Multiple Human Identification and Cosegmentation: A Human-Oriented CRF Approach with Poselets

Hongyuan Zhu, *Member, IEEE,* Jiangbo Lu, *Senior Member, IEEE,* Jianfei Cai, *Senior Member, IEEE,* Jianmin Zheng, Shijian Lu, and Nadia M. Thalmann

***Abstract*— Localizing, identifying and extracting humans with consistent appearance jointly from a personal photo stream is an important problem and has wide applications. The strong variations in foreground and background and irregularly occurring foreground humans make this realistic problem challenging. Inspired by the advance in object detection, scene understanding and image cosegmentation, in this paper we explore explicit constraints to label and segment human objects rather than other non-human objects and "stuff". We refer to such a problem as Multiple Human Identification and Cosegmentation (MHIC). To identify specific human subjects, we propose an efficient human instance detector by combining an extended color line model with a poselet-based human detector. Moreover, to capture high level human shape information, a novel soft shape cue is proposed. It is initialized by the human detector, then further enhanced through a generalized geodesic distance transform, and refined finally with a joint bilateral filter. We also propose to capture the rich feature context around each pixel by using an adaptive cross region data structure, which gives a higher discriminative power than a single pixel-based estimation. The high-level object cues from the detector and the shape are then integrated with the low-level pixel cues and mid-level contour cues into a principled conditional random field (CRF) framework, which can be efficiently solved by using fast graph cut algorithms. We evaluate our method over a newly created NTU-MHIC human dataset, which contains 351 images with manually annotated ground-truth segmentation. Both visual and quantitative results demonstrate that our method achieves state-of-the-art performance for the MHIC task.**

I. Introduction

THE popularity of digital cameras and smart phones allows people to record their daily life in visually rich way with ease. Human activity understanding [1]–[3] has become a core task to manage and exploit this large volume of photo streams which often focus on humans. Toward such high level understanding, accurate human recognition and pixel wise segmentation can add potential to many related applications. For example, action recognition can be improved by recovering human shapes and structures [1], [4]; users can also group their photos based on consistent appearance patterns of the persons-of-interest or propagate editing operations across intended human instances. Many other exciting multimedia applications can be further developed. Moreover, most photos are not captured in isolation, but they come as an album that records an event for certain moments. Therefore, this paper focuses on the problem of localizing, identifying and segmenting multiple humans jointly in a personal photo album, which we refer to as a *Multiple Human Identification and Cosegmentation* (MHIC) problem.

The MHIC problem is challenging due to its unique irregular object-occurring patterns, strong variations in foreground/background and feature sharing among different classes, which inherit from a similar *Multiple Foreground Cosegmentation* (MFC) problem [5]. It is actually very different from the classic cosegmentation task studied by many existing algorithms [6]–[13], which assume object-of-interests to appear in all input images. Although some recent works [5], [14], [15] achieve certain progress in the MFC scenario, their performance for the "human" class is still far from satisfactory due to the large variation of human bodies, self-occlusion, etc.

Inspired by the impressive recent advances in scene understanding [16]–[18], human detection [19], [20], tracking [21] and segmentation [1], [22], [23], we solve the MHIC problem with a conditional random fields (CRFs) framework in a principled manner. At the heart of our approach is the integration of the human notion into a probabilistic CRF model, which is implemented with a few innovative human object cues proposed in this paper. Our key observation is that the essential goal of MHIC is to segment out and annotate "humans" rather than other objects or "stuff" (e.g. sky, grass). Such a human-centric constraint has not been explored in the previous cosegmentation works [5], [14], [15], and we propose an effective and efficient framework to address the MHIC task with significantly improved performance. Similar ideas of incorporating object-like proposals [24] or object detectors [20], [25] in a conventional CRF framework have been successfully applied before to other visual computing tasks such as large-scale image segmentation [12], [13], scene understanding [26] and co-segmentation [15], [27] according to a recent survey by Zhu *et al*. [28]. However, the MHIC task considered here is unique and very challenging – the user only gives a minimal amount of annotations on just a few example photos, while possible geometric and photometric variations that irregularly occurring human instances exhibit across the photo set can be quite large. This paper is hence triggered to answer how far we can achieve for the challenging MHIC task, by employing recent advances from human detection and tracking [19], [21]

H. Zhu and S. Lu are with Institute for Infocomm Research, A*Star, Singapore (email: {zhuh, slu}@i2r.a-star.edu.sg).

J. Lu is with the Advanced Digital Sciences Center, Singapore, 138632 (e-mail: jiangbo.lu@adsc.com.sg). (*Corresponding author: J. Lu.)*

J. Cai, J. Zheng, and N. M. Thalmann are with School of Computer Engineering, Nanyang Technological University, Singapore. (email: {asjfcai, asjmzheng, nadiathalmann}@ntu.edu.sg).

This study is supported by the research grant for the Human-Centered Cyber-physical Systems Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR). This work was mainly done when H. Zhu was with the Advanced Digital Sciences Center and Nanyang Technological University, Singapore.

and robust higher-order CRFs inference [18].

In this paper, we make the following contributions to address the MHIC problem systematically and effectively:
1) We extend the CRF framework of our earlier MFRC work [15] to the MHIC problem by incorporating the human cues which greatly improve the performance for the 'human' class, where only a small fraction of input images are weakly labelled with bounding boxes. Specifically, 2) we propose an efficient human instance detector to localize human instances by extending a previously proposed multi-class color line model [15] with the poselet-based human part detector [19]. 3) We also propose to capture a long range feature context by introducing adaptive cross regions [29] as basic spatial supports for the first time to evaluate dense pixel wise histogram features, which has proved its effectiveness in other visual computing tasks such as stereo matching and saliency detection. 4) To capture human shape information, we propose an effective high-level edge-aware human shape cue by enhancing the detector response map with the help of recently proposed geodesic distance transform [30] and joint bilateral filtering [31]. The shape cue is further used to weight the detector's higher-order potential, which improves visual results and inference speed. 5) We create a novel *NTU-MHIC* dataset to facilitate benchmarking the performance of various algorithms on the MHIC task. This dataset consists of 22 subsets and 351 images with pixel wise ground-truth for human instance labelling, featuring multiple human instances in the MHIC scenario with various poses and scale change. We will make the dataset publicly available in the near future.

II. Related Works

Human Detection and Segmentation: Human detection and grouping is a fundamental task with many real applications. The tree-structured pictorial framework [20], [32] is well-known for human detection and achieves leading performance on several PASCAL VOC challenges. However, this framework is not good at handling fore-shortening and partial occlusion. Bourdev *et al*.'s work [19] eschewed the pictorial structure by learning poselets for human parts which are tightly clustered in the appearance and configuration space and achieved more accurate localization. It also provides richer information from training data, e.g. by transferring the binary segmentation mask from the training data to the test image, so it can generate a rough, aggregated belief mask which indicates the location of certain human parts. Using larger spatial support, human detection is widely applied in recent human grouping tasks [33]–[37] in image and video tracking, and achieves better results than previous works using features extracted from human face detection [38]–[40]. Face detection-based approaches still have difficulties with profile poses and positions. However, these recent works based on human detection do not perform any pixel wise labelling or background modelling to boost human identification.

Human segmentation, on the other hand, intends to assign each pixel a label to indicate whether the pixel belongs to a human or background. Traditional semantic segmentation [16] cannot be directly applied to the MHIC task as all the pixels will be assigned the same label "person". Recently, Ladicky *et al*. [22] and Vineet *et al*. [1] proposed to segment human instances from a single image or video using conditional random filed and human detection. However, these methods essentially made prior decisions on human hypotheses, because they are initialized by human detectors and are sensitive to false detections. Different from these methods, our method makes a joint decision on human instances by taking account of the cues from various levels, which is more robust to false and duplicate human detections. Finally, they also do not perform any identification on human instances.

Cosegmentation: There is a vast amount of prior work on cosegmentation [5]–[10], [41]–[43]. Most of the existing works focus on handling the binary cases, separating foreground(s) from the background, but few of them are designed for joint multi-class object recognition and segmentation. The unsupervised methods such as DC [9] and Cosand [6] used low-level bottom-up features, so they cannot distinguish "stuff" from "objects" in presence of background clutter and sharing features among classes. To overcome the ill-defined nature of unsupervised methods, some user inputs are hence desired and also often necessary, and one notable work is iCoseg [42].

The aforementioned methods, however, require the user to carefully sort out a given event photo set manually to group images containing the same objects together. Recently, Kim and Xing [5] propose the first method to handle irregularly occurred multiple objects cosegmentation problem – *Multiple Foreground Cosegmentation (MFC)*. They used an combinatorial auction approach with spanning tree-based pruning, which is an over-simplified model and produces sub-optimal results. Ma and Latecki [14] proposed to solve the MFC problem using a semi-supervised graph transduction framework which enforces connectivity in the labelling result, but this method is weak in scalability due to the reliance on dense pair-wise image analysis. Both of the aforementioned methods did not model the concept of "objects" explicitly, and they frequently label regions belonging to "stuff" such as "sky" and "grass" as foreground objects. To overcome this problem, our previous work [15] includes higher level, non-local object cues into a probabilistic inference and optimization framework. The object cues derive from an object detector based on color-line modelling without any shape information. Although our previous work can handle objects exhibiting certain degrees of rotation, scale, and illumination changes and produce state-of-the-art performance for MFC tasks, it still generates unsatisfactory results for the "human" class due to strong appearance variations of human bodies as well as background clutters.

To detect specific human instances and better handle body variations, we propose to tackle the challenge by extending our previous work [15] , and combine it with the poselet-based human detector [19]. In addition, a soft shape mask map for humans is newly generated for each input image, which captures the spatial distribution of articulated human bodies probabilistically. Employing a shape prior has been proved to be very useful for scene understanding [44], but usually a rigid shape model is used [16], [44]. Finally, we propose to compute pixel wise features using larger spatial

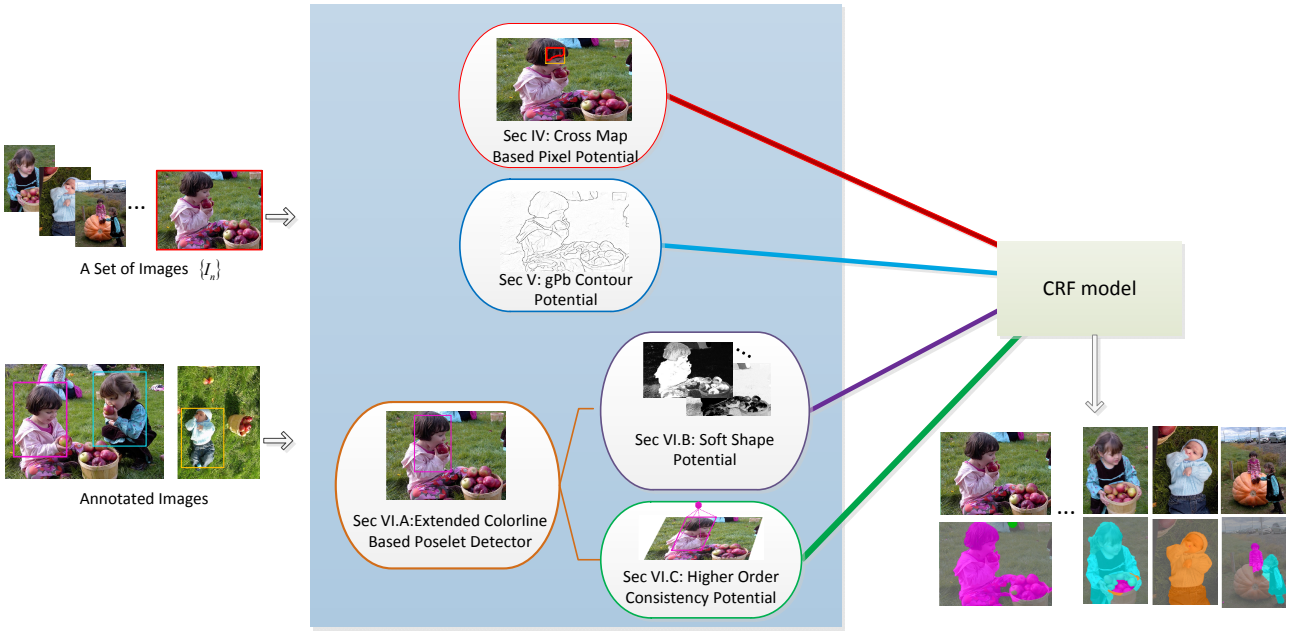

Fig. 1: Overview of our MHIC algorithm. Given a set of event photos, a user annotates only a few of them to indicate human instances of interest with bounding-boxes in different colors. Our algorithm then jointly localizes, identifies and segments out all the human instances for the given image set. It tackles the MHIC task by integrating object-level cues with mid-level and low-level cues in a probabilistic CRF framework. The proposed algorithm yields the segmentation of human instances of interest in each image, and identifies them by taking on the same color annotated earlier for the corresponding human class.

supports provided by the cross region structure [29], [45], which have been successfully applied to dense stereo [29] and saliency detection tasks [46], achieving robustness to noise and providing a better discriminative power. Beyond these novel algorithmic designs, this paper also has the scalability and long-range object modelling advantages of our previous work [15], but greatly advances its performance in differentiating human classes from other non-human objects and "stuff".

Scene Understanding: The last few years have seen impressive progress in combining multi-class object segmentation and recognition techniques to address the grand challenge of complete scene understanding [16], [26]. Ladicky *et al*. [26] proposed to incorporate object detector-induced potentials into a CRF energy optimization framework as a soft constraint, which clearly improved the standard object class segmentation models that tend to under-perform on the "things" classes for complex scenes. Recently, Tighe and Lazebnik [44] utilized the rigid shape information transferred by Exemplar SVM [47] in scene understanding tasks, which achieved state-of-the-art performance. Inspired by these nice existing techniques, our work, however, also differs from them in several aspects. First of all, as argued in our previous work [15], the MHIC task is very unique and challenging due to the high variability of foreground objects across the given set of photos and the minimal supervision that is available. Second, geared towards this MHIC task, our algorithm is designed with some novel and critical technical modules that explicitly model and handle "human" classes, whose non-rigid motions and complex interactions among them and also with the background create several real challenges.

III. Problem Formulation

Given a set of N input images $\mathcal{I} = \{I_1, ..., I_N\}$, assume m $(m \ll N)$ of them $\mathcal{I}_t = \{I_t^1, ..., I_t^m\} \subset \mathcal{I}$ are annotated with bounding boxes or polylines to delineate the spatial extents of certain humans of interest. Each image from this training set $\mathcal{I}_t$ contains a subset of annotated humans belonging to K different humans $\mathcal{H} = \{H_1, ..., H_K\}$. Each human H_i is associated with a label $l_i \in \mathcal{L} = \{0, 1, ..., K\}$, where 0 is used to denote the background. The MHIC problem is formulated in terms of a global energy function defined on a conditional random field (CRF), for which the goal is to assign a random variable x_i for each pixel i in each image a label from $\mathcal{L}$. Our framework learns hierarchical, complementary "human" cues from the trained adaptive spatial support classifier, contour detectors and the human detectors. The proposed framework is generic and flexible, and also allows to integrate other multi-class object detectors and classifiers.

Fig. 1 illustrates the proposed framework, which consists of two main stages and several specific modules. During the preprocessing stage, various cues such as a cross region based pixel classifier, a color line based human instance detector and gPb contour are modelled and generated. Pixel and object detectors are trained with user-drawn bounding boxes. After all the cues are computed, we integrate them into a global energy function which enforces the labelling consistency between various level cues and finally produces the solution with fast expansion/move solvers. Once the initial segmentation is generated, our framework supports iteratively updating the learned models and performing the recognition and segmentation tasks to further improve the results.

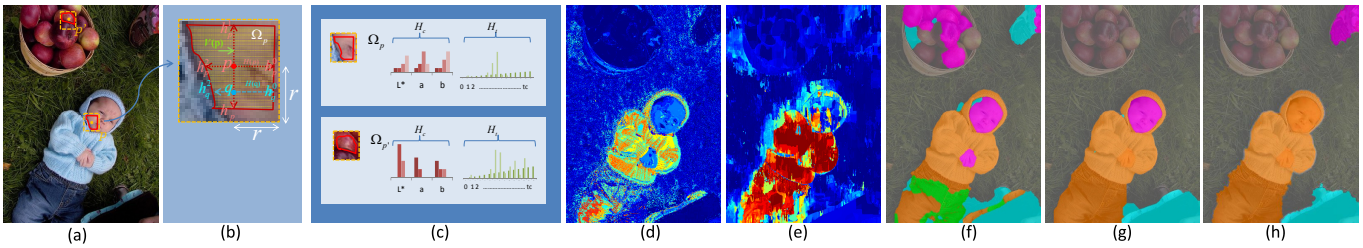

Fig. 2: The proposed cross region based pixel wise unary potential outperforms the conventional single pixel based counterpart [15] for the MHIC task. (a) Two example adaptive cross regions defined at pixel p and p'. (b) A close-up view of the adaptive cross region Ω_p defined at pixel p. (c) The color histograms H_c and texton histograms H_t extracted from the respective cross support regions at p and p'. (d) The conventional pixel-level classifier response map for the human class "boy-in-blue". (e) The proposed adaptive region-based classifier response map for the same "boy-in-blue" class. (f) The final multi-class labelling result based on (d). (g) The final multi-class labelling result based on (e). (h) The ground truth label map. (All the figures in this paper are best viewed electronically.)

A. Proposed CRF Framework for MHIC

To make the MHIC problem tractable, we make two practical assumptions: i) each image I_n contains a subset of human $\mathcal{H}$, and ii) persons with consistent appearance in terms of color and shape should be assigned the same human identity label. If one dresses differently in each image, it is nearly unlikely to label them consistently across input images.

The MHIC task is formulated as a multi-labelling problem within a CRF framework on a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V}$ is the set of all image pixels of image I_n, while $\mathcal{E}$ corresponds to the set of all edges defined by an eight-connected neighborhood system. The proposed energy function is:

$$\begin{aligned} \mathbf{E}(\mathbf{x}) = & \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j) \\ & \sum_{i \in \mathcal{V}} \psi_i^s(x_i) + \sum_{d \in D} \psi_d(\mathbf{x}_d) \ . \end{aligned} \tag{1}$$

In Eq. (1), $\mathbf{x}$ denotes the valid label map assigned to the random variables $\{x_i\}$, which takes a value from the label set $\mathcal{L}$. We denote the set of human detections with D, which are returned bounding boxes enclosing potential human instances. The energy function consists of four terms: **(1)** the pixel-based unary potential $\psi_i(x_i)$, which is trained by using the features extracted from a pixel wise adaptive cross support region [29]; **(2)** the pairwise smoothness potential $\psi_{ij}(x_i, x_j)$ based on a gPb contour detector [48]; **(3)** the soft shape potential $\psi_i^s(x_i)$ that evaluates the likelihood of each pixel to lie within each potential human shape, where the shape is adapted according to the internal image structure; **(4)** the object detector potential $\psi_d(\mathbf{x}_d)$, where $\mathbf{x}_d$ is the clique defined at the bounding box d, charging the label inconsistency cost robustly with the number of variables in the bounding box not taking the detector label. These terms collectively capture the information for human instances in a complementary way. We will elaborate the four terms in following sections.

IV. Cross Region Based Unary Potential

The first term $\psi_i(x_i)$ is a unary potential defined on each pixel which indicates its cost of being assigned a label $l \in \mathcal{L}$:

$$\psi_i(x_i) = -\omega_{pix} \log P(x_i | \mathcal{C}_{pix}) \ , \tag{2}$$

where ω_{pix} is the weighting factor. $P(x_i|\mathcal{C}_{pix})$ is the normalized probability evaluated by a Random Forest (RF) classifier $\mathcal{C}_{pix}$ [49]. Given the human class label provided in the form of user-drawn bounding boxes, a RF classifier is typically trained using color and texton features extracted from a single pixel, which is also the scheme used in our previous work [15]. However, the pixel-level features are usually too local to capture the change of neighborhood patterns, often resulting in a noisy and weak classifier response (see Fig. 2(d)). As shown in Fig. 2(f), this weak response signal leads to an unsatisfactory segmentation result. This observation motivates us to make use of a larger spatial context for each pixel when training a RF classifier. Though superpixels appear to be an option here, partitioning an image into non-overlapping local regions suffers from the superpixel quantization artifacts. To produce accurate pixel segmentation, Kohli *et al.* [18] and Zhu *et al.* [15] formulate the multi-level superpixel cues as higher order term, causing an increased inference cost.

Recently, pixel wise adaptive spatial supports–cross regions [29], [45]–have been successfully used in stereo matching cost aggregation [45], image filtering [29] and saliency detection [46], [50], which achieved more robust and accurate results than single pixel or superpixel based estimation. Another advantage of adaptive cross regions is that they can be very efficiently computed for each pixel densely. In this paper, we investigate incorporating this flexible data structure to evaluate the pixel wise classification cost.

To make the manuscript self-contained, we first give a brief introduction to the construction of cross regions, and more technical details can be found in [29]. An adaptive cross support region is constructed with the following steps. First, for each pixel p, four varying support arm lengths $\{h_p^0, h_p^1, h_p^2, h_p^3\}$ are decided based on the guidance image I, which is called a cross skeleton in [45]. An improved strategy for adaptive scale selection is proposed in [29]. The arm lengths record the largest up/down vertical span and the left/right horizontal span, s.t. $|I_c(s) - I_c(p)| \leq t, c \in \{R, G, B\}, s \in W_p$, where W_p is the window of size $(2r+1) \times (2r+1)$ centered at

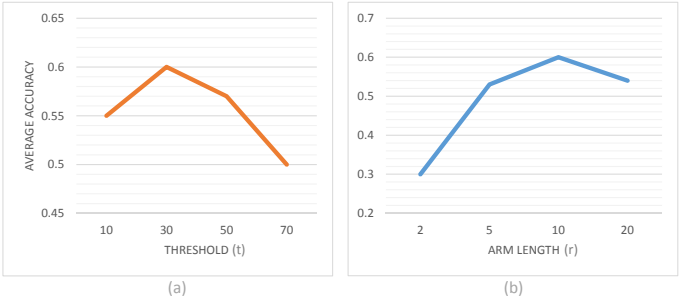

Fig. 3: Parameter sensitivity analysis for (a) the color similarity threshold t and (b) the preset maximum arm length r. The y-axis shows the average segmentation accuracy, which will be detailed in Sect. VIII.

p, and s is the pixel on the tip of the left(/right or up/down) arm. The influence of the color similarity threshold t and the maximum allowable arm length r is shown in Fig. 3(a) and Fig. 3(b), respectively. The findings are generally similar to those observed in related early works [45], [46]: the optimal maximum arm length r typically takes values from $[5, 15]$, and the color threshold t varies in $[20, 40]$. For our task, we empirically set $r = 10$ and $t = 30$. When r is set too small, the performance degrades due to the insufficient spatial support to pool discriminative features reliably. However, if r and t are set too large, the performance also gets much worse, because the noise from an excessively large spatial support contaminates local feature observations.

Once a pixel wise cross skeleton is adequately decided, a shape-adaptive full support region Ω_p is readily available as an area integral of multiple horizontal segments $H(q)$ spanned by pixel q [29]. Specifically, $\Omega_p = \bigcup_{q \in V(p)} H(q)$, where q is a support pixel located on the vertical segment $V(p)$ defined for pixel p. Two example cross regions are shown in Fig. 2(a).

With the cross map constructed, we extract a pixel wise color histogram H_c and a texton histogram H_t from the adaptive support region around each pixel, as shown in Fig. 2(c). The color histogram H_c is generated by quantizing each channel of L^*ab to 8 bins. The texton histogram H_t is generated by convolving the image with 17-dimensional filter banks at different scales, and then the responses are clustered using the Euclidean-distance K-means algorithm into $T_c = 92$ code words. Therefore, each pixel will be represented by a 116-dimensional feature vector. Based on these densely computed region-level features, we train a multi-class human classifier. An example response map (color-coded as a heat map) for the "boy-in-blue" class is shown in Fig. 2(e), which gives a much stronger and reliable response for those true pixels covered by the boy in blue. Such an improved pixel wise unary potential also leads to a much better identification and segmentation result as shown in Fig. 2(g).

V. Contour Based Pairwise Smoothness Potential

Conventional contour detectors typically capture part transitions by finding local extrema, which usually produce a high-recall but low-precision contour detection result. Recently,

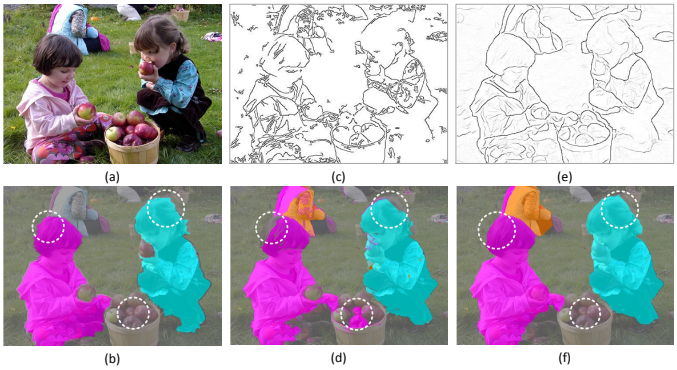

Fig. 4: Comparison between the two different contrast-sensitive smoothness measures. Left column: (a) Input color image. (b) Ground-truth label map. Middle column: (c) Canny edge detection map (intensity inverted) for the input image. (d) The final labelling result based on (c). Right column: (e) gPb contour detection map (intensity inverted) for the input image. (f) The final labelling result based on (e). The white dashed circles highlight three places where using a gPb based pairwise potential yields much better labelling results than the Canny edge based potential.

Arbelaez *et al.* [48] proposed a new method called gPb, which combines the local contour with the contour signal from eigen vectors that considers the region size and contour strength. The gPb method achieves the state-of-the-art contour detection results.

As the gPb contour map [48] provides more reliable and higher-level reasoning of salient contours, we replace the classical color contrast based pairwise potential with a gPb based potential $\psi_{ij}(x_i, x_j)$, which is defined as follows,

$$\psi_{ij}(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j \\ \omega_a(1 - \|\nabla C(i,j)\|^2) & \text{otherwise .} \end{cases} \quad (3)$$

where ω_a gives the weight of the pairwise potential. $\nabla C(i, j)$ measures the gPb signal contrast between two adjacent pixels i and j. We observe using gPb as the base edge map makes the labelling snap to salient object boundaries, and an example is demonstrated in Fig. 4.

VI. Incorporating Human Detector Based Cues

The information carried by an image patch or segment by itself is often too local and hence ambiguous, which can be easily interfered when background contains similar local patterns, as it is incapable to capture the global configuration of object instances. This motivates us to address the MHIC challenge with higher and longer range grouping cues which have been proved to be useful in recent image summarization and scene understanding research [17], [18], [44]. A popular approach is to reason about the objects of interest with the help from rectangular bounding boxes generated by some detection methods [19], [20], [25], [51] or using rigid shape templates [44]. Both methods have limitations. Firstly, most object detectors have little or no information of the shape that the detected objects cover, so existing methods often

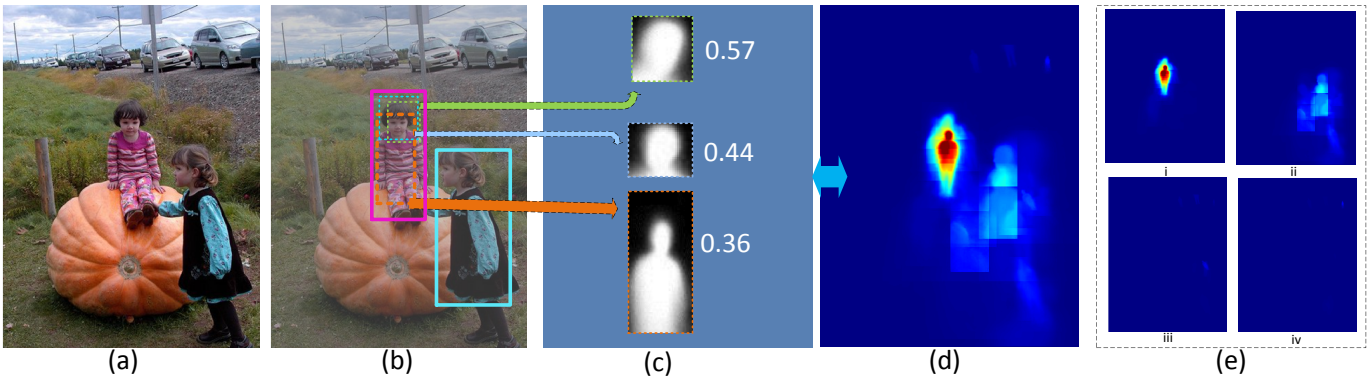

Fig. 5: The proposed human instance detector allows to utilize the soft segmentation mask associated with each poselet template to define a human-sensitive shape prior. (a) Input image. (b) The final detected human bounding boxes (solid boxes in pink and blue) for different human instances. They are computed by running all the poselet detectors (three example dashed boxes) at various positions and scales as in [19]. (c) The detection scores for the three example poselet detectors. (d) The human class-agnostic response map computed by fusing the soft masks transferred from all poselet templates, with the fusion weights defined by the respective poselet detection scores. (e) The human class-specific response map (four classes in this example) by further considering the detection score evaluated by our proposed multi-class color line texton classifier.

rely on a heuristic Grabcut approach [52] to generate the shape mask. However, Grabcut itself is sensitive to the background/foreground color modelling when an image contains clutter. On the other hand, the shape template previously learned or human labelled [44] exhibits strong rigidity: though the high template response area can hit part of the real object, the response map produced by this approach still does not overlap with real object locations sufficiently well, and is also not aligned to object boundaries due to the sliding window step size. This section will introduce a novel technique to generate an edge-aligned soft shape mask based on the poselet human detector [19] and a multi-class multiple color line model [15] for the MHIC task.

A. Efficient Color Line and Texton Histogram Based Poselet Human Detector

To detect humans of a certain identity with a desired degree of invariance to e.g. scale and rotation changes, we train two classifiers. One is the poselet human detector trained with the annotated H3D dataset [19], and the other is the multi-class interactive off-line color and texton histogram based object detector [15], [21].

The poselet detector is trained by finding patches with similar key-point configurations in the training objects to guarantee the semantic consistency of detected parts. Each poselet comes with a soft mask by averaging all aligned masks of training examples. In our paper, a pre-trained human model and its associated soft masks from [19] is used in our implementation. Example poselet soft masks are shown in Fig. 5(c).

Given a user-drawn bounding box, to train the interactive off-line color and texton histogram based object detector, we project all pixel colors onto a set of one-dimensional (1D) lines in the RGB color space [15], [21]. These lines are evenly sampled in 13 directions which pass through $(128, 128, 128)$ and then a 1D (normalized) histogram of the projected values is calculated on each line. Through an empirical comparison in train/validation sets, we use eight bins for each line and treat all 13×8 color bins as the final color line feature, which can be efficiently extracted by using integral histogram [53]. To better handle background clutter, we additionally extract a texton histogram for each bounding box as the texture features, whose dimension is the same as the histogram used in Sec. IV, and the final feature dimension used for training is 196 dimensions. A JointBoosting classifier [54] is used to train a multi-class bounding box classifier. The details of our learning procedure resemble closely with those described in [15]. The positive training samples are provided by the user annotated bounding boxes with multi-class labels. To generate more positive examples and also be robust to variations across images, the same appearance perturbation scheme [15], [21] is employed, which perturbs the position and lighting scale of the object rectangles randomly by a small amount. The negative examples are randomly sampled around the non-selected foreground regions using the bounding boxes of the same size as the user-specified ones, also with simulated scale and lighting variations.

The detection proposals are generated by first sweeping different poselet templates in the test image and then the activations are merged into the final bounding boxes as in [19]. Then each bounding box is evaluated by both the trained multi-class color line texton classifier and poselet detector, the corresponding detection scores s_c^l and s_p are linearly combined using an empirically set trade-off factor $\varepsilon = 0.9$ to form the final score $R_d = (1 - \varepsilon) \times s_c^l + \varepsilon \times s_p$. The parameter ε is selected through grid search in the range of $[0.1, 0.9]$ with a step-size of 0.1. We find that the performance is generally good in the range of $[0.6, 0.9]$, with $\varepsilon = 0.9$ performing slightly better than other settings. This can be explained by the challenging background clutter often seen in our image dataset, which makes the shape cue more important than the color cue.

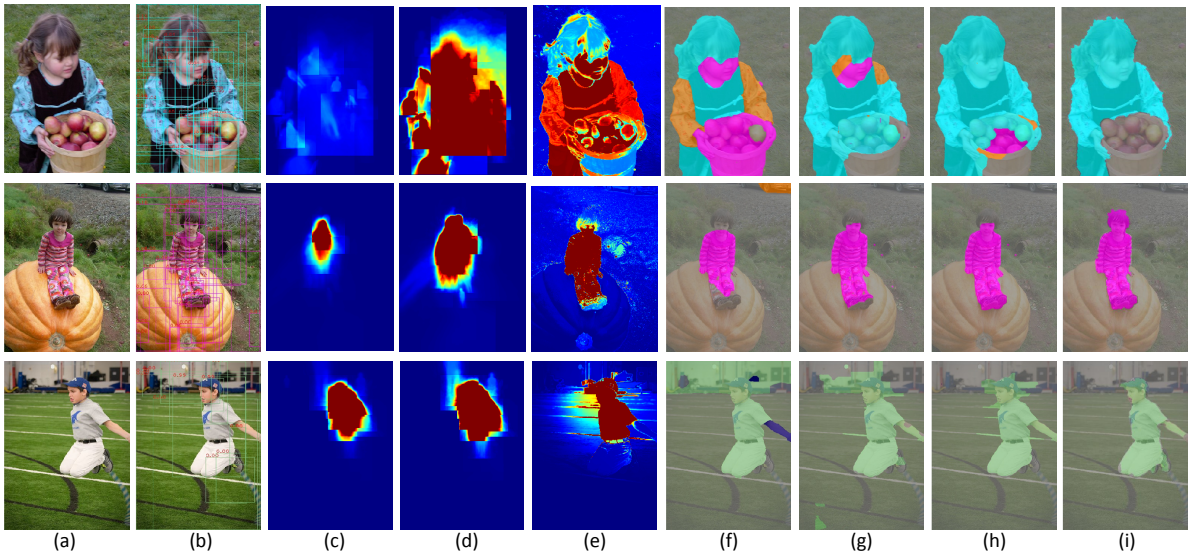

Fig. 6: Effects of the soft shape cue and the higher order object detector potential. (a) Input images. (b) Detected bounding boxes for corresponding human classes. (c) The initial transferred poselet masks. (d) The enhanced shape masks using generalized geodesic distance transform [30]. (e) The final soft shape maps generated by using an edge-aware joint bilateral filter [31] to filter (d). (f) The labelling results without using soft shape cues, where the pixels with different labelling colors from the ground truth are the labelling errors. (g) The labelling results by combining shape cues and the cross region based unary term. (h) The results using higher order detector potentials. (i) The ground-truth label maps. By incrementally including the shape potential and the higher order detector potential, one can observe the improvements in labelling accuracy.

The bounding boxes for each class l are retained, which form the final bounding boxes hypotheses set D in Eq. (1).

B. Edge-Aware Soft Shape Map with GGDT and JBF

As rigid shape templates have proved to greatly improve the performance of scene understanding tasks, it is interesting to explore whether a shape template can improve the more challenging MHIC task concerning non-rigid humans. Since the training dataset provided by [19] includes foreground/background annotation, each poselet template t comes with a soft mask $M_t \in [0,1]$ by averaging the binary segmentation annotations among all example patches used for training the respective template. Therefore, for each bounding box of each class, we overlay the soft masks of the merged poselets on the test image, which are weighted by their corresponding poselet template detection score p_t and then further weighted by their corresponding bounding box score R_d. Iterating over all the detected top human bounding boxes, this process produces a pixel wise soft belief map $M^l : \Omega \rightarrow [0,1]$ for each human class l, where Ω is the discrete image 2D domain. Fig. 5 illustrates the process how the soft masks are generated. Fig. 6(c) shows the rough hit maps of different human classes.

Despite that the proposed detector succeeds in eliminating most non-object regions, the belief map M^l (see Fig. 6(c)) for each class l is still blurry and contains some mistakes caused by sliding window offsets and false detection. Based on this observation, we propose to use the color image's internal structure to refine and enhance the initial belief map. In this paper, we apply two efficient approaches to enhance the derived belief map M, and we explain them next.

As objects are often compactly clustered in space, such connectivity can be captured by recent *Generalized Geodesic Distance Transform* (GGDT) filter [30]. Given a guidance image J, GGDT filtering can efficiently assign each pixel a shortest distance $\mathcal{Q}$ from the non-object region defined by the soft belief map M,

$$\mathcal{Q}(\mathbf{m}; M, \nabla J) = \min_{\mathbf{m}' \in \Omega} (G(\mathbf{m}, \mathbf{m}') + \nu M(\mathbf{m}')) \,, \qquad (4)$$

where Ω is the image 2D domain, and ν is an amplification parameter. The geodesic distance $G(\mathbf{m}, \mathbf{n})$ between pixels $\mathbf{m}$ and $\mathbf{n}$ is given as:

$$G(\mathbf{m}, \mathbf{n}) = \inf_{\Gamma \in \mathcal{P}_{\mathbf{m},\mathbf{n}}} \int_0^{l(\Gamma)} \sqrt{1 + \gamma^2 (\nabla J(s) \cdot \Gamma'(s))^2} ds \,, \qquad (5)$$

where Γ is a sub-path of all the paths $\mathcal{P}_{\mathbf{m},\mathbf{n}}$ connecting two points $\mathbf{m}$ and $\mathbf{n}$. The parameter γ controls the relative importance of the spatial distance to the image gradient.

By applying the GGDT to the initial belief map M, this weak signal is clearly enhanced, as shown in Fig. 6(d). The reason is that most background regions tend to be more spatially connected and homogeneous than the foreground regions, so the distances assigned to the foreground regions are usually larger than those on the background. This fact effectively helps to enhance the initial shape mask map.

Despite this improvement, the obtained belief map is still spatially inaccurate with respect to the true object locations. To produce an edge-aware soft belief map, we propose to use an efficient high dimensional joint bilateral filter [31] to filter the result produced by GGDT to yield the final soft spatial mask, see Fig. 6(e). This result highlights the pixel wise extent of

objects. In our current formulation in Eq. (1), we choose to integrate the soft shape mask as an additional unary potential $\psi_i^s(x_i)$ to compete with other hypothesis evaluations. As the belief map is produced from shapes, it is complementary to the color and texton based cues. Denoting by M_s the final soft shape maps for all the intended human classes, we introduce and compute the shape potential as follows,

$$\psi_i^s(x_i) = -\omega_{shape} \log P(x_i|M_s) \ , \tag{6}$$

where ω_{shape} is the weighting factor. With the help of such a soft shape term, the segmentation result is significantly improved. Fig. 6(f) and Fig. 6(g) provide some visual comparison between the results with and without the shape cue.

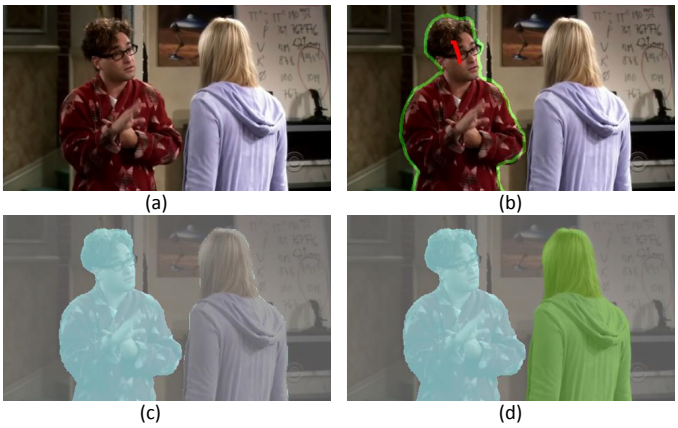

Fig. 7: Ground-truth annotation for the NTU-MHIC dataset. (a) Input image. (b) A user segments out the first human (the green contour) and indicates him as the human of interest by placing a red stroke. (c) The generated ground-truth label for the first human class. (d) The second human in the image is annotated following the same procedure. The whole process to annotate an image takes $3 \sim 4$ minutes.

C. Detector-Based Robust Consistency Potential

Although the proposed shape cue can greatly improve the result, to make our system more robust, we further incorporate higher order label consistency constraints from the detected bounding boxes. With the help of such a higher order constraint, we can revolve ambiguities which would otherwise be too hard to solve at a local level. The bounding box proposals are designed as a kind of soft constraint which works jointly with other hypotheses to overcome false positives and over-counting of object instances. Given the d-th detection bounding box with a confidence score R_d belonging to a certain class as presented in Sect. VI-A, the clique potential $\psi_{detector}(\mathbf{x}_d)$ defined on the clique $\mathbf{x}_d$ (i.e. all the pixels enclosed in the d-th detected bounding box) is defined as:

$$\psi_{detector}(\mathbf{x}_d) = \begin{cases} N_d \frac{1}{Q_d} \gamma_{max} & \text{if } N_d \leq Q_d \\ \gamma_{max} & \text{otherwise} \ , \end{cases} \tag{7}$$

where $N_d = \sum_{i \in \mathbf{x}_d} \delta(x_i \neq l_d)$ is the number of variables in $\mathbf{x}_d$ not taking the dominant label l_d . The truncation parameter $Q_d = p_d|\mathbf{x}_d|$ controls the maximum number of inconsistent pixels in the enclosed bounding box area, where p_d defines the percentage of the inconsistent pixels. The cost γ_{max} is defined as a linear truncated function $f(\cdot)$:

$$f(\mathbf{x}_d, R_d) = \omega_d \varepsilon_d |\mathbf{x}_d| \max(0, R_d - R_t) \ , \tag{8}$$

where R_t is a threshold. In Eq. (8), ω_d defines the detector potential weight, and ε_d is the aggregation of the weights w_p^i of the inconsistent pixels i in $\mathbf{x}_d$.

The proposed object potential $\psi_d(\mathbf{x}_d)$ can be transformed to take the robust P^n form [18], [26]:

$$\psi_d(\mathbf{x}_d) = -f(\mathbf{x}_d, R_d) + \min(f(\mathbf{x}_d, R_d), k_d \cdot \sum_{i \in \mathbf{x}_d} w_p^i \delta(x_i \neq l_d)) \ , \tag{9}$$

where k_d is a slope parameter defined in the same way as in [26], and w_p^i is the same as in ε_d of Eq. (8). This detector-based label consistency constraint is similar to the object detector term used in [26] for scene understanding. If $R_d < R_t$, $\psi_d(\mathbf{x}_d)$ will be zero, therefore automatically exclude the weak hypotheses. If a detector response is strong i.e. $R_d \geq R_t$, the higher-order potential will charge a penalty which is increased with the underlying inconsistent pixel number, and the penalty is increased until the truncation threshold Q_d. Minimizing the functional will amount to reducing the inconsistent pixels and therefore encouraging as many pixels belonging to the bounding box $\mathbf{x}_d$ to take the dominant bounding box label l_d. Such a soft higher-order constraint produces better labelling results than the standard P^n Potts model [55], as the robust potential allows a certain number of pixels within the clique to take different labels than the others in $\mathbf{x}_d$.

In our previous work [15], the weight w_p^i is assumed uniform across all the pixels within the bounding box. This uniform setting may produce visual artifacts and increase the inference cost, because it does not model the spatial extent of the underlying object and cannot treat the foreground human and the background differently. Given the soft shape maps produced in the preceding step, we reuse the maps to impose a spatial weighting to bias the graph cut to expand the label where it is more likely to be the potential human. The basic idea is to sum up the background belief maps of all the human instances $M_{bg} = \Sigma M_{bg}^l \backslash |L|, l \in L$, and then use the pixel wise weighting map $M_w = 1 - M_{bg}$ to replace the constant w_p^i.

Including the detector term to a CRF model is implemented by adding two auxiliary nodes into the graph, and the augmented energy function can be efficiently minimized with the graph cut algorithms. Interested readers are referred to [18] for the graph optimization details. Fig. 6(h) demonstrates the strength of the object detector-based potential when integrated into our CRF framework. Without using the detector-based potential, the girl's face in Fig. 6(g) can only be partly segmented due to the competition between the hypotheses on pixels and shapes. The object detector potential provides a complementary high-level evidence, and integrating it into the CRF model results in a more accurate result of recognizing the missing face part.

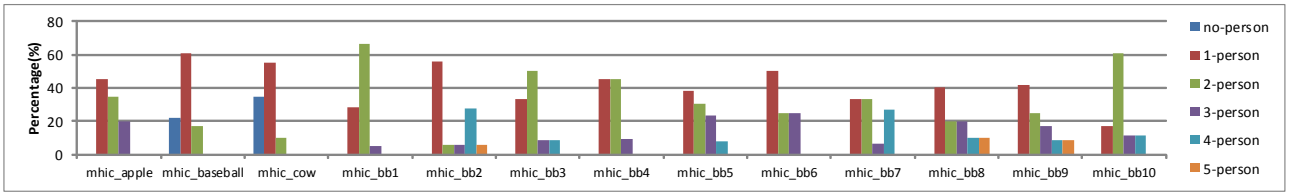

Fig. 8: The percentage of the images containing $0 \sim 5$ human instances in each test subset of the NTU-MHIC dataset.

VII. The NTU-MHIC Dataset

To evaluate our proposed approach and to create a new benchmark dataset for future work, we introduce the first and the largest multiple human foreground cosegmentation and identification dataset – the NTU-MHIC dataset. In fact, till now there exists no benchmark dataset to evaluate a method's performance for the MHIC task. The FlickrMFC dataset [5] contains some subsets with human classes, but some of them are not so challenging for the MHIC task. In addition, other object classes are also mixed with the human classes in this dataset. The CoDel dataset [33] is intended for a multiple foreground human detection task by sampling representative video frames in "Big Bang Theory", but it only provides bounding box labels.

As one of our main contributions, we create a new dataset by collecting image subsets which match the MHIC scenario from both the FlickrMFC and CoDel datasets. We manually annotate the images with pixel wise class labels, which serve as the ground-truth to evaluate various MHIC algorithms. The obtained NTU-MHIC dataset contains 22 subsets with a total of 351 images, and each subset includes around $10 \sim 20$ images for the same event. The dataset is split into two sets. One set is for training/validation and the other is for testing. The training/validation set consists of 10 subsets with 166 images. The test set consists of 13 subsets with 185 images. This dataset is challenging, because it contains both indoor and outdoor human activities (e.g. sports, child play, and group chat) with large viewpoint change and displacement, background/lighting variations and occlusion, and also only a finite number of repeating subjects are present in each image (see Fig. 10∼12). Figure 8 illustrates the statistics of the percentage of the images that contain $0 \sim 5$ human instances in each test subset of the NTU-MHIC dataset. For different image subsets, there is clearly a high diversity in terms of the distribution of the human instance number per image. For those subsets sampled from the CoDel dataset [33], more images containing at least 2 or more human instances.

Dataset Annotation: The ground-truth pixel wise labels for each image in the dataset are generated by an annotator using a labelling tool. To provide high-precision ground-truth labeling maps for benchmarking different algorithms, the human labeller first needs to delineate a closed contour around the object-of-interest, and then a foreground stroke is marked over the object to conduct a final object cutout. The labelling tool allows the user to add/delete a stroke to further refine the annotation. Fig. 7 shows the labelling process. The average time to annotate an image is around $3 \sim 4$ minutes. We make this dataset and annotation tool publicly available at our website http://hongyuanzhu.github.io/mhic/ to facilitate future work.

VIII. Experimental Results and Discussions

To tune the weights of our model, we apply a piecewise training approach. We first randomly sample 20% images of each training/validation subset as the training images as MFC [5] and MFRC [15]. Then the potentials for each subset are trained with the selected images. Finally, the weights are manually tuned on each potential to achieve the highest performance in the rest 80% validation images with ground truth. These processes repeat for five times. The final estimated weights are $\omega_{pix} = 1, \omega_a = 10, \omega_s = 0.2, \omega_d = 0.3, R_t = 0.3, \rho_{max} = -\log(0.8)$, $p_d = 0.2$, $\gamma = 1$ and $\nu = 2$. Then the parameters are fixed across all the tests. The overall time (including preprocessing, detection and segmentation) to process an image is around $10 \sim 20$ seconds on a laptop with Intel Core i7 Q740 1.7GHZ and 22GB RAM.

Quantitative Results: We first compare our method with some baselines: MFRC [15], MFC [5], CoSand (COS) [6], and Discriminative Clustering (DC) [9]. We adopt the procedure introduced in MFC [5] for evaluation. For supervised methods such as our method, MFRC and MFC's supervised version (MFC-S), we randomly pick 20% of the input images (2∼4 images) of each subset to annotate, which covers all of the objects-of-interest. For unsupervised MFC (MFC-U), we run it by changing the foreground number K from two to eight, and report the best scores for each subset. For the unsupervised binary class methods COS [6] and DC [9], the dataset is divided into several subgroups such that the images in each subgroup contain the same objects of interest, and the methods are applied to each subgroup individually.

Fig. 9 summarizes the segmentation accuracy on the 13 groups of the MHIC test dataset. We evaluate the segmentation accuracy by the standard intersection-over-union metric $\frac{(GT_i \cap R_i)}{(GT_i \cup R_i)}$. The leftmost bar set presents the average segmentation accuracy on 13 groups. Some interesting results can be observed from the chart: 1) For unsupervised methods, COS and DC's accuracies are better than the unsupervised version of MFC. The underlying reason is that DC and COS enforce a strong assumption that the set of images provided must have one common foreground in each image. 2) The methods with some supervision from a human, such as supervised MFC, MFRC and our proposed method are better than unsupervised methods in most of the cases, which on one hand proves that supervision can be beneficial for the task. 3) The higher

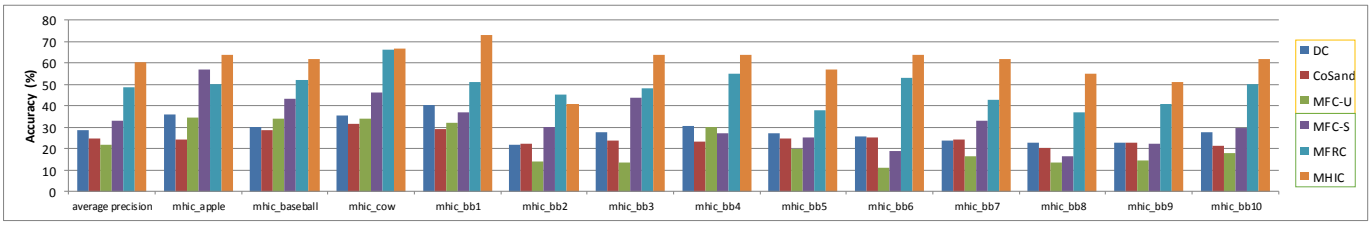

Fig. 9: Segmentation accuracy comparison between our method (MHIC) and other baselines (MFRC [15], MFC [5], CoS [6], and DC [9]) for the MHIC test dataset. In the legend, we group the unsupervised methods into the yellow box, while using the green box to group those supervised methods together.

accuracy of MFRC and our newly proposed method in comparison with supervised MFC demonstrates the clear benefits of explicit modelling of the objectness constraint. On average, MFRC and MHIC achieves 16% and 28% improvements over the supervised-MFC (MFC-S). In some subset, such as $bb6$ (where bb stands for "Big Bang Theory"), we achieve nearly 45% improvement.

We have also evaluated the average accuracy gain contributed by including the proposed shape and higher-order detector potentials into the CRF model in Table I, and also compared these models with our previous MFRC method [15]. The MFRC method [15] adopts the unary term which is derived from a single pixel estimate and a higher order term that needs multiple oversegmentations. From Table I, one can observe that by using a cross region-based unary potential alone, we achieve comparable performance with our previously proposed MFRC, which includes more complex modelling and has a much slower inference speed. This demonstrates that by using a more reliable and discriminative unary term can already bring a substantial improvement than those models which use more complex modelling. The incorporation of the shape cue contributes nearly 4% improvement in accuracy, which suggests that the shape cue and the cross region based unary term are complementary to each other. The improvement also proves that shape template indeed improves the human class segmentation, as suggested in recent computer vision literature [44]. From the visual results in Fig. 6, one can observe that such edge-aware high level shape cues greatly resolve the ambiguity still faced by the cross region based unary term.

While including the higher-order detector term brings an additional 3% improvement, the small numerical gain by the higher order detector term has also been observed in Shotton *et al.* [16] and Kohli *et al.* [18] in scene understanding research. As also indicated in [16], [18], we observe that including this potential often brings a pronounced increase in perceived accuracy in boundary areas. Moreover, a similar phenomenon can also be observed in introducing gPb contours and the weighted higher order term. A potential interpretation is that most human subjects in daily event photos have enough contrasts and the unary term provided by cross regions is often strong, therefore the improvement in boundary areas which is significant in visual perception is not that easy to be reflected in quantitative measure. In addition, the inclusion of the weighted higher order term reduces the runtime by $1\sim2$ seconds. As a result, we choose to include these two modules as they can benefit applications which require highly accurate pixel wise labelling (e.g. object cutout).

TABLE I: Components evaluation of the proposed MHIC algorithm (column 3–5) in comparison with previous MFRC [15].

Subset	MFRC ($\times100\%$)	Unary + Pairwise ($\times100\%$)	Unary + Pairwise + Shape ($\times100\%$)	Full Model ($\times100\%$)
apple	0.50	0.58	0.61	0.64
baseball	0.52	0.5	0.61	0.62
cow	0.66	0.65	0.66	0.67
bb1	0.51	0.59	0.67	0.73
bb2	0.45	0.37	0.4	0.41
bb3	0.48	0.56	0.58	0.64
bb4	0.55	0.6	0.62	0.64
bb5	0.38	0.42	0.53	0.57
bb6	0.53	0.63	0.59	0.64
bb7	0.43	0.59	0.62	0.62
bb8	0.37	0.45	0.48	0.55
bb9	0.41	0.42	0.44	0.51
bb10	0.5	0.53	0.57	0.62
average precision	0.48	0.53	0.57	0.60

On the other hand, we should also notice that introducing the shape cue occasionally reduces the accuracy. However, this only happens for one subset $bb6$, and the loss is recovered by using the higher order term. One potential reason for such a phenomenon is as follows. If the shape potential is derived from a wrong human detection result, which is not consistent with other hypotheses, our method may choose to label it as the background. Further human intervention should improve the result, which is left as one future direction.

Visual Results: Fig. 10$\sim$12 show some visual results from seven groups of the MHIC dataset. For each set, the input images and color-coded ground-truth segmentation results are displayed in the first two rows. We also show visual results from supervised MFC (MFC-S) [5], MFRC [15] and our proposed MHIC method. The regions which are labelled with the same color in each set indicate they belong to the same category. The tags below each set explain the meaning of each color. From these images, one can observe that MFC-S produced quite obvious segmentation and human identification errors. Its reliance on a coarse superpixel segmentation prevents it from correcting the errors made in the initial superpixel generation process. Without including human-specific modelling (or potentials), both MFC-S and MFRC frequently

misclassified some non-human regions as human instances, though MFRC performs much better than MFC-S due to the inclusion of object notions/terms in the CRF model. Thanks to the novel human-centric cues as well as our region-based unary terms, MHIC achieves superior human segmentation and identification quality over other competing methods for this challenging MHIC dataset. Our method can handle irregularly appearing humans and produce coherent and more accurate segmentation results. The images with no foregrounds can also be identified, e.g., the *baseball* dataset. On the other hand, our current model still cannot very well handle the cases of humans in small size or from a highly profiled view (e.g. *bb3* and *bb5* image sets in Fig. 11), or the cases with a great appearance overlap between foreground and background or the combination of these factors (e.g. the third image in the last row of Fig. 11. This remains as a future research direction.

IX. Conclusion

This paper studies the challenging problem of multiple human identification and cosegmentation, and proposes to solve the problem with a principled CRF framework using a few weakly labelled images. A novel human instance detector is proposed by combining an extended multiclass multiple color line model [15], [21] with a poselet-based human part detector [19]. We also proposed a more robust unary potential based on pixel wise adaptive cross regions, and an effective high-level human shape cue generated by applying enhancement filtering to an initial human instance detection response map. The proposed framework is flexible and can be generalized to tackle other objects. The experiments on a newly created MHIC dataset show the state-of-the-art performance on our proposed algorithm. We have released the MHIC dataset together with all ground-truth annotations to the research community to facilitate more future work on this important yet challenging topic. We also plan to explore applying recent dense correspondence methods for non-rigid object/human matching [56], [57] in this MHIC task.

References

[1] V. Vineet, J. Warrell, L. Ladicky, and P. H. S. Torr, "Human instance segmentation from video using detector-based conditional random fields," in *British Machine Vision Conference (BMVC)*, 2011.

[2] G. Yu, N. A. Goussies, J. Yuan, and Z. Liu, "Fast action detection via discriminative random forest voting and top-k subvolume search," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 507–517, 2011.

[3] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 66–75, 2012.

[4] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *IEEE International Conference on Computer Vision (ICCV)*, 2003.

[5] G. Kim and E. P. Xing, "On multiple foreground cosegmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 837–844.

[6] G. Kim, E. P. Xing, F.-F. Li, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[7] A. Joulin, F. Bach, and J. Ponce, "Multi-class cosegmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[8] J. C. Rubio, J. Serrat, A. M. López, and N. Paragios, "Unsupervised co-segmentation through region matching," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[9] A. Joulin, F. R. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[10] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[11] H. Zhu, J. Cai, J. Zheng, J. Wu, and N. Magnenat-Thalmann, "Salient object cutout using google images," in *IEEE International symposium on circuits and systems (ISCAS)*, 2013.

[12] H. Li, F. Meng, and K. N. Ngan, "Co-salient object detection from multiple images," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1896–1909, 2013.

[13] F. Meng, H. Li, G. Liu, and K. N. Ngan, "Object co-segmentation based on shortest path algorithm and saliency model," *IEEE Transactions on Multimedia*, vol. 14, no. 5, pp. 1429–1441, 2012.

[14] T. Ma and L. J. Latecki, "Graph transduction learning with connectivity constraints with application to multiple foreground cosegmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[15] H. Zhu, J. Lu, J. Cai, J. Zheng, and N. Thalmann, "Multiple foreground recognition and cosegmentation: An object-oriented crf model with robust higher-order potentials," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

[16] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *International Journal of Computer Vision (IJCV)*, vol. 81, no. 1, pp. 2–23, 2009.

[17] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr, "Graph cut based inference with co-occurrence statistics," in *European Conference on Computer Vision (European Conference on Computer Vision (ECCV))*, 2010.

[18] P. Kohli, L. Ladicky, and P. H. S. Torr, "Robust higher order potentials for enforcing label consistency," *International Journal of Computer Vision (IJCV)*, vol. 82, no. 3, pp. 302–324, 2009.

[19] L. D. Bourdev, S. Maji, T. Brox, and J. Malik, "Detecting people using mutually consistent poselet activations," in *European Conference on Computer Vision (European Conference on Computer Vision (ECCV))*, 2010.

[20] P. F. Felzenszwalb, R. B. Girshick, and D. A. McAllester, "Cascade object detection with deformable part models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[21] Y. Wei, J. Sun, X. Tang, and H.-Y. Shum, "Interactive offline tracking for color objects," in *IEEE International Conference on Computer Vision (ICCV)*, 2007.

[22] L. Ladicky, P. H. S. Torr, and A. Zisserman, "Human pose estimation using a joint pixel-wise and part-wise formulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[23] H. Zhu, J. Zheng, J. Cai, and N. Magnenat-Thalmann, "Object-level image segmentation using low level cues," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 4019–4027, 2013.

[24] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[25] R. Fergus, P. Perona, and A. Zisserman, "Weakly supervised scale-invariant learning of models for visual recognition," *International Journal of Computer Vision (IJCV)*, vol. 71, no. 3, pp. 273–303, 2007.

[26] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr, "What, where and how many? combining object detectors and crfs," in *European Conference on Computer Vision (European Conference on Computer Vision (ECCV))*, 2010.

[27] H. Zhu, J. Lu, J. Cai, J. Zheng, and N. Magnenat-Thalmann, "Poselet-based multiple human identification and cosegmentation," in *IEEE International conference on image proccessing (ICIP)*, 2014.

[28] H. Zhu, F. Meng, J. Cai, and S. Lu, "Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation," *J. Visual Communication and Image Representation*, vol. 34, pp. 12–27, 2016.

[29] J. Lu, K. Shi, D. Min, L. Lin, and M. N. Do, "Cross-based local multipoint filtering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[30] A. Criminisi, T. Sharp, C. Rother, and P. Pérez, "Geodesic image and video editing," *ACM Transactions on Graphics*, vol. 29, no. 5, p. 134, 2010.

[31] A. Adams, J. Baek, and M. A. Davis, "Fast high-dimensional filtering using the permutohedral lattice," *Computer Graphics Forum*, vol. 29, no. 2, pp. 753–762, 2010.

Fig. 10: Some randomly drawn examples from seven groups of the MHIC dataset. From top to bottom, each set presents its input images, ground-truth, color-labelled segmentation results for supervised MFC [5], MFRC [15] and our MHIC method. The colored tag below each set indicates which category each region is assigned to.

[32] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35, no. 12, 2013.

[33] J. Shi, R. Liao, and J. Jia, "Codel: An human co-detection and labeling framework," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[34] R. Garg, S. M. Seitz, D. Ramanan, and N. Snavely, "Where's waldo: Matching people in images of crowds," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[35] M. Tapaswi, M. Bäuml, and R. Stiefelhagen, "'knock! knock! who is it?' probabilistic person identification in tv-series," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[36] J. Sivic, M. Everingham, and A. Zisserman, "'who are you?' - learning person specific classifiers from video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[37] J. Sivic, C. L. Zitnick, and R. Szeliski, "Finding people in repeated shots of the same scene," in *British Machine Vision Conference (BMVC)*, 2006.

[38] D. Anguelov, K. chih Lee, S. B. Göktürk, and B. Sumengen, "Contextual identity recognition in personal photo albums," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[39] Y. Song and T. Leung, "Context-aided human recognition - clustering," in *European Conference on Computer Vision (European Conference on Computer Vision (ECCV))*, 2006.

[40] L. Zhang, L. Chen, M. Li, and H. Zhang, "Automated annotation of human faces in family albums," in *ACM Multimedia*, 2003.

[41] L. Mukherjee, V. Singh, and J. Peng, "Scale invariant cosegmentation for image groups," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[42] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "Interactively co-segmentating topically related images with intelligent scribble guidance," *International Journal of Computer Vision (IJCV)*, vol. 93, no. 3, pp. 273–292, 2011.

[43] M. D. Collins, J. Xu, L. Grady, and V. Singh, "Random walks based multi-image segmentation: Quasiconvexity results and gpu-based solu-

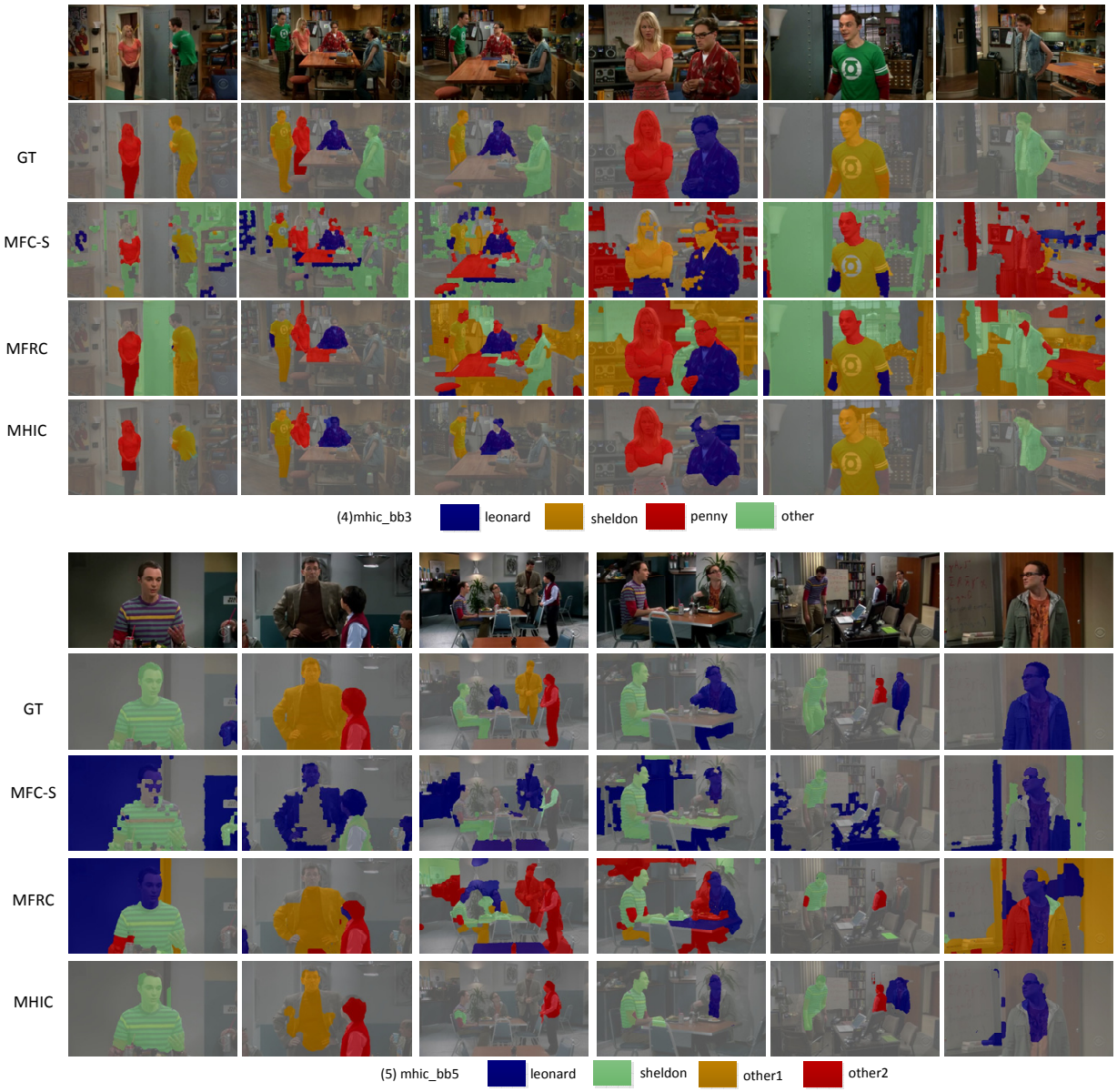

Fig. 11: Some randomly drawn examples from seven groups of the MHIC dataset. From top to bottom, each set presents its input images, ground-truth, color-labelled segmentation results for supervised MFC [5], MFRC [15] and our MHIC method. The colored tag below each set indicates which category each region is assigned to.

tions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[44] J. Tighe and S. Lazebnik, "Finding things: Image parsing with regions and per-exemplar detectors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[45] K. Zhang, J. Lu, and G. Lafruit, "Cross-based local stereo matching using orthogonal integral images," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 19, no. 7, pp. 1073–1079, 2009.

[46] K. Shi, K. Wang, J. Lu, and L. Lin, "PISA: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[47] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," in *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[48] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 33, no. 5, pp. 898–916, 2011.

[49] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," *Neural Computation*, vol. 9, no. 7, pp. 1545–1588, 1997.

[50] K. Wang, L. Lin, J. Lu, C. Li, and K. Shi, "PISA: Pixelwise image saliency by aggregating complementary appearance contrast measures with edge-preserving coherence," *IEEE Trans. Image Processing*, vol. 24, no. 10, pp. 3019–3033, Oct. 2015.

[51] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[52] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 309–314, 2004.

[53] F. M. Porikli, "Integral histogram: A fast way to extract histograms in cartesian spaces," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[54] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing features:

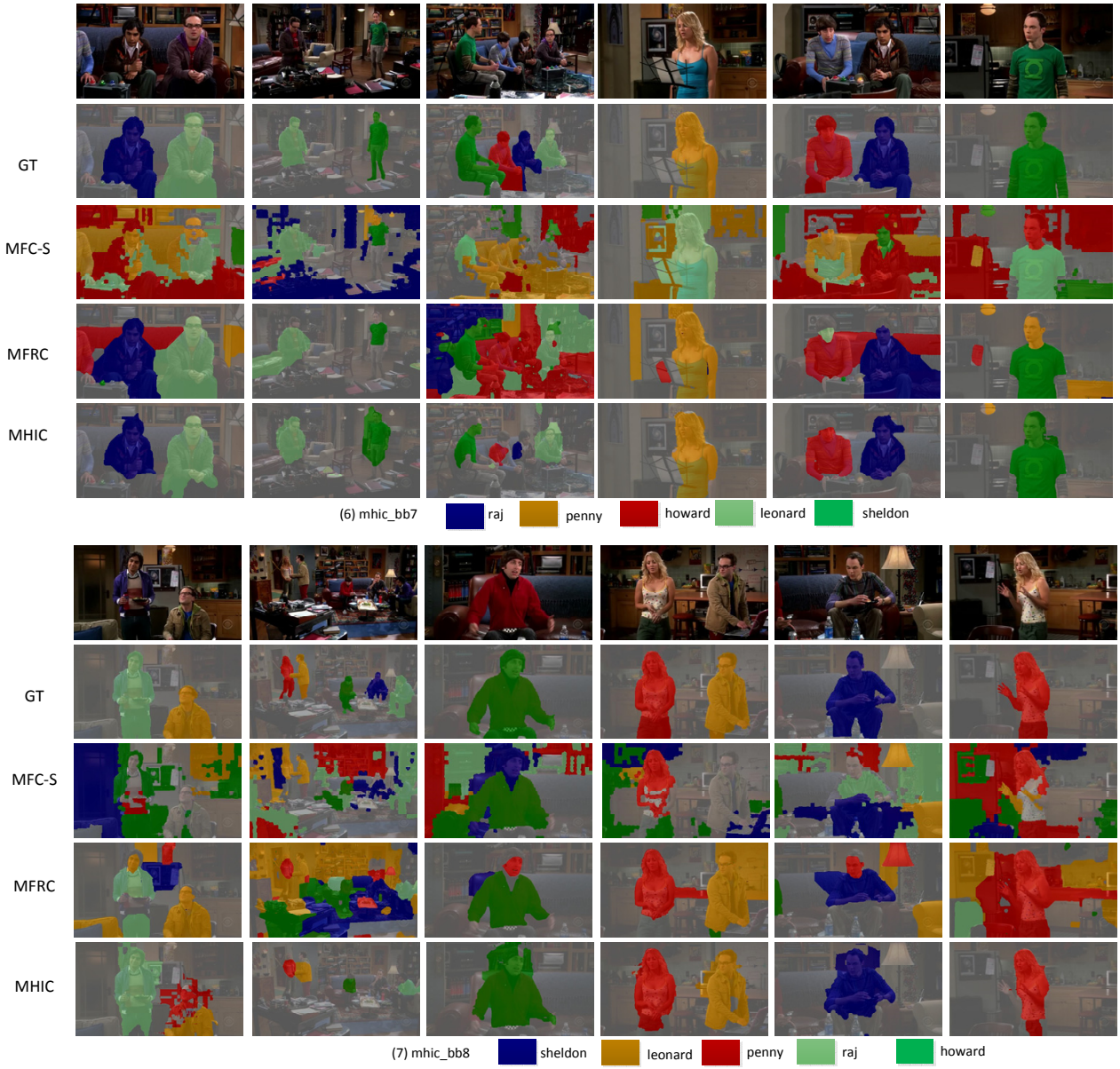

Fig. 12: Some randomly drawn examples from seven groups of the MHIC dataset. From top to bottom, each set presents its input images, ground-truth, color-labelled segmentation results for supervised MFC [5], MFRC [15] and our MHIC method. The colored tag below each set indicates which category each region is assigned to.

Efficient boosting procedures for multiclass object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[55] P. Kohli, M. P. Kumar, and P. H. S. Torr, "P3 & beyond: Move making algorithms for solving higher order functions," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 31, no. 9, pp. 1645–1656, 2009.

[56] Y. HaCohen, E. Shechtman, D. B. Goldman, and D. Lischinski, "Non-rigid dense correspondence with applications for image enhancement," in *ACM SIGGRAGH*, 2011.

[57] H. Yang, W.-Y. Lin, and J. Lu, "DAISY filter flow: A generalized discrete approach to dense correspondences," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Hongyuan Zhu (S'12-M'16) is a research scientist at the Institute for Infocomm Research, A*STAR, Singapore. Before that, he received his PhD in computer engineering from Nanyang Technological University in 2014, and B.S in software engineering from University of Macau in 2010. His research interests include multimedia content analysis and segmentation, specially image segmentation/cosegmentation, object detection, scene recognition and saliency detection. He received the top-10% paper award at the IEEE International Conference on Image Processing (ICIP 2014). He is the reviewer for IEEE Transactions on Image Processing (TIP), IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) and Springer the Visual Computer. He is a member of IEEE.

Jiangbo Lu (M'09-SM'15) received the B.S. and M.S. degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering, Katholieke Universiteit Leuven, Leuven, Belgium, in 2009.

Since September 2009, he has been working with the Advanced Digital Sciences Center, Singapore, which is a joint research center between the University of Illinois at Urbana-Champaign, Urbana, and the Agency for Science, Technology and Research (A*STAR), Singapore, where he is leading a few research projects as a Senior Research Scientist. His research interests include computer vision, visual computing, image and video processing, robotics, interactive multimedia applications and systems, and efficient algorithms for various architectures.

Dr. Lu served as an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) in 2012-2016. He received the 2012 TCSVT Best Associate Editor Award.

Jianfei Cai (S'98-M'02-SM'07) received his PhD degree from the University of Missouri-Columbia. He is currently an Associate Professor and has served as the Head of Visual & Interactive Computing Division and the Head of Computer Communication Division at the School of Computer Engineering, Nanyang Technological University, Singapore. His major research interests include computer vision, visual computing and multimedia networking. He has published more than 170 technical papers in international journals and conferences. He has been actively participating in program committees of various conferences. He has served as the leading Technical Program Chair for IEEE International Conference on Multimedia & Expo (ICME) 2012 and the leading General Chair for Pacific-rim Conference on Multimedia (PCM) 2012. Since 2013, he has been serving as an Associate Editor for IEEE Trans on Image Processing (T-IP). He has also served as an Associate Editor for IEEE Trans on Circuits and Systems for Video Technology (T-CSVT) from 2006 to 2013.

Jianmin Zheng Jianmin Zheng is an associate professor in the School of Computer Science and Engineering at Nanyang Technological University, Singapore. He received the BS and PhD degrees from Zhejiang University, China. His recent research focuses on T-spline technologies, digital geometric processing, virtual reality, visualization, interactive digital media and applications. He has published more than 150 technical papers in international conferences and journals. He was the conference co-chair of Geometric Modeling and Processing 2014 and has served on the program committee of many international conferences.

Shijian Lu received Ph.D. in electrical and computer engineering from National University of Singapore in 2005. He is currently the head of Visual Attention Lab at the Institute for Infocomm Research (I2R), A*STAR, Singapore, and an Adjunct Assistant Professor at the School of Computer Engineering, Nanyang Technological University, Singapore. He has published more than 100 international refereed journal and conference papers and co-authored 6 patents in his research areas of document image analysis and understanding, medical image analysis, computer vision, visual attention, and pattern recognition.

Nadia Magnenat Thalmann Professor Thalmann is the Director of the interdisciplinary Institute for Media Innovation in NTU and the Director of MIRALab, at the University of Geneva. She has authored dozens of books, published with her team more than 600 papers on virtual humans/virtual worlds and social robots, organised major conferences as CGI, CASA, and delivered more than 300 keynote addresses, some of them at global events such as the World Economic Forum in Davos. In NTU, Singapore, recently, she revolutionized social robotics by unveiling the first social robot Nadine that can have mood and emotions and remember people and actions. (See https://en.wikipedia.org/wiki/Nadine_Social_Robot). Besides having bachelor's and master's degrees in disciplines such as psychology, biology, chemistry and computer science, Professor Thalmann completed her PhD in quantum physics at the University of Geneva. She has received honorary doctorates from Leibniz University of Hannover and the University of Ottawa in Canada and several prestigious other awards as the Humboldt Research Award in Germany. She is Editor-in-Chief of The Visual Computer, co-Editor-in-Chief of Computer Animation and Virtual Worlds, and editor of many other scientific journals. She is a life member of the Swiss Academy of Engineering Sciences. (See http://en.wikipedia.org/wiki/Nadia_Magnenat_Thalmann)