these algorithms are quite sensitive to the distortions of the given shape. However, this problem can be tackled through recent methods such as [21]. Second, there are also some traditional image processing methods [28, 29, 11] which can generate skeletons by exploiting gradient intensity maps. They generate skeletons even for stuffs like sky, sea, etc, which usually need some object prior to be suppressed. Third, there are also some supervised learning based methods which require groundtruth skeletons of training images for learning. This class of methods includes both traditional machine learning based methods [25, 20] and the recent deep learning based methods [27, 22]. The performance of the traditional machine learning based methods is not satisfactory due to the limited feature learning capability in homogeneous regions. On the other hand, the recent deep learning based methods have made great progress in the skeletonization process as reported in [22] at the cost of requiring complex training process on a substantial amount of annotated data. In contrast, our method is a weakly supervised one, although it can utilize the annotated data as well, if available.

Segmentation: Image segmentation is a classical problem, and there are many types of approaches like interactive segmentation [15, 24], image co-segmentation [4, 7, 5], semantic segmentation [18], etc. While interactive segmentation needs human efforts, image co-segmentation exploits weak supervision in the form of requiring the association of same category images and uses an inter-image prior to help segment each individual image. Semantic image segmenta-

tion not only segments objects but also provides a label for each pixel. In the past few years, deep learning based methods such as fully convolution networks (FCN) have greatly advanced the performance of semantic image segmentation. Recently, [10] proposed a joint framework to combine interactive segmentation with FCN based semantic segmentation [18] so as to help each other. In a similar spirit, in this work, we propose coupling of co-skeletonization and co-segmentation to assist each other.

3. Proposed Method

In this section, we discuss our joint framework of co-skeletonization and co-segmentation in detail.

3.1. Overview of Our Approach

Given a set of m similar images belonging to the same category, denoted by $\mathcal{I} = \{I_1, I_2, \cdots, I_m\}$, we aim to provide two output sets: $\mathcal{K} = \{K_1, K_2, \cdots, K_m\}$ and $\mathcal{O} = \{O_1, O_2, \cdots, O_m\}$, comprising skeleton masks and segmentation masks, respectively, where $K_i(p), O_i(p) \in \{0, 1\}$ indicating whether a pixel p is a skeleton pixel ($K_i(p) = 1$) and whether it is a foreground pixel ($O_i(p) = 1$).

Our overall objective function for an image I_i is defined as

$$\min_{K_i, O_i} \lambda\psi_{pr}(K_i, O_i|\mathcal{N}_i) + \psi_{in}(K_i, O_i|I_i) + \psi_{sm}(K_i, O_i|I_i) \quad s.t.\ K_i \subseteq \mathbf{ma}(O_i) \tag{1}$$

where the first term ψ_{pr} accounts for the priors from the set of neighbor images denoted as $\mathcal{N}_i$, the second term ψ_{in} is to enforce the interdependence between the skeleton K_i and the shape / segmentation O_i in image I_i, the third term ψ_{sm} is the smoothness term to enforce smoothness, and λ is a parameter to control the influence of the inter-image prior term. The constraint in (1) means the skeleton must be a subset of medial axis ($\mathbf{ma}$) [3] of the shape.

We resort to the typical alternative optimization strategy to solve (1), i.e., dividing (1) into two sub-problems and solve them iteratively. In particular, one sub-problem is as follows. Given the shape O_i, we solve co-skeletonization by

$$\min_{K_i} \lambda\psi^k_{pr}(K_i|\mathcal{N}_i) + \psi^k_{in}(K_i|O_i) + \psi^k_{sm}(K_i) \quad s.t.\ K_i \subseteq \mathbf{ma}(O_i). \tag{2}$$

The other sub-problem is that given the skeleton K_i, we solve co-segmentation by

$$\min_{O_i} \lambda\psi^o_{pr}(O_i|\mathcal{N}_i) + \psi^o_{in}(O_i|K_i, I_i) + \psi^o_{sm}(O_i|I_i). \tag{3}$$

If we treat both the inter-image prior term ψ^k_{pr} and the shape prior term ψ^k_{in} as a combined prior, (2) turns out to be a skeleton pruning problem and can be solved using the approach similar to [21], where branches in the skeleton are iteratively removed as long as it reduces the energy. Similarly, if we combine both the inter-image prior ψ^o_{pr} and the skeleton prior ψ^o_{in} as the data term, (3) become a standard MRF-based segmentation formulation, which can be solved using GrabCut [15]. Thus, compared with the existing works, the key differences of our formulation lie in the designed inter-image prior terms as well as the interdependence terms, which link the co-skeletonization and co-segmentation together.

Iteratively solving (2) and (3) requires a good initialization. We propose to initialize $\mathcal{O}$ by Otsu thresholded saliency maps and $\mathcal{K}$ by the medial axis mask [3]. Alg. 1 summarizes our approach, where $(\psi_{pr} + \psi_{in} + \psi_{sm})^{(t)}$ denotes the objective function value of (1) at the t^{th} iteration and $\psi_{pr} = \psi^k_{pr} + \psi^o_{pr}$, $\psi_{in} = \psi^k_{in} + \psi^o_{in}$, $\psi_{sm} = \psi^k_{sm} + \psi^o_{sm}$.

Algorithm 1: Our approach for solving (1)

Data: An image set $\mathcal{I}$ containg images of the same category

Result: Sets $\mathcal{O}$ and $\mathcal{K}$ containing segmentations and skeletons of images in $\mathcal{I}$

Initialization: $\forall I_i \in \mathcal{I}$, $O_i^{(0)}$ = Otsu thresholded saliency map and $K_i^{(0)} = \mathbf{ma}(O_i^{(0)})$;

Process: $\forall I_i \in \mathcal{I}$,

do

1) Obtain $O_i^{(t+1)}$ by solving (3) using [15] with $\mathcal{O}^{(t)}$ and $K_i^{(t)}$.

2) Obtain $K_i^{(t+1)}$ by solving (2) using [21] with $\mathcal{K}^{(t)}$ and $O_i^{(t+1)}$, $s.t.\ K_i^{(t+1)} \in \mathbf{ma}(O_i^{(t+1)})$.

while $(\lambda\psi_{pr} + \psi_{in} + \psi_{sm})^{(t+1)} \leq (\lambda\psi_{pr} + \psi_{in} + \psi_{sm})^{(t)}$;

$\mathcal{O} \leftarrow \mathcal{O}^{(t)}$ and $\mathcal{K} \leftarrow \mathcal{K}^{(t)}$

3.2. Object Co-skeletonization

As shown in Alg. 1, the step of object co-skeletonization is to obtain $K^{(t+1)}$ by minimizing (2), given the shape $O^{(t+1)}$ and the previous skeleton set $\mathcal{K}^t$. Considering the constraint of $K_i^{(t+1)} \in \mathbf{ma}(O_i^{(t+1)})$, we only need to search skeleton pixels from the medial axis pixels. We build up our solution based on [21], but with our carefully designed individual terms for (2) as explained below.

Prior Term (ψ^k_{pr}): In the object co-skeletonization, a good skeleton pixel will be the one which is repetitive across images. To account for this repetitiveness, we need to find corresponding skeleton pixels in other images. However, skeleton pixels usually lie on homogeneous regions (see Fig. 2(d)&(e)) and are thus difficult to match. Thus, instead of trying to match sparse skeleton pixels, we make

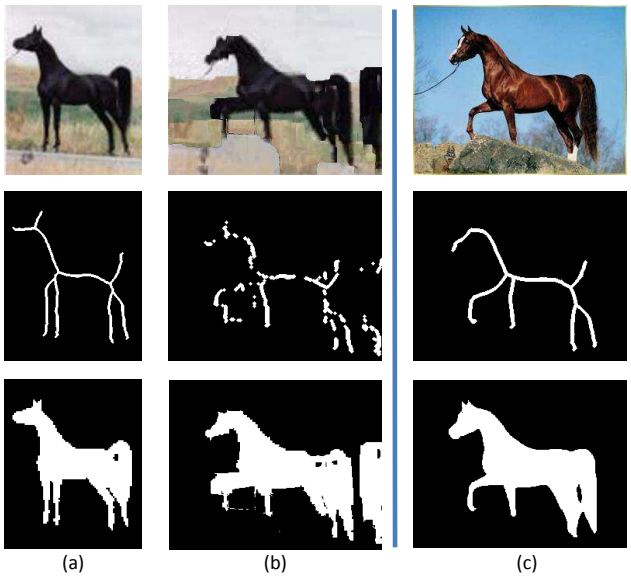

Figure 4. Dense correspondences preserve the skeleton and segmentation structures roughly. Here (a) is warped to generate (b) to be used as a prior for (c).

use of dense correspondences using SIFT Flow [12], which preserve the skeleton and segmentation structures well, as shown in Fig. 4.

Once correspondence is established, we utilize the warped skeleton pixels from neighboring images to develop the prior term. Particularly, we align all the neighboring images' t^{th} iteration's skeleton maps to the concerned image I_i, and generate a co-skeleton prior at the $(t+1)^{th}$ iteration as

$$\widetilde{K}_i^{(t+1)} = \frac{K_i^{(t)} + \sum_{I_j \in \mathcal{N}_i} \mathbf{W}_j^i(K_j^{(t)})}{|\mathcal{N}_i| + 1} \tag{4}$$

where we align other skeleton maps using a warping function $\mathbf{W}_j^i$ [12] and then average them with I_i's own skeleton map. Note that the neighborhood $\mathcal{N}_i$ is developed simply based on the GIST distance [14]. For simplicity, we drop the superscriptions such as $(t+1)$ in all the following derivations.

Considering that the corresponding skeleton pixels from other images may not exactly align with the skeleton pixels of the considered image, we define our inter-image prior term as

$$\psi_{pr}^k(K_i|\mathcal{N}_i) = \sum_{p \in \mathbf{ma}(O_i)} -K_i(p) \log\Big(1 + \sum_{q \in \mathbb{N}(p)} \widetilde{K}_i(q)\Big). \tag{5}$$

(5) essentially measures the consistency among image I_i's own skeleton mask and the recommended skeleton mask from its neighbor images. Note that we accumulate the co-skeleton prior scores in a certain neighborhood $\mathbb{N}(p)$ for each pixel p to account for the rough skeleton alignment across the images.

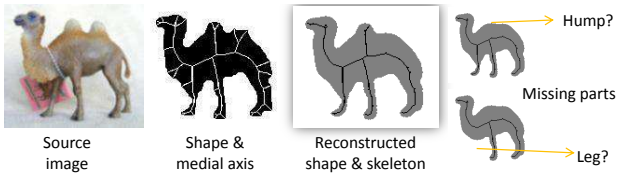

Figure 5. Shape reconstruction from skeleton. Compared to the reconstructed shape from the medial axis (2nd column), the reconstructed shape (3rd column) from our simplified skeleton is simpler and smoother while still preserving the main structure. Nevertheless, we do not want an over-simplified skeleton, which will result in missing important parts in the corresponding shape reconstruction (4th column).

Interdependence Term (ψ_{in}^k): Our interdependence term is similar to the traditional data term in skeleton pruning, i.e., it enforces that the skeleton should provide a good reconstruction of the given shape, which medial axis already does well. However, a medial axis often contains spurious branches, while the noisy shapes obtained from imperfect co-segmentation only make this worse. To avoid spurious branches, we prefer a simplified skeleton, whose reconstructed shape is expected to be smooth while still preserving the main structure of the given shape (see Fig. 5 for example). On the other hand, we do not want an over-simplified skeleton, whose reconstructed shape is likely to miss some important parts (see the 4th column of Fig. 5).

Therefore, we expect the reconstructed shape from the skeleton to match the given shape, but not necessary to be exactly the same as the given shape. In this spirit, we define our interdependence term ψ_{in}^k as

$$\psi_{in}^k(K_i|O_i) = -\alpha \log \frac{|\mathbf{R}(K_i, O_i) \cap O_i|}{|\mathbf{R}(K_i, O_i) \cup O_i|} \tag{6}$$

where we use IoU to measure the closeness between the reconstructed shape $\mathbf{R}(K_i, O_i)$ and the given shape O_i, and α is the normalization factor as defined in [21]. The reconstructed shape $\mathbf{R}(K_i, O_i)$ is basically the union of maximal disks at skeleton pixels [21], i.e.,

$$\mathbf{R}(K_i, O_i) = \bigcup_{p \in \mathbf{ma}(O_i)} d(p, O_i) \tag{7}$$

where $d(p, O_i)$ denotes the maximal disk at skeleton pixel p for the given O_i, and the maximal disk is the disk that exactly fits within O_i with skeleton pixel p as the center.

Smoothness Term (ψ_{sm}^k): To ensure a smoother and simpler skeleton, we aim for a skeleton whose: (i) branches are less in number and (ii) branches are long. Our criteria discourage skeletons with spurious branches while at the same time encouraging skeletons with structure-defining

branches. This is different from the criteria in [21] which only aims for less number of skeleton pixels. Specifically, we define the smoothness term ψ^k_{sm} as

$$\psi^k_{sm}(K_i) = |\mathbf{b}(K_i)| \times \sum_{u=1}^{|\mathbf{b}(K_i)|} \frac{1}{length\Big(b_u(K_i)\Big)} \tag{8}$$

where $\mathbf{b}(K_i) = \{b_1(K_i), \cdots, b_{|\mathbf{b}(K_i)|}(K_i)\}$ denotes the set of branches of the skeleton K_i. In this way, we punish skeletons with either large number of branches or short-length branches.

3.3. Object Co-segmentation

The object co-segmentation problem here is as follows. Given the skeleton K_i, find the optimal O_i that minimizes the objective function defined in (3). The individual terms in (3) are defined in the following manner.

Prior Term (ψ^o_{pr}): We generate an inter-image co-segment prior, similar to that for co-skeletonization. In particular, we align segmentation masks of neighboring images and fuse them with that of the concerned image, i.e.,

$$\widetilde{O}_i = \frac{O_i + \sum_{I_j \in \mathcal{N}_i} \mathbf{W}^i_j(O_j)}{|\mathcal{N}_i| + 1} \tag{9}$$

where $\mathbf{W}^i_j$ is the same warping function from image j to image i. Then, with the help of $\widetilde{O}_i$, we define our inter-image prior term as

$$\psi^o_{pr}(O_i|\mathcal{N}_i) = \sum_{p \in D_i} -\Bigg(O_i(p) \log\Big(\frac{1}{|\mathbb{N}(p)|} \sum_{q \in \mathbb{N}(p)} \widetilde{O}_i(q)\Big) + \Big(1 - O_i(p)\Big) \log\Big(1 - \frac{1}{|\mathbb{N}(p)|} \sum_{q \in \mathbb{N}(p)} \widetilde{O}_i(q)\Big)\Bigg) \tag{10}$$

which encourages the shape to be consistent with $\widetilde{O}_i$. Here again we account for pixel correspondence errors by neighborhood $\mathbb{N}(p)$ (in the pixel domain D_i) averaging.

Interdependence Term (ψ^o_{in}): For the co-segmentation process to benefit from co-skeletonization, our basic idea is to build up foreground and background appearance models based on the given skeleton K_i. Particularly, we use GMM for appearance models. The foreground GMM model is learned using K_i (i.e., treating skeleton pixels as foreground seeds), whereas the background GMM is learned using the background part of K_i's reconstructed shape $\mathbf{R}(K_i, O_i)$. In this manner, the appearance model is developed entirely using the skeleton. Note that at the beginning it is not robust to build up the GMM appearance models in this manner since the initial skeleton extracted based on saliency is not reliable at all. Thus, at initialization, we develop the foreground and background appearance models based on the inter-image priors $\widetilde{K}_i$ and $\widetilde{O}_i$, respectively.

Denoting $\theta(K_i, I_i)$ as the developed appearance models, we define the interdependence term ψ^o_{in} as

$$\psi^o_{in}(O_i|K_i, I_i) = \sum_{p \in D_i} -\log\Bigg(P\Big(O_i(p) \,\Big|\, \theta(K_i, I_i), I_i(p)\Big)\Bigg) \tag{11}$$

where $P\Big(O_i(p) \,|\, \theta(K_i, I_i), I_i(p)\Big)$ denotes how likely a pixel of color $I(p)$ will take the label $O_i(p)$ given $\theta(K_i, I_i)$. ψ^o_{in} is similar to the data term in the interactive segmentation method [15].

Smoothness Term (ψ^o_{sm}): For ensuring smooth foreground and background segments, we simply adopt the smoothness term of GrabCut [15], i.e.,

$$\psi^o_{sm}(O_i|I_i) = \gamma \sum_{(p,q) \in E_i} [O_i(p) \neq O_i(q)] e^{(-\beta||I_i(p) - I_i(q)||^2)} \tag{12}$$

where E_i denotes the set of neighboring pixel pairs in the image I_i, and γ and β are segmentation smoothness related parameters as discussed in [15].

3.4. Implementation Details

We use the saliency extraction method [2] for initialization of our framework in our experiments. We use the same default setting as that in [15] for the segmentation parameters γ and β in (12) throughout our experiments. For the parameters of SIFT flow [12], we follow the setting in [16] in order to handle the possible matching of different semantic objects. The parameter λ in both (2) and (3), which controls the influence of joint processing, is set to 0.1.

4. Experimental Results

4.1. Datasets and Evaluation Metrics

Datasets: There is only one publicly available dataset, i.e. WH-SYMMAX dataset [20], on which weakly supervised co-skeletonization can be performed, but it contains only the horse category of images. In order to evaluate the co-skeletonization task extensively, we develop a new benchmark dataset called the CO-SKEL dataset. It consists of 26 categories with total 353 images of animals, birds, flowers and humans. These images are collected from the MSRC dataset, CosegRep, Weizmann Horses and iCoseg datasets along with their groundtruth segmentation masks. Then, we apply [21] (with our improved terms) on these groundtruth masks, in the same manner as the WH-SYMMAX dataset has been generated from the Weizmann Horses dataset [1]. Fig. 6 shows some example images, and their skeletons using [21] and our improvement of [21][1]. It can be seen that our skeletons are much smoother and better in representing the shapes.

[1]We will make our dataset with groundtruths and code publicly available.

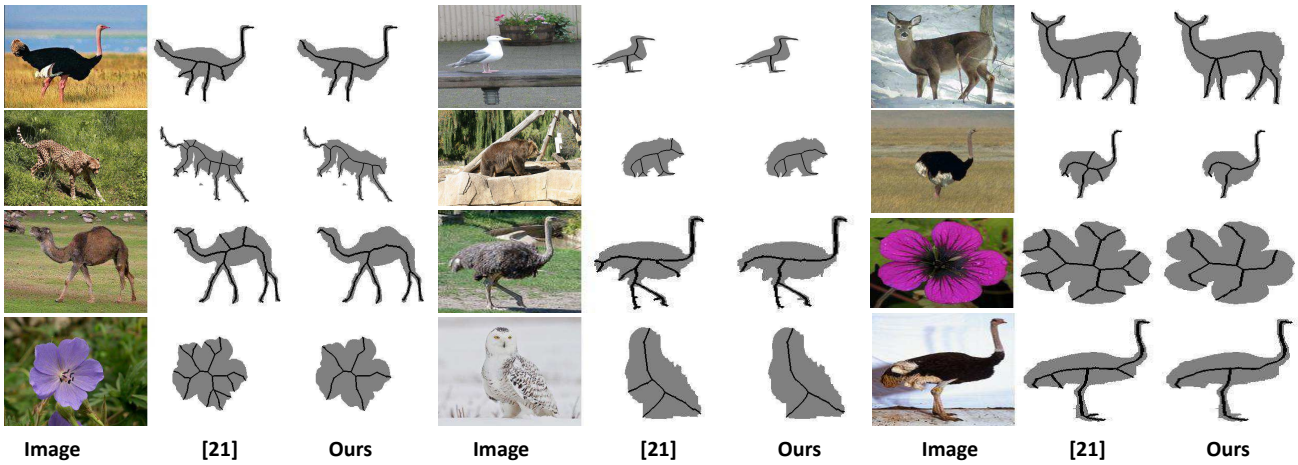

Figure 6. Given the shape, we improve skeletonization method [21] using our improved terms in their objective function. It can be seen that our skeletons are much smoother and better in representing the shape. We use these improved results as groundtruths in our CO-SKEL dataset.

Since our method searches for k-nearest neighbors first and then performs joint processing, our method can also work in an unsupervised way as long as there are a sufficient number of images of same category objects or visually similar objects. Thus, our method can also be applied to datasets like the SK506 dataset [22], which consists of many uncategorized images.

Metrics: For evaluation of skeletonization and segmentation, we calculate F-measure (including precision and recall) and Jaccard Similarity, respectively. Considering it is very difficult to get a resultant skeleton mask exactly aligned with the groundtruth, if a resultant skeleton pixel is nearby a groundtruth skeleton pixel, it should be considered as a hit. Therefore, we consider a resultant skeleton pixel as correct if it is at a distance of $d \in \{0, 1, 2, 3\}$ pixels from a groundtruth skeleton pixel, for which we denote F^d as the corresponding F-measure. Jaccard Similarity (denoted as J) is basically the IoU of groundtruth and our segmentation result.

4.2. Weakly Supervised Results

We report our overall co-skeletonization and co-segmentation results on WH-SYMMAX and our CO-SKEL datasets in Tables 1 and 2, respectively. Note that since we do not perform any kind of training, we combine both training and test images of the WH-SYMMAX dataset, and then obtain the results. It can be seen that our method greatly improves over our initialization baseline. To demonstrate the importance of considering the interdependence between co-segmentation and co-skeletonization, we also compare the proposed method with another baseline, Ours (w/o ψ_{in}), where we remove the interdependence, i.e., running co-segmentation first and then performing skeletonization from the resultant foreground segments.

Method	F^0	F^1	F^2	F^3	J
Ours$^{(0)}$	0.095	0.229	0.282	0.319	0.412
Ours (w/o ψ_{in})	0.168	0.337	0.391	0.434	0.649
Ours	**0.189**	**0.405**	**0.464**	**0.506**	**0.721**

Table 1. Comparisons of the co-skeletonization and co-segmentation results of our method and its two baselines on the WH-SYMMAX dataset. Ours$^{(0)}$: our initialization baseline using Otsu thresholded saliency maps [2] for segmentation and [21] for skeleton. Ours (w/o ψ_{in}): our method without the interdependence terms, i.e. running co-segmentation followed by skeletonization.

	F^0	F^1	F^2	F^3	J
Ours$^{(0)}$	0.129	0.306	0.371	0.416	0.600
Ours (w/o ψ_{in})	0.236	0.426	0.484	0.522	0.725
Ours	**0.237**	**0.435**	**0.495**	**0.535**	**0.741**

Table 2. Comparisons of the co-skeletonization and co-segmentation results of our method and its two baselines on our CO-SKEL dataset.

It can be seen that our method outperforms this baseline on both the datasets. Marginal improvement on the CO-SKEL dataset may be due to already good initialization. Specifically, it can be seen that J for initialization is already 0.600 in the CO-SKEL dataset compared to 0.412 in the WH-SYMMAX dataset, suggesting that there is relatively less room for improvement.

We also evaluate how our method performs at different iterations in Fig. 8 on the WH-SYMMAX dataset. It can be seen that our method first improves the performance swiftly and then it becomes somewhat steady. This suggests that 2-3 iterations are good enough for our method.

Figure 7. Some examples of steadily improving skeletonization and segmentation after each iteration. The top-right example shows that our model continues to reproduce similar results once the optimal shape and skeleton are obtained.

	m	F^0	F^1	F^2	F^3	J
bear	4	0.075	0.1714	0.213	0.246	0.846
iris	10	0.363	0.600	0.658	0.698	0.837
camel	10	0.224	0.353	0.395	0.432	0.674
cat	8	0.118	0.360	0.469	0.523	0.733
cheetah	10	0.078	0.221	0.287	0.335	0.735
cormorant	8	0.351	0.545	0.606	0.642	0.768
cow	28	0.142	0.437	0.580	0.669	0.789
cranesbill	7	0.315	0.619	0.670	0.696	0.935
deer	6	0.214	0.366	0.407	0.449	0.644
desertrose	15	0.360	0.662	0.721	0.759	0.934
dog	11	0.122	0.356	0.457	0.522	0.746
egret	14	0.470	0.642	0.669	0.693	0.760
firepink	6	0.416	0.685	0.756	0.805	0.918
frog	7	0.163	0.358	0.418	0.471	0.734
geranium	17	0.299	0.633	0.716	0.764	0.940
horse	31	0.217	0.435	0.490	0.529	0.726
man	20	0.144	0.246	0.274	0.295	0.385
ostrich	11	0.298	0.530	0.592	0.634	0.752
panda	15	0.037	0.102	0.140	0.174	0.696
pigeon	16	0.181	0.326	0.361	0.382	0.590
seagull	13	0.257	0.461	0.520	0.562	0.662
seastar	9	0.440	0.649	0.681	0.702	0.750
sheep	10	0.078	0.249	0.342	0.401	0.769
snowowl	10	0.089	0.222	0.268	0.306	0.543
statue	29	0.306	0.506	0.542	0.564	0.681
woman	23	0.305	0.463	0.503	0.533	0.674
variance		0.015	0.028	0.029	0.030	0.016

Table 3. Categorywise number of images and our weakly supervised results on the CO-SKEL dataset.

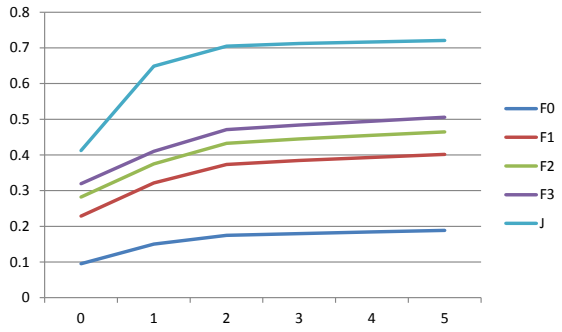

Figure 8. Performance v/s Iteration plot. It can be seen that the performance improves swiftly at first and then becomes steady.

Please refer to Fig. 7 for examples where the results improve steadily with each iteration. Fig. 9 shows some sample results of our method along with groundtruths from the WH-SYMMMAX and CO-SKEL datasets.

We also show our results on individual categories and the variance in performance across the categories of our CO-SKEL dataset in Table 3. Low variances for both F^d and J metrics suggest that our method is quite reliable.

4.3. Supervised Results

In the literature, since only the fully supervised skeletonization methods are available, for fair comparison, we follow the original process but with a change in the initialization. We replace the saliency initialization with ground truth initialization for training images. This will help develop better joint processing priors for remaining images which are the test images. We do the comparisons on test images of the WH-SYMMAX and SK506 datasets in Ta-

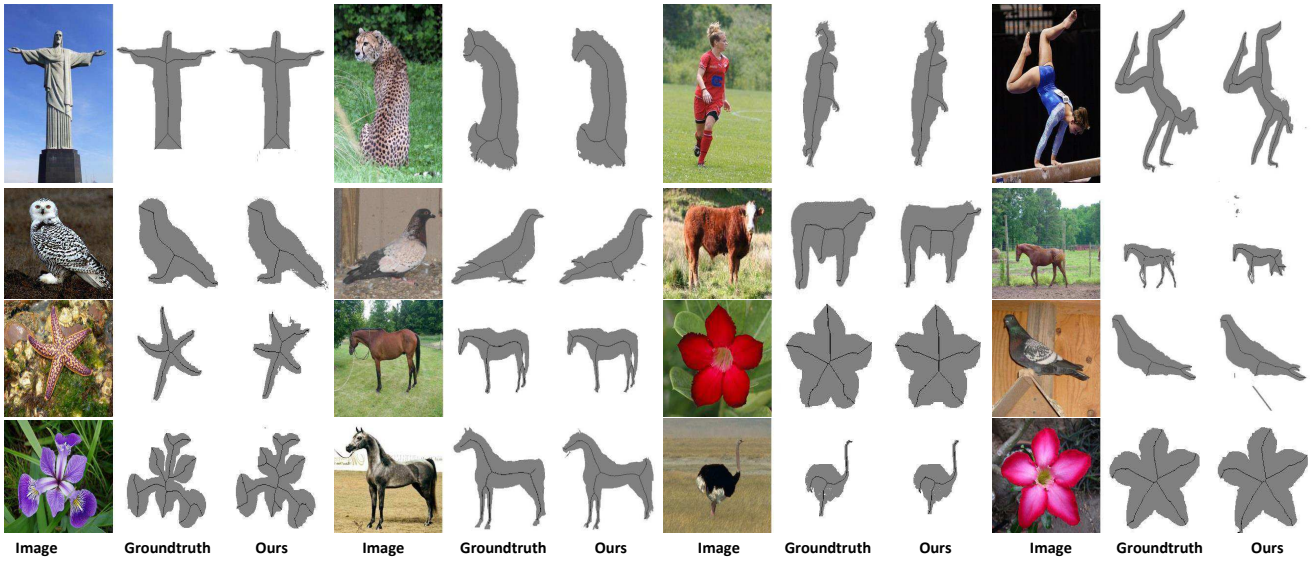

Figure 9. Sample co-skeletonization results along with our final shape masks. It can be seen that both are quite close to the groundtruths.

Methods	WH-SYMMAX	SK506
[9]	0.174	0.218
[8]	0.223	0.252
[26]	0.334	0.226
[23]	0.103	-
[25]	0.365	0.392
[29]	0.402	-
Ours$^{(0)}$	0.322	0.261
Ours	0.530	0.483
Ours (S)	**0.594**	**0.523**

Table 4. Comparisons of the results of F^d of our methods with supervised methods. Ours$^{(0)}$: our initialization baseline. Ours (S): our method with groundtruth initialization on training images. Note that here $d = 0.0075 \times \sqrt{width^2 + height^2}$ following [22].

ble 4. Note that to make the distinction between our supervised method (groundtruth initialization) and our weakly supervised method (with saliency initialization), we denote the results of our supervised approach as "Ours (S)". It can be seen that not only our supervised method comfortably outperforms all the traditional supervised methods, but also our weakly supervised (unsupervised for SK506) approach is able to do so. Note that the other performance values reported here are directly taken from [22]. We would like to point out that the recently developed deep learning based supervised method [22] reports much better performance. We did not compare with it since our method essentially is a weakly supervised approach.

4.4. Limitations

Our method has some limitations. First, for initialization, our method requires common object parts to be salient in general across the neighboring images if not in all. Therefore, it depends on the quality of the neighboring images. The second limitation lies in the difficulty during warping process. For example, when the neighboring images contain objects at different sizes or at different viewpoints, the warping processing will have difficulty in aligning the images. Such a situation will not be crucial when there is a large number of images to select from. Another issue is that smoothing the skeleton may cause missing out some important short branches.

5. Conclusion

The major contributions of this paper lie in the newly defined co-skeletonization problem and the proposed joint co-skeletonization and co-segmentation framework, which effectively exploits inherent interdependencies between the two to assist each other synerergistically. Extensive experiments demonstrate that the proposed method achieves very competitive results on a few benchmark datasets.

Acknowledgements This research is supported by the National Research Foundation, Prime Ministers Office, Singapore, under its IDM Futures Funding Initiative. It is also supported by the HCCS research grant at the ADSC[2] from Singapore's A*STAR.

[2]This work was partly done when Koteswar and Jiangbo were interning and working in ADSC

References

[1] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *European Conference on Computer Vision (ECCV)*, pages 109–122. Springer Berlin Heidelberg, 2002.

[2] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *Computer Vision and Pattern Recognition (CVPR)*, pages 409–416. IEEE, 2011.

[3] W.-P. Choi, K.-M. Lam, and W.-C. Siu. Extraction of the euclidean skeleton based on a connectivity criterion. *Pattern Recognition*, 36(3):721 – 729, 2003.

[4] J. Dai, Y. N. Wu, J. Zhou, and S.-C. Zhu. Cosegmentation and cosketch by unsupervised learning. In *International Conference on Computer Vision (ICCV)*. IEEE, 2013.

[5] K. R. Jerripothula, J. Cai, and J. Yuan. Group saliency propagation for large scale and quick image co-segmentation. In *International Conference on Image Processing (ICIP)*, pages 4639–4643. IEEE, 2015.

[6] K. R. Jerripothula, J. Cai, and J. Yuan. Cats: Co-saliency activated tracklet selection for video co-localization. In *European Conference on Computer vision (ECCV)*, pages 187–202. Springer, 2016.

[7] K. R. Jerripothula, J. Cai, and J. Yuan. Image co-segmentation via saliency co-fusion. *IEEE Transactions on Multimedia (T-MM)*, 18(9):1896–1909, Sept 2016.

[8] T. S. H. Lee, S. Fidler, and S. Dickinson. Detecting curved symmetric parts using a deformable disc model. In *International Conference on Computer Vision (ICCV)*, pages 1753–1760. IEEE, 2013.

[9] A. Levinshtein, S. Dickinson, and C. Sminchisescu. Multiscale symmetric part detection and grouping. In *International Conference on Computer Vision (ICCV)*, pages 2162–2169. IEEE, 2009.

[10] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[11] T. Lindeberg. Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):117–156, 1998.

[12] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 33(5):978–994, 2011.

[13] F. Meng, J. Cai, and H. Li. Cosegmentation of multiple image groups. *Computer Vision and Image Understanding (CVIU)*, 146:67 – 76, 2016.

[14] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision (IJCV)*, 42(3):145–175, 2001.

[15] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *Transactions on Graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.

[16] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1939–1946. IEEE, 2013.

[17] P. K. Saha, G. Borgefors, and G. S. di Baja. A survey on skeletonization algorithms and their applications. *Pattern Recognition Letters*, 76:3 – 12, 2016. Special Issue on Skeletonization and its Application.

[18] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 39(4):640–651, 2017.

[19] W. Shen, X. Bai, R. Hu, H. Wang, and L. J. Latecki. Skeleton growing and pruning with bending potential ratio. *Pattern Recognition*, 44(2):196 – 209, 2011.

[20] W. Shen, X. Bai, Z. Hu, and Z. Zhang. Multiple instance subspace learning via partial random projection tree for local reflection symmetry in natural images. *Pattern Recognition*, 52:306 – 316, 2016.

[21] W. Shen, X. Bai, X. Yang, and L. J. Latecki. Skeleton pruning as trade-off between skeleton simplicity and reconstruction error. *Science China Information Sciences*, 56(4):1–14, 2013.

[22] W. Shen, K. Zhao, Y. Jiang, Y. Wang, Z. Zhang, and X. Bai. Object skeleton extraction in natural images by fusing scale-associated deep side outputs. In *Computer Vision and Pattern Recognition (CVPR)*, pages 222–230. IEEE, 2016.

[23] A. Sironi, V. Lepetit, and P. Fua. Multiscale centerline detection by learning a scale-space distance transform. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2697–2704. IEEE, 2014.

[24] M. Tang, L. Gorelick, O. Veksler, and Y. Boykov. Grabcut in one cut. In *International Conference on Computer Vision (ICCV)*, pages 1769–1776, 2013.

[25] S. Tsogkas and I. Kokkinos. Learning-based symmetry detection in natural images. In *European Conference on Computer Vision (ECCV)*, pages 41–54. Springer Berlin Heidelberg, 2012.

[26] N. Widynski, A. Moevus, and M. Mignotte. Local symmetry detection in natural images using a particle filtering approach. *IEEE Transactions on Image Processing (T-IP)*, 23(12):5309–5322, 2014.

[27] S. Xie and Z. Tu. Holistically-nested edge detection. In *International Conference on Computer Vision (ICCV)*, pages 1395–1403. IEEE, 2015.

[28] Z. Yu and C. Bajaj. A segmentation-free approach for skeletonization of gray-scale images via anisotropic vector diffusion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 415 – 420. IEEE, 2004.

[29] Q. Zhang and I. Couloigner. Accurate centerline detection and line width estimation of thick lines using the radon transform. *IEEE Transactions on Image Processing (T-IP)*, 16(2):310–316, 2007.

[30] H. Zhu, F. Meng, J. Cai, and S. Lu. Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. *Journal of Visual Communication and Image Representation (JVCIR)*, 34:12 – 27, 2016.